

OpenCourseWare

STOCHASTIC PROCESSES: Theory for Applications

Prof. Robert Gallager

STOCHASTIC PROCESSES:
Theory for Applications

Draft

R. G. Gallager

December 2, 2013

Preface

This text has evolved over some 20 years, starting as lecture notes for two first-year graduate subjects at M.I.T., namely, *Discrete Stochastic Processes* (6.262) and *Random Processes, Detection, and Estimation* (6.432). The two sets of notes are closely related and have been integrated into one text. Instructors and students can pick and choose the topics that meet their needs, and suggestions for doing this follow this preface.

These subjects originally had an application emphasis, the first on queueing and congestion in data networks and the second on modulation and detection of signals in the presence of noise. As the notes have evolved, it has become increasingly clear that the mathematical development (with minor enhancements) is applicable to a much broader set of applications in engineering, operations research, physics, biology, economics, finance, statistics, etc.

The field of stochastic processes is essentially a branch of probability theory, treating probabilistic models that evolve in time. It is best viewed as a branch of mathematics, starting with the axioms of probability and containing a rich and fascinating set of results following from those axioms. Although the results are applicable to many areas, they are best understood initially in terms of their mathematical structure and interrelationships.

Applying axiomatic probability results to a real-world area requires creating a probabiity model for the given area. Mathematically precise results can then be derived within the model and translated back to the real world. If the model fits the area sufficiently well, real problems can be solved by analysis within the model. However, since models are almost always simplified approximations of reality, precise results within the model become approximations in the real world.

Choosing an appropriate probability model is an essential part of this process. Sometimes an application area will have customary choices of models, or at least structured ways of selecting them. For example, there is a well developed taxonomy of queueing models. A sound knowledge of the application area, combined with a sound knowledge of the behavior of these queueing models, often lets one choose a suitable model for a given issue within the application area. In other cases, one can start with a particularly simple model and use the behavior of that model to gain insight about the application, and use this to iteratively guide the selection of more general models.

An important aspect of choosing a probability model for a real-world area is that a prospective choice depends heavily on prior understanding, at both an intuitive and mathematical level, of results from the range of mathematical models that might be involved. This partly explains the title of the text — Theory for applications. The aim is to guide the reader in both the mathematical and intuitive understanding necessary in developing and using stochastic process models in studying application areas.

Application-oriented students often ask why it is important to understand axioms, theorems, and proofs in mathematical models when the precise results in the model become approximations in the real-world system being modeled. One answer is that a deeper understanding of the mathematics leads to the required intuition for understanding the differences between model and reality. Another answer is that theorems are transferable between applications, and thus enable insights from one application area to be transferred to another.

Given the need for precision in the theory, however, why is an axiomatic approach needed? Engineering and science students learn to use calculus, linear algebra and undergraduate probability effectively without axioms or rigor. Why doesn't this work for more advanced probability and stochastic processes?

Probability theory has more than its share of apparent paradoxes, and these show up in very elementary arguments. Undergraduates are content with this, since they can postpone these questions to later study. For the more complex issues in graduate work, however, reasoning without a foundation becomes increasingly frustrating, and the axioms provide the foundation needed for sound reasoning without paradoxes.

I have tried to avoid the concise and formal proofs of pure mathematics, and instead use explanations that are longer but more intuitive while still being precise. This is partly to help students with limited exposure to pure math, and partly because intuition is vital when going back and forth between a mathematical model and a real-world problem. In doing research, we grope toward results, and successful groping requires both a strong intuition and precise reasoning.

The text neither uses nor develops measure theory. Measure theory is undoubtedly important in understanding probability at a deep level, but most of the topics useful in many applications can be understood without measure theory. I believe that the level of precision here provides a good background for a later study of measure theory.

The text does require some background in probability at an undergraduate level. Chapter 1 presents this background material as review, but it is too concentrated and deep for most students without prior background. Some exposure to linear algebra and analysis (especially concrete topics like vectors, matrices, and limits) is helpful, but the text develops the necessary results. The most important prerequisite is the mathematical maturity and patience to couple precise reasoning with intuition.

The organization of the text, after the review in Chapter 1 is as follows: Chapters 2, 3, and 4 treat three of the simplest and most important classes of stochastic processes, first Poisson processes, next Gaussian processes, and finally finite-state Markov chains. These are beautiful processes where almost everything is known, and they contribute insights, examples, and initial approaches for almost all other processes. Chapter 5 then treats renewal processes, which generalize Poisson processes and provide the foundation for the rest of the text.

Chapters 6 and 7 use renewal theory to generalize Markov chains to countable state spaces and continuous time. Chapters 8 and 10 then study decision making and estimation, which in a sense gets us out of the world of theory and back to using the theory. Finally Chapter 9 treats random walks, large deviations, and martingales and illustrates many of their applications.

Most results here are quite old and well established, so I have not made any effort to attribute results to investigators. My treatment of the material is indebted to the texts by Bertsekas and Tsitsiklis [2], Sheldon Ross [22] and William Feller [8] and [9].

Contents

1 INTRODUCTION AND REVIEW OF PROBABILITY **1**

1.1 Probability models 1

1.1.1 The sample space of a probability model 3

1.1.2 Assigning probabilities for finite sample spaces 4

1.2 The axioms of probability theory 5

1.2.1 Axioms for events 7

1.2.2 Axioms of probability 8

1.3 Probability review 9

1.3.1 Conditional probabilities and statistical independence 9

1.3.2 Repeated idealized experiments 11

1.3.3 Random variables 12

1.3.4 Multiple random variables and conditional probabilities 15

1.4 Stochastic processes 16

1.4.1 The Bernoulli process 17

1.5 Expectations and more probability review 20

1.5.1 Random variables as functions of other random variables 23

1.5.2 Conditional expectations 25

1.5.3 Typical values of rv's; mean and median 28

1.5.4 Indicator random variables 29

1.5.5 Moment generating functions and other transforms 30

1.6 Basic inequalities 32

1.6.1 The Markov inequality 32

1.6.2 The Chebyshev inequality 33

1.6.3 Chernoff bounds . . . 33
1.7 The laws of large numbers . . . 37
1.7.1 Weak law of large numbers with a finite variance . . . 37
1.7.2 Relative frequency . . . 40
1.7.3 The central limit theorem . . . 40
1.7.4 Weak law with an infinite variance . . . 45
1.7.5 Convergence of random variables . . . 46
1.7.6 Convergence with probability 1 . . . 49
1.8 Relation of probability models to the real world . . . 51
1.8.1 Relative frequencies in a probability model . . . 52
1.8.2 Relative frequencies in the real world . . . 53
1.8.3 Statistical independence of real-world experiments . . . 55
1.8.4 Limitations of relative frequencies . . . 56
1.8.5 Subjective probability . . . 57
1.9 Summary . . . 57
1.10 Exercises . . . 59

2 POISSON PROCESSES 74
2.1 Introduction . . . 74
2.1.1 Arrival processes . . . 74
2.2 Definition and properties of a Poisson process . . . 76
2.2.1 Memoryless property . . . 77
2.2.2 Probability density of S_n and joint density of $S_1, \ldots, S_n$ . . . 80
2.2.3 The PMF for $N(t)$ . . . 81
2.2.4 Alternate definitions of Poisson processes . . . 82
2.2.5 The Poisson process as a limit of shrinking Bernoulli processes . . . 84
2.3 Combining and splitting Poisson processes . . . 86
2.3.1 Subdividing a Poisson process . . . 88
2.3.2 Examples using independent Poisson processes . . . 89
2.4 Non-homogeneous Poisson processes . . . 91
2.5 Conditional arrival densities and order statistics . . . 94
2.6 Summary . . . 98
2.7 Exercises . . . 100

3 GAUSSIAN RANDOM VECTORS AND PROCESSES 109

3.1 Introduction 109

3.2 Gaussian random variables 109

3.3 Gaussian random vectors 111

3.3.1 Generating functions of Gaussian random vectors 112

3.3.2 IID normalized Gaussian random vectors 112

3.3.3 Jointly-Gaussian random vectors 113

3.3.4 Joint probability density for Gaussian n-r**v**'s (special case) 116

3.4 Properties of covariance matrices 118

3.4.1 Symmetric matrices 118

3.4.2 Positive definite matrices and covariance matrices 119

3.4.3 Joint probability density for Gaussian n-r**v**'s (general case) 121

3.4.4 Geometry and principal axes for Gaussian densities 123

3.5 Conditional PDF's for Gaussian random vectors 125

3.6 Gaussian processes 129

3.6.1 Stationarity and related concepts: 131

3.6.2 Orthonormal expansions 132

3.6.3 Continuous-time Gaussian processes 135

3.6.4 Gaussian sinc processes 136

3.6.5 Filtered Gaussian sinc processes 139

3.6.6 Filtered continuous-time stochastic processes 141

3.6.7 Interpretation of spectral density and covariance 142

3.6.8 White Gaussian noise 144

3.6.9 The Wiener process / Brownian motion 146

3.7 Circularly-symmetric complex random vectors 149

3.7.1 Circular symmetry and complex Gaussian rv's 149

3.7.2 Covariance and pseudo-covariance of complex n-r**v**'s 150

3.7.3 Covariance matrices of complex n-r**v** 152

3.7.4 Linear transformations of $\boldsymbol{W} \sim \mathcal{CN}(0, [I_\ell])$ 153

3.7.5 Linear transformations of $\boldsymbol{Z} \sim \mathcal{CN}(0, [K])$ 154

3.7.6 The PDF of circularly-symmetric Gaussian n-r**v**'s 155

3.7.7 Conditional PDF's for circularly symmetric Gaussian r**v**'s 157
3.7.8 Circularly-symmetric Gaussian processes 158
3.8 Summary 160
3.9 Exercises 161

4 FINITE-STATE MARKOV CHAINS 167
4.1 Introduction 167
4.2 Classification of states 169
4.3 The matrix representation 173
4.3.1 Steady state and $[P^n]$ for large n 174
4.3.2 Steady state assuming $[P] > 0$ 177
4.3.3 Ergodic Markov chains 178
4.3.4 Ergodic unichains 180
4.3.5 Arbitrary finite-state Markov chains 182
4.4 The eigenvalues and eigenvectors of stochastic matrices 182
4.4.1 Eigenvalues and eigenvectors for $\mathsf{M} = 2$ states 183
4.4.2 Eigenvalues and eigenvectors for $\mathsf{M} > 2$ states 184
4.5 Markov chains with rewards 187
4.5.1 Expected first-passage times 187
4.5.2 The expected aggregate reward over multiple transitions 191
4.5.3 The expected aggregate reward with an additional final reward . . . 194
4.6 Markov decision theory and dynamic programming 196
4.6.1 Dynamic programming algorithm 197
4.6.2 Optimal stationary policies 201
4.6.3 Policy improvement and the seach for optimal stationary policies . . 204
4.7 Summary 208
4.8 Exercises 210

5 RENEWAL PROCESSES 223
5.1 Introduction 223
5.2 The strong law of large numbers and convergence WP1 226
5.2.1 Convergence with probability 1 (WP1) 226

5.2.2 Strong law of large numbers (SLLN) 228
5.3 Strong law for renewal processes 230
5.4 Renewal-reward processes; time averages 235
5.4.1 General renewal-reward processes 238
5.5 Stopping times for repeated experiments 242
5.5.1 Wald's equality 244
5.5.2 Applying Wald's equality to $\mathsf{E}[N(t)]$ 247
5.5.3 Generalized stopping trials, embedded renewals, and G/G/1 queues 247
5.5.4 Little's theorem 251
5.5.5 M/G/1 queues 254
5.6 Expected number of renewals; ensemble averages 257
5.6.1 Laplace transform approach 259
5.6.2 The elementary renewal theorem 260
5.7 Renewal-reward processes; ensemble averages 262
5.7.1 Age and duration for arithmetic processes 264
5.7.2 Joint age and duration: non-arithmetic case 267
5.7.3 Age $Z(t)$ for finite t: non-arithmetic case 269
5.7.4 Age $Z(t)$ as $t \to \infty$; non-arithmetic case 271
5.7.5 Arbitrary renewal-reward functions: non-arithmetic case 273
5.8 Delayed renewal processes 276
5.8.1 Delayed renewal-reward processes 278
5.8.2 Transient behavior of delayed renewal processes 278
5.8.3 The equilibrium process 279
5.9 Summary 280
5.10 Exercises 282

6 COUNTABLE-STATE MARKOV CHAINS 300

6.1 Introductory examples 300
6.2 First passage times and recurrent states 303
6.3 Renewal theory applied to Markov chains 307
6.3.1 Renewal theory and positive recurrence 308

6.3.2 Steady state . . . 310
6.3.3 Blackwell's theorem applied to Markov chains . . . 313
6.3.4 Age of a renewal process . . . 314
6.4 Birth-death Markov chains . . . 315
6.5 Reversible Markov chains . . . 317
6.6 The M/M/1 sample-time Markov chain . . . 320
6.7 Branching processes . . . 323
6.8 Round-robin and processor sharing . . . 326
6.9 Summary . . . 331
6.10 Exercises . . . 333

7 MARKOV PROCESSES WITH COUNTABLE STATE SPACES 338

7.1 Introduction . . . 338
7.1.1 The sampled-time approximation to a Markov process . . . 342
7.2 Steady-state behavior of irreducible Markov processes . . . 343
7.2.1 Renewals on successive entries to a given state . . . 345
7.2.2 The limiting fraction of time in each state . . . 345
7.2.3 Finding $\{p_j(i);\, j \geq 0\}$ in terms of $\{\pi_j;\, j \geq 0\}$ . . . 347
7.2.4 Solving for the steady-state process probabilities directly . . . 349
7.2.5 The sampled-time approximation again . . . 350
7.2.6 Pathological cases . . . 351
7.3 The Kolmogorov differential equations . . . 351
7.4 Uniformization . . . 355
7.5 Birth-death processes . . . 356
7.5.1 The M/M/1 queue again . . . 357
7.5.2 Other birth/death systems . . . 358
7.6 Reversibility for Markov processes . . . 358
7.7 Jackson networks . . . 364
7.7.1 Closed Jackson networks . . . 370
7.8 Semi-Markov processes . . . 371
7.8.1 Example — the M/G/1 queue . . . 375
7.9 Summary . . . 376
7.10 Exercises . . . 378

8 DETECTION, DECISIONS, AND HYPOTHESIS TESTING 391

8.1 Decision criteria and the MAP criterion 392

8.2 Binary MAP detection 395

8.2.1 Sufficient statistics I 397

8.2.2 Binary detection with a one-dimensional observation 398

8.2.3 Binary MAP detection with vector observations 402

8.2.4 Sufficient statistics II 407

8.3 Binary detection with a minimum-cost criterion 412

8.4 The error curve and the Neyman-Pearson rule 413

8.4.1 The Neyman-Pearson detection rule 419

8.4.2 The min-max detection rule 420

8.5 Finitely many hypotheses 420

8.5.1 Sufficient statistics with $\mathsf{M} \geq 2$ hypotheses 423

8.5.2 More general min-cost tests 425

8.6 Summary 426

8.7 Exercises 428

9 RANDOM WALKS, LARGE DEVIATIONS, AND MARTINGALES 435

9.1 Introduction 435

9.1.1 Simple random walks 436

9.1.2 Integer-valued random walks 437

9.1.3 Renewal processes as special cases of random walks 437

9.2 The queueing delay in a G/G/1 queue: 438

9.3 Threshold crossing probabilities in random walks 441

9.3.1 The Chernoff bound 441

9.3.2 Tilted probabilities 443

9.3.3 Large deviations and compositions 446

9.3.4 Back to threshold crossings 449

9.4 Thresholds, stopping rules, and Wald's identity 451

9.4.1 Wald's identity for two thresholds 452

9.4.2 The relationship of Wald's identity to Wald's equality 453

9.4.3 Zero-mean random walks 454
9.4.4 Exponential bounds on the probability of threshold crossing 455
9.5 Binary hypotheses with IID observations 456
9.5.1 Binary hypotheses with a fixed number of observations 457
9.5.2 Sequential decisions for binary hypotheses 460
9.6 Martingales 462
9.6.1 Simple examples of martingales 463
9.6.2 Scaled branching processes 465
9.6.3 Partial isolation of past and future in martingales 465
9.7 Submartingales and supermartingales 466
9.8 Stopped processes and stopping trials 468
9.8.1 The Wald identity 471
9.9 The Kolmogorov inequalities 472
9.9.1 The strong law of large numbers (SLLN) 474
9.9.2 The martingale convergence theorem 475
9.10 A simple model for investments 477
9.10.1 Portfolios with constant fractional allocations 480
9.10.2 Portfolios with time-varying allocations 485
9.11 Markov modulated random walks 487
9.11.1 Generating functions for Markov random walks 489
9.11.2 Stopping trials for martingales relative to a process 490
9.11.3 Markov modulated random walks with thresholds 490
9.12 Summary 492
9.13 Exercises 494

10 ESTIMATION **510**

10.1 Introduction 510
10.1.1 The squared-cost function 511
10.1.2 Other cost functions 512
10.2 MMSE estimation for Gaussian random vectors 514
10.2.1 Scalar iterative estimation 517

10.2.2 Scalar Kalman filter 518
10.3 Linear least-squares error estimation 520
10.4 Filtered vector signal plus noise 522
10.4.1 Estimate of a single rv in IID vector noise 524
10.4.2 Estimate of a single rv in arbitrary vector noise 524
10.4.3 Vector iterative estimation 525
10.4.4 Vector Kalman filter 526
10.5 Estimation for circularly-symmetric Gaussian r**v**'s 528
10.6 The vector space of rv's; orthogonality 530
10.7 MAP estimation and sufficient statistics 535
10.8 Parameter estimation 537
10.8.1 Fisher information and the Cramer-Rao bound 539
10.8.2 Vector observations 542
10.8.3 Information 543
10.9 Summary 545
10.10Exercises 547

Chapter 1

INTRODUCTION AND REVIEW OF PROBABILITY

1.1 Probability models

Stochastic processes is a branch of probability theory treating probabilistic systems tha t evolve in time. There seems to be no very good reason for trying to define stochastic processes precisely, but as we hope will become evident in this chapter, there is a very good reason for trying to be precise about probability itself. Those particular topics in which evolution in time is important will then unfold naturally. Section 1.5 gives a brief introduction to one of the very simplest stochastic processes, the Bernoulli process, and then Chapters 2, 3, and 4 develop three basic stochastic process models which serve as simple examples and starting points for the other processes to be discussed later.

Probability theory is a central field of mathematics, widely applicable to scientific, technological, and human situations involving uncertainty. The most obvious applications are to situations, such as games of chance, in which repeated trials of essentially the same procedure lead to differing outcomes. For example, when we flip a coin, roll a die, pick a card from a shuffled deck, or spin a ball onto a roulette wheel, the procedure is the same from one trial to the next, but the outcome (heads (H) or tails (T) in the case of a coin, one to six in the case of a die, etc.) varies from one trial to another in a seemingly random fashion.

For the case of flipping a coin, the outcome of the flip could be predicted from the initial position, velocity, and angular momentum of the coin and from the nature of the surface on which it lands. Thus, in one sense, a coin flip is deterministic rather than random and the same can be said for the other examples above. When these initial conditions are unspecified, however, as when playing these games, the outcome can again be viewed as random in some intuitive sense.

Many scientific experiments are similar to games of chance in the sense that multiple trials of apparently the *same* procedure lead to results that *vary* from one trial to another. In some cases, this variation is due to slight variations in the experimental procedure, in some it is due to noise, and in some, such as in quantum mechanics, the randomness is generally

believed to be fundamental. Similar situations occur in many types of systems, especially those in which noise and random delays are important. Some of these systems, rather than being repetitions of a common basic procedure, are systems that evolve over time while still containing a sequence of underlying similar random occurrences.

This intuitive notion of randomness, as described above, is a very special kind of uncertainty. Rather than involving a lack of understanding, it involves a type of uncertainty that can lead to probabilistic models with precise results. As in any scientific field, the models might or might not correspond to reality very well, but when they do correspond to reality, there is the sense that the situation is completely understood, while still being random.

For example, we all feel that we understand flipping a coin or rolling a die, but still accept randomness in each outcome. The theory of probability was initially developed particularly to give precise and quantitative understanding to these types of situations. The remainder of this section introduces this relationship between the precise view of probability theory and the intuitive view as used in applications and everyday language.

After this introduction, the following sections of this chapter review probability theory as a mathematical discipline, with a special emphasis on the laws of large numbers. In the final section, we use the theory and the laws of large numbers to obtain a fuller understanding of the relationship between theory and the real world.[1]

Probability theory, as a mathematical discipline, started to evolve in the 17th century and was initially focused on games of chance. The importance of the theory grew rapidly, particularly in the 20th century, and it now plays a central role in risk assessment, statistics, data networks, operations research, information theory, control theory, theoretical computer science, quantum theory, game theory, neurophysiology, and many other fields.

The core concept in probability theory is that of a *probability model*. Given the extent of the theory, both in mathematics and in applications, the simplicity of probability models is surprising. The first component of a probability model is a *sample space*, which is a *set* whose elements are called *sample points* or *outcomes*. Probability models are particularly simple in the special case where the sample space is finite, and we consider only this case in the remainder of this section. The second component of a probability model is a class of *events*, which can be considered for now simply as the class of all subsets of the sample space. The third component is a *probability measure*, which can be regarded for now as the assignment of a nonnegative number to each outcome, with the restriction that these numbers must sum to one over the sample space. The probability of an event is the sum of the probabilities of the outcomes comprising that event.

These probability models play a dual role. In the first, the many known results about various classes of models, and the many known relationships between models, constitute the essence of probability theory. Thus one often studies a model not because of any relationship to the real world, but simply because the model provides a building block or example useful for

[1]It would be appealing to show how probability theory evolved from real-world random situations, but probability theory, like most mathematical theories, has evolved from complex interactions between theoretical developments and initially over-simplified models of real situations. The successes and flaws of such models lead to refinements of the models and the theory, which in turn suggest applications to totally different fields.

the theory and thus ultimately useful for other models. In the other role, when probability theory is applied to some game, experiment, or some other situation involving randomness, a probability model is used to represent the experiment (in what follows, we refer to all of these random situations as experiments).

For example, the standard probability model for rolling a die uses $\{1, 2, 3, 4, 5, 6\}$ as the sample space, with each possible outcome having probability 1/6. An *odd* result, *i.e.*, the subset $\{1, 3, 5\}$, is an example of an event in this sample space, and this event has probability 1/2. The correspondence between model and actual experiment seems straightforward here. Both have the same set of outcomes and, given the symmetry between faces of the die, the choice of equal probabilities seems natural. Closer inspection, however, reveals an important difference between the model and the actual rolling of a die.

The model above corresponds to a single roll of a die, with a probability defined for each possible outcome. In a real-world experiment where a single die is rolled, one of the six faces, say face k comes up, but there is no *observable* probability for k.

Our intuitive notion of rolling dice, however, involves an experiment with n consecutive rolls of a die. There are then 6^n possible outcomes, one for each possible n-tuple of individual die outcomes. As reviewed in subsequent sections, the standard probability model for this repeated-roll experiment is to assign probability 6^{-n} to each possible n-tuple, which leads to a probability $\binom{n}{m}(1/6)^m(5/6)^{n-m}$ that the face k comes up on m of the n rolls, *i.e.*, that the relative frequency of face k is m/n. The distribution of these relative frequencies is increasingly clustered around 1/6 as n is increased. Thus if a real-world experiment for tossing n dice is reasonably modeled by this probability model, we would also expect the relative frequency to be close to 1/6 for large n. This relationship through relative frequencies in a repeated experiment helps overcome the non-observable nature of probabilities in the real world.

1.1.1 The sample space of a probability model

An *outcome* or *sample point* in a probability model corresponds to a complete result (with all detail specified) of the experiment being modeled. For example, a game of cards is often appropriately modeled by the arrangement of cards within a shuffled 52 card deck, thus giving rise to a set of 52! outcomes (incredibly detailed, but trivially simple in structure), even though the entire deck might not be played in one trial of the game. A poker hand with 4 aces is an *event* rather than an *outcome* in this model, since many arrangements of the cards can give rise to 4 aces in a given hand. The possible outcomes in a probability model (and in the experiment being modeled) are mutually exclusive and collectively constitute the entire sample space (space of possible outcomes). An outcome ω is often called a *finest grain* result of the model in the sense that a singleton event $\{\omega\}$ containing only ω clearly contains no proper subsets. Thus events (other than singleton events) typically give only partial information about the result of the experiment, whereas an outcome fully specifies the result.

In choosing the sample space for a probability model of an experiment, we often omit details that appear irrelevant for the purpose at hand. Thus in modeling the set of outcomes for a

coin toss as $\{H, T\}$, we ignore the type of coin, the initial velocity and angular momentum of the toss, etc. We also omit the rare possibility that the coin comes to rest on its edge. Sometimes, conversely, the sample space is enlarged beyond what is relevant in the interest of structural simplicity. An example is the above use of a shuffled deck of 52 cards.

The choice of the sample space in a probability model is similar to the choice of a mathematical model in any branch of science. That is, one simplifies the physical situation by eliminating detail of little apparent relevance. One often does this in an iterative way, using a very simple model to acquire initial understanding, and then successively choosing more detailed models based on the understanding from earlier models.

The mathematical theory of probability views the sample space simply as an abstract set of elements, and from a strictly mathematical point of view, the idea of doing an experiment and getting an outcome is a distraction. For visualizing the correspondence between the theory and applications, however, it is better to view the abstract set of elements as the set of possible outcomes of an idealized experiment in which, when the idealized experiment is performed, one and only one of those outcomes occurs. The two views are mathematically identical, but it will be helpful to refer to the first view as a probability model and the second as an idealized experiment. In applied probability texts and technical articles, these idealized experiments, rather than real-world situations, are often the primary topic of discussion.[2]

1.1.2 Assigning probabilities for finite sample spaces

The word *probability* is widely used in everyday language, and most of us attach various intuitive meanings[3] to the word. For example, everyone would agree that something virtually impossible should be assigned a probability close to 0 and something virtually certain should be assigned a probability close to 1. For these special cases, this provides a good rationale for choosing probabilities. The meaning of *virtually* and *close to* are slightly unclear at the moment, but if there is some implied limiting process, we would all agree that, in the limit, certainty and impossibility correspond to probabilities 1 and 0 respectively.

Between virtual impossibility and certainty, if one outcome appears to be closer to certainty than another, its probability should be correspondingly greater. This intuitive notion is imprecise and highly subjective; it provides little rationale for choosing numerical probabilities for different outcomes, and, even worse, little rationale justifying that probability models bear any precise relation to real-world situations.

Symmetry can often provide a better rationale for choosing probabilities. For example, the symmetry between H and T for a coin, or the symmetry between the the six faces of a die, motivates assigning equal probabilities, $1/2$ each for H and T and $1/6$ each for the six faces

[2]This is not intended as criticism, since we will see that there are good reasons to concentrate initially on such idealized experiments. However, readers should always be aware that modeling errors are the major cause of misleading results in applications of probability, and thus modeling must be seriously considered before using the results.

[3]It is popular to try to define probability by likelihood, but this is unhelpful since the words are essentially synonyms.

of a die. This is reasonable and extremely useful, but there is no completely convincing reason for choosing probabilities based on symmetry.

Another approach is to perform the experiment many times and choose the probability of each outcome as the relative frequency of that outcome (*i.e.*, the number of occurrences of that outcome divided by the total number of trials). Experience shows that the relative frequency of an outcome often approaches a limiting value with an increasing number of trials. Associating the probability of an outcome with that limiting relative frequency is certainly close to our intuition and also appears to provide a testable criterion between model and real world. This criterion is discussed in Sections 1.8.1 and 1.8.2 and provides a very concrete way to use probabilities, since it suggests that the randomness in a single trial tends to disappear in the aggregate of many trials. Other approaches to choosing probability models will be discussed later.

1.2 The axioms of probability theory

As the applications of probability theory became increasingly varied and complex during the 20th century, the need arose to put the theory on a firm mathematical footing. This was accomplished by an axiomatization of the theory, successfully carried out by the great Russian mathematician A. N. Kolmogorov [17] in 1932. Before stating and explaining these axioms of probability theory, the following two examples explain why the simple approach of the last section, assigning a probability to each sample point, often fails with infinite sample spaces.

Example 1.2.1. Suppose we want to model the phase of a sine wave, where the phase is viewed as being "uniformly distributed" between 0 and 2π. If this phase is the only quantity of interest, it is reasonable to choose a sample space consisting of the set of real numbers between 0 and 2π. There are uncountably[4] many possible phases between 0 and 2π, and with any reasonable interpretation of uniform distribution, one must conclude that each sample point has probability zero. Thus, the simple approach of the last section leads us to conclude that any event in this space with a finite or countably infinite set of sample points should have probability zero. That simple approach does not help in finding the probability, say, of the interval $(0, \pi)$.

For this example, the appropriate view is the one taken in all elementary probability texts, namely to assign a *probability density* $\frac{1}{2\pi}$ to the phase. The probability of an event can then usually be found by integrating the density over that event. Useful as densities are, however, they do not lead to a general approach over arbitrary sample spaces.[5]

[4] A set is uncountably infinite if it is infinite and its members cannot be put into one-to-one correspondence with the positive integers. For example the set of real numbers over some interval such as $(0, 2\pi)$ is uncountably infinite. The Wikipedia article on countable sets provides a friendly introduction to the concepts of countability and uncountability.

[5] It is possible to avoid the consideration of infinite sample spaces here by quantizing the possible phases. This is analogous to avoiding calculus by working only with discrete functions. Both usually result in both artificiality and added complexity.

Example 1.2.2. Consider an infinite sequence of coin tosses. The usual probability model is to assign probability 2^{-n} to each possible initial n-tuple of individual outcomes. Then in the limit $n \to \infty$, the probability of any given sequence is 0. Again, expressing the probability of an event involving infinitely many tosses as a sum of individual sample-point probabilities does not work. The obvious approach (which we often adopt for this and similar situations) is to evaluate the probability of any given event as an appropriate limit, as $n \to \infty$, of the outcome from the first n tosses.

We will later find a number of situations, even for this almost trivial example, where working with a finite number of elementary experiments and then going to the limit is very awkward. One example, to be discussed in detail later, is the strong law of large numbers (SLLN). This law looks directly at events consisting of infinite length sequences and is best considered in the context of the axioms to follow.

Although appropriate probability models can be generated for simple examples such as those above, there is a need for a consistent and general approach. In such an approach, rather than assigning probabilities to sample points, which are then used to assign probabilities to events, *probabilities must be associated directly with events.* The axioms to follow establish consistency requirements between the probabilities of different events. The axioms, and the corollaries derived from them, are consistent with one's intuition, and, for finite sample spaces, are consistent with our earlier approach. Dealing with the countable unions of events in the axioms will be unfamiliar to some students, but will soon become both familiar and consistent with intuition.

The strange part of the axioms comes from the fact that defining the class of events as the collection of *all* subsets of the sample space is usually inappropriate when the sample space is uncountably infinite. What is needed is a class of events that is large enough that we can almost forget that some very strange subsets are excluded. This is accomplished by having two simple sets of axioms, one defining the class of events,[6] and the other defining the relations between the probabilities assigned to these events. In this theory, all events have probabilities, but those truly weird subsets that are not events do not have probabilities. This will be discussed more after giving the axioms for events.

The axioms for events use the standard notation of set theory. Let Ω be the sample space, *i.e.*, the set of all sample points for a given experiment. It is assumed throughout that Ω is nonempty. The events are subsets of the sample space. The union of n subsets (events) $A_1, A_2, \cdots, A_n$ is denoted by either $\bigcup_{i=1}^n A_i$ or $A_1 \bigcup \cdots \bigcup A_n$, and consists of all points in at least one of $A_1, A_2 \ldots, A_n$. Similarly, the intersection of these subsets is denoted by either $\bigcap_{i=1}^n A_i$ or[7] $A_1 A_2 \cdots A_n$ and consists of all points in all of $A_1, A_2 \ldots, A_n$.

A *sequence* of events is a collection of events in one-to-one correspondence with the positive integers, *i.e.*, $A_1, A_2, \ldots$ ad infinitum. A countable union, $\bigcup_{i=1}^\infty A_i$ is the set of points in one or more of $A_1, A_2, \ldots$. Similarly, a countable intersection $\bigcap_{i=1}^\infty A_i$ is the set of points in all of $A_1, A_2, \ldots$. Finally, the complement A^c of a subset (event) A is the set of points in Ω but not A.

[6] A class of elements satisfying these axioms is called a σ-algebra or, less commonly, a σ-field.

[7] Intersection is also sometimes denoted as $A_1 \bigcap \cdots \bigcap A_n$, but is usually abbreviated as $A_1 A_2 \cdots A_n$.

1.2.1 Axioms for events

Given a sample space Ω, the class of subsets of Ω that constitute the set of events satisfies the following axioms:

1. Ω is an event.
2. For every sequence of events $A_1, A_2, \ldots$, the union $\bigcup_{n=1}^{\infty} A_n$ is an event.
3. For every event A, the complement A^c is an event.

There are a number of important corollaries of these axioms. First, the empty set $\emptyset$ is an event. This follows from Axioms 1 and 3, since $\emptyset = \Omega^c$. The empty set does not correspond to our intuition about events, but the theory would be extremely awkward if it were omitted. Second, every finite union of events is an event. This follows by expressing $A_1 \bigcup \cdots \bigcup A_n$ as $\bigcup_{i=1}^{\infty} A_i$ where $A_i = \emptyset$ for all $i > n$. Third, every finite or countable intersection of events is an event. This follows from De Morgan's law,

$$\left[\bigcup\nolimits_n A_n\right]^c = \bigcap\nolimits_n A_n^c.$$

Although we will not make a big fuss about these axioms in the rest of the text, we will be careful to use only complements and countable unions and intersections in our analysis. Thus subsets that are not events will not arise.

Note that the axioms do not say that all subsets of Ω are events. In fact, there are many rather silly ways to define classes of events that obey the axioms. For example, the axioms are satisfied by choosing only the universal set Ω and the empty set $\emptyset$ to be events. We shall avoid such trivialities by assuming that for each sample point ω, the singleton subset $\{\omega\}$ is an event. For finite sample spaces, this assumption, plus the axioms above, imply that all subsets are events.

For uncountably infinite sample spaces, such as the sinusoidal phase above, this assumption, plus the axioms above, still leaves considerable freedom in choosing a class of events. As an example, the class of all subsets of Ω satisfies the axioms but surprisingly does not allow the probability axioms to be satisfied in any sensible way. How to choose an appropriate class of events requires an understanding of measure theory which would take us too far afield for our purposes. Thus we neither assume nor develop measure theory here.[8]

From a pragmatic standpoint, we start with the class of events of interest, such as those required to define the random variables needed in the problem. That class is then extended so as to be closed under complementation and countable unions. Measure theory shows that this extension is possible.

[8]There is no doubt that measure theory is useful in probability theory, and serious students of probability should certainly learn measure theory at some point. For application-oriented people, however, it seems advisable to acquire more insight and understanding of probability, at a graduate level, before concentrating on the abstractions and subtleties of measure theory.

1.2.2 Axioms of probability

Given any sample space Ω and any class of events $\mathcal{E}$ satisfying the axioms of events, a probability rule is a function $\Pr\{\cdot\}$ mapping each $A \in \mathcal{E}$ to a (finite[9]) real number in such a way that the following three probability axioms[10] hold:

1. $\Pr\{\Omega\} = 1$.
2. For every event A, $\Pr\{A\} \geq 0$.
3. The probability of the union of any sequence $A_1, A_2, \ldots$ of disjoint[11] events is given by

$$\Pr\Big\{\bigcup_{n=1}^{\infty} A_n\Big\} = \sum_{n=1}^{\infty} \Pr\{A_n\}\,, \tag{1.1}$$

   where $\sum_{n=1}^{\infty} \Pr\{A_n\}$ is shorthand for $\lim_{m\to\infty} \sum_{n=1}^{m} \Pr\{A_n\}$.

The axioms imply the following useful corollaries:

$$\begin{aligned}
\Pr\{\emptyset\} &= 0 && && (1.2)\\
\Pr\Big\{\bigcup_{n=1}^{m} A_n\Big\} &= \sum_{n=1}^{m} \Pr\{A_n\} && \text{for } A_1, \ldots, A_m \text{ disjoint} && (1.3)\\
\Pr\{A^c\} &= 1 - \Pr\{A\} && \text{for all } A && (1.4)\\
\Pr\{A\} &\leq \Pr\{B\} && \text{for all } A \subseteq B && (1.5)\\
\Pr\{A\} &\leq 1 && \text{for all } A && (1.6)\\
\sum_n \Pr\{A_n\} &\leq 1 && \text{for } A_1, \ldots \text{ disjoint} && (1.7)\\
\Pr\Big\{\bigcup_{n=1}^{\infty} A_n\Big\} &= \lim_{m\to\infty} \Pr\Big\{\bigcup_{n=1}^{m} A_n\Big\} && && (1.8)\\
\Pr\Big\{\bigcup_{n=1}^{\infty} A_n\Big\} &= \lim_{n\to\infty} \Pr\{A_n\} && \text{for } A_1 \subseteq A_2 \subseteq \cdots && (1.9)\\
\Pr\Big\{\bigcap_{n=1}^{\infty} A_n\Big\} &= \lim_{n\to\infty} \Pr\{A_n\} && \text{for } A_1 \supseteq A_2 \supseteq \cdots\,. && (1.10)
\end{aligned}$$

To verify (1.2), consider a sequence of events, $A_1, A_2, \ldots$ for which $A_n = \emptyset$ for each n. These events are disjoint since $\emptyset$ contains no outcomes, and thus has no outcomes in common with itself or any other event. Also, $\bigcup_n A_n = \emptyset$ since this union contains no outcomes. Axiom 3 then says that

$$\Pr\{\emptyset\} = \lim_{m\to\infty} \sum_{n=1}^{m} \Pr\{A_n\} = \lim_{m\to\infty} m\Pr\{\emptyset\}\,.$$

Since $\Pr\{\emptyset\}$ is a real number, this implies that $\Pr\{\emptyset\} = 0$.

[9] The word *finite* is redundant here, since the set of real numbers, by definition, does not include $\pm\infty$. The set of real numbers with $\pm\infty$ appended, is called the *extended* set of real numbers

[10] Sometimes finite additivity, (1.3), is added as an additional axiom. This addition is quite intuitive and avoids the technical and somewhat peculiar proofs given for (1.2) and (1.3).

[11] Two sets or events A_1, A_2 are disjoint if they contain no common events, *i.e.*, if $A_1 A_2 = \emptyset$. A collection of sets or events are disjoint if all pairs are disjoint.

To verify (1.3), apply Axiom 3 to the disjoint sequence $A_1, \ldots, A_m, \emptyset, \emptyset, \ldots$.

To verify (1.4), note that $\Omega = A \bigcup A^c$. Then apply (1.3) to the disjoint sets A and A^c.

To verify (1.5), note that if $A \subseteq B$, then $B = A \bigcup (B - A)$ where $B - A$ is an alternate way to write $B \bigcap A^c$. We see then that A and $B - A$ are disjoint, so from (1.3),

$$\Pr\{B\} = \Pr\Big\{A \bigcup (B - A)\Big\} = \Pr\{A\} + \Pr\{B - A\} \geq \Pr\{A\},$$

where we have used Axiom 2 in the last step.

To verify (1.6) and (1.7), first substitute Ω for B in (1.5) and then substitute $\bigcup_n A_n$ for A.

Finally, (1.8) is established in Exercise 1.2, (e), and (1.9) and (1.10) are simple consequences of (1.8).

The axioms specify the probability of any *disjoint* union of events in terms of the individual event probabilities, but what about a finite or countable union of arbitrary events? Exercise 1.2 (c) shows that in this case, (1.3) can be generalized to

$$\Pr\Big\{\bigcup_{n=1}^{m} A_n\Big\} = \sum_{n=1}^{m} \Pr\{B_n\}, \tag{1.11}$$

where $B_1 = A_1$ and for each $n > 1$, $B_n = A_n - \bigcup_{m=1}^{n-1} A_m$ is the set of points in A_n but not in any of the sets $A_1, \ldots, A_{n-1}$. That is, the sets B_n are disjoint. The probability of a countable union of disjoint sets is then given by (1.8). In order to use this, one must know not only the event probabilities for $A_1, A_2 \ldots$, but also the probabilities of their intersections. The union bound, which is derived in Exercise 1.2 (d), depends only on the individual event probabilities, and gives the following frequently useful upper bound on the union probability.

$$\Pr\Big\{\bigcup_{n} A_n\Big\} \leq \sum_{n} \Pr\{A_n\} \qquad \text{(Union bound)}. \tag{1.12}$$

1.3 Probability review

1.3.1 Conditional probabilities and statistical independence

Definition 1.3.1. *For any two events A and B in a probability model, the conditional probability of A, conditional on B, is defined if $\Pr\{B\} > 0$ by*

$$\Pr\{A|B\} = \Pr\{AB\} / \Pr\{B\}. \tag{1.13}$$

To motivate this definition, consider a discrete experiment in which we make a partial observation B (such as the result of a given medical test on a patient) but do not observe the complete outcome (such as whether the patient is sick and the outcome of other tests). The event B consists of all the sample points with the given outcome of the given test. Now let A be an arbitrary event (such as the event that the patient is sick). The conditional

probability, $\Pr\{A|B\}$ is intended to represent the probability of A from the observer's viewpoint.

For the observer, the sample space can now be viewed as the set of sample points in B, since only those sample points are now possible. For any event A, only the event AB, *i.e.*, the original set of sample points in A that are also in B, is relevant, but the probability of A in this new sample space should be scaled up from $\Pr\{AB\}$ to $\Pr\{AB\}/\Pr\{B\}$, *i.e.*, to $\Pr\{A|B\}$.

With this scaling, the set of events conditional on B becomes a probability space, and it is easily verified that all the axioms of probability theory are satisfied for this conditional probability space. Thus all known results about probability can also be applied to such conditional probability spaces.

Another important aspect of the definition in (1.13) is that it maintains consistency between the original probability space and this new conditional space in the sense that for any events A_1, A_2 and any scalars α_1, α_2, we have

$$\Pr\{\alpha_1 A_1 + \alpha_2 A_2 \mid B\} = \alpha_1 \Pr\{A_1|B\} + \alpha_2 \Pr\{A_2|B\}\,.$$

This means that we can easily move back and forth between unconditional and conditional probability spaces.

The intuitive statements about partial observations and probabilities from the standpoint of an observer are helpful in reasoning probabilistically, but sometimes cause confusion. For example, Bayes law, in the form

$$\Pr\{A|B\}\Pr\{B\} = \Pr\{B|A\}\Pr\{A\}$$

is an immediate consequence of the definition of conditional probability in (1.13). However, if we can only interpret $\Pr\{A|B\}$ when B is 'observed' or occurs 'before' A, then we cannot interpret $\Pr\{B|A\}$ and $\Pr\{A|B\}$ together. This caused immense confusion in probabilistic arguments before the axiomatic theory and clean definitions based on axioms were developed.

Definition 1.3.2. *Two events, A and B, are statistically independent (or, more briefly,* independent*) if*

$$\Pr\{AB\} = \Pr\{A\}\Pr\{B\}\,.$$

For $\Pr\{B\} > 0$, this is equivalent to $\Pr\{A|B\} = \Pr\{A\}$. This latter form often corresponds to a more intuitive view of independence, since it says that A and B are independent if the observation of B does not change the observer's probability of A.

The notion of independence is of vital importance in defining, and reasoning about, probability models. We will see many examples where very complex systems become very simple, both in terms of intuition and analysis, when appropriate quantities are modeled as statistically independent. An example will be given in the next subsection where repeated independent experiments are used to understand arguments about relative frequencies.

Often, when the assumption of independence is unreasonable, it is reasonable to assume conditional independence, where A and B are said to be *conditionally independent* given C if $\Pr\{AB|C\} = \Pr\{A|C\}\Pr\{B|C\}$. Most of the stochastic processes to be studied here are characterized by various forms of independence or conditional independence.

For more than two events, the definition of statistical independence is a little more complicated.

Definition 1.3.3. *The events $A_1, \ldots, A_n$, $n > 2$ are statistically independent if for each collection S of two or more of the integers 1 to n.*

$$\Pr\Big\{\bigcap\nolimits_{i\in S} A_i\Big\} = \prod\nolimits_{i\in S} \Pr\{A_i\}\,. \tag{1.14}$$

This includes the entire collection $\{1, \ldots, n\}$, so one necessary condition for independence is that

$$\Pr\Big\{\bigcap\nolimits_{i=1}^{n} A_i\Big\} = \prod\nolimits_{i=1}^{n} \Pr\{A_i\}\,. \tag{1.15}$$

It might be surprising that (1.15) does not imply (1.14), but the example in Exercise 1.4 will help clarify this. This definition will become clearer (and simpler) when we see how to view independence of events as a special case of independence of random variables.

1.3.2 Repeated idealized experiments

Much of our intuitive understanding of probability comes from the notion of repeating the same idealized experiment many times (*i.e.*, performing multiple trials of the same experiment). However, the axioms of probability contain no explicit recognition of such repetitions. The appropriate way to handle n repetitions of an idealized experiment is through an extended experiment whose sample points are n-tuples of sample points from the original experiment. Such an extended experiment is viewed as n *trials* of the original experiment. The notion of multiple trials of a given experiment is so common that one sometimes fails to distinguish between the original experiment and an extended experiment with multiple trials of the original experiment.

To be more specific, given an original sample space Ω, the sample space of an n-repetition model is the Cartesian product

$$\Omega^n = \{(\omega_1, \omega_2, \ldots, \omega_n) \,:\, \omega_i \in \Omega \text{ for each } i, 1 \le i \le n\}, \tag{1.16}$$

i.e., the set of all n-tuples for which each of the n components of the n-tuple is an element of the original sample space Ω. Since each sample point in the n-repetition model is an n-tuple of points from the original Ω, it follows that an event in the n-repetition model is a subset of Ω^n, *i.e.*, a collection of n-tuples $(\omega_1, \ldots, \omega_n)$, where each ω_i is a sample point from Ω. This class of events in Ω^n should include each event of the form $\{(A_1 A_2 \cdots A_n)\}$, where $\{(A_1 A_2 \cdots A_n)\}$ denotes the collection of n-tuples $(\omega_1, \ldots, \omega_n)$ where $\omega_i \in A_i$ for $1 \le i \le n$. The set of events (for n-repetitions) must also be extended to be closed under complementation and countable unions and intersections.

The simplest and most natural way of creating a probability model for this extended sample space and class of events is through the assumption that the n-trials are statistically independent. More precisely, we assume that for each extended event $\{A_1 \times A_2 \times \cdots \times A_n\}$ contained in Ω^n, we have

$$\Pr\{A_1 \times A_2 \times \cdots \times A_n\} = \prod_{i=1}^{n} \Pr\{A_i\}, \tag{1.17}$$

where $\Pr\{A_i\}$ is the probability of event A_i in the original model. Note that since Ω can be substituted for any A_i in this formula, the subset condition of (1.14) is automatically satisfied. In fact, the Kolmogorov extension theorem asserts that for any probability model, there is an extended independent n-repetition model for which the events in each trial are independent of those in the other trials. In what follows, we refer to this as the probability model for n independent identically distributed (IID) trials of a given experiment.

The niceties of how to create this model for n IID arbitrary experiments depend on measure theory, but we simply rely on the existence of such a model and the independence of events in different repetitions. What we have done here is very important conceptually. A probability model for an experiment does not say anything directly about repeated experiments. However, questions about independent repeated experiments can be handled directly within this extended model of n IID repetitions. This can also be extended to a countable number of IID trials.

1.3.3 Random variables

The outcome of a probabilistic experiment often specifies a collection of numerical values such as temperatures, voltages, numbers of arrivals or departures in various time intervals, etc. Each such numerical value varies, depending on the particular outcome of the experiment, and thus can be viewed as a mapping from the set Ω of sample points to the set $\mathbb{R}$ of real numbers (note that $\mathbb{R}$ does not include $\pm\infty$). These mappings from sample points to real numbers are called random variables.

Definition 1.3.4. *A random variable (rv) is essentially a function X from the sample space Ω of a probability model to the set of real numbers $\mathbb{R}$. Three modifications are needed to make this precise. First, X might be undefined or $\pm\infty$ for a subset of Ω that has 0 probability.*[12] *Second, the mapping $X(\omega)$ must have the property that $\{\omega \in \Omega : X(\omega) \leq x\}$ is an event*[13] *for each $x \in \mathbb{R}$. Third, every finite set of rv's $X_1, \ldots, X_n$ has the property that for each $x_1 \in \mathbb{R}, \ldots, x_n \in \mathbb{R}$, the set $\{\omega : X_1(\omega) \leq x_1, \ldots, X_n(\omega) \leq x_n\}$ is an event .*

As with any function, there is often confusion between the function itself, which is called X in the definition above, and the value $X(\omega)$ taken on for a sample point ω. This is

[12] For example, consider a probability model in which Ω is the closed interval $[0, 1]$ and the probability is uniformly distributed over Ω. If $X(\omega) = 1/\omega$, then the sample point 0 maps to ∞ but X is still regarded as a rv. These subsets of 0 probability are usually ignored, both by engineers and mathematicians.

[13] These last two modifications are technical limitations connected with measure theory. They can usually be ignored, since they are satisfied in all but the most bizarre conditions. However, just as it is important to know that not all subsets in a probability space are events, one should know that not all functions from Ω to $\mathbb{R}$ are rv's.

particularly prevalent with random variables (rv's) since we intuitively associate a rv with its sample value when an experiment is performed. We try to control that confusion here by using X, $X(\omega)$, and x, respectively, to refer to the rv, the sample value taken for a given sample point ω, and a generic sample value.

Definition 1.3.5. *The cumulative distribution function (CDF)*[14] *of a random variable (rv) X is a function $\mathsf{F}_X(x)$ mapping $\mathbb{R} \to \mathbb{R}$ and defined by $\mathsf{F}_X(x) = \Pr\{\omega \in \Omega : X(\omega) \leq x\}$. The argument ω is usually omitted for brevity, so $\mathsf{F}_X(x) = \Pr\{X \leq x\}$.*

Note that x is the argument of $\mathsf{F}_X(x)$ and the subscript X denotes the particular rv under consideration. As illustrated in Figure 1.1, the CDF $\mathsf{F}_X(x)$ is non-decreasing with x and must satisfy $\lim_{x\to-\infty} \mathsf{F}_X(x) = 0$ and $\lim_{x\to\infty} \mathsf{F}_X(x) = 1$. Exercise 1.5 proves that $\mathsf{F}_X(x)$ is continuous from the right (*i.e.*, that for every $x \in \mathbb{R}$, $\lim_{\epsilon\downarrow 0} \mathsf{F}_X(x+\epsilon) = \mathsf{F}_X(x)$).

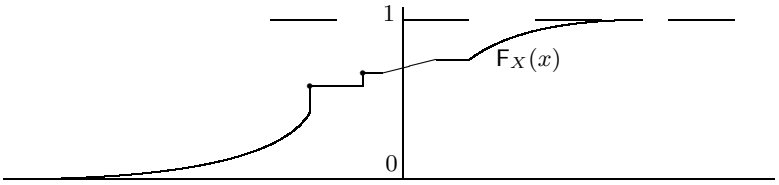


Figure 1.1: Example of a CDF for a rv that is neither continuous nor discrete. If $\mathsf{F}_X(x)$ has a discontinuity at some x_o, it means that there is a discrete probability at x_o equal to the magnitude of the discontinuity. In this case $\mathsf{F}_X(x_o)$ is given by the height of the upper point at the discontinuity.

Because of the definition of a rv, the set $\{X \leq x\}$ for any rv X and any real number x must be an event, and thus $\Pr\{X \leq x\}$ must be defined for all real x.

The concept of a rv is often extended to complex random variables (rv's) and vector rv's. A *complex random variable* is a mapping from the sample space to the set of finite complex numbers, and a *vector random variable (rv)* is a mapping from the sample space to the finite vectors in some finite-dimensional vector space. Another extension is that of defective rv's. A *defective* rv X is a mapping from the sample space to the extended real numbers, which satisfies the conditions of a rv except that the set of sample points mapped into $\pm\infty$ has positive probability.

When rv's are referred to (without any modifier such as complex, vector, or defective), the original definition, i.e., a function from Ω to $\mathbb{R}$, is intended.

If X has only a finite or countable number of possible sample values, say $x_1, x_2, \ldots$, the probability $\Pr\{X = x_i\}$ of each sample value x_i is called the probability mass function (PMF) at x_i and denoted by $\mathsf{p}_X(x_i)$; such a random variable is called *discrete*. The CDF of a discrete rv is a 'staircase function,' staying constant between the possible sample values and having a jump of magnitude $\mathsf{p}_X(x_i)$ at each sample value x_i. Thus the PMF and the CDF each specify the other for discrete rv's.

[14] The cumulative distribution function is often referred to simply as the distribution function. We will use the word distribution in a generic sense to refer to any probabilistic characterization from which the CDF, PMF, PDF can in principle be found.

If the CDF $\mathsf{F}_X(x)$ of a rv X has a (finite) derivative at x, the derivative is called the *probability density* (PDF) of X at x and denoted by $\mathsf{f}_X(x)$; for $\delta > 0$ sufficiently small, $\mathsf{f}_X(x)\delta$ then approximates the probability that X is mapped to a value between x and $x+\delta$. A rv is said to be *continuous* if there is a function $\mathsf{f}_X(x)$ such that, for each $x \in \mathbb{R}$, the CDF satisfies $\mathsf{F}_X(x) = \int_{-\infty}^{x} \mathsf{f}_X(y)\, dy$. If such a $\mathsf{f}_X(x)$ exists, it is called the PDF. Essentially this means that $\mathsf{f}_X(x)$ is the derivative of $\mathsf{F}_X(x)$, but it is slightly more general in that it permits $\mathsf{f}_X(x)$ to be discontinuous.

Elementary probability courses work primarily with the PMF and the PDF, since they are convenient for computational exercises. We will work more often with the CDF here. This is partly because it is always defined, partly to avoid saying everything thrice, for discrete, continuous, and other rv's, and partly because the CDF is often most important in limiting arguments such as steady-state time-average arguments. For CDF's, PDF's and PMF's, the subscript denoting the rv is often omitted if the rv is clear from the context. The same convention is used for complex or vector rv's.

The following tables list some widely used rv's. If the PDF or PMF is given only in a limited region, it is zero outside of that region. The moment generating function (MGF), of a rv X is defined as $\mathsf{E}\left[e^{rX}\right]$ and will be discussed in Section 1.5.5.

Name	**PDF** $\mathsf{f}_X(x)$	**Mean**	**Variance**	**MGF** $\mathsf{g}_X(r)$
Exponential:	$\lambda \exp(-\lambda x); \quad x{\geq}0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda - r}; \quad \text{for } r < \lambda$
Erlang:	$\frac{\lambda^n x^{n-1} \exp(-\lambda x)}{(n-1)!}; \quad x{\geq}0$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$	$\left(\frac{\lambda}{\lambda - r}\right)^n; \quad \text{for } r < \lambda$
Gaussian:	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-a)^2}{2\sigma^2}\right)$	a	σ^2	$\exp(ra + r^2\sigma^2/2)$
Uniform:	$\frac{1}{a}; \quad 0{\leq}x{\leq}a$	$\frac{a}{2}$	$\frac{a^2}{12}$	$\frac{\exp(ra)-1}{ra}$

Table 1.1: The PDF, mean, variance and MGF for some common continuous rv's.

Name	**PMF** $\mathsf{p}_M(m)$	**Mean**	**Variance**	**MGF** $\mathsf{g}_M(r)$
Binary:	$\mathsf{p}_M(1) = p;\ \mathsf{p}_M(0) = 1 - p$	p	$p(1-p)$	$1 - p + pe^r$
Binomial:	$\binom{n}{m} p^m (1-p)^{n-m};\ 0{\leq}m{\leq}n$	np	$np(1-p)$	$[1 - p + pe^r]^n$
Geometric:	$p(1-p)^{m-1};\ m{\geq}1$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^r}{1-(1-p)e^r};\ \text{for } r < \ln\frac{1}{1-p}$
Poisson:	$\frac{\lambda^n \exp(-\lambda)}{n!};\ n{\geq}0$	λ	λ	$\exp[\lambda(e^r - 1)]$

Table 1.2: The PMF, mean, variance and MGF for some common discrete rv's.

1.3.4 Multiple random variables and conditional probabilities

Often we must deal with multiple random variables (rv's) in a single probability experiment. If $X_1, X_2, \ldots, X_n$ are rv's or the components of a vector rv, their joint CDF is defined by

$$\mathsf{F}_{X_1 \cdots X_n}(x_1, \ldots, x_n) = \Pr\{\omega \in \Omega : X_1(\omega) \leq x_1, X_2(\omega) \leq x_2, \ldots, X_n(\omega) \leq x_n\}. \quad (1.18)$$

This definition goes a long way toward explaining why we need the notion of a sample space Ω when all we want to talk about is a set of rv's. The CDF of a rv fully describes the individual behavior of that rv (and gives rise to its name, as in Tables 1.1 and 1.2), but Ω and the above mappings are needed to describe how the rv's interact.

For a vector rv $\boldsymbol{X}$ with components $X_1, \ldots, X_n$, or a complex rv X with real and imaginary parts X_1, X_2, the CDF is also defined by (1.18). Note that $\{X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\}$ is an event and the corresponding probability is non-decreasing in each argument x_i. Also the CDF of any subset of random variables is obtained by setting the other arguments to $+\infty$. For example, the CDF of a single rv (called a *marginal* CDF for a given joint CDF) is given by

$$\mathsf{F}_{X_i}(x_i) = \mathsf{F}_{X_1 \cdots X_{i-1} X_i X_{i+1} \cdots X_n}(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty).$$

If the rv's are all discrete, there is a joint PMF which specifies and is specified by the joint CDF. It is given by

$$\mathsf{p}_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \Pr\{X_1 = x_1, \ldots, X_n = x_n\}.$$

Similarly, if the joint CDF can be differentiated as below, then it specifies and is specified by the joint PDF,

$$\mathsf{f}_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \frac{\partial^n \mathsf{F}(x_1, \ldots, x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n}.$$

Two rv's, say X and Y, are *statistically independent* (or, more briefly, *independent*) if

$$\mathsf{F}_{XY}(x, y) = \mathsf{F}_X(x)\mathsf{F}_Y(y) \qquad \text{for each } x \in \mathbb{R}, y \in \mathbb{R}. \quad (1.19)$$

If X and Y are discrete rv's, then the definition of independence in (1.19) is equivalent to the corresponding statement for PMF's,

$$\mathsf{p}_{XY}(x_i, y_j) = \mathsf{p}_X(x_i)\mathsf{p}_Y(y_j) \qquad \text{for each value } x_i \text{ of } X \text{ and } y_j \text{ of } Y.$$

Since $\{X = x_i\}$ and $\{Y = y_j\}$ are events, the conditional probability of $\{X = x_i\}$ conditional on $\{Y = y_j\}$ (assuming $\mathsf{p}_Y(y_j) > 0$) is given by (1.13) to be

$$\mathsf{p}_{X|Y}(x_i \mid y_j) = \frac{\mathsf{p}_{XY}(x_i, y_j)}{\mathsf{p}_Y(y_j)}.$$

If $\mathsf{p}_{X|Y}(x_i \mid y_j) = \mathsf{p}_X(x_i)$ for all i, j, then it is seen that X and Y are independent. This captures the intuitive notion of independence better than (1.19) for discrete rv's , since it can be viewed as saying that the PMF of X is not affected by the sample value of Y.

If X and Y have a joint density, then (1.19) is equivalent to

$$\mathsf{f}_{XY}(x,y) = \mathsf{f}_X(x)\mathsf{f}_Y(y) \qquad \text{for each } x \in \mathbb{R}, y \in \mathbb{R}. \tag{1.20}$$

If the joint density exists and the marginal density $\mathsf{f}_Y(y)$ is positive, the conditional density can be defined as $\mathsf{f}_{X|Y}(x|y) = \frac{\mathsf{f}_{XY}(x,y)}{\mathsf{f}_Y(y)}$. In essence, $\mathsf{f}_{X|Y}(x|y)$ is the density of X conditional on $Y = y$, but, being more precise, it is a limiting conditional density as $\delta \to 0$ of X conditional on $Y \in [y, y+\delta)$.

If X and Y have a joint density, then statistical independence can also be expressed as

$$\mathsf{f}_{X|Y}(x|y) = \mathsf{f}_X(x) \qquad \text{for each } x \in \mathbb{R},\ y \in \mathbb{R} \text{ such that } \mathsf{f}_Y(y) > 0. \tag{1.21}$$

This often captures the intuitive notion of statistical independence for continuous rv's better than (1.20)

More generally, the probability of an arbitrary event A, conditional on a given value of a continuous rv Y, is given by

$$\Pr\{A \mid Y = y\} = \lim_{\delta \to 0} \frac{\Pr\{A, Y \in [y, y+\delta]\}}{\Pr\{Y \in [y, y+\delta]\}}.$$

We next generalize the above results about two rv's to the case of n rv's $\boldsymbol{X} = X_1, \ldots, X_n$. Statistical independence is then defined by the equation

$$\mathsf{F}_{\boldsymbol{X}}(x_1, \ldots, x_n) = \prod_{i=1}^n \Pr\{X_i \le x_i\} = \prod_{i=1}^n \mathsf{F}_{X_i}(x_i) \qquad \text{for all } x_1, \ldots, x_n \in \mathbb{R}. \tag{1.22}$$

In other words, $X_1, \ldots, X_n$ are independent if the events $X_i \le x_i$ for $1 \le i \le n$ are independent for all choices of $x_1, \ldots, x_n$. If the density or PMF exists, (1.22) is equivalent to a product form for the density or mass function. A set of rv's is said to be pairwise independent if each pair of rv's in the set is independent. As shown in Exercise 1.23, pairwise independence does not imply that the entire set is independent.

Independent rv's are very often also identically distributed, *i.e.*, they all have the same CDF. These cases arise so often that we abbreviate independent identically distributed by IID. For the IID case (1.22) becomes

$$\mathsf{F}_{\boldsymbol{X}}(x_1, \ldots, x_n) = \prod_{i=1}^n \mathsf{F}_X(x_i).$$

1.4 Stochastic processes

A stochastic process (or random process[15]) is an infinite collection of rv's defined on a common probability model. These rv's are usually indexed by an integer or a real number

[15]Stochastic and random are synonyms, but *random* has become more popular for random variables and *stochastic* for stochastic processes. The reason for the author's choice is that the common-sense intuition associated with randomness appears more important than mathematical precision in reasoning about rv's, whereas for stochastic processes, common-sense intuition causes confusion much more frequently than with rv's. The less familiar word *stochastic* warns the reader to be more careful.

often interpreted as time. Thus each sample point of the probability model maps to an infinite collection of sample values of rv's. If the index is regarded as time, then each sample point maps to a function of time called a sample path or sample function. These sample paths might vary continuously with time or might vary only at discrete times, and if they vary at discrete times, those times might be deterministic or random.

In many cases, this collection of rv's comprising the stochastic process is the only thing of interest. In this case, the sample points of the probability model can be taken to be the sample paths of the process. Conceptually, then, each event is a collection of sample paths. Often the most important of these events can be defined in terms of a finite set of rv's.

As an example of sample paths that change at only discrete times, we might be concerned with the times at which customers arrive at some facility. These 'customers' might be customers entering a store, incoming jobs for a computer system, arriving packets to a communication system, or orders for a merchandising warehouse.

The Bernoulli process is an example of how such customers could be modeled and is perhaps the simplest non-trivial stochastic process. We now define this process and develop a few of its many properties. We will return to it frequently as an example.

1.4.1 The Bernoulli process

A *Bernoulli process* is a sequence, $Z_1, Z_2, \ldots$, of IID binary random variables.[16] Let $p = \Pr\{Z_i = 1\}$ and $q = 1-p = \Pr\{Z_i = 0\}$. We often visualize a Bernoulli process as evolving in discrete time with the event $\{Z_i = 1\}$ representing an arriving customer at time i and $\{Z_i = 0\}$ representing no arrival. Thus at most one arrival occurs at each integer time.

When viewed as arrivals in time, it is interesting to understand something about the intervals between successive arrivals and about the aggregate number of arrivals up to any given time (see Figure 1.2). By convention, the first interarrival time is the time of the first arrival. These interarrival times and aggregate numbers of arrivals are rv's that are functions of the underlying sequence $Z_1, Z_2, \ldots$. The topic of rv's that are defined as functions of other rv's (*i.e.*, whose sample values are functions of the sample values of the other rv's) is taken up in more generality in Section 1.5.1, but the interarrival times and aggregate arrivals for Bernoulli processes are so specialized and simple that it is better to treat them from first principles.

The first interarrival time, X_1, is 1 if and only if the first binary rv Z_1 is 1, and thus $\mathsf{p}_{X_1}(1) = p$. Next, $X_1 = 2$ if and only $Z_1 = 0$ and $Z_2 = 1$, so $\mathsf{p}_{X_1}(2) = p(1-p)$. Continuing, we see that X_1 has the *geometric* PMF,

$$\mathsf{p}_{X_1}(j) = p(1-p)^{j-1} \qquad \text{where } j \geq 1.$$

[16] We say that a sequence $Z_1, Z_2, \ldots$, of rv's are IID if for each integer n, the rv's $Z_1, \ldots, Z_n$ are IID. There are some subtleties in going to the limit $n \to \infty$, but we can avoid most such subtleties by working with finite n-tuples and going to the limit at the end.

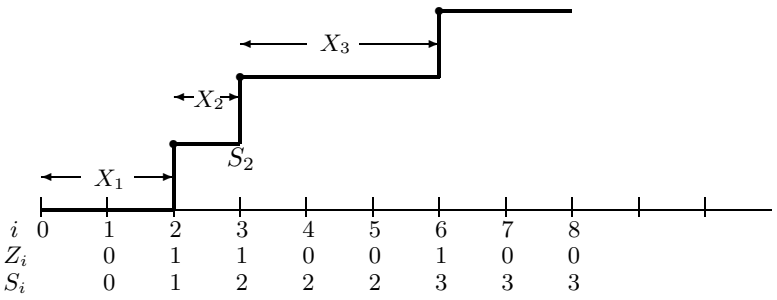


Figure 1.2: Illustration of a sample path for a Bernoulli process: The sample values of the binary rv's Z_i are shown below the time instants. The sample value of the aggregate number of arrivals, $S_n = \sum_{i=1}^{n} Z_i$, is the illustrated step function, and the interarrival intervals are the intervals between steps.

Each subsequent interarrival time, $X_2, X_3, \ldots$, can be found in this same way.[17] Each has the same geometric PMF and is statistically independent of $X_1, \ldots, X_{k-1}$. Thus the sequence of interarrival times is an IID sequence of geometric rv's.

It can be seen from Figure 1.2 that a sample path of interarrival times also determines a sample path of the binary arrival rv's, $\{Z_i;\ i \geq 1\}$. Thus the Bernoulli process can also be characterized in terms of a sequence of IID geometric rv's.

For our present purposes, the most important rv's in a Bernoulli process are the partial sums $S_n = \sum_{i=1}^{n} Z_i$. Each rv S_n is the number of arrivals up to and including time n, *i.e.*, S_n is simply the sum of n binary IID rv's and its distribution is known as the binomial distribution. That is, $\mathsf{p}_{S_n}(k)$ is the probability that k out of n of the Z_i's have the value 1. There are $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ arrangements of a binary n-tuple with k 1's, and each has probability $p^k q^{n-k}$. Thus

$$\mathsf{p}_{S_n}(k) = \binom{n}{k} p^k q^{n-k}. \tag{1.23}$$

We will use the binomial PMF extensively as an example in explaining the laws of large numbers and the central limit theorem later in this chapter, and will often use it in later chapters as an example of a sum of IID rv's. For these examples, we need to know how $\mathsf{p}_{S_n}(k)$ behaves asymptotically as $n \to \infty$ and $k \to \infty$. We make a short digression here to state and develop tight upper and lower bounds to the binomial PMF that makes this asymptotic behavior clear. The ratio of k to n in these bounds is denoted $\tilde{p} = k/n$ and the bounds are given in terms of a quantity $D(\tilde{p}\|p)$ called the binary Kullback-Liebler divergence (or the binary relative entropy) and defined by

$$\mathsf{D}(\tilde{p}\|p) = \tilde{p}\ln(\frac{\tilde{p}}{p}) + (1-\tilde{p})\ln(\frac{1-\tilde{p}}{1-p}) \ \geq\ 0. \tag{1.24}$$

[17] This is one of those maddening arguments that, while intuitively obvious, requires some careful reasoning to be completely convincing. We go through several similar arguments with great care in Chapter 2, and suggest that skeptical readers wait until then to think this through carefully .

Theorem 1.4.1. *Let $\mathsf{p}_{S_n}(k)$ be the PMF of the binomial distribution for an underlying binary PMF $\mathsf{p}_Z(1) = p > 0$, $\mathsf{p}_Z(0) = 1 - p > 0$ and let $\tilde{p} = k/n$. Then for each integer $\tilde{p}n$, $1 \leq \tilde{p}n \leq n-1$,*

$$\mathsf{p}_{S_n}(\tilde{p}n) < \sqrt{\frac{1}{2\pi n\tilde{p}(1-\tilde{p})}}\, \exp\big[-n\mathsf{D}(\tilde{p}\|p)\big] \tag{1.25}$$

$$\mathsf{p}_{S_n}(\tilde{p}n) > \left(1 - \frac{1}{12n\tilde{p}(1-\tilde{p})}\right) \sqrt{\frac{1}{2\pi n\tilde{p}(1-\tilde{p})}}\, \exp\big[-n\mathsf{D}(\tilde{p}\|p)\big]. \tag{1.26}$$

Also, $\mathsf{D}(\tilde{p}\|p) \geq 0$ with strict inequality for all $\tilde{p} \neq p$.

Discussion: The parameter $\tilde{p} = k/n$ is the relative frequency of 1's in the n-tuple $Z_1, \ldots, Z_n$. The upper bound in (1.25), as a function of n for fixed p and $\tilde{p}$, decreases exponentially with n (aside from a slowly varying coefficient proportional to $1/\sqrt{n}$).

The lower bound in (1.26) is the same as the upper bound in (1.25) except for the initial term of (1.26). For fixed $\tilde{p}$ satisfying $0 < \tilde{p} < 1$, this initial term approaches 1 with increasing n, so that the upper and lower bounds are asymptotically the same in the sense that their ratio goes to 1 as $n \to \infty$. An upper and lower bound of this type is said to be asymptotically tight, and the result is denoted as

$$\mathsf{p}_{S_n}(\tilde{p}n)) \sim \sqrt{\frac{1}{2\pi n\tilde{p}(1-\tilde{p})}}\, \exp\big[-n\mathsf{D}(\tilde{p}\|p)\big] \qquad \text{for } 0 < \tilde{p} < 1, \tag{1.27}$$

where the symbol $\sim$ here means that the ratio of the upper to lower bound approaches 1 uniformly in p and $\widetilde{p}$ as $n \to \infty$ for $\tilde{p}$ and p in any interval bounded away from 0 and 1,

The reason for fussing about uniform convergence here is that $\widetilde{p}n$ must be an integer, so that one cannot simply pick any fixed $\widetilde{p}$ and go to the limit $n \to \infty$. The reason for requiring $\tilde{p}$ to be bounded away from 0 and 1 for this asymptotic result can be seen by looking at $\mathsf{p}_{S_n}(1) = np(1-p)^{n-1}$. This corresponds to $\tilde{p} = 1/n$. The extra factor in the lower bound, $(1 - 1/(12n\tilde{p}(1-\tilde{p}))$ then does not quite approach 1 with increasing n.

The divergence, $\mathsf{D}(\tilde{p}\|p)$, provides a measure of the difference between $\tilde{p}$ and p that is also useful in detection and the study of large deviations. This same type of quantity appears for arbitrary discrete rv's in Chapter 9.

A guided proof of the theorem is given in Exercise 1.9

We saw earlier that the Bernoulli process can also be characterized as a sequence of IID geometrically distributed interarrival intervals. An interesting generalization of this arises by allowing the interarrival intervals to be arbitrary discrete or continuous nonnegative IID rv's rather than geometric rv's. These processes are known as *renewal processes* and are the topic of Chapter 5. Poisson processes are special cases of renewal processes in which the interarrival intervals have an exponential PDF. These are treated in Chapter 2 and have many connections to Bernoulli processes.

Renewal processes are examples of *discrete stochastic processes*. The distinguishing characteristic of such processes is that interesting things (arrivals, departures, changes of state)

occur at discrete instants of time separated by deterministic or random intervals. Discrete stochastic processes are to be distinguished from noise-like stochastic processes in which changes are continuously occurring and the sample paths are continuously varying functions of time. The description of discrete stochastic processes above is not intended to be precise, but Chapters 2, 4, and 5 are restricted to discrete stochastic processes in this sense, whereas Chapter 3 is restricted to continuous processes.

1.5 Expectations and more probability review

The *expected value* $\mathsf{E}[X]$ of a random variable X is also called the *expectation* or the *mean* and is frequently denoted as $\overline{X}$. Before giving a general definition, we discuss several special cases. First consider nonnegative discrete rv's. The expected value $\mathsf{E}[X]$ is then given by

$$\mathsf{E}[X] = \sum_x x\, \mathsf{p}_X(x). \tag{1.28}$$

If X has a finite number of possible sample values, the above sum must be finite since each sample value must be finite. On the other hand, if X has a countable number of nonnegative sample values, the sum in (1.28) might be either finite or infinite. Example 1.5.1 illustrates a case in which the sum is infinite. The expectation is said to *exist* only if the sum is finite (*i.e.*, if the sum converges to a real number), and in this case $\mathsf{E}[X]$ is given by (1.28). If the sum is infinite, we say that $\mathsf{E}[X]$ does not exist, but also say[18] that $\mathsf{E}[X] = \infty$. In other words, (1.28) can be used in both cases, but $\mathsf{E}[X]$ is said to *exist* only if the sum is finite.

Example 1.5.1. This example will be useful frequently in illustrating rv's that have an infinite expectation. Let N be a positive integer-valued rv with the CDF $\mathsf{F}_N(n) = n/(n+1)$ for each integer $n \geq 1$. Then N is clearly a positive rv since $\mathsf{F}_N(0) = 0$ and $\lim_{n\to\infty} \mathsf{F}_N(n) = 1$. For each $n \geq 1$, the PMF is given by

$$\mathsf{p}_N(n) = \mathsf{F}_N(n) - \mathsf{F}_N(n-1) = \frac{n}{n+1} - \frac{n-1}{n} = \frac{1}{n(n+1)}. \tag{1.29}$$

Since $\mathsf{p}_N(n)$ is a PMF, we see that $\sum_{n=1}^{\infty} 1/[n(n+1)] = 1$, which is a frequently useful fact. The following equation, however, shows that $\mathsf{E}[N]$ does not exist and has infinite value.

$$\mathsf{E}[N] = \sum_{n=1}^{\infty} n\, \mathsf{p}_N(n) = \sum_{n=1}^{\infty} \frac{n}{n(n+1)} = \sum_{n=1}^{\infty} \frac{1}{n+1} = \infty,$$

where we have used the fact that the harmonic series diverges.

We next derive an alternative expression for the expected value of a nonnegative discrete rv. This new expression is given directly in terms of the CDF. We then use this new expression as a general definition of expectation which applies to all nonnegative rv's, whether discrete, continuous, or arbitrary. It contains none of the convergence questions that could cause confusion for arbitrary rv's or for continuous rv's with wild densities.

[18] It seems metaphysical to say that something doesn't exist but has infinite value. However, the word *exist* here is shorthand for *exist as a real number,* which makes it quite reasonable to also consider the value in the extended real number system, which includes $\pm\infty$.

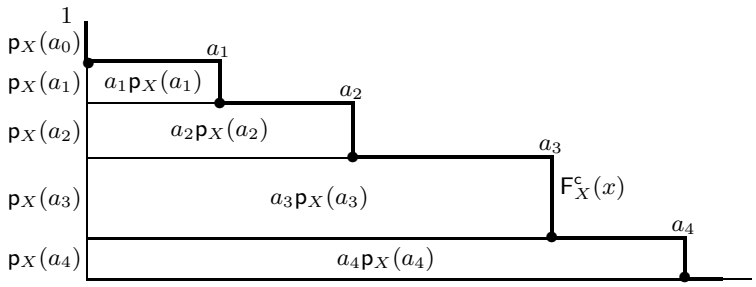


Figure 1.3: The figure shows the complementary CDF $\mathsf{F}^{\mathsf{c}}_X$ of a nonnegative discrete rv X. For this example, X takes on five possible values, $0 = a_0 < a_1 < a_2 < a_3 < a_4$. Thus $\mathsf{F}^{\mathsf{c}}_X(x) = \Pr\{X > x\} = 1-p_X(a_0)$ for $x < a_1$. For $a_1 \le x < a_2$, $\Pr\{X > x\} = 1-\mathsf{p}_X(a_0)-\mathsf{p}_X(a_1)$, and $\Pr\{X > x\}$ has similar drops as x reaches a_2, a_3, and a_4. $\mathsf{E}[X]$, from (1.28), is $\sum_i a_i \mathsf{p}_X(a_i)$, which is the sum of the rectangles in the figure. This is also the area under the curve $\mathsf{F}^{\mathsf{c}}_X(x)$, *i.e.*, $\int_0^\infty \mathsf{F}^{\mathsf{c}}_X(x)\,dx$. It can be seen that this argument applies to any nonnegative rv, thus verifying (1.30).

For a nonnegative discrete rv X, Figure 1.3 illustrates that (1.28) is simply the integral of the complementary CDF, where the *complementary CDF* F^{c} of a rv is defined as $\mathsf{F}^{\mathsf{c}}_X(x) = \Pr\{X > x\} = 1 - \mathsf{F}_X(x)$.

$$\mathsf{E}[X] \;=\; \int_0^\infty \mathsf{F}^{\mathsf{c}}_X\,dx = \int_0^\infty \Pr\{X > x\}\,dx. \tag{1.30}$$

Although Figure 1.3 only illustrates the equality of (1.28) and (1.30) for one special case, one easily sees that the argument applies to any nonnegative discrete rv, including those with countably many values, by equating the sum of the indicated rectangles with the integral.

For a nonnegative integer-vaued rv X, (1.30) reduces to a simpler form that is often convenient when X has a countable set of sample values.

$$\mathsf{E}[X] \;=\; \sum_{n=0}^{\infty} \Pr\{X > n\} \;=\; \sum_{n=1}^{\infty} \Pr\{X \ge n\} \tag{1.31}$$

For an arbitrary nonnegative rv X, we can visualize quantizing the rv, finding the expected value of the quantized rv, and then going to the limit of arbitrarily fine quantizations. Each quantized rv is discrete, so its expectation is given by (1.30) applied to the quantized rv. Each such expectation can be viewed as a Riemann sum for the integral $\int_0^\infty F^c_X(x)\,dx$ of the original rv.

There are no mathematical subtleties in integrating an arbitrary nonnegative non-increasing function, so $\int_0^\infty F^c_X(x)\,dx$ must have either a finite or infinite limit. This leads us to the following fundamental definition of expectation for arbitrary nonnegative rv's:

Definition 1.5.1. *The expectation $\mathsf{E}[X]$ of a nonnegative rv X is defined by (1.30). The expectation is said to exist if and only if the integral is finite. Otherwise the expectation is said to not exist and is also said to be infinite.*

Exercise 1.6 shows that this definition is consistent with the conventional definition of expectation for the case of continuous rv's, *i.e.*,

$$\mathsf{E}[X] = \lim_{b\to\infty} \int_0^b x \mathsf{f}_X(x)\, dx. \tag{1.32}$$

This can also be seen using integration by parts.

Next consider rv's with both positive and negative sample values. If X has a finite number of positive and negative sample values, say $a_1, a_2, \ldots, a_n$ the expectation $\mathsf{E}[X]$ is given by

$$\begin{aligned} \mathsf{E}[X] &= \sum_i a_i \mathsf{p}_X(a_i) \\ &= \sum_{a_i \le 0} a_i\, \mathsf{p}_X(a_i) + \sum_{a_i > 0} a_i\, \mathsf{p}_X(a_i). \end{aligned} \tag{1.33}$$

If X has a countably infinite set of sample values, then (1.33) can still be used if each of the sums in (1.33) converges to a finite value, and otherwise the expectation does not exist (as a real number). It can be seen that each sum in (1.33) converges to a finite value if and only if $\mathsf{E}[|X|]$ exists (*i.e.*, converges to a finite value) for the nonnegative rv $|X|$.

If $\mathsf{E}[X]$ does not exist (as a real number), it still might have the value ∞ if the first sum converges and the second does not, or the value $-\infty$ if the second sum converges and the first does not. If both sums diverge, then $\mathsf{E}[X]$ is undefined, even as $\pm\infty$. In this latter case, the partial sums can be arbitrarily small or large depending on the order in which the terms of (1.33) are summed (see Exercise 1.8).

As illustrated for a finite number of sample values in Figure 1.4, the expression in (1.33) can also be expressed directly in terms of the CDF and complementary CDF as

$$\mathsf{E}[X] = -\int_{-\infty}^{0} \mathsf{F}_X(x)\, dx + \int_0^\infty \mathsf{F}^{\mathsf{c}}_X(x)\, dx. \tag{1.34}$$

Since $\mathsf{F}^{\mathsf{c}}_X(x) = 1 - \mathsf{F}_X(x)$, this can also be expressed as

$$\mathsf{E}[X] = \int_{-\infty}^{\infty} \big[u(x) - \mathsf{F}_X(x)\big]\, dx,$$

where $u(x)$ is the unit step, $u(x) = 1$ for $x \ge 0$ and $u(x) = 0$ otherwise.

The first integral in (1.34) corresponds to the negative sample values and the second to the positive sample values, and $\mathsf{E}[X]$ exists if and only if both integrals are finite (*i.e.*, if $\mathsf{E}[|X|]$ is finite).

For continuous-valued rv's with positive and negative sample values, the conventional definition of expectation (assuming that $\mathsf{E}[|X|]$ exists) is given by

$$\mathsf{E}[X] = \int_{-\infty}^{\infty} x\, \mathsf{f}_X(x)\, dx. \tag{1.35}$$

This is equal to (1.34) by the same argument as with nonnegative rv's. Also, as with nonnegative rv's, (1.34) also applies to arbitrary rv's. We thus have the following fundamental definition of expectation:

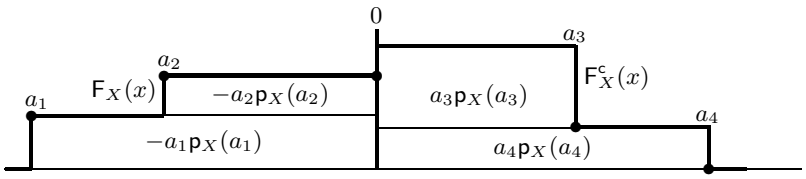


Figure 1.4: For this example, X takes on four possible sample values, $a_1 < a_2 < 0 < a_3 < a_4$. The figure plots $\mathsf{F}_X(x)$ for $x \leq 0$ and $\mathsf{F}^{\mathsf{c}}_X(x)$ for $x > 0$. As in Figure 1.3, $\int_{x\geq 0} \mathsf{F}^{\mathsf{c}}_X(x)\,dx = a_3\mathsf{p}_X(a_3) + a_4\mathsf{p}_X(a_4)$. Similarly, $\int_{x<0} \mathsf{F}_X(x)\,dx = -a_1\mathsf{p}_X(a_1) - a_2\mathsf{p}_X(a_2)$.

Definition 1.5.2. *The expectation $\mathsf{E}[X]$ of a rv X exists, with the value given in (1.34), if each of the two terms in (1.34) is finite. The expectation does not exist, but has value ∞ ($-\infty$), if the first term is finite (infinite) and the second infinite (finite). The expectation does not exist and is undefined if both terms are infinite.*

We should not view the general expression in (1.34) for expectation as replacing the need for the conventional expressions in (1.35) and (1.33). We will use all of these expressions frequently, using whichever is most convenient. The main advantages of (1.34) are that it applies equally to all rv's, it poses no questions about convergence, and it is frequently useful, especially in limiting arguments.

Example 1.5.2. The *Cauchy* rv X is the classic example of a rv whose expectation does not exist and is undefined. The probability density is $\mathsf{f}_X(x) = \frac{1}{\pi(1+x^2)}$. Thus $x\mathsf{f}_X(x)$ is proportional to $1/x$ both as $x \to \infty$ and as $x \to -\infty$. It follows that $\int_0^\infty x\mathsf{f}_X(x)\,dx$ and $\int_{-\infty}^0 -x\mathsf{f}_X(x)\,dx$ are both infinite. On the other hand, we see from symmetry that the Cauchy principal value of the integral in (1.35) is given by

$$\lim_{A\to\infty} \int_{-A}^{A} \frac{x}{\pi(1+x^2)}\,dx = 0.$$

There is usually little motivation for considering the upper and lower limits of the integration to have the same magnitude, and the Cauchy principal value usually has little significance for expectations.

1.5.1 Random variables as functions of other random variables

Random variables (rv's) are often defined in terms of each other. For example, if h is a function from $\mathbb{R}$ to $\mathbb{R}$ and X is a rv, then $Y = h(X)$ is the random variable that maps each sample point ω to the composite function $h(X(\omega))$. The CDF of Y can be found from this, and the expected value of Y can then be evaluated by (1.34).

It is often more convenient to find $\mathsf{E}[Y]$ directly using the CDF of X. Exercise 1.19 indicates that $\mathsf{E}[Y]$ is given by $\int h(x)\mathsf{f}_X(x)\,dx$ for continuous rv's and by $\sum_x h(x)\mathsf{p}_X(x)$ for discrete

rv's. In order to avoid continuing to use separate expressions for continuous and discrete rv's, we express both of these relations by

$$\mathsf{E}[Y] = \int_{-\infty}^{\infty} h(x)\, d\mathsf{F}_X(x), \tag{1.36}$$

This is known as a Stieltjes integral, which can be used as a generalization of both the continuous and discrete cases. For most purposes, we use Stieltjes integrals[19] as a notational shorthand for either $\int h(x)\mathsf{f}_X(x)\,dx$ or $\sum_x h(x)\mathsf{p}_X(x)$.

The existence of $\mathsf{E}[X]$ does not guarantee the existence of $\mathsf{E}[Y]$, but we will treat the question of existence as it arises rather than attempting to establish any general rules.

Particularly important examples of such expected values are the moments $\mathsf{E}[X^n]$ of a rv X and the central moments $\mathsf{E}\left[(X-\overline{X})^n\right]$ of X, where $\overline{X}$ is the mean $\mathsf{E}[X]$. The second central moment is called the *variance*, denoted by σ_X^2 or $\mathsf{VAR}[X]$. It is given by

$$\sigma_X^2 = \mathsf{E}\left[(X-\overline{X})^2\right] = \mathsf{E}\left[X^2\right] - \overline{X}^2. \tag{1.37}$$

The *standard deviation* σ_X of X is the square root of the variance and provides a measure of dispersion of the rv around the mean. Thus the mean is often viewed as a 'typical value' for the outcome of the rv (see Section 1.5.3) and σ_X is similarly viewed as a typical difference between X and $\overline{X}$. An important connection between the mean and standard deviation is that $\mathsf{E}\left[(X-x)^2\right]$ is minimized over x by choosing x to be $\mathsf{E}[X]$ (see Exercise 1.24).

Next suppose X and Y are rv's and consider the rv[20] $Z = X + Y$. If we assume that X and Y are independent, then the CDF of $Z = X + Y$ is given by[21]

$$\mathsf{F}_Z(z) = \int_{-\infty}^{\infty} \mathsf{F}_X(z-y)\, d\mathsf{F}_Y(y) = \int_{-\infty}^{\infty} \mathsf{F}_Y(z-x)\, d\mathsf{F}_X(x). \tag{1.38}$$

If X and Y both have densities, this can be rewritten as

$$\mathsf{f}_Z(z) = \int_{-\infty}^{\infty} \mathsf{f}_X(z-y)\mathsf{f}_Y(y)\, dy = \int_{-\infty}^{\infty} \mathsf{f}_Y(z-x)\mathsf{f}_X(x)\, dx. \tag{1.39}$$

Eq. (1.39) is the familiar convolution equation from linear systems, and we similarly refer to (1.38) as the convolution of CDF's (although it has a different functional form from (1.39)).

[19] More specifically, the Riemann-Stieltjes integral, abbreviated here as the Stieltjes integral, is denoted as $\int_a^b h(x)dF_X(x)$. This integral is defined as the limit of a generalized Riemann sum, $\lim_{\delta\to 0}\sum_n h(x_n)[\mathsf{F}(y_n) - \mathsf{F}(y_{n-1})]$ where $\{y_n; n \geq 1\}$ is a sequence of increasing numbers from a to b satisfying $y_n - y_{n-1} \leq \delta$ and $y_{n-1} < x_n \leq y_n$ for all n. The Stieltjes integral is defined to exist over finite limits if the limit exists and is independent of the choices of $\{y_n\}$ and $\{x_n\}$ as $\delta \to 0$. It exists over infinite limits if it exists over finite lengths and a limit over the integration limits can be taken. See Rudin [24] for an excellent elementary treatment of Stieltjes integration, and see Exercise 1.15 for some examples.

[20] The question whether a real-valued function of a rv is itself a rv is usually addressed by the use of measure theory, and since we neither use nor develop measure theoary in this text, we usually simply assume (within the limits of common sense) that any such function is itself a rv. However, the sum $X + Y$ of rv's is so important that Exercise 1.13 provides a guided derivation of this result for $X + Y$. In the same way, the sum $S_n = X_1 + \cdots + X_n$ of any finite collection of rv's is also a rv.

[21] See Exercise 1.15 for some peculiarities about this definition.

If X and Y are nonnegative random variables, then the integrands in (1.38) and (1.39) are non-zero only between 0 and z, so we often use 0 and z as the limits in (1.38) and (1.39).

If $X_1, X_2, \ldots, X_n$ are independent rv's, then the distribution of the rv $S_n = X_1 + X_2 + \cdots + X_n$ can be found by first convolving the CDF's of X_1 and X_2 to get the CDF of S_2 and then, for each $n \geq 2$, convolving the CDF of S_n and X_{n+1} to get the CDF of S_{n+1}. The CDF's can be convolved in any order to get the same resulting CDF.

Whether or not $X_1, X_2, \ldots, X_n$ are independent, the expected value of $S_n = X_1 + X_2 + \cdots + X_n$ satisfies

$$\mathsf{E}[S_n] = \mathsf{E}[X_1 + X_2 + \cdots + X_n] = \mathsf{E}[X_1] + \mathsf{E}[X_2] + \cdots + \mathsf{E}[X_n]. \tag{1.40}$$

This says that the expected value of a sum is equal to the sum of the expected values, whether or not the rv's are independent (see exercise 1.14). The following example shows how this can be a valuable problem solving aid with an appropriate choice of rv's.

Example 1.5.3. Consider a switch with n input nodes and n output nodes. Suppose each input is randomly connected to a single output in such a way that each output is also connected to a single input. That is, each output is connected to input 1 with probability $1/n$. Given this connection, each of the remaining outputs are connected to input 2 with probability $1/(n-1)$, and so forth.

An input node is said to be *matched* if it is connected to the output of the same number. We want to show that the expected number of matches (for any given n) is 1. Note that the first node is matched with probability $1/n$, and therefore the expectation of a match for node 1 is $1/n$. Whether or not the second input node is matched depends on the choice of output for the first input node, but it can be seen from symmetry that the *marginal CDF* for the output node connected to input 2 is $1/n$ for each output. Thus the expectation of a match for node 2 is also $1/n$. In the same way, the expectation of a match for each input node is $1/n$. From (1.40), the expected total number of matches is the sum over the expected number for each input, and is thus equal to 1. This exercise would be much more difficult without the use of (1.40).

If the rv's $X_1, \ldots, X_n$ are independent, then, as shown in exercises 1.14 and 1.21, the variance of $S_n = X_1 + \cdots + X_n$ is given by

$$\sigma^2_{S_n} = \sum\nolimits_{i=1}^{n} \sigma^2_{X_i}. \tag{1.41}$$

If $X_1, \ldots, X_n$ are also identically distributed (*i.e.*, $X_1, \ldots, X_n$ are IID) with variance σ^2_X, then $\sigma^2_{S_n} = n\sigma^2_X$. Thus the standard deviation of S_n is $\sigma_{S_n} = \sqrt{n}\sigma_X$. Sums of IID rv's appear everywhere in probability theory and play an especially central role in the laws of large numbers. It is important to remember that the mean of S_n is linear in n but the standard deviation increases only with the square root of n. Figure 1.5 illustrates this behavior.

1.5.2 Conditional expectations

Just as the *conditional CDF* of one rv conditioned on a sample value of another rv is important, the *conditional expectation* of one rv based on the sample value of another is

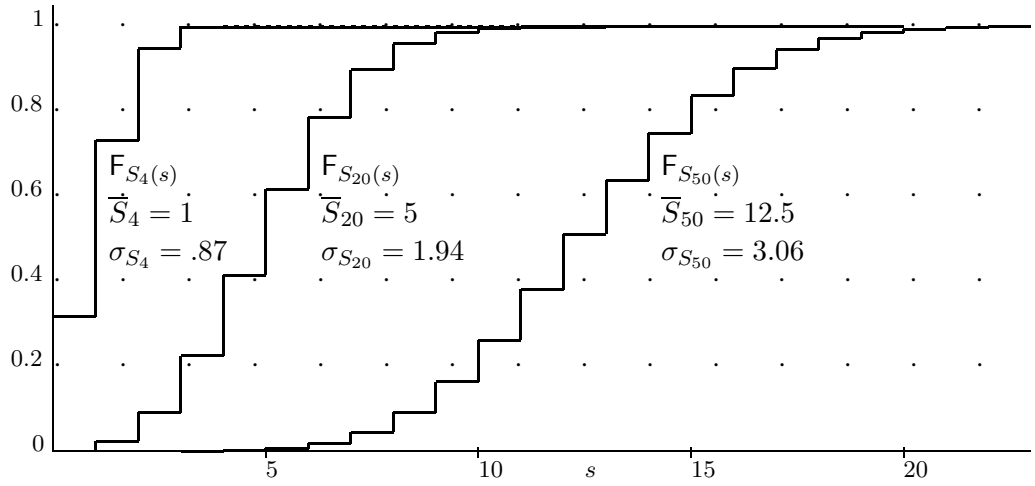


Figure 1.5: The CDF $\mathsf{F}_{S_n}(s)$ of $S_n = X_1 + \cdots + X_n$ where $X_1, \ldots, X_n$ are typical IID rv's and n takes the values 4, 20, and 50. The particular rv X in the figure is binary with $\mathsf{p}_X(1) = 1/4$, $\mathsf{p}_X(0) = 3/4$. Note that the mean of S_n is proportional to n and the standard deviation to $\sqrt{n}$.

equally important. Initially let X be a positive discrete rv and let y be a sample value of another discrete rv Y such that $\mathsf{p}_Y(y) > 0$. Then the conditional expectation of X given $Y = y$ is defined to be

$$\mathsf{E}\,[X \mid Y{=}y] = \sum_x x\, \mathsf{p}_{X|Y}(x \mid y). \tag{1.42}$$

This is simply the ordinary expected value of X using the conditional probabilities in the reduced sample space corresponding to $Y = y$. This value can be finite or infinite as before. More generally, if X can take on positive or negative values, then there is the possibility that the conditional expectation is undefined. In other words, for discrete rv's, the conditional expectation is exactly the same as the ordinary expectation, except that it is taken using conditional probabilities over the reduced sample space.

More generally yet, let X be an arbitrary rv and let y be a sample value of a discrete rv Y with $\mathsf{p}_Y(y) > 0$. The conditional CDF of X conditional on $Y = y$ is defined as

$$\mathsf{F}_{X|Y}(x \mid y) = \frac{\Pr\{X \le x, Y = y\}}{\Pr\{Y = y\}}.$$

Since this is an ordinary CDF in the reduced sample space where $Y = y$, (1.34) expresses the expectation of X conditional on $Y = y$ as

$$\mathsf{E}\,[X \mid Y = y] = -\int_{-\infty}^{0} \mathsf{F}_{X|Y}(x \mid y)\, dx + \int_0^{\infty} \mathsf{F}^{\mathsf{c}}_{X|Y}(x \mid y)\, dx. \tag{1.43}$$

The forms of conditional expectation in (1.42) and (1.43) are given for individual sample values of Y for which $\mathsf{p}_Y(y) > 0$.

We next show that the conditional expectation of X conditional on a discrete rv Y can also be viewed as a rv. With the possible exception of a set of zero probability, each $\omega \in \Omega$ maps to $\{Y = y\}$ for some y with $\mathsf{p}_Y(y) > 0$ and $\mathsf{E}[X \mid Y = y]$ is defined for that y. Thus we can define $\mathsf{E}[X \mid Y]$ as[22] a rv that is a function of Y, mapping ω to a sample value, say y of Y, and mapping that y to $\mathsf{E}[X \mid Y = y]$. Regarding a conditional expectation as a rv that is a function of the conditioning rv is a powerful tool both in problem solving and in advanced work. We now explain how this can be used to express the unconditional mean of X as

$$\mathsf{E}[X] = \mathsf{E}\big[\mathsf{E}[X \mid Y]\big], \tag{1.44}$$

where the inner expectation is over X for each sample value of Y and the outer expectation is over the rv $\mathsf{E}[X \mid Y]$, which is a function of Y.

Example 1.5.4. Consider rolling two dice, say a red die and a black die. Let X_1 be the number on the top face of the red die, and X_2 that for the black die. Let $S = X_1 + X_2$. Thus X_1 and X_2 are IID integer rv's, each uniformly distributed from 1 to 6. Conditional on $S = j$, X_1 is uniformly distributed between 1 and $j - 1$ for $j \leq 7$ and between $j - 6$ and 6 for $j \geq 7$. For each $j \leq 7$, it follows that $\mathsf{E}[X_1 \mid S = j] = j/2$. Similarly, for $j \geq 7$, $\mathsf{E}[X_1 \mid S = j] = j/2$. This can also be seen by the symmetry between X_1 and X_2.

The rv $\mathsf{E}[X_1 \mid S]$ is thus a discrete rv taking on values from 1 to 6 in steps of 1/2 as the sample value of S goes from 2 to 12. The PMF of $\mathsf{E}[X_1 \mid S]$ is given by $\mathsf{p}_{\mathsf{E}[X_1|S]}(j/2) = \mathsf{p}_S(j)$. Using (1.44), we can then calculate $\mathsf{E}[X_1]$ as

$$\mathsf{E}[X_1] = \mathsf{E}\big[\mathsf{E}[X_1 \mid S]\big] = \sum_{j=2}^{12} \frac{j}{2}\, \mathsf{p}_S(j) = \frac{\mathsf{E}[S]}{2} = \frac{7}{2}.$$

This example is not intended to show the value of (1.44) in calculating expectation, since $\mathsf{E}[X_1] = 7/2$ is initially obvious from the uniform integer distribution of X_1. The purpose is simply to illustrate what the rv $\mathsf{E}[X_1 \mid S]$ means.

To illustrate (1.44) in a more general way, while still assuming X to be discrete, we can write out this expectation by using (1.42) for $\mathsf{E}[X \mid Y = y]$.

$$\begin{aligned} \mathsf{E}[X] &= \mathsf{E}\big[\mathsf{E}[X \mid Y]\big] = \sum_y \mathsf{p}_Y(y)\mathsf{E}[X \mid Y = y] \\ &= \sum_y \mathsf{p}_Y(y) \sum_x x\, \mathsf{p}_{X|Y}(x|y). \end{aligned} \tag{1.45}$$

Operationally, there is nothing very fancy in the example or in (1.44). Combining the sums, (1.45) simply says that $\mathsf{E}[X] = \sum_{y,x} x\, \mathsf{p}_{YX}(y, x)$. As a concept, however, viewing the conditional expectation $\mathsf{E}[X \mid Y]$ as a rv based on the conditioning rv Y is often a useful theoretical tool. This approach is equally useful as a tool in problem solving, since there are many problems where it is easy to find conditional expectations, and then to find the

[22] This assumes that $\mathsf{E}[X \mid Y = y]$ is finite for each y, which is one of the reasons that expectations are said to exist only if they are finite.

total expectation by averaging over the conditioning variable. For this reason, this result is sometimes called either the total expectation theorem or the iterated expectation theorem. Exercise 1.20 illustrates the advantages of this approach, particularly where it is initially clear that the expectation is finite. The following cautionary example, however, shows that this approach can sometimes hide convergence questions and give the wrong answer.

Example 1.5.5. Let Y be a geometric rv with the PMF $\mathsf{p}_Y(y) = 2^{-y}$ for integer $y \geq 1$. Let X be an integer rv that, conditional on Y, is binary with equiprobable values $\pm 2^y$ given $Y = y$. We then see that $\mathsf{E}[X \mid Y = y] = 0$ for all y, and thus, (1.45) indicates that $\mathsf{E}[X] = 0$. On the other hand, it is easy to see that $\mathsf{p}_X(2^k) = \mathsf{p}_X(-2^k) = 2^{-k-1}$ for each integer $k \geq 1$. Thus the expectation over positive values of X is ∞ and that over negative values is $-\infty$. In other words, the expected value of X is undefined and (1.45) is incorrect.

The difficulty in the above example cannot occur if X is a nonnegative rv. Then (1.45) is simply a sum of a countable number of nonnegative terms, and thus it either converges to a finite sum independent of the order of summation, or it diverges to ∞, again independent of the order of summation.

If X has both positive and negative components, we can separate it into $X = X^+ + X^-$ where $X^+ = \max(0, X)$ and $X^- = \min(X, 0)$. Then (1.45) applies to X^+ and $-X^-$ separately. If at most one is infinite, then (1.45) applies to X, and otherwise X is undefined. This is summarized in the following theorem:

Theorem 1.5.1 (Total expectation). *Let X and Y be discrete rv's. If X is nonnegative, then $\mathsf{E}[X] = \mathsf{E}\big[\mathsf{E}[X \mid Y]\big] = \sum_y \mathsf{p}_Y(y)\mathsf{E}[X \mid Y = y]$. If X has both positive and negative values, and if at most one of $\mathsf{E}[X^+]$ and $\mathsf{E}[-X^-]$ is infinite, then $\mathsf{E}[X] = \mathsf{E}\big[\mathsf{E}[X \mid Y]\big] = \sum_y \mathsf{p}_Y(y)\mathsf{E}[X \mid Y = y]$.*

We have seen above that if Y is a discrete rv, then the conditional expectation $\mathsf{E}[X|Y = y]$ is only a little more complicated than the unconditional expectation, and this is true whether X is discrete, continuous, or arbitrary. If X and Y are continuous, we can essentially extend these results to probability densities. In particular, defining $\mathsf{E}[X \mid Y = y]$ as

$$\mathsf{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x\, \mathsf{f}_{X|Y}(x \mid y)\, dx, \tag{1.46}$$

we have

$$\mathsf{E}[X] = \int_{-\infty}^{\infty} \mathsf{f}_Y(y)\mathsf{E}[X \mid Y{=}y]\; dy = \int_{-\infty}^{\infty} \mathsf{f}_Y(y) \int_{-\infty}^{\infty} x\, \mathsf{f}_{X|Y}(x \mid y)\, dx\, dy. \tag{1.47}$$

We do not state this as a theorem because the details about the integration do not seem necessary for the places where it is useful.

1.5.3 Typical values of rv's; mean and median

The CDF of a rv often has more detail than we are interested in, and the mean is often taken as a 'typical value.' Similarly, in statistics, the average of a set of numerical data

values is often taken to be representative of the entire set. For example, students always want to know the average of the scores on an exam, and investors always want to know the Dow-Jones average. Economists are also interested, for example, in such averages as the average annual household income over various geographical regions. These averages often take on an importance and a life of their own, particularly in terms of how they vary in time.

The median of a rv (or set of data values) is often an alternate choice of a single number to serve as a typical value. We say that α is a median of X if $\Pr\{X \leq \alpha\} \geq 1/2$ and $\Pr\{X \geq \alpha\} \geq 1/2$. It is possible for the median to be nonunique, with all values in an interval satisfying the definition. Exercise 1.10 illustrates what this definition means. In addition, Exercise 1.11 shows that if the mean exists, then the median is an x that minimizes $\mathsf{E}[|X - x|]$.

Another interesting property of the median, suggested in Exercise 1.34 is that in essence a median of a large number of IID sample values of a rv is close to a median of the distribution with high probability. Another property, relating the median α to the mean $\overline{X}$ of a rv with standard deviation σ, is (see Exercise 1.33)

$$| \overline{X} - \alpha | \leq \sigma \tag{1.48}$$

The question now arises whether the mean or the median is preferable as a single number describing a rv. The question is too vague to be answered in any generality, but the answer depends heavily on what the single number is to be used for. To illustrate this, consider a rv whose equiprobable sample values are the yearly household incomes of a large society (or, almost equivalently, consider a large data set consisting of these yearly household incomes).

For the mean, the probability of each sample value is weighted by the household income, so that a household income of $\$10^9$ is weighted the same as 100,000 household incomes of $\$10^4$ each. For the median, this weighting disappears, and if our billionaire has a truly awful year with only $\$10^6$ income, the median is unchanged. If one is interested in the total purchasing power of the society, then the mean might be the more appropriate value. On the other hand, if one is interested in the well-being of the society, the median is the more appropriate value.[23]

1.5.4 Indicator random variables

For any event A, the *indicator random variable* of A, denoted $\mathbb{I}_A$, is a binary rv that has the value 1 for all $\omega \in A$ and the value 0 otherwise. It then has the PMF $\mathsf{p}_{\mathbb{I}_A}(1) = \Pr\{A\}$ and $\mathsf{p}_{\mathbb{I}_A}(0) = 1 - \Pr\{A\}$. The corresponding CDF $\mathsf{F}_{\mathbb{I}_A}$ is then illustrated in Figure 1.6. It is easily seen that $\mathsf{E}[\mathbb{I}_A] = \Pr\{A\}$.

Indicator rv's are useful because they allow us to apply the many known results about rv's and particularly binary rv's to events. For example, the laws of large numbers are expressed in terms of sums of rv's. If these rv's are taken to be the indicaor functions for

[23] Unfortunately, the choice between median and mean (and many similar choices) is often made for commercial or political expediency rather than scientific or common-sense appropriateness.

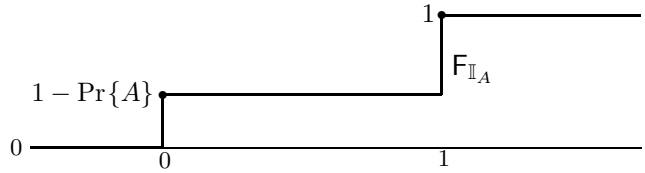


Figure 1.6: The CDF $\mathsf{F}_{\mathbb{I}_A}$ of an indicator random variable $\mathbb{I}_A$.

the occurencees of an event over successive trials, then the law of large numbers applies to the relative frequency of that event.

1.5.5 Moment generating functions and other transforms

The *moment generating function* (MGF) for a rv X is given by

$$\mathsf{g}_X(r) = \mathsf{E}\left[e^{rX}\right] = \int_{-\infty}^{\infty} e^{rx}\, d\mathsf{F}_X(x). \tag{1.49}$$

where r is a real variable. The integrand is nonnegative, and we can study where the integral exists (*i.e.*, where it is finite) by separating it as follows:

$$\mathsf{g}_X(r) = \int_{0}^{\infty} e^{rx}\, d\mathsf{F}_X(x) + \int_{-\infty}^{0} e^{rx}\, d\mathsf{F}_X(x). \tag{1.50}$$

Both integrals exist for $r = 0$, since the first is $\Pr\{X > 0\}$ and the second is $\Pr\{X \leq 0\}$. The first integral is increasing in r, and thus if it exists for one value of r, it also exists for all smaller values. For example, if X is a nonnegative exponential rv with the density $\mathsf{f}_X(x) = e^{-x}$, then the first integral exists if and only if $r < 1$, and it then has the value $\frac{1}{1-r}$. As another example, if X satisfies $\Pr\{X > A\} = 0$ for some finite A, then the first integral is at most e^{rA}, which is finite for all real r.

Let $r_+(X)$ be the supremum of values of r for which the first integral exists. Then $0 \leq r_+(X) \leq \infty$ and the first integral exists for all $r < r_+(X)$. In the same way, let $r_-(X)$ be the infimum of values of r for which the the second integral exists. Then $0 \geq r_-(X) \geq -\infty$ and the second integral exists for all $r > r_-(X)$.

Combining the two integrals, the region of r over which $\mathsf{g}_X(r)$ exists is an interval $I(X)$ from $r_-(X) \leq 0$ to $r_+(X) \geq 0$. Either or both of the end points, $r_-(X)$ and $r_+(X)$, might be included in $I(X)$; *i.e.*, each end of the interval $I(X)$ can be open or closed. We denote these quantities as I, r_-, and r_+ when the rv X is clear from the context. Exercise 1.25 illustrates r_- and r_+ further.

If $\mathsf{g}_X(r)$ exists in an open region of r around 0 (*i.e.*, if $r_- < 0 < r_+$), then derivatives[24] of

[24]This result depends on interchanging the order of differentiation (with respect to r) and integration (with respect to x). This can be shown to be permissible because $\mathsf{g}_X(r)$ exists for r both greater and smaller than 0, which in turn implies, first, that $1 - \mathsf{F}_X(x)$ must approach 0 at least exponentially as $x \to \infty$ and, second, that $\mathsf{F}_X(x)$ must approach 0 at least exponentially as $x \to -\infty$.

all orders exist in that region. They are given by

$$\frac{d^k \mathsf{g}_X(r)}{dr^k} = \int_{-\infty}^{\infty} x^k e^{rx} d\mathsf{F}_X(x) \quad ; \qquad \left.\frac{d^k \mathsf{g}_X(r)}{dr^k}\right|_{r=0} = \mathsf{E}\left[X^k\right]. \tag{1.51}$$

This shows that finding the moment generating function often provides a convenient way to calculate the moments of a random variable (see Exercise 3.2 for an example). If any moment of a rv fails to exist, however, then the MGF is infinite either throughout $(0, \infty)$ or $(\infty, 0)$. (see Exercise 1.37).

Another important feature of moment generating functions is their usefulness in treating sums of independent rv's. For example, let $S_n = X_1 + X_2 + \cdots + X_n$. Then

$$\begin{aligned} \mathsf{g}_{S_n}(r) &= \mathsf{E}\left[e^{rS_n}\right] = \mathsf{E}\left[\exp\left(\textstyle\sum_{i=1}^n rX_i\right)\right] \\ &= \mathsf{E}\left[\textstyle\prod_{i=1}^n \exp(rX_i)\right] = \textstyle\prod_{i=1}^n \mathsf{g}_{X_i}(r). \end{aligned} \tag{1.52}$$

In the last step, we have used a result of Exercise 1.14, which shows that for independent rv's, the mean of the product is equal to the product of the means. If $X_1, \ldots, X_n$ are also IID, then

$$\mathsf{g}_{S_n}(r) = [\mathsf{g}_X(r)]^n. \tag{1.53}$$

We will use this property frequently in treating sums of IID rv's. Note that this also implies that the region over which the MGF's of S_n and X exist are the same, *i.e.*, $I(S_n) = I(X)$.

The real variable r in the MGF can be replaced by a complex variable, giving rise to a number of other transforms. A particularly important case is to view r as a pure imaginary variable, say $i\theta$ where $i = \sqrt{-1}$ and θ is real. Then[25] $\mathsf{g}_X(i\theta) = \mathsf{E}\left[e^{i\theta x}\right]$ is called the *characteristic function* of X. Since $|e^{i\theta x}|$ is 1 for all x, $\mathsf{g}_X(i\theta)$ exists for all rv's X and all real θ, and its magnitude is at most one.

A minor but important variation on the characteristic function of X is the Fourier transform of the probability density of X. If X has a density $\mathsf{f}_X(x)$, then the Fourier transform of $\mathsf{f}_X(x)$ is given by

$$\mathsf{g}_X(-i2\pi\theta) = \int_{-\infty}^{\infty} \mathsf{f}_X(x) \exp(-i2\pi\theta)\, dx \tag{1.54}$$

The major advantage of the Fourier transform (aside from its familiarity) is that $\mathsf{f}_X(x)$ can usually be found from $\mathsf{g}_X(-i2\pi\theta)$ as the inverse Fourier transform,[26]

$$\mathsf{f}_X(x) = \int_{-\infty}^{\infty} \mathsf{g}_X(-i2\pi\theta) \exp(i2\pi\theta x)\, dx, \tag{1.55}$$

[25]The notation here can be slightly dangerous, since one cannot necessarily take an expression for $\mathsf{g}_X(r)$, valid for real r, and replace r by $i\theta$ with real θ to get the characteristic function.

[26]This integral does not necessarily converge, particularly if X does not have a PDF. However, it can be shown (see [27] Chap. 2.12, or [9], Chap. 15) that the characteristic function/ Fourier transform of an arbitrary rv does uniquely specify the CDF.)

The Z-transform is the result of replacing e^r with z in $\mathsf{g}_X(r)$. This is useful primarily for integer-valued rv's, but if one transform can be evaluated, the other can be found immediately. Finally, if we use $-s$, viewed as a complex variable, in place of r, we get the two sided Laplace transform of the density of the random variable. Note that for all of these transforms, multiplication in the transform domain corresponds to convolution of the CDF's or densities, and summation of independent rv's. The simplicity of taking products of transforms is a major reason that transforms are so useful in probability theory.

1.6 Basic inequalities

Inequalities play a particularly fundamental role in probability, partly because many important models are too complex to find exact answers, and partly because many of the most useful theorems establish limiting rather than exact results. In this section, we study three related inequalities, the Markov, Chebyshev, and Chernoff bounds. These are used repeatedly both in the next section and in the remainder of the text.

1.6.1 The Markov inequality

This is the simplest and most basic of these inequalities. It states that if a nonnegative random variable Y has a mean $\mathsf{E}[Y]$, then, for every $y > 0$, $\Pr\{Y \geq y\}$ satisfies[27]

$$\Pr\{Y \geq y\} \leq \frac{\mathsf{E}[Y]}{y} \qquad \text{for every } y > 0 \quad \text{(Markov Inequality)}. \tag{1.56}$$

Figure 1.7 derives this result using the fact (see Figure 1.3) that the mean of a nonnegative rv is the integral of its complementary CDF, i.e., of the area under the curve $\Pr\{Y > z\}$. Exercise 1.30 gives another simple proof using an indicator random variable.

As an example of this inequality, assume that the average height of a population of people is 1.6 meters. Then the Markov inequality states that at most half of the population have a height exceeding 3.2 meters. We see from this example that the Markov inequality is often very weak. However, for any $y > 0$, we can consider a rv that takes on the value y with probability ϵ and the value 0 with probability $1 - \epsilon$; this rv satisfies the Markov inequality at the point y with equality. Figure 1.7 (as elaborated in Exercise 1.45) also shows that, for any nonnegative rv Y with a finite mean,

$$\lim_{y\to\infty} y \Pr\{Y \geq y\} = 0. \tag{1.57}$$

This will be useful shortly in the proof of Theorem 1.7.4.

[27]The CDF of any given rv Y is known (at least in principle), and thus one might question why an upper bound is ever preferable to the exact value. One answer is that Y might be given as a function of many other rv's and that the parameters (such as the mean) used in a bound are often much easier to find than the CDF. Another answer is that such inequalities are often used in theorems which state results in terms of simple statistics such as the mean rather than the entire CDF. This will be evident as we use these bounds.

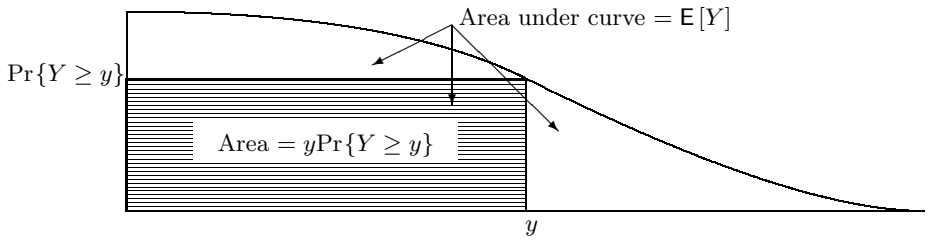


Figure 1.7: Demonstration that $y\Pr\{Y \geq y\} \leq \mathsf{E}[Y]$. By letting $y \to \infty$, it can also be seen that the shaded area becomes a negligible portion of the area $\mathsf{E}[Y]$, so that $\lim_{y\to\infty} y\Pr\{Y > y\} = 0$ if $\mathsf{E}[Y] \leq \infty$.

1.6.2 The Chebyshev inequality

We now use the Markov inequality to establish the well-known Chebyshev inequality. Let Z be an arbitrary rv with finite mean $\mathsf{E}[Z]$ and finite variance σ_Z^2, and define Y as the nonnegative rv $Y = (Z - \mathsf{E}[Z])^2$. Thus $\mathsf{E}[Y] = \sigma_Z^2$. Applying (1.56),

$$\Pr\big\{(Z - \mathsf{E}[Z])^2 \geq y\big\} \leq \frac{\sigma_Z^2}{y} \qquad \text{for every } y > 0.$$

Replacing y with ϵ^2 and noting that the event $\{(Z-\mathsf{E}[Z])^2 \geq \epsilon^2\}$ is the same as $|Z-\mathsf{E}[Z]| \geq \epsilon$, this becomes

$$\Pr\{|Z - \mathsf{E}[Z]| \geq \epsilon\} \leq \frac{\sigma_Z^2}{\epsilon^2} \qquad \text{for every } \epsilon > 0 \quad \text{(Chebyshev inequality).} \tag{1.58}$$

Note that the Markov inequality bounds just the upper tail of the CDF and applies only to nonnegative rv's, whereas the Chebyshev inequality bounds both tails of the CDF. The more important differences, however, are that the Chebyshev bound requires a finite variance and approaches zero inversely with the square of the distance from the mean, whereas the Markov bound does not require a finite variance and goes to zero inversely with the distance from 0 (and thus asymptotically with distance from the mean).

The Chebyshev inequality is particularly useful when Z is the sample average, $(X_1 + X_2 + \cdots + X_n)/n$, of a set of IID rv's. This will be used shortly in proving the weak law of large numbers.

1.6.3 Chernoff bounds

We saw that the Chebyshev inequality on a rv Z results from applying the Markov inequality to $(Z - \overline{Z})^2$. In the same way, the Chernoff (or exponential) bound on a rv Z results from applying the Markov inequality to e^{rZ} for a given r. The Chernoff bound, when it applies, approaches 0 exponentially with increasing distance from the mean.

For any given rv Z, let $I(Z)$ be the interval over which the MGF $\mathsf{g}_Z(r) = \mathsf{E}\left[e^{rZ}\right]$ exists. Letting $Y = e^{rZ}$ for any $r \in I(Z)$, the Markov inequality (1.56) applied to Y is

$$\Pr\{\exp(rZ) \geq y\} \leq \frac{\mathsf{g}_Z(r)}{y} \qquad \text{for every } y > 0 \quad \text{(Chernoff bound)}.$$

This takes on a more meaningful form if y is replaced by e^{rb}. Note that $\exp(rZ) \geq \exp(rb)$ is equivalent to $Z \geq b$ for $r > 0$ and to $Z \leq b$ for $r < 0$. Thus, for any real b, we get the following two bounds, one for $r > 0$ and the other for $r < 0$:

$$\Pr\{Z \geq b\} \leq \mathsf{g}_Z(r)\exp(-rb) \quad ; \qquad \text{(Chernoff bound for } r > 0,\ r \in I(Z)) \tag{1.59}$$
$$\Pr\{Z \leq b\} \leq \mathsf{g}_Z(r)\exp(-rb) \quad ; \qquad \text{(Chernoff bound for } r < 0,\ r \in I(Z)). \tag{1.60}$$

Note that $g_Z(r) = 1$ for $r = 0$, so each bound has the value 1 for $r = 0$ and is thus valid (but uninteresting) for $r = 0$. It is somewhat simpler, however, when optimizing over r to include $r = 0$ in each bound. We will soon see that for $b \geq \overline{Z}$, the bound in (1.60) is at least 1 over the range of r where it is valid, and thus only (1.59) can be useful for $b \geq \overline{Z}$. Similarly, only (1.60) can be useful for $b \leq \overline{Z}$ (and thus neither is useful for $b = \overline{Z}$).

The most important application of Chernoff bounds is to sums of IID rv's. Let $S_n = X_1 + \cdots + X_n$ where $X_1, \ldots, X_n$ are IID with the MGF $\mathsf{g}_X(r)$. Then $\mathsf{g}_{S_n}(r) = [\mathsf{g}_X(r)]^n$, so (1.59) and (1.60) (with b replaced by na) become

$$\Pr\{S_n \geq na\} \leq [\mathsf{g}_X(r)]^n \exp(-rna); \qquad \text{for } r > 0,\ r \in I(X) \tag{1.61}$$
$$\Pr\{S_n \leq na\} \leq [\mathsf{g}_X(r)]^n \exp(-rna); \qquad \text{for } r < 0,\ r \in I(X). \tag{1.62}$$

These equations are easier to understand if we define the *semi-invariant MGF*, $\gamma_X(r)$, as

$$\gamma_X(r) = \ln \mathsf{g}_X(r). \tag{1.63}$$

The semi-invariant MGF for a typical rv X is sketched in Figure 1.8. Important features of the figure, as explained in the caption, are, first, that $\gamma'_X(0) = \mathsf{E}[X]$ and, second, that $\gamma''_X(r) \geq 0$ for r in the interior of $I(X)$.

In terms of $\gamma_X(r)$, (1.61) and (1.62) become

$$\Pr\{S_n \geq na\} \leq \exp(n[\gamma_X(r) - ra]); \qquad \text{for } r \geq 0,\ r \in I(X) \tag{1.64}$$
$$\Pr\{S_n \leq na\} \leq \exp(n[\gamma_X(r) - ra]); \qquad \text{for } r \leq 0,\ r \in I(X). \tag{1.65}$$

These bounds are exponential in n for fixed a and r, and for any given a, we can optimize (1.64), simultaneously for all n, by minimizing $\gamma(r) - ra$ over $r \geq 0$, $r \in I(X)$. Similarly, (1.65) can be minimized over $r \leq 0$. We will explore these minimizations in detail in Section 9.3, but for now, consider the case where $a > \overline{X}$. Note that $[\gamma(r) - ra]_{r=0} = 0$ and $\frac{d}{dr}[\gamma(r) - ra]|_{r=0} = \overline{X} - a < 0$. This means that $\gamma_X(r) - ra$ must be negative for sufficiently small $r > 0$ and thus means that the bound in (1.64) is decreasing exponentially in n for small enough $r > 0$.

For $r < 0$, $\gamma(r) - ra$ is positive for r close to 0, so the bound in (1.65) exceeds 1 and is worthless. Since $\gamma''(r) > 0$ over $I(X)$, $\frac{d}{dr}[\gamma(r) - ra] < 0$ for all negative $r \in I(X)$, so the

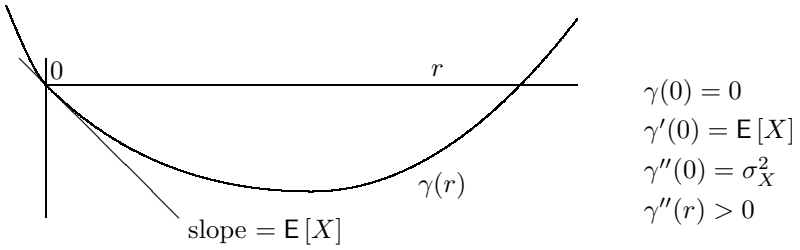


Figure 1.8: Semi-invariant moment-generating function $\gamma(r)$ for a typical rv X assuming that 0 is in the interior of $I(X)$. Since $\gamma(r) = \ln \mathsf{g}(r)$, we see that $\frac{d}{dr}\gamma(r) = \frac{1}{\mathsf{g}(r)}\frac{d}{dr}\mathsf{g}(r)$. Thus $\gamma'(0) = \mathsf{E}[X]$. Also, for r in the interior of $I(X)$, Exercise 1.26 shows that $\gamma''(r) \geq 0$, and in fact, $\gamma''(r)$ is strictly positive except in the uninteresting case where X is deterministic (takes on a single value with probability 1). As indicated in the figure, the straight line of slope $\mathsf{E}[X]$ through the origin is tangent to $\gamma(r)$ at $r = 0$.

bound in (1.65) is worthless for all r in its range when $a > \overline{X}$. This is not surprising, since $\Pr\{S_n \geq na\} = 1 - \Pr\{S_n \leq na\}$, so both bounds can not be exponentially decreasing in n for the same a.

The same argument can be reversed for $a < \overline{X}$, and in this case the bound (1.65) is exponentially decreasing in n for negative r sufficiently close to 0 and (1.64) is worthless. We also see that both bounds are worthless for $a = \overline{X}$, and this will be more obvious after discussing the central limit theorem.

The roles of (1.64) and (1.65) can be reversed by replacing $\{X_n;\, n \geq 1\}$ with $\{-X_n;\, n \geq 1\}$ and therefore we consider only (1.64) with $a > \overline{X}$ as the Chernoff bound in what follows.

The results of this section can be summarized in the following lemma.

Lemma 1.6.1. *Let $\{X_i;\, i \geq 1\}$ be IID rv's and let $S_n = X_1 + \cdots + X_n$ for each $n \geq 1$. Let $I(X)$ be the interval over which the semi-invariant MGF $\gamma_X(r)$ is finite and assume 0 is in the interior of $I(X)$. Then, for each $n \geq 1$ and each real number a,*

$$\Pr\{S_n \geq na\} \leq \exp(n\mu_X(a)) \;; \qquad \text{where } \mu_X(a) = \inf_{r\geq 0,\, r\in I(X)} \gamma_X(r) - ra. \tag{1.66}$$

Furthermore, $\mu_X(a) < 0$ for $a > \overline{X}$ and $\mu_X(a) = 0$ for $a \leq \overline{X}$.

These Chernoff bounds will be used in the next section to help understand several laws of large numbers. They will also be used and further developed in Chapter 9. They are useful for detection, random walks, information theory, and other areas where large deviations are important.

The following example evaluates these bounds for the case where the IID rv's are binary. We will see that in this case the bounds are exponentially tight in a sense to be described.

Example 1.6.1. Let X be binary with $\mathsf{p}_X(1) = p$ and $\mathsf{p}_X(0) = q = 1 - p$. Then $\mathsf{g}_X(r) = q + pe^r$ for $-\infty < r < \infty$. Also, $\gamma_X(r) = \ln(q + pe^r)$. To be consistent with the expression

for the binomial PMF in (1.25), we will find the Chernoff bound on $\Pr\{S_n \geq \tilde{p}n\}$ for $\tilde{p} > p$. Thus, according to Lemma 1.6.1, we first evaluate

$$\mu_X(\tilde{p}) = \inf_{r\geq 0}[\gamma_X(r) - \tilde{p}r].$$

The minimum over all r occurs at that r for which $\gamma'_X(r) = \tilde{p}$, *i.e.*, at

$$\frac{pe^r}{q + pe^r} = \tilde{p}.$$

Rearranging terms,

$$e^r = \frac{\tilde{p}q}{p\tilde{q}} \qquad \text{where } \tilde{q} = 1 - \tilde{p}. \tag{1.67}$$

For $\tilde{p} > p$, this r is positive, achieving the minimization over $r \geq 0$. Substituting this r into $\ln(q + pe^r) - r\tilde{p}$ and rearranging terms,

$$\mu_X(\tilde{p}) = \tilde{p}\ln\frac{p}{\tilde{p}} + \tilde{q}\ln\frac{\tilde{q}}{q}. \tag{1.68}$$

Substituting this into (1.66),

$$\Pr\{S_n{\geq}n\tilde{p}\} \leq \exp\left\{n\left[\tilde{p}\ln\frac{p}{\tilde{p}} + \tilde{q}\ln\frac{q}{\tilde{q}}\right]\right\}\ ; \qquad \text{for } \tilde{p} > p. \tag{1.69}$$

So far, it seems that we have simply developed another upper bound on the tails of the CDF for the binomial. It will then perhaps be surprising to compare this bound with the asymptotically correct value (repeated below) for the binomial PMF in (1.27) with $\tilde{p} = k/n$.

$$\mathsf{p}_{S_n}(k) \sim \sqrt{\frac{1}{2\pi n\tilde{p}\tilde{q}}}\,\exp\{-n\mathsf{D}(\tilde{p}\|p)\} \qquad \text{where } \mathsf{D}(\tilde{p}\|p) = \left[\tilde{p}\ln\frac{\tilde{p}}{p} + \tilde{q}\ln\frac{\tilde{q}}{q}\right]. \tag{1.70}$$

For any integer value of $n\tilde{p}$ with $\tilde{p} > p$, we can lower bound $\Pr\{S_n{\geq}n\tilde{p}\}$ by the single term $\mathsf{p}_{S_n}(n\tilde{p})$. Thus $\Pr\{S_n{\geq}n\tilde{p}\}$ is both upper and lower bounded by quantities that decrease exponentially with n at the same rate. In a sense, then, the upper bound and the lower bound are essentially the same for large n. We can express the sense in which they are the same by considering the log of the upper bound in (1.69) and the lower bound arising from (1.70).

$$\lim_{n\to\infty}\frac{\ln\Pr\{S_n \geq n\tilde{p}\}}{n} = -\mathsf{D}(\tilde{p}\|p) \qquad \text{where } \tilde{p} > p. \tag{1.71}$$

In other words, the Chernoff bound when optimized over r is not only an upper bound, but is also exponentially tight in the sense of (1.71).

In Section 9.3 we will show that this property is typical for sums of IID rv's. Thus we see that the Chernoff bounds are not 'just bounds,' but rather are bounds that when optimized provide the correct asymptotic exponents for the tails of the distribution of sums of IID rv's. In this sense these bounds are quite different from the Markov and Chebyshev bounds.

1.7 The laws of large numbers

The laws of large numbers are a collection of results in probability theory that describe the behavior of the arithmetic average of n rv's for large n. For any n rv's, $X_1, \ldots, X_n$, the *arithmetic average* is the rv $(1/n)\sum_{i=1}^{n} X_i$. Since in any outcome of the experiment, the sample value of this rv is the arithmetic average of the sample values of $X_1, \ldots, X_n$, this random variable is usually called the *sample average*. If $X_1, \ldots, X_n$ are viewed as successive variables in time, this sample average is called the time average. Under fairly general assumptions, the standard deviation of the sample average goes to 0 with increasing n, and, in various ways depending on the assumptions, the sample average approaches the mean.

These results are central to the study of stochastic processes because they allow us to relate time averages (i.e., the average over time of individual sample paths) to ensemble averages (i.e., the mean of the value of the process at a given time). In this section, we develop and discuss one of these results, the weak law of large numbers for IID rv's. We also briefly discuss another of these results, the strong law of large numbers. The strong law requires considerable patience to understand, and its derivation and fuller discussion are postponed to Chapter 5 where it is first needed. We also discuss the central limit theorem, partly because it enhances our understanding of the weak law, and partly because of its importance in its own right.

1.7.1 Weak law of large numbers with a finite variance

Let $X_1, X_2, \ldots, X_n$ be IID rv's with a finite mean $\overline{X}$ and finite variance σ^2. Let $S_n = X_1 + \cdots + X_n$, and consider the sample average S_n/n. We saw in (1.41) that $\sigma^2_{S_n} = n\sigma^2$. Thus the variance of S_n/n is

$$\mathsf{VAR}\left[\frac{S_n}{n}\right] = \mathsf{E}\left[\left(\frac{S_n - n\overline{X}}{n}\right)^2\right] = \frac{1}{n^2}\,\mathsf{E}\left[\left(S_n - n\overline{X}\right)^2\right] = \frac{\sigma^2}{n}. \tag{1.72}$$

This says that the variance of the sample average S_n/n is σ^2/n, which approaches 0 as n increases. Figure 1.9 illustrates this decrease in the variance of S_n/n with increasing n. In contrast, recall that Figure 1.5 illustrated how the variance of S_n increases with n. From (1.72), we see that

$$\lim_{n\to\infty} \mathsf{E}\left[\left(\frac{S_n}{n} - \overline{X}\right)^2\right] = 0. \tag{1.73}$$

As a result, we say that S_n/n *converges in mean square* to $\overline{X}$.

This convergence in mean square says that the sample average, S_n/n, differs from the mean, $\overline{X}$, by a random variable whose variance approaches 0 with increasing n. This convergence in mean square is one sense in which S_n/n approaches $\overline{X}$, but the idea of a sequence of rv's (*i.e.*, a sequence of functions) approaching a constant is clearly much more involved than a sequence of numbers approaching a constant. The laws of large numbers bring out this

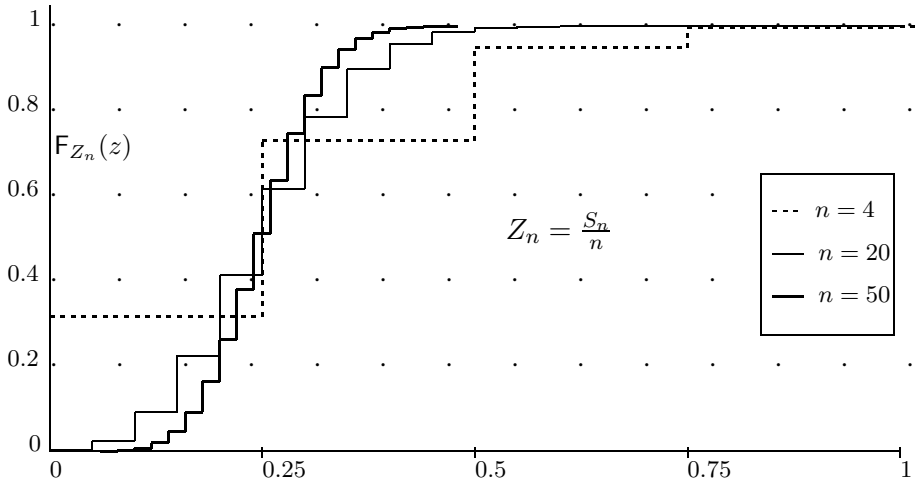


Figure 1.9: The same CDF as Figure 1.5, scaled differently to give the CDF of the sample average Z_n. It can be visualized that as n increases, the CDF of Z_n becomes increasingly close to a unit step at the mean, 0.25, of the variables X being summed.

central idea in a more fundamental, and usually more useful, way. We start the development by applying the Chebyshev inequality (1.58) to the sample average,

$$\Pr\left\{\left|\frac{S_n}{n} - \overline{X}\right| > \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}. \tag{1.74}$$

This is an upper bound on the probability that S_n/n differs by more than ϵ from its mean, $\overline{X}$. This is illustrated in Figure 1.9 which shows the CDF of S_n/n for various n. The figure suggests that $\lim_{n\to\infty} \mathsf{F}_{S_n/n}(z) = 0$ for all $z < \overline{X}$ and $\lim_{n\to\infty} \mathsf{F}_{S_n/n}(z) = 1$ for all $z > \overline{X}$. This is stated more cleanly in the following weak law of large numbers, abbreviated WLLN

Theorem 1.7.1 (WLLN with finite variance). *For each integer $n \geq 1$, let $S_n = X_1 + \cdots + X_n$ be the sum of n IID rv's with a finite variance. Then*

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{S_n}{n} - \overline{X}\right| > \epsilon\right\} = 0 \qquad \textit{for every } \epsilon > 0. \tag{1.75}$$

Proof: For every $\epsilon > 0$, $\Pr\{|S_n/n - \overline{X}| > \epsilon\}$ is bounded between 0 and $\sigma^2/n\epsilon^2$. Since the upper bound goes to 0 with increasing n, the theorem is proved. □

Discussion: The algebraic proof above is both simple and rigorous. However, the graphical description in Figure 1.10 probably provides more intuition about how the limit takes place. It is important to understand both.

We refer to (1.75) as saying that S_n/n converges to $\overline{X}$ in probability. To make sense out of this, we should view $\overline{X}$ as a deterministic random variable, *i.e.*, a rv that takes the value $\overline{X}$ for each sample point of the space. Then (1.75) says that the probability that the absolute

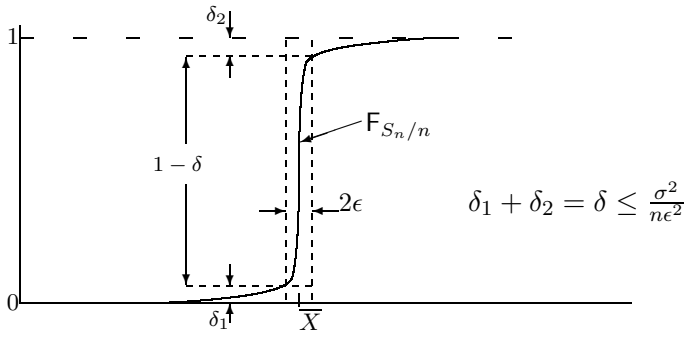


Figure 1.10: Approximation of the CDF $\mathsf{F}_{S_n/n}$ of a sample average by a step function at the mean: From (1.74), the probability δ that S_n/n differs from $\overline{X}$ by more than ϵ (*i.e.*, $\Pr\{|S_n/n-\overline{X}|\ge\epsilon\}$) is at most $\sigma^2/n\epsilon^2$. The complementary event, where $|S_n/n-\overline{X}|<\epsilon$, has probability $1-\delta\ge 1-\sigma^2/n\epsilon^2$. This means that we can construct a rectangle of width 2ϵ centered on $\overline{X}$ and of height $1-\delta$ such that $\mathsf{F}_{S_n/n}$ enters the rectangle at the lower left (say at $(\overline{X}-\epsilon,\delta_1)$) and exits at the upper right, say at $(\overline{X}+\epsilon,1-\delta_2)$). Now visualize increasing n while holding ϵ fixed. In the limit, $1-\delta\to 1$ so $\Pr\{|S_n/n-\overline{X}|\ge\epsilon\}\to 0$. Since this is true for every $\epsilon>0$ (usually with slower convergence as ϵ gets smaller), $\mathsf{F}_{S_n/n}(z)$ approaches 0 for every $z<\overline{X}$ and approaches 1 for every $z>\overline{X}$, *i.e.*, $\mathsf{F}_{S_n/n}$ approaches a unit step at $\overline{X}$. Note that there are two 'fudge factors' here, ϵ and δ and, since we are approximating an entire CDF, neither can be omitted, except by directly going to a limit as $n\to\infty$.

difference, $|S_n/n-\overline{X}|$, exceeds any given $\epsilon>0$ goes to 0 as $n\to\infty$.[28]

One should ask at this point what (1.75) adds to the more specific bound in (1.74), which in fact provides an upper bound on the rate of convergence for the limit in (1.75). The answer is that (1.75) remains valid when the theorem is generalized. For variables that are not IID or have an infinite variance, (1.74) is no longer necessarily valid. In some situations, as we see later, it is valuable to know that (1.75) holds, even if the rate of convergence is extremely slow or unknown.

One difficulty with the bound in (1.74) is that it is extremely loose in most cases. If S_n/n actually approached $\overline{X}$ this slowly, the weak law of large numbers would often be more a mathematical curiosity than a highly useful result. If we assume that the MGF of X exists in an open interval around 0, then (1.74) can be strengthened considerably. Note that the event $\{|S_n/n-\overline{X}|\ge\epsilon\}$ can be expressed as the union of the events $\{S_n/n\ge\overline{X}+\epsilon\}$ and $\{-S_n/n\ge\mathsf{E}[-X]+\epsilon\}$. Recall from Lemma 1.6.1 that for any $\epsilon>0$,

$$\Pr\{S_n/n\ge\mathsf{E}[X]+\epsilon\} \le \exp(n\mu_X(\mathsf{E}[X]+\epsilon)) \tag{1.76}$$

$$\Pr\{-S_n/n\ge\mathsf{E}[-X]+\epsilon\} \le \exp(n\mu_{(-X)}(\mathsf{E}[-X]+\epsilon)), \tag{1.77}$$

where both these bounds are exponentially decreasing in n. The two probabilities being bounded here are the quantities marked δ_2 and δ_1 in Figure 1.10. We see then that when

[28]Saying this in words gives one added respect for mathematical notation, and perhaps in this case, it is preferable to simply understand the mathematical statement (1.75).

the MGF exists in a region around $\overline{X}$, the convergence in δ for fixed ϵ is exponential in n rather than harmonic as suggested in Figure 1.10.

Each of these upper bounds are exponentially decreasing in n where from Lemma 1.6.1, $\mu_X(a) = \inf_r\{\gamma_X(r) - ra\} < 0$ for $a \neq \overline{X}$. Thus, for any $\epsilon > 0$,

$$\Pr\big\{|S_n/n - \overline{X}| \geq \epsilon\big\} \leq \exp[n\mu_X(\overline{X} + \epsilon)] + \exp[n\mu_X(\overline{X} - \epsilon)]. \tag{1.78}$$

The bound here, for any fixed $\epsilon > 0$, decreases geometrically in n rather than harmonically. In terms of Figure 1.10, the height of the rectangle must approach 1 at least geometrically in n.

1.7.2 Relative frequency

We next show that the WLLN applies to the relative frequency of an event as well as to the sample average of a random variable. Suppose that A is some event in a single experiment, and that the experiment is independently repeated n times. Then, in the probability model for the n repetitions, let A_i be the event that A occurs at the ith trial, $1 \leq i \leq n$. The events $A_1, A_2, \ldots, A_n$ are then IID.

If we let $\mathbb{I}_{A_i}$ be the indicator rv for A on the ith trial, then the rv $S_n = \mathbb{I}_{A_1} + \mathbb{I}_{A_2} + \cdots + \mathbb{I}_{A_n}$ is the number of occurrences of A over the n trials. It follows that

$$\text{relative frequency of } A = \frac{S_n}{n} = \frac{\sum_{i=1}^{n} \mathbb{I}_{A_i}}{n}. \tag{1.79}$$

Thus the relative frequency of A is the sample average of the binary rv's $\mathbb{I}_{A_i}$, and everything we know about the sum of IID rv's applies equally to the relative frequency of an event. In fact, everything we know about sums of IID *binary* rv's applies to relative frequency.

1.7.3 The central limit theorem

The weak law of large numbers says that with high probability, S_n/n is close to $\overline{X}$ for large n, but it establishes this via an upper bound on the tail probabilities rather than an estimate of what $\mathsf{F}_{S_n/n}$ looks like. If we look at the shape of $\mathsf{F}_{S_n/n}$ for various values of n in the example of Figure 1.9, we see that the function $\mathsf{F}_{S_n/n}$ becomes increasingly compressed around $\overline{X}$ as n increases (in fact, this is the essence of what the weak law is saying). If we normalize the random variable S_n/n to 0 mean and unit variance, we get a normalized rv, $Z_n = (S_n/n - \overline{X})\sqrt{n}\,/\sigma$. The CDF of Z_n is illustrated in Figure 1.11 for the same underlying X as used for S_n/n in Figure 1.9. The curves in the two figures are the same except that each curve has been horizontally scaled by $\sqrt{n}$ in Figure 1.11.

Inspection of Figure 1.11 shows that the normalized CDF's there seem to be approaching a limiting distribution. The critically important *central limit theorem* states that there is indeed such a limit, and it is the normalized Gaussian CDF.

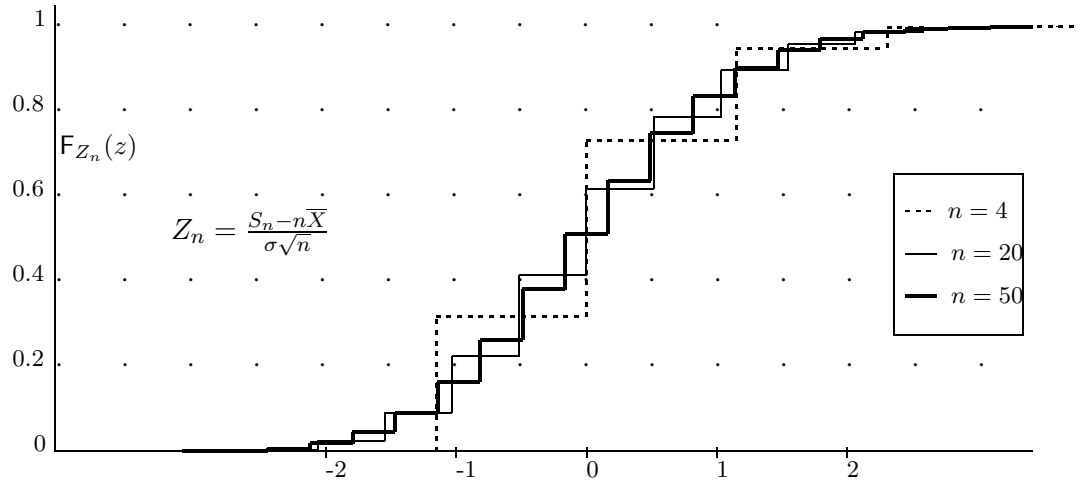


Figure 1.11: The same CDF's as Figure 1.5 normalized to 0 mean and unit standard deviation, *i.e.*, the CDF's of $Z_n = (S_n/n - \overline{X})\frac{\sqrt{n}}{\sigma}$ for $n = 4$, 20, 50. Note that as n increases, the CDF of Z_n slowly starts to resemble the normal CDF.

Theorem 1.7.2 (Central limit theorem (CLT)). *Let $X_1, X_2, \ldots$ be IID rv's with finite mean $\overline{X}$ and finite variance σ^2. Then for every real number z,*

$$\lim_{n\to\infty} \Pr\left\{\frac{S_n - n\overline{X}}{\sigma\sqrt{n}} \leq z\right\} \;=\; \Phi(z), \tag{1.80}$$

where $\Phi(z)$ is the normal CDF, i.e.*, the Gaussian distribution with mean 0 and variance 1,*

$$\Phi(z) \;=\; \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\frac{y^2}{2})\, dy.$$

Discussion: The rv $Z_n = (S_n - n\overline{X})/(\sigma\sqrt{n})$, for each $n \geq 1$ on the left side of (1.80), has mean 0 and variance 1. The central limit theorem (CLT), as expressed in (1.80), says that the sequence of CDF's, $\mathsf{F}_{Z_1}(z), \mathsf{F}_{Z_2}(z), \ldots$ converges at each value of z to $\Phi(z)$ as $n \to \infty$. In other words, $\lim_{n\to\infty} \mathsf{F}_{Z_n}(z) = \Phi(z)$ for each $z \in \mathbb{R}$. This is called *convergence in distribution*, since it is the sequence of CDF's, rather than the sequence of rv's that is converging. The theorem is illustrated by Figure 1.11.

The CLT tells us quite a bit about how $\mathsf{F}_{S_n/n}$ converges to a step function at $\overline{X}$. To see this, rewrite (1.80) in the form

$$\lim_{n\to\infty} \Pr\left\{\frac{S_n}{n} - \overline{X} \leq \frac{\sigma z}{\sqrt{n}}\right\} \;=\; \Phi(z). \tag{1.81}$$

This is illustrated in Figure 1.12 where we have used the Gaussian distribution as an approximation for the probability on the left.

The reason why the word *central* appears in the CLT can also be seen from (1.81). Asymptotically, we are looking at a limit (as $n \to \infty$) of the probability that the sample average

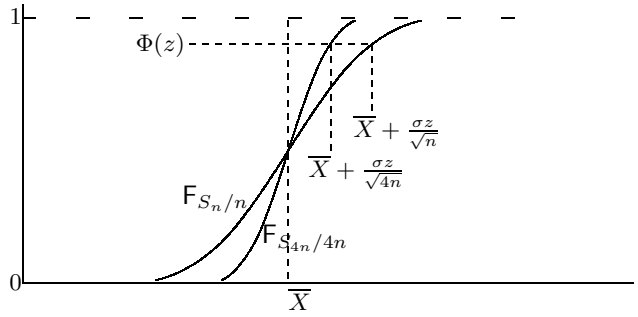


Figure 1.12: Approximation of the CDF $\mathsf{F}_{S_n/n}$ of a sample average by a Gaussian distribution of the same mean and variance. The WLLN says that $F_{S_n/n}$ approaches a step function at $\overline{X}$ as $n \to \infty$. The CLT says that the shape of S_n/n approximates a Gaussian distribution $\mathcal{N}(\overline{X}, \sigma^2/n)$. If n is increased by a factor of 4, the Gaussian distribution curve is horizontally scaled inward toward $\overline{X}$ by a factor of 2. The CLT says both that the CDF of S_n/n is scaled horizontally as $1/\sqrt{n}$ and also that it is better approximated by the Gaussian of the given mean and variance as n increases.

differs from the mean by at most a quantity going to 0 as $1/\sqrt{n}$. This should be contrasted with the corresponding optimized Chernoff bound in (1.78) which looks at the limit of the probability that the sample average differs from the mean by at most a constant amount. Those latter results are exponentially decreasing in n and are known as large deviation results.

Theorem 1.7.2 says nothing about the rate of convergence to the normal distribution. The Berry-Esseen theorem (see, for example, Feller, [9]) provides some guidance about this for cases in which the third central moment $\mathsf{E}\left[|X - \overline{X}|^3\right]$ exists. This theorem states that

$$\left| \Pr\left\{ \frac{(S_n - n\overline{X})}{\sigma\sqrt{n}} \leq z \right\} - \Phi(z) \right| \;\leq\; \frac{C\mathsf{E}\left[|X - \overline{X}|^3\right]}{\sigma^3\sqrt{n}}. \tag{1.82}$$

where C can be upper bounded by 0.766 (later improved to 0.4784). We will come back shortly to discuss convergence in greater detail.

The CLT helps explain why Gaussian rv's play such a central role in probability theory. In fact, many of the cookbook formulas of elementary statistics are based on the tacit assumption that the underlying variables are Gaussian, and the CLT helps explain why these formulas often give reasonable results.

One should be careful to avoid reading more into the CLT than it says. For example, the normalized sum, $(S_n - n\overline{X})/\sigma\sqrt{n}$ need not have a density that is approximately Gaussian. In fact, if the underlying variables are discrete, the normalized sum is discrete and has no density. The PMF of the normalized sum might have very detailed and wild fine structure; this does not disappear as n increases, but becomes "integrated out" in the CDF.

A proof of the CLT requires mathematical tools that will not be needed subsequently.[29] Thus we give a proof only for the binomial case. Before doing this, however, we will show

[29] Many elementary texts provide 'simple proofs,' using transform techniques. These proofs can be correct

that the PMF for S_n in the binomial approaches a sampled form of the Gaussian density. This detailed form of the PMF does not follow from the CLT and is often valuable in its own right.

Theorem 1.7.3. *Let $\{X_i; i \geq 1\}$ be a sequence of IID binary rv's with $p = \mathsf{p}_X(1) > 0$ and $q = 1-p = \mathsf{p}_X(0) > 0$. Let $S_n = X_1 + \cdots + X_n$ for each $n \geq 1$ and let α be a fixed constant satisfying $1/2 < \alpha < 2/3$. Then constants C and n_o exist such that for all integers k such that $|k - np| \leq n^\alpha$,*

$$p_{S_n}(k) \;=\; \frac{1 \pm Cn^{3\alpha-2}}{\sqrt{2\pi npq}} \exp\left(\frac{-(k-np)^2}{2npq}\right) \qquad \text{for } n \geq n_o, \tag{1.83}$$

where this 'equation' is to be interpreted as an upper bound when the $\pm$ sign is replaced with $+$ and a lower bound when $\pm$ is replaced with $-$.

Note that $n^{3\alpha-2}$ goes to 0 with increasing n for $\alpha < 2/3$, so the ratio between the upper and lower bound in (1.83) approaches 1 at the rate $n^{3\alpha-2}$. This is independent of k within the range $|k-n| \leq n^\alpha$. We will see why this fussiness is required when we go from the PMF to the CDF.

Proof: Let $\tilde{p} = k/n$ and $\tilde{q} = 1 - \tilde{p}$. From (1.27) we have $\mathsf{p}_{S_n}(\tilde{p}n) \sim \frac{1}{\sqrt{2\pi n\tilde{p}\tilde{q}}} \exp -n\mathsf{D}(\tilde{p}\|p)$ where $\mathsf{D}(\tilde{p}\|p) = \tilde{p}\ln(\tilde{p}/p) + \tilde{q}\ln(\tilde{q}/q)$. The ratio between $\mathsf{p}_{S_n}(k)$ and its asymptotic value is bounded in (1.25) and (1.26) by $1 - 1/(12n\tilde{p}(1-\tilde{p}))$. Since $0 < p < 1$ and $\tilde{p}$ approaches p with increasing n, we can replace this asymptotic form by

$$\mathsf{p}_{S_n}(\tilde{p}n) = \frac{1 \pm C_1 n^{-1}}{\sqrt{2\pi n\tilde{p}\tilde{q}}} \exp -n\mathsf{D}(\tilde{p}\|p) \qquad \text{for large enough } C_1 \text{ and } n \tag{1.84}$$

The remainder of the proof shows that $\mathsf{D}(\tilde{p}\|p)$ approaches the exponential term in (1.83) as $n \to \infty$. Let $\epsilon = \tilde{p} - p$. Then

$$\begin{aligned}
\mathsf{D}(\tilde{p}\|p) &= (p+\epsilon)\ln\left(1 + \frac{\epsilon}{p}\right) + (q-\epsilon)\ln\left(1 - \frac{\epsilon}{q}\right) \\
&= \frac{\epsilon^2}{2p} - \frac{\epsilon^3}{6p^2} + \cdots + \frac{\epsilon^2}{2q} + \frac{\epsilon^3}{6q^2}\cdots \qquad (1.85) \\
&= \frac{\epsilon^2}{2pq} - \frac{\epsilon^3}{6p^2} + \cdots + \frac{\epsilon^3}{6q^2}\cdots, \qquad (1.86)
\end{aligned}$$

where in (1.85) we have used the power series expansion, $\ln(1+u) = u - u^2/2 + u^3/3\cdots$. In (1.86), we used the fact that $1/p + 1/q = 1/pq$. Substituting this into the exponential term of (1.84),

$$\begin{aligned}
\exp -n\mathsf{D}(\tilde{p}\|p) &= \exp\left(-\frac{n\epsilon^2}{2pq}\right)\exp\left(\frac{n\epsilon^3}{6p^2} + \cdots - \frac{n\epsilon^3}{6q^2} + \cdots\right) \qquad (1.87) \\
&= \exp\left(-\frac{n\epsilon^2}{2pq}\right)(1 \pm C_2 n^{-3\alpha-2}), \qquad (1.88)
\end{aligned}$$

if they quote powerful enough transform results, but mny of them indicate that the normalized sum has a density that approaches the Gaussian density; this is incorrect for all discrete rv's. A direct proof is given by Feller ([8] and [9]).

where in (1.88) we used, first, the fact that the neglected terms in (1.87) are decreasing at least geometrically with ϵ, second, the condition in the theorem that $|\epsilon| = |k-np|/n \leq n^{\alpha-1}$, and third, the expansion $e^x = 1 + x + x^2/2 + \cdots$. Substituting this into (1.84), we get

$$\mathsf{p}_{S_n}(\tilde{p}n) = \frac{(1 \pm C_1 n^{-1})(1 \pm C_2 n^{-3\alpha+2})}{\sqrt{2\pi n \tilde{p}\tilde{q}}} \exp\left(-\frac{n\epsilon^2}{2pq}\right)$$

Finally, in the square root term in the denominator, we can represent $\tilde{p}\tilde{q}$ as $pq(1 \pm n^{-1+\alpha})$. Then, combining all the approximation terms, we get (1.83). □

Proof of Theorem 1.7.2 (binomial case): The central limit theorem (for the binary case) in the form of Theorem 1.7.2 simply converts the PMF of (1.83) into a CDF. To do this carefully, we first choose arbitrary real numbers $z' < z$ and show that

$$\lim_{n\to\infty} \Pr\left\{z' \leq \frac{S_n - n\overline{X}}{\sigma\sqrt{n}} \leq z\right\} = \int_{z'}^{z} \frac{1}{\sqrt{2\pi}} \exp\frac{-y^2}{2}\, dy\,. \tag{1.89}$$

Choose $\alpha \in (\frac{1}{2}, \frac{2}{3})$ and for that α, choose n_o and C to satisfy (1.83). Then choose m to satisfy $m \geq n_o$ and $-m^\alpha \leq z'\sqrt{mpq} < z\sqrt{mpq} \leq m^\alpha$. Since $\alpha > 1/2$, this guarantees that

$$-n^\alpha \leq z'\sqrt{npq} < z\sqrt{npq} \leq n^\alpha \qquad \text{for } n \geq m. \tag{1.90}$$

For $n \geq m$, we then have

$$\begin{aligned}
\Pr\left\{z' \leq \frac{S_n - np}{\sqrt{npq}} \leq z\right\} &= \sum_{k=\lceil np+z'\sqrt{npq}\rceil}^{\lfloor np+z\sqrt{npq}\rfloor} \mathsf{p}_{S_n}(k) \qquad (1.91)\\
&= \left(1 \pm Cn^{3\alpha-2}\right) \left(\sum_{k=\lceil np+z'\sqrt{npq}\rceil}^{\lfloor np+z\sqrt{npq}\rfloor} \frac{1}{\sqrt{2\pi npq}} \exp\left(\frac{-(k-np)^2}{2npq}\right)\right).
\end{aligned}$$

As seen from (1.90), each term in (1.91) satisfies $|k - np| \leq n^\alpha$, which justifies the bounds in the following sum. That following sum can be viewed as a Riemann sum for the integral in (1.80). Thus the sum approaches the integral as $n^{-1/2}$. Taking the limit as $n \to \infty$ in (1.91), the term $Cn^{3\alpha-2}$ approaches 0, justifying (1.89). The theorem follows by taking the limit $z' \to -\infty$. □

Since the CLT provides such explicit information about the convergence of S_n/n to $\overline{X}$, it is reasonable to ask why the weak law of large numbers (WLLN) is so important. The first reason is that the WLLN is so simple that it can be used to give clear insights into situations where the CLT could confuse the issue. A second reason is that the CLT requires a variance, where as we see next, the WLLN does not. A third reason is that the WLLN can be extended to many situations in which the variables are not independent and/or not identically distributed.[30] A final reason is that the WLLN provides an upper bound on the tails of $\mathsf{F}_{S_n/n}$, whereas the CLT provides only an approximation.

[30]Central limit theorems also hold in many of these more general situations, but they do not hold as widely as the WLLN.

1.7.4 Weak law with an infinite variance

We now establish the WLLN without assuming a finite variance.

Theorem 1.7.4 (WLLN). *For each integer $n \geq 1$, let $S_n = X_1 + \cdots + X_n$ where $X_1, X_2, \ldots$ are IID rv's satisfying $\mathsf{E}[|X|] < \infty$. Then for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}[X]\right| > \epsilon\right\} = 0. \tag{1.92}$$

Proof:[31] We use a truncation argument; such arguments are used frequently in dealing with rv's that have infinite variance. The underlying idea in these arguments is important, but some less important details are treated in Exercise 1.41. Let b be a positive number (which we later take to be increasing with n), and for each variable X_i, define a new rv $\breve{X}_i$ (see Figure 1.13) by

$$\breve{X}_i = \begin{cases} X_i & \text{for} \quad \mathsf{E}[X] - b \leq X_i \leq \mathsf{E}[X] + b \\ \mathsf{E}[X] + b & \text{for} \quad X_i > \mathsf{E}[X] + b \\ \mathsf{E}[X] - b & \text{for} \quad X_i < \mathsf{E}[X] - b. \end{cases} \tag{1.93}$$

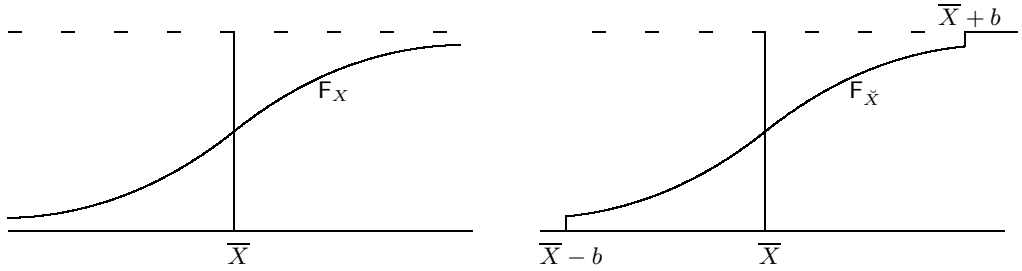


Figure 1.13: The truncated rv $\breve{X}$ for a given rv X has a CDF which is truncated at $\overline{X} \pm b$.

The truncated variables $\breve{X}_i$ are IID and, because of the truncation, must have a finite second moment. Thus the WLLN applies to the sample average $\breve{S}_n = \breve{X}_1 + \cdots \breve{X}_n$. More particularly, using the Chebshev inequality in the form of (1.74) on $\breve{S}_n/n$, we get

$$\Pr\left\{\left|\frac{\breve{S}_n}{n} - \mathsf{E}\left[\breve{X}\right]\right| > \frac{\epsilon}{2}\right\} \leq \frac{4\sigma^2_{\breve{X}}}{n\epsilon^2} \leq \frac{8b\mathsf{E}[|X|]}{n\epsilon^2},$$

where Exercise 1.41 demonstrates the final inequality. Exercise 1.41 also shows that $\mathsf{E}\big[\breve{X}\big]$ approaches $\mathsf{E}[X]$ as $b \to \infty$ and thus that

$$\Pr\left\{\left|\frac{\breve{S}_n}{n} - \mathsf{E}[X]\right| > \epsilon\right\} \leq \frac{8b\mathsf{E}[|X|]}{n\epsilon^2}, \tag{1.94}$$

[31] The details of this proof can be omitted without loss of continuity. However, truncation arguments are important in many places and should be understood at some point.

for all sufficiently large b. This bound also applies to S_n/n in the case where $S_n = \breve{S}_n$, so we have the following bound (see Exercise 1.41 for further details):

$$\Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}[X]\right| > \epsilon\right\} \leq \Pr\left\{\left|\frac{\breve{S}_n}{n} - \mathsf{E}[X]\right| > \epsilon\right\} + \Pr\left\{S_n \neq \breve{S}_n\right\}. \tag{1.95}$$

The original sum S_n is the same as $\breve{S}_n$ unless one of the X_i has an outage, *i.e.*, $|X_i - \overline{X}| > b$. Thus, using the union bound, $\Pr\left\{S_n \neq \breve{S}_n\right\} \leq n\Pr\{|X_i - \overline{X}| > b\}$. Substituting this and (1.94) into (1.95),

$$\Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}[X]\right| > \epsilon\right\} \leq \frac{8b\mathsf{E}[|X|]}{n\epsilon^2} + \frac{n}{b}\left[b\Pr\{\,|X - \mathsf{E}[X]\,| > b\}\right]. \tag{1.96}$$

We now show that for any $\epsilon > 0$ and $\delta > 0$, $\Pr\{|S_n/n - \overline{X}| \geq \epsilon\} \leq \delta$ for all sufficiently large n. We do this, for given ϵ, δ, by choosing $b(n)$ for each n so that the first term in (1.96) is equal to $\delta/2$. Thus $b(n) = n\delta\epsilon^2/16\mathsf{E}[|X|]$. This means that $n/b(n)$ in the second term is independent of n. Now from (1.57), $\lim_{b\to\infty} b\Pr\{|X - \overline{X}| > b\} = 0$, so by choosing $b(n)$ sufficiently large (and thus n sufficiently large), the second term in (1.96) is also at most $\delta/2$. □

1.7.5 Convergence of random variables

This section has developed a number of results about how the sequence of sample averages, $\{S_n/n;\ n \geq 1\}$, for a sequence of IID rv's $\{X_i;\ i \geq 1\}$, approaches the mean $\overline{X}$. In the case of the CLT, the limiting distribution around the mean is also specified to be Gaussian. At the outermost intuitive level, *i.e.*, at the level most useful when first looking at some very complicated set of issues, viewing the limit of the sample averages as being essentially equal to the mean is highly appropriate.

At the next intuitive level down, the meaning of the word *essentially* becomes important and thus involves the details of the above laws. All of the results involve how the rv's S_n/n change with n and become better and better approximated by $\overline{X}$. When we talk about a sequence of rv's (namely a sequence of functions on the sample space) being approximated by a rv or numerical constant, we are talking about some kind of *convergence*, but it clearly is not as simple as a sequence of real numbers (such as $1/n$ for example) converging to some given number (0 for example).

The purpose of this section, is to give names and definitions to these various forms of convergence. This will give us increased understanding of the laws of large numbers already developed, but, equally important, it will allow us to develop another law of large numbers called the *strong law of large numbers* (SLLN). Finally, it will put us in a position to use these convergence results later for sequences of rv's other than the sample averages of IID rv's.

We discuss four types of convergence in what follows, convergence in distribution, in probability, in mean square, and with probability 1. For the first three, we first recall the type of large-number result with that type of convergence and then give the general definition.

For convergence with probability 1 (WP1), we will define this type of convergence and then provide some understanding of what it means. This will then be used in Chapter 5 to state and prove the SLLN.

We start with the central limit theorem, which, from (1.80) says

$$\lim_{n\to\infty} \Pr\left\{\frac{S_n - n\overline{X}}{\sqrt{n}\,\sigma} \le z\right\} = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) dx \qquad \text{for every } z \in \mathbb{R}.$$

This is illustrated in Figure 1.11 and says that the sequence (in n) of CDF's $\Pr\left\{\frac{S_n - n\overline{X}}{\sqrt{n}\sigma} \le z\right\}$ converges at every z to the normal CDF at z. This is an example of *convergence in distribution*.

Definition 1.7.1. *A sequence of random variables, $Z_1, Z_2, \ldots$, converges in distribution to a random variable Z if $\lim_{n\to\infty} \mathsf{F}_{Z_n}(z) = \mathsf{F}_Z(z)$ at each z for which $\mathsf{F}_Z(z)$ is continuous.*

For the CLT example, the rv's that converge in distribution are $\{\frac{S_n - n\overline{X}}{\sqrt{n}\sigma};\ n \ge 1\}$, and they converge in distribution to the normal Gaussian rv.

Convergence in distribution does not say that the rv's themselves converge in any reasonable sense, but only that their CDF's converge. For example, let $Y_1, Y_2, \ldots$ be IID rv's with the CDF F_Y. For each $n \ge 1$, if we let let $Z_n = Y_n + 1/n$, then it is easy to see that $\{Z_n;\ n \ge 1\}$ converges in distribution to Y. However (assuming Y has variance σ_Y^2 and is independent of each Z_n), we see that $Z_n - Y$ has variance $2\sigma_Y^2$. Thus Z_n does not get close to Y as $n \to \infty$ in any reasonable sense, and $Z_n - Z_m$ does not get small as n and m both get large.[32] As an even more trivial example, the IID sequence $\{Y_n;\ n \ge 1\}$ converges in distribution to Y.

For the CLT, it is the rv's $\frac{S_n - n\overline{X}}{\sqrt{n}\sigma}$ that converge in distribution to the normal. As shown in Exercise 1.44, however, the rv $\frac{S_n - n\overline{X}}{\sqrt{n}\sigma} - \frac{S_{2n} - 2n\overline{X}}{\sqrt{2n}\sigma}$ is not close to 0 in any reasonable sense, even though the two terms have CDF's that are very close for large n.

For the next type of convergence of rv's, the WLLN, in the form of (1.92), says that

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{S_n}{n} - \overline{X}\right| > \epsilon\right\} = 0 \qquad \text{for every } \epsilon > 0.$$

This is an example of *convergence in probability,* as defined below:

Definition 1.7.2. *A sequence of random variables $Z_1, Z_2, \ldots$ converges in probability to a rv Z if $\lim_{n\to\infty} \Pr\{|Z_n - Z| > \epsilon\} = 0$ for every $\epsilon > 0$.*

For the WLLN example, Z_n in the definition is the sample average S_n/n and Z is the constant rv $\overline{X}$. It is probably simpler and more intuitive in thinking about convergence of rv's to think of the sequence of rv's $\{Y_n = Z_n - Z;\ n \ge 1\}$ as converging to 0 in some

[32] In fact, saying that a sequence of rv's converges in distribution is unfortunate but standard terminology. It would be just as concise, and far less confusing, to say that a sequence of CDF's converge rather than saying that a sequence of rv's converge in distribution.

sense.[33] As illustrated in Figure 1.9, convergence in probability means that $\{Y_n; n \geq 1\}$ converges in distribution to a unit step function at 0.

An equivalent statement, as illustrated in Figure 1.10, is that $\{Y_n; n \geq 1\}$ converges in probability to 0 if $\lim_{n\to\infty} \mathsf{F}_{Y_n}(y) = 0$ for all $y < 0$ and $\lim_{n\to\infty} \mathsf{F}_{Y_n}(y) = 1$ for all $y > 0$. This shows that convergence in probability to a constant is a special case of convergence in distribution, since with convergence in probability, the sequence F_{Y_n} of CDF's converges to a unit step at 0. Note that $\lim_{n\to\infty} \mathsf{F}_{Y_n}(y)$ is not specified at $y = 0$. However, the step function is not continuous at 0, so the limit there need not be specified for convergence in distribution.

Convergence in probability says quite a bit more than convergence in distribution. As an important example of this, consider the difference $Y_n - Y_m$ for n and m both large. If $\{Y_n; n \geq 1\}$ converges in probability to 0, then Y_n and Y_m are both close to 0 with high probability for large n and m, and thus close to each other. More precisely, $\lim_{m\to\infty,n\to\infty} \Pr\{|Y_n - Y_m| > \epsilon\} = 0$ for every $\epsilon > 0$. If the sequence $\{Y_n; n \geq 1\}$ merely converges in distribution to some arbitrary distribution, then, as we saw, $Y_n - Y_m$ can be large with high probability, even when n and m are large. Another example of this is given in Exercise 1.44.

It appears paradoxical that the CLT is more explicit about the convergence of S_n/n to $\overline{X}$ than the weak law, but it corresponds to a weaker type of convergence. The resolution of this paradox is that the sequence of rv's in the CLT is $\{\frac{S_n - n\overline{X}}{\sqrt{n}\sigma}; n \geq 1\}$. The presence of $\sqrt{n}$ in the denominator of this sequence provides much more detailed information about how S_n/n approaches $\overline{X}$ with increasing n than the limiting unit step of $\mathsf{F}_{S_n/n}$ itself. For example, it is easy to see from the CLT that $\lim_{n\to\infty} \mathsf{F}_{S_n/n}(\overline{X}) = 1/2$, which can't be derived directly from the weak law.

Yet another kind of convergence is *convergence in mean square* (MS). An example of this, for the sample average S_n/n of IID rv's with a variance, is given in (1.73), repeated below:

$$\lim_{n\to\infty} \mathsf{E}\left[\left(\frac{S_n}{n} - \overline{X}\right)^2\right] = 0.$$

The general definition is as follows:

Definition 1.7.3. *A sequence of rv's $Z_1, Z_2, \ldots$ converges in mean square (MS) to a rv Z if $\lim_{n\to\infty} \mathsf{E}\left[(Z_n - Z)^2\right] = 0$.*

Our derivation of the weak law of large numbers (Theorem 1.7.1) was essentially based on the MS convergence of (1.73). Using the same approach, Exercise 1.43 shows in general that convergence in MS implies convergence in probability. Convergence in probability does not imply MS convergence, since as shown in Theorem 1.7.4, the weak law of large numbers holds without the need for a variance.

Figure 1.14 illustrates the relationship between these forms of convergence, *i.e.*, mean-squared convergence implies convergence in probability, which in turn implies convergence

[33] Definition 1.7.2 gives the impression that convergence to a rv Z is more general than convergence to a constant or convergence to 0, but converting the rv's to $Y_n = Z_n - Z$ makes it clear that this added generality is quite superficial.

in distribution. The figure also shows convergence with probability 1 (WP1), which is the next form of convergence to be discussed.

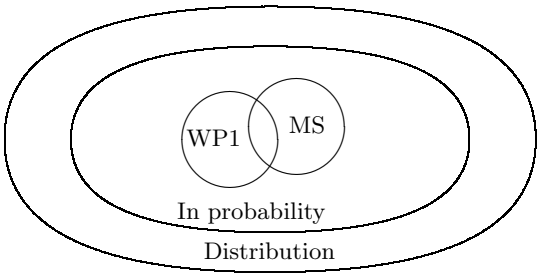


Figure 1.14: Relationship between different kinds of convergence: Convergence in distribution is the most general and is implied by all the others. Convergence in probability is the next most general and is implied both by convergence with probability 1 (WP1) and by mean square (MS) convergence, neither of which implies the other.

1.7.6 Convergence with probability 1

Convergence with probability 1, abbreviated as convergence WP1, is often referred to as convergence a.s. (almost surely) and convergence a.e. (almost everywhere). The strong law of large numbers, which is discussed briefly in this section and further discussed and proven in various forms in Chapters 5 and 9, provides an extremely important example of convergence WP1. The general definition is as follows:

Definition 1.7.4. *Let $Z_1, Z_2, \ldots$ be a sequence of rv's in a sample space Ω and let Z be another rv in Ω. Then $\{Z_n; n \geq 1\}$ is defined to converge to Z with probability 1 (WP1) if*

$$\Pr\Big\{\omega \in \Omega \;:\; \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\Big\} = 1. \tag{1.97}$$

The condition $\Pr\{\omega \in \Omega : \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\} = 1$ is often stated more compactly as $\Pr\{\lim_n Z_n = Z\} = 1$, and even more compactly as $\lim_n Z_n = Z$ WP1, but the form here is the simplest for initial understanding. As discussed in Chapter 5. the SLLN says that if $X_1, X_2, \ldots$ are IID with $\mathsf{E}[|X|] < \infty$, then the sequence of sample averages, $\{S_n/n; n \geq 1\}$ converges WP1 to $\overline{X}$.

In trying to understand (1.97), note that each sample point ω of the underlying sample space Ω maps to a sample value $Z_n(\omega)$ of each rv Z_n, and thus maps to a sample path $\{Z_n(\omega);\ n \geq 1\}$. For any given ω, such a sample path is simply a sequence of real numbers. That sequence of real numbers might converge to $Z(\omega)$ (which is a real number for the given ω), it might converge to something else, or it might not converge at all. Thus a set of ω exists for which the corresponding sample path $\{Z_n(\omega); n \geq 1\}$ converges to $Z(\omega)$, and a second set for which the sample path converges to something else or does not converge at all. Convergence WP1 of the sequence of rv's is thus defined to occur when the first set of sample paths above is an event that has probability 1.

For each ω, the sequence $\{Z_n(\omega);\ n \geq 1\}$ is simply a sequence of real numbers, so we briefly review what the limit of such a sequence is. A sequence of real numbers $b_1, b_2, \ldots$ is said to have a limit b if, for every $\epsilon > 0$, there is an integer m_ϵ such that $|b_n - b| \leq \epsilon$ for all $n \geq m_\epsilon$. An equivalent statement is that $b_1, b_2, \ldots$ has a limit b if, for every integer $k \geq 1$, there is an integer $m(k)$ such that $|b_n - b| \leq 1/k$ for all $n \geq m(k)$.

Figure 1.15 illustrates this definition for those, like the author, whose eyes blur on the second or third 'there exists', 'such that', etc. in a statement. As illustrated, an important aspect of convergence of a sequence $\{b_n;\ n \geq 1\}$ of real numbers is that b_n becomes close to b for large n and stays close for all sufficiently large values of n.

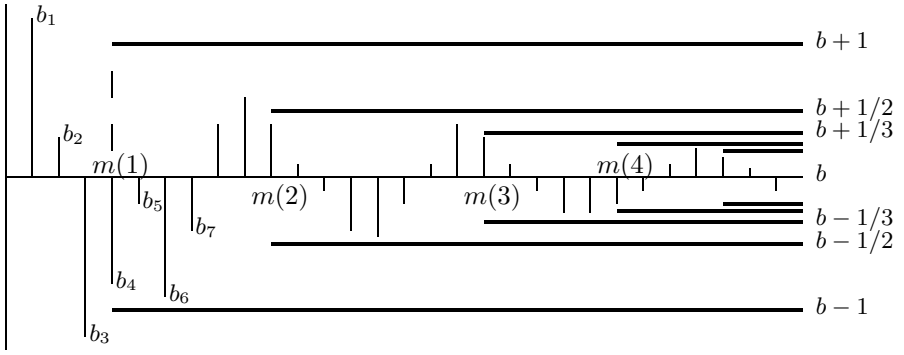


Figure 1.15: Illustration of a sequence of real numbers $b_1, b_2, \ldots$ that converge to a number b. The figure illustrates an integer $m(1)$ such that for all $n \geq m(1)$, b_n lies in the interval $b \pm 1$. Similarly, for each $k \geq 1$, there is an integer $m(k)$ such that b_n lies in $b \pm 1/k$ for all $n \geq m(k)$. Thus $\lim_{n\to\infty} b_n = b$ means that for a sequence of ever tighter constraints, the kth constraint can be met for all sufficiently large n, (*i.e.*, all $n \geq m(k)$). Intuitively, convergence means that the elements $b_1, b_2, \ldots$ get close to b *and stay close*. The sequence of positive integers $m(1), m(2), \ldots$ is non-decreasing, but otherwise arbitrary, depending only on the sequence $\{b_n;\ n \geq 1\}$. For sequences that converge very slowly, the integers $m(1), m(2), \ldots$ are simply correspondingly larger.

Figure 1.16 gives an example of a sequence of real numbers that does not converge. Intuitively, this sequence is close to 0 (and in fact identically equal to 0) for most large n, but it doesn't stay close, because of ever more rare outages.

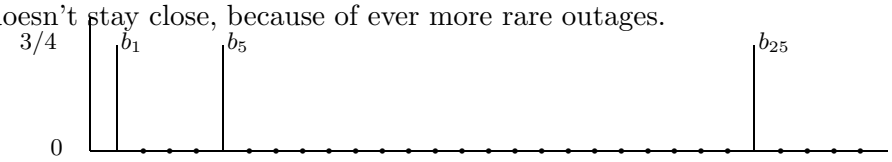


Figure 1.16: Illustration of a non-convergent sequence of real numbers $b_1, b_2, \ldots$. The sequence is defined by $b_n = 3/4$ for $n = 1,\ 5,\ 25, \ldots, 5^j, \ldots$ for all integer $j \geq 0$. For all other n, $b_n = 0$. The terms for which $b_n \neq 0$ become increasingly rare as $n \to \infty$. Note that $b_n \in [-1, 1]$ for all n, but there is no $m(2)$ such that $b_n \in [-\frac{1}{2}, \frac{1}{2}]$ for all $n \geq m(2)$. Thus the sequence does not converge.

The following example illustrates how a sequence of rv's can converge in probability but

not converge WP1. The example also provides some clues as to why convergence WP1 is important.

Example 1.7.1. Consider a sequence $\{Y_n;\, n \geq 1\}$ of rv's for which the sample paths constitute the following slight variation of the sequence of real numbers in Figure 1.16. In particular, as illustrated in Figure 1.17, the non-zero term at $n = 5^j$ in Figure 1.16 is replaced by a non-zero term at a randomly chosen n in the interval[34] $[5^j,\, 5^{j+1})$.

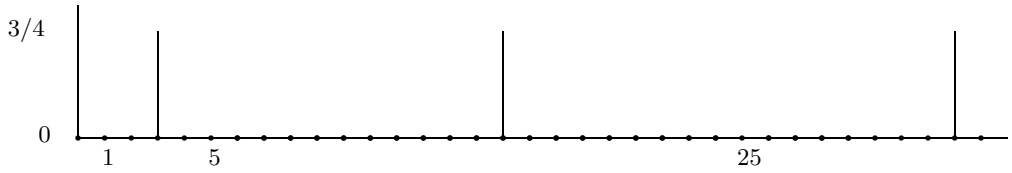


Figure 1.17: Illustration of a sample path of a sequence of rv's $\{Y_n; n \geq 0\}$ where, for each $j \geq 0$, $Y_n = 1$ for an equiprobable choice of $n \in [5^j, 5^{j+1})$ and $Y_n = 0$ otherwise.

Since each sample path contains a single one in each segment $[5^j,\, 5^{j+1})$, and contains zeroes elsewhere, none of the sample paths converge. In other words, $\Pr\{\omega : \lim Y_n(\omega) = 0\} = 0$. On the other hand $\Pr\{Y_n = 0\} = 1 - 5^{-j}$ for $5^j \leq n < 5^{j+1}$, so $\lim_{n\to\infty} \Pr\{Y_n = 0\} = 1$.

Thus this sequence of rv's converges to 0 in probability, but does not converge to 0 WP1. This sequence also converges in mean square and (since it converges in probability) in distribution. Thus we have shown (by example) that convergence WP1 is not implied by any of the other types of convergence we have discussed. We will show in Section 5.2 that convergence WP1 does imply convergence in probability and in distribution but not in mean square (as illustrated in Figure 1.14).

The interesting point in this example is that this sequence of rv's is not bizarre (although it is somewhat specialized to make the analysis simple). Another important point is that this definition of convergence has a long history of being accepted as the 'useful,' 'natural,' and 'correct' way to define convergence for a sequence of real numbers. Thus it is not surprising that convergence WP1 will turn out to be similarly useful for sequences of rv's.

There is a price to be paid in using the concept of convergence WP1. We must then look at the entire sequence of rv's and can no longer analyze finite n-tuples and then go to the limit as $n \to \infty$. This requires a significant additional layer of abstraction, which involves additional mathematical precision and initial loss of intuition. For this reason we put off further discussion of convergence WP1 and the SLLN until Chapter 5 where it is needed.

1.8 Relation of probability models to the real world

Whenever experienced and competent engineers or scientists construct a probability model to represent aspects of some system that either exists or is being designed for some application, they must acquire a deep knowledge of the system and its surrounding circumstances,

[34] There is no special significance to the number 5 here other than making the figure easy to visualize. We could replace 5 by 2 or 3 etc.

and concurrently consider various types of probability models used in probabilistic analyses of the same or similar systems. Usually very simple probability models help in understanding the real-world system, and knowledge about the real-world system helps in understanding what aspects of the system are well-modeled by a given probability model. For a text such as this, there is insufficient space to understand the real-world aspects of each system that might be of interest. We must use the language of various canonical real-world systems for motivation and insight when studying probability models for various classes of systems, but such models must necessarily be chosen more for their tutorial than practical value.

There is a danger, then, that readers will come away with the impression that analysis is more challenging and important than modeling. To the contrary, for work on real-world systems, modeling is almost always more difficult, more challenging, and more important than analysis. The objective here is to provide the necessary knowledge and insight about probabilistic models so that the reader can later combine this with a deep understanding of particular real application areas. This will result in a useful interactive use of models, analysis, and experimentation.

In this section, our purpose is not to learn how to model real-world problems, since, as said above, this requires deep and specialized knowledge of whatever application area is of interest. Rather it is to understand the following conceptual problem that was posed in Section 1.1. Suppose we have a probability model of some real-world experiment involving randomness in the sense expressed there. When the real-world experiment being modeled is performed, there is an outcome, which presumably is one of the outcomes of the probability model, but there is no observable probability.

It appears to be intuitively natural, for experiments that can be carried out repeatedly under essentially the same conditions, to associate the probability of a given event with the relative frequency of that event over many repetitions. We now have the background to understand this approach. We first look at relative frequencies within the probability model, and then within the real world.

1.8.1 Relative frequencies in a probability model

We have seen that for any probability model, an extended probability model exists for n IID idealized experiments of the original model. For any event A in the original model, the indicator function $\mathbb{I}_A$ is a random variable, and the relative frequency of A over n IID experiments is the sample average of n IID rv's each with the distribution of $\mathbb{I}_A$. From the weak law of large numbers, this relative frequency converges in probability to $\mathsf{E}[\mathbb{I}_A] = \Pr\{A\}$. By taking the limit $n \to \infty$, the strong law of large numbers says that the relative frequency of A converges with probability 1 to $\Pr\{A\}$.

In plain English, this says that for large n, the relative frequency of an event (in the n-repetition IID model) is essentially the same as the probability of that event. The word *essentially* is carrying a great deal of hidden baggage. For the weak law, for any $\epsilon, \delta > 0$, the relative frequency is within some ϵ of $\Pr\{A\}$ with a confidence level $1 - \delta$ whenever n is sufficiently large. For the strong law, the ϵ and δ are avoided, but only by looking directly at the limit $n \to \infty$. Despite the hidden baggage, though, relative frequency and

probability are related as indicated.

1.8.2 Relative frequencies in the real world

In trying to sort out if and when the laws of large numbers have much to do with real-world experiments, we should ignore the mathematical details for the moment and agree that for large n, the relative frequency of an event A over n IID trials of an idealized experiment is essentially $\Pr\{A\}$. We can certainly visualize a real-world experiment that has the same set of possible outcomes as the idealized experiment and we can visualize evaluating the relative frequency of A over n repetitions with large n. If that real-world relative frequency is essentially equal to $\Pr\{A\}$, and this is true for the various events A of greatest interest, then it is reasonable to hypothesize that the idealized experiment is a reasonable model for the real-world experiment, at least so far as those given events of interest are concerned.

One problem with this comparison of relative frequencies is that we have carefully specified a model for n IID repetitions of the idealized experiment, but have said nothing about how the real-world experiments are repeated. The IID idealized experiments specify that the conditional probability of A at one trial is the same no matter what the results of the other trials are. Intuitively, we would then try to isolate the n real-world trials so they don't affect each other, but this is a little vague. The following examples help explain this problem and several others in comparing idealized and real-world relative frequencies.

Example 1.8.1. Coin tossing: Tossing coins is widely used as a way to choose the first player in various games, and is also sometimes used as a primitive form of gambling. Its importance, however, and the reason for its frequent use, is its simplicity. When tossing a coin, we would argue from the symmetry between the two sides of the coin that each should be equally probable (since any procedure for evaluating the probability of one side should apply equally to the other). Thus since H and T are the only outcomes (the remote possibility of the coin balancing on its edge is omitted from the model), the reasonable and universally accepted model for coin tossing is that H and T each have probability $1/2$.

On the other hand, the two sides of a coin are embossed in different ways, so that the mass is not uniformly distributed. Also the two sides do not behave in quite the same way when bouncing off a surface. Each denomination of each currency behaves slightly differently in this respect. Thus, not only do coins violate symmetry in small ways, but different coins violate it in different ways.

How do we test whether this effect is significant? If we assume for the moment that successive tosses of the coin are well-modeled by the idealized experiment of n IID trials, we can essentially find the probability of H for a particular coin as the relative frequency of H in a sufficiently large number of independent tosses of that coin. This gives us slightly different relative frequencies for different coins, and thus slightly different probability models for different coins.

The assumption of independent tosses is also questionable. Consider building a carefully engineered machine for tossing coins and using it in a vibration-free environment. A standard coin is inserted into the machine in the same way for each toss and we count the number of heads and tails. Since the machine has essentially eliminated the randomness, we would

expect all the coins, or almost all the coins, to come up the same way — the more precise the machine, the less independent the results. By inserting the original coin in a random way, a single trial might have equiprobable results, but successive tosses are certainly not independent. The successive trials would be closer to independent if the tosses were done by a slightly inebriated individual who tossed the coins high in the air.

The point of this example is that there are many different coins and many ways of tossing them, and the idea that one model fits all is reasonable under some conditions and not under others. Rather than retreating into the comfortable world of theory, however, note that we can now find the relative frequency of heads for any given coin and essentially for any given way of tossing that coin.[35]

Example 1.8.2. Binary data: Consider the binary data transmitted over a communication link or stored in a data facility. The data is often a mixture of encoded voice, video, graphics, text, etc., with relatively long runs of each, interspersed with various protocols for retrieving the original non-binary data.

The simplest (and most common) model for this is to assume that each binary digit is 0 or 1 with equal probability and that successive digits are statistically independent. This is the same as the model for coin tossing after the trivial modification of converting $\{H, T\}$ into $\{0, 1\}$. This is also a rather appropriate model for designing a communication or storage facility, since all n-tuples are then equiprobable (in the model) for each n, and thus the facilities need not rely on any special characteristics of the data. On the other hand, if one wants to compress the data, reducing the required number of transmitted or stored bits per incoming bit, then a more elaborate model is needed.

Developing such an improved model would require finding out more about where the data is coming from — a naive application of calculating relative frequencies of n-tuples would probably not be the best choice. On the other hand, there are well-known data compression schemes that in essence track dependencies in the data and use them for compression in a coordinated way. These schemes are called *universal data-compression* schemes since they don't rely on a probability model. At the same time, they are best analyzed by looking at how they perform for various idealized probability models.

The point of this example is that choosing probability models often depends heavily on how the model is to be used. Models more complex than IID binary digits are usually based on what is known about the input processes. Measuring relative frequencies and associating them with probabilities is the basic underlying conceptual connection between real-world and models, but in practice this is essentially the relationship of last resort. For most of the applications we will consider, there is a long history of modeling to build on, with experiments as needed.

Example 1.8.3. Fable: In the year 2008, the financial structure of the USA failed and the world economy was brought to its knees. Much has been written about the role of greed

[35] We are not suggesting that distinguishing different coins for the sake of coin tossing is an important problem. Rather, we are illustrating that even in such a simple situation, the assumption of identically prepared experiments is questionable and the assumption of independent experiments is questionable. The extension to n repetitions of IID experiments is not necessarily a good model for coin tossing. In other words, one has to question not only the original model but also the n-repetition model.

on Wall Street and incompetence in Washington. Another aspect of the collapse, however, was a widespread faith in stochastic models for limiting risk. These models encouraged people to engage in investments that turned out to be far riskier than the models predicted. These models were created by some of the brightest PhD's from the best universities, but they failed miserably because they modeled everyday events very well, but modeled the rare events and the interconnection of events poorly. They failed badly by not understanding their application, and in particular, by trying to extrapolate typical behavior when their primary goal was to protect against highly atypical situations. The moral of the fable is that brilliant analysis is not helpful when the modeling is poor; as computer engineers say, "garbage in, garbage out."

The examples above show that the problems of modeling a real-world experiment are often connected with the question of creating a model for a set of experiments that are not exactly the same and do not necessarily correspond to the notion of independent repetitions within the model. In other words, the question is not only whether the probability model is reasonable for a single experiment, but also whether the IID repetition model is appropriate for multiple copies of the real-world experiment.

At least we have seen, however, that if a real-world experiment can be performed many times with a physical isolation between performances that is well modeled by the IID repetition model, then the relative frequencies of events in the real-world experiment correspond to relative frequencies in the idealized IID repetition model, which correspond to probabilities in the original model. In other words, under appropriate circumstances, the probabilities in a model become essentially observable over many repetitions.

We will see later that our emphasis on IID repetitions was done for simplicity. There are other models for repetitions of a basic model, such as Markov models, that we study later. These will also lead to relative frequencies approaching probabilities within the repetition model. Thus, for repeated real-world experiments that are well modeled by these repetition models, the real world relative frequencies approximate the probabilities in the model.

1.8.3 Statistical independence of real-world experiments

We have been discussing the use of relative frequencies of an event A in a repeated real-world experiment to test $\Pr\{A\}$ in a probability model of that experiment. This can be done essentially successfully if the repeated trials correpond to IID trials in the idealized experiment. However, the statement about IID trials in the idealized experiment is a statement about probabilities in the extended n-trial model. Thus, just as we tested $\Pr\{A\}$ by repeated real-world trials of a single experiment, we should be able to test $\Pr\{A_1, \ldots, A_n\}$ in the n-repetition model by a much larger number of real-world repetitions of n-tuples rather than single trials.

To be more specific, choose two large integers, m and n, and perform the underlying real-world experiment mn times. Partition the mn trials into m runs of n trials each. For any given n-tuple $A_1, \ldots, A_n$ of successive events, find the relative frequency (over m trials of n tuples) of the n-tuple event $A_1, \ldots, A_n$. This can then be used essentially to test the

probability $\Pr\{A_1, \ldots, A_n\}$ in the model for n IID trials. The individual event probabilities can also be tested, so the condition for independence can be tested.

The observant reader will note that there is a tacit assumption above that successive n tuples can be modeled as independent, so it seems that we are simply replacing a big problem with a bigger problem. This is not quite true, since if the trials are dependent with some given probability model for dependent trials, then this test for independence will essentially reject the independence hypothesis for large enough n. In other words, we can not completely verify the correctness of an independence hypothesis for the n-trial model, although in principle we could eventually falsify it if it is false.

Choosing models for real-world experiments is primarily a subject for statistics, and we will not pursue it further except for brief discussions when treating particular application areas. The purpose here has been to treat a fundamental issue in probability theory. As stated before, probabilities are non-observables — they exist in the theory but are not directly measurable in real-world experiments. We have shown that probabilities essentially become observable in the real-world via relative frequencies over repeated trials.

1.8.4 Limitations of relative frequencies

Most real-world applications that are modeled by probability models have such a large sample space that it is impractical to conduct enough trials to choose probabilities from relative frequencies. Even a shuffled deck of 52 cards would require many more than $52! \approx 8 \times 10^{67}$ trials for most of the outcomes to appear even once. Thus relative frequencies can be used to test the probability of given individual events of importance, but are usually impractical for choosing the entire model and even more impractical for choosing a model for repeated trials.

Since relative frequencies give us a concrete interpretation of what probability means, however, we can now rely on other approaches, such as symmetry, for modeling. From symmetry, for example, it is clear that all 52! possible arrangements of a card deck should be equiprobable after shuffling. This leads, for example, to the ability to calculate probabilities of different poker hands, etc., which are such popular exercises in elementary probability classes.

Another valuable modeling procedure is that of constructing a probability model where the possible outcomes are independently chosen n-tuples of outcomes in a simpler model. More generally, most of the random processes to be studied in this text are defined as various ways of combining simpler idealized experiments.

What is really happening as we look at modeling increasingly sophisticated systems and studying increasingly sophisticated models is that we are developing mathematical results for simple idealized models and relating those results to real-world results (such as relating idealized statistically independent trials to real-world independent trials). The association of relative frequencies to probabilities forms the basis for this, but is usually exercised only in the simplest cases.

The way one selects probability models of real-world experiments in practice is to use

scientific knowledge and experience, plus simple experiments, to choose a reasonable model. The results from the model (such as the law of large numbers) are then used both to hypothesize results about the real-world experiment and to provisionally reject the model when further experiments show it to be highly questionable. Although the results about the model are mathematically precise, the corresponding results about the real-world are at best insightful hypotheses whose most important aspects must be validated in practice.

1.8.5 Subjective probability

There are many useful applications of probability theory to situations other than repeated trials of a given experiment. When designing a new system in which randomness (of the type used in probability models) is hypothesized, one would like to analyze the system before actually building it. In such cases, the real-world system does not exist, so indirect means must be used to construct a probability model. Often some sources of randomness, such as noise, can be modeled in the absence of the system. Often similar systems or simulation can be used to help understand the system and help in formulating appropriate probability models. However, the choice of probabilities is to a certain extent subjective.

Another type of situation (such as risk analysis for nuclear reactors) deals with a large number of very unlikely outcomes, each catastrophic in nature. Experimentation clearly cannot be used to establish probabilities, and it is not clear that probabilities have any real meaning here. It can be helpful, however, to choose a probability model on the basis of subjective beliefs which can be used as a basis for reasoning about the problem.

Choosing a probability model on the basis of subjective beliefs can be helpful here if used as a basis for reasoning about the problem. When handled well, this can at least make the subjective biases clear, leading to a more rational approach. When handled poorly (as for example in some risk analyses of large financial systems) it can hide both the real risks and the arbitrary nature of possibly poor decisions.

We will not discuss the various, often ingenious, methods for choosing subjective probabilities. The reason is that subjective beliefs should be based on intensive and long term exposure to the particular problem involved; discussing these problems in abstract probability terms weakens this link. We will focus instead on the analysis of idealized models. These can be used to provide insights for subjective models, and more refined and precise results for objective models.

1.9 Summary

This chapter started with an introduction into the correspondence between probability theory and real-world experiments involving randomness. While almost all work in probability theory works with established probability models, it is important to think through what these probabilities mean in the real world, and elementary subjects rarely address these questions seriously.

The next section discussed the axioms of probability theory, along with some insights about

why these particular axioms were chosen. This was followed by a review of conditional probabilities, statistical independence, and random variables. Stochastic processes in the guise of the Bernoulli process was then introduced with an emphasis on asymptotic behavior. The review of probability was then resumed with an emphasis on the underlying structure of the field rather than reviewing details and problem solving techniques.

Next the laws of large numbers were developed and discussed at a somewhat deeper level than most elementary courses. This involved a fair amount of abstraction, combined with mathematical analysis. The central idea is that the sample average of n IID rv's approaches the mean with increasing n. As a special case, the relative frequency of an event A approaches $\Pr\{A\}$. What the word *approaches* means here is both tricky and vital in understanding probability theory. The strong law of large numbers and convergence WP1 requires mathematical maturity, and was postponed to Chapter 5 where it is first used.

The final section came back to the fundamental problem of understanding the relation between probability theory and randomness in the real-world. It was shown, via the laws of large numbers, that probabilities become essentially observable via relative frequencies calculated over repeated experiments.

There are too many texts on elementary probability to mention here, and most of them serve to give added understanding and background to the material in this chapter. We recommend Bertsekas and Tsitsiklis [2], both for a careful statement of the fundamentals and for a wealth of well-chosen and carefully explained examples.

Texts that cover similar material to that here are [22] and [14]. Kolmogorov [17] is readable for the mathematically mature and is also of historical interest as the translation of the 1933 book that first put probability on a firm mathematical basis. Feller [8] is the classic extended and elegant treatment of elementary material from a mature point of view. Rudin [23] is an excellent text on measure theory for those with advanced mathematical preparation.

1.10 Exercises

Exercise 1.1. Let A_1 and A_2 be arbitrary events and show that $\Pr\{A_1 \bigcup A_2\}+\Pr\{A_1A_2\} = \Pr\{A_1\} + \Pr\{A_2\}$. Explain which parts of the sample space are being double counted on both sides of this equation and which parts are being counted once.

Exercise 1.2. This exercise derives the probability of an arbitrary (non-disjoint) union of events, derives the union bound, and derives some useful limit expressions.

a) For 2 arbitrary events A_1 and A_2, show that

$$A_1 \bigcup A_2 = A_1 \bigcup (A_2 - A_1) \qquad \text{where } A_2 - A_1 = A_2A_1^c.$$

Show that A_1 and $A_2 - A_1$ are disjoint Hint: This is what Venn diagrams were invented for.

b) For an arbitrary sequence of events, $\{A_n; n \geq 1\}$, let $B_1 = A_1$ and for each $n \geq 2$ define $B_n = A_n - \bigcup_{m=1}^{n-1} A_m$. Show that $B_1, B_2, \ldots$, are disjoint events and show that for each $n \geq 2$, $\bigcup_{m=1}^{n} A_m = \bigcup_{m=1}^{n} B_m$. Hint: Use induction.

c) Show that

$$\Pr\Big\{\bigcup_{n=1}^{\infty} A_n\Big\} \;=\; \Pr\Big\{\bigcup_{n=1}^{\infty} B_n\Big\} \;=\; \sum_{n=1}^{\infty} \Pr\{B_n\}\,.$$

Hint: Use the axioms of probability for the second equality.

d) Show that for each n, $\Pr\{B_n\} \leq \Pr\{A_n\}$. Use this to show that

$$\Pr\Big\{\bigcup_{n=1}^{\infty} A_n\Big\} \;\leq\; \sum_{n=1}^{\infty} \Pr\{A_n\}\,.$$

e) Show that $\Pr\{\bigcup_{n=1}^{\infty} A_n\} = \lim_{m\to\infty} \Pr\{\bigcup_{n=1}^{m} A_n\}$. Hint: Combine parts c) and b). Note that this says that the probability of a limit of unions is equal to the limit of the probabilities. This might well appear to be obvious without a proof, but you will see situations later where similar appearing interchanges cannot be made.

f) Show that $\Pr\{\bigcap_{n=1}^{\infty} A_n\} = \lim_{n\to\infty} \Pr\{\bigcap_{i=1}^{n} A_i\}$. Hint: Remember De Morgan's equalities.

Exercise 1.3. Find the probability that a five card poker hand, chosen randomly from a 52 card deck, contains 4 aces. That is, if all 52! arrangements of a deck of cards are equally likely, what is the probability that all 4 aces are in the first 5 cards of the deck.

Exercise 1.4. Consider a sample space of 8 equiprobable sample points and let A_1, A_2, A_3 be three events each of probability 1/2 such that $\Pr\{A_1A_2A_3\} = \Pr\{A_1\}\Pr\{A_2\}\Pr\{A_3\}$.

a) Create an example where $\Pr\{A_1A_2\} = \Pr\{A_1A_3\} = \frac{1}{4}$ but $\Pr\{A_2A_3\} = \frac{1}{8}$. Hint: Make a table with a row for each sample point and a column for each of the above 3 events and try different ways of assigning sample points to events (the answer is not unique).

b) Show that, for your example, A_2 and A_3 are not independent. Note that the definition of statistical independence would be very strange if it allowed A_1, A_2, A_3 to be independent while A_2 and A_3 are dependent. This illustrates why the definition of independence requires (1.14) rather than just (1.15).

Exercise 1.5. This exercise shows that for all rv's X, $\mathsf{F}_X(x)$ is continuous from the right.

a) For any given rv X, any real number x, and each integer $n \geq 1$, let $A_n = \{\omega : X > x + 1/n\}$, and show that $A_1 \subseteq A_2 \subseteq \cdots$. Use this and the corollaries to the axioms of probability to show that $\Pr\left\{\bigcup_{n\geq 1} A_n\right\} = \lim_{n\to\infty} \Pr\{A_n\}$.

b) Show that $\Pr\left\{\bigcup_{n\geq 1} A_n\right\} = \Pr\{X > x\}$ and that $\Pr\{X > x\} = \lim_{n\to\infty} \Pr\{X > x + 1/n\}$.

c) Show that for $\epsilon > 0$, $\lim_{\epsilon\to 0} \Pr\{X \leq x + \epsilon\} = \Pr\{X \leq x\}$.

d) Define $\widetilde{\mathsf{F}}_X(x) = \Pr\{X < x\}$. Show that $\widetilde{\mathsf{F}}_X(x)$ is continuous from the left. In other words, the continuity from the right for the CDF arises from the almost arbitrary (but universally accepted) choice in defining the CDF as $\Pr\{X \leq x\}$ rather than $\Pr\{X < x\}$.

Exercise 1.6. Show that for a continuous nonnegative rv X,

$$\int_0^\infty \Pr\{X > x\}\, dx = \int_0^\infty x \mathsf{f}_X(x)\, dx \tag{1.98}$$

Hint 1: First rewrite $\Pr\{X > x\}$ on the left side of (1.98) as $\int_x^\infty \mathsf{f}_X(y)\, dy$. Then think through, to your level of comfort, how and why the order of integration can be interchnged in the resulting expression.

Hint 2: As an alternate approach, derive (1.98) using integration by parts.

Exercise 1.7. Suppose X and Y are discrete rv's with the PMF $\mathsf{p}_{XY}(x_i, y_j)$. Show (a picture will help) that this is related to the joint CDF by

$$\mathsf{p}_{XY}(x_i, y_j) = \lim_{\delta>0,\delta\to 0} \left[\mathsf{F}(x_i, y_j) - \mathsf{F}(x_i - \delta, y_j) - \mathsf{F}(x_i, y_j - \delta) + \mathsf{F}(x_i - \delta, y_j - \delta)\right].$$

Exercise 1.8. A variation of Example 1.5.1 is to let M be a random variable that takes on both positive and negative values with the PMF

$$\mathsf{p}_M(m) = \frac{1}{2|m|\,(|m|+1)}.$$

In other words, M is symmetric around 0 and $|M|$ has the same PMF as the nonnegative rv N of Example 1.5.1.

a) Show that $\sum_{m\geq 0} m\mathsf{p}_M(m) = \infty$ and $\sum_{m<0} m\mathsf{p}_M(m) = -\infty$. (Thus show that the expectation of M not only does not exist but is undefined even in the extended real number system.)

b) Suppose that the terms in $\sum_{m=-\infty}^{\infty} m\mathsf{p}_M(m)$ are summed in the order of 2 positive terms for each negative term (*i.e.*, in the order $1, 2, -1, 3, 4, -2, 5, \cdots$). Find the limiting value of the partial sums in this series. Hint: You may find it helpful to know that

$$\lim_{n\to\infty}\left[\sum_{i=1}^{n}\frac{1}{i}-\int_1^n\frac{1}{x}\,dx\right]=\gamma,$$

where γ is the Euler-Mascheroni constant, $\gamma = 0.57721\cdots$.

c) Repeat (b) where, for any given integer $k > 0$, the order of summation is k positive terms for each negative term.

Exercise 1.9. (Proof of Theorem 1.4.1) The bounds on the binomial in this theorem are based on the *Stirling bounds*. These say that for all $n \geq 1$, $n!$ is upper and lower bounded by

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n < n! < \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{1/12n}. \tag{1.99}$$

The ratio $\sqrt{2\pi n}(n/e)^n/n!$ of the first two terms is monotonically increasing with n toward the limit 1, and the ratio $\sqrt{2\pi n}(n/e)^n \exp(1/12n)/n!$ is monotonically decreasing toward 1. The upper bound is more accurate, but the lower bound is simpler and known as the Stirling approximation. See [8] for proofs and further discussion of the above facts.

a) Show from (1.99) and from the above monotone property that

$$\binom{n}{k} < \sqrt{\frac{n}{2\pi k(n-k)}}\;\frac{n^n}{k^k(n-k)^{n-k}}.$$

Hint: First show that $n!/k! < \sqrt{n/k}\, n^n k^{-k} e^{-n+k}$ for $k < n$.

b) Use the result of (a) to upper bound $\mathsf{p}_{S_n}(k)$ by

$$\mathsf{p}_{S_n}(k) < \sqrt{\frac{n}{2\pi k(n-k)}}\;\frac{p^k(1-p)^{n-k}n^n}{k^k(n-k)^{n-k}}.$$

Show that this is equivalent to the upper bound in Theorem 1.4.1.

c) Show that

$$\binom{n}{k} > \sqrt{\frac{n}{2\pi k(n-k)}}\;\frac{n^n}{k^k(n-k)^{n-k}}\left[1-\frac{n}{12k(n-k)}\right].$$

d) Derive the lower bound in Theorem 1.4.1.

e) Show that $\mathsf{D}(\tilde{p}\|p) = \tilde{p}\ln(\frac{\tilde{p}}{p}) + (1-\tilde{p})\ln(\frac{1-\tilde{p}}{1-p})$ is 0 at $\tilde{p} = p$ and nonnegative elsewhere.

Exercise 1.10. Let X be a ternary rv taking on the 3 values 0, 1, 2 with probabilities p_0, p_1, p_2 respectively. Find the median of X for each of the cases below.

a) $p_0 = 0.2,\ p_1 = 0.4,\ p_2 = 0.4.$

b) $p_0 = 0.2,\ p_1 = 0.2,\ p_2 = 0.6.$

c) $p_0 = 0.2,\ p_1 = 0.3,\ p_2 = 0.5.$

Note 1: The median is not unique in (c). Find the interval of values that are medians. Note 2: Some people force the median to be distinct by defining it as the midpoint of the interval satisfying the definition given here.

d) Now suppose that X is nonnegative and continuous with the density $\mathsf{f}_X(x) = 1$ for $0 \leq x \leq 0.5$ and $\mathsf{f}_X(x) = 0$ for $0.5 < x \leq 1$. We know that $\mathsf{f}_X(x)$ is positive for all $x > 1$, but it is otherwise unknown. Find the median or interval of medians.

The median is sometimes (incorrectly) defined as that α for which $\Pr\{X > \alpha\} = \Pr\{X < \alpha\}$. Show that it is possible for no such α to exist. Hint: Look at the examples above.

Exercise 1.11. a) For any given rv Y, express $\mathsf{E}[|Y|]$ in terms of $\int_{y<0} \mathsf{F}_Y(y)\,dy$ and $\int_{y\geq 0} \mathsf{F}^{\mathsf{c}}_Y(y)\,dy$. Hint: Review the argument in Figure 1.4.

b) For some given rv X with $\mathsf{E}[|X|] < \infty$, let $Y = X - \alpha$. Using (a), show that

$$\mathsf{E}[|X-\alpha|] = \int_{-\infty}^{\alpha} \mathsf{F}_X(x)\,dx + \int_{\alpha}^{\infty} \mathsf{F}^{\mathsf{c}}_X(x)\,dx.$$

c) Show that $\mathsf{E}[|X-\alpha|]$ is minimized over α by choosing α to be a median of X. Hint: Both the easy way and the most instructive way to do this is to use a graphical argument illustrating the above integrals. Be careful to show that when the median is an interval, all points in this interval achieve the minimum.

Exercise 1.12. Let X be a rv with CDF $\mathsf{F}_X(x)$. Find the CDF of the following rv's.

a) The maximum of n IID rv's, each with CDF $\mathsf{F}_X(x)$.

b) The minimum of n IID rv's, each with CDF $\mathsf{F}_X(x)$.

c) The difference of the rv's defined in a) and b); assume X has a density $\mathsf{f}_X(x)$.

Exercise 1.13. Let X and Y be rv's in some sample space Ω and let $Z = X + Y$, *i.e.*, for each $\omega \in \Omega$, $Z(\omega) = X(\omega) + Y(\omega)$. The purpose of this exercise is to show that Z is a rv. This is a mathematical fine point that many readers may prefer to simply accept without proof.

a) Show that the set of ω for which $Z(\omega)$ is infinite or undefined has probability 0.

b) We must show that $\{\omega \in \Omega : Z(\omega) \leq \alpha\}$ is an event for each real α, and we start by approximating that event. To show that $Z = X + Y$ is a rv, we must show that for each real number α, the set $\{\omega \in \Omega : X(\omega) + Y(\omega) \leq \alpha\}$ is an event. Let $B(n,k) = \{\omega : X(\omega) \leq k/n\} \bigcap \{Y(\omega) \leq \alpha + (1-k)/n\}$ Let $D(n) = \bigcup_k B(n,k)$ and show that $D(n)$ is an event.

c) On a 2 dimensional sketch for a given α, show the values of $X(\omega)$ and $Y(\omega)$ for which $\omega \in D(n)$. Hint: This set of values should be bounded by a staircase function.

d) Show that

$$\{\omega : X(\omega) + Y(\omega) \leq \alpha\} = \bigcap\nolimits_{n\geq 1} D(n). \tag{1.100}$$

Explain why this shows that $Z = X + Y$ is a rv.

e) Explain why this implies that if $Y = X_1 + X_2 + \cdots + X_n$ and if $X_1, X_2, \ldots, X_n$ are rv's, then Y is a rv. Hint: Only one or two lines of explanation are needed.

Exercise 1.14. a) Let $X_1, X_2, \ldots, X_n$ be rv's with expected values $\overline{X}_1, \ldots, \overline{X}_n$. Show that $\mathsf{E}[X_1 + \cdots + X_n] = \overline{X}_1 + \cdots + \overline{X}_n$. You may assume that the rv's have a joint density function, but do not assume that the rv's are independent.

b) Now assume that $X_1, \ldots, X_n$ are statistically independent and show that the expected value of the product is equal to the product of the expected values.

c) Again assuming that $X_1, \ldots, X_n$ are statistically independent, show that the variance of the sum is equal to the sum of the variances.

Exercise 1.15. (Stieltjes integration) **a)** Let $h(x) = u(x)$ and $\mathsf{F}_X(x) = u(x)$ where $u(x)$ is the unit step, *i.e.*, $u(x) = 0$ for $-\infty < x < 0$ and $u(x) = 1$ for $x \geq 0$. Using the definition of the Stieltjes integral in Footnote 19, show that $\int_{-1}^{1} h(x) d\mathsf{F}_X(x)$ does not exist. Hint: Look at the term in the Riemann sum including $x = 0$ and look at the range of choices for $h(x)$ in that interval. Intuitively, it might help initially to view $d\mathsf{F}_X(x)$ as a unit impulse at $x = 0$.

b) Let $h(x) = u(x - a)$ and $\mathsf{F}_X(x) = u(x - b)$ where a and b are in $(-1, +1)$. Show that $\int_{-1}^{1} h(x) d\mathsf{F}_X(x)$ exists if and only if $a \neq b$. Show that the integral has the value 1 for $a < b$ and the value 0 for $a > b$. Argue that this result is still valid in the limit of integration over $(-\infty, \infty)$.

c) Let X and Y be independent discrete rv's, each with a finite set of possible values. Show that $\int_{-\infty}^{\infty} \mathsf{F}_X(z - y) d\mathsf{F}_Y(y)$, defined as a Stieltjes integral, is equal to the distribution of $Z = X + Y$ at each z other than the possible sample values of Z, and is undefined at each sample value of Z. Hint: Express F_X and F_Y as sums of unit steps. Note: This failure of Stieltjes integration is not a serious problem; $\mathsf{F}_Z(z)$ is a step function, and the integral is undefined at its points of discontinuity. We automatically define $\mathsf{F}_Z(z)$ at those step values so that F_Z is a CDF (*i.e.*, is continuous from the right). This problem does not arise if either X or Y is continuous.

Exercise 1.16. Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of IID continuous rv's with the common probability density function $\mathsf{f}_X(x)$; note that $\Pr\{X{=}\alpha\} = 0$ for all α and that $\Pr\{X_i{=}X_j\} = 0$ for all $i \neq j$. For $n \geq 2$, define X_n as a *record-to-date* of the sequence if $X_n > X_i$ for all $i < n$.

a) Find the probability that X_2 is a record-to-date. Use symmetry to obtain a numerical answer without computation. A one or two line explanation should be adequate.

b) Find the probability that X_n is a record-to-date, as a function of $n \geq 1$. Again use symmetry.

c) Find a simple expression for the expected number of records-to-date that occur over the first m trials for any given integer m. Hint: Use indicator functions. Show that this expected number is infinite in the limit $m \to \infty$.

Exercise 1.17. (Continuation of Exercise 1.16)

a) Let N_1 be the index of the *first* record-to-date in the sequence. Find $\Pr\{N_1 > n\}$ for each $n \geq 2$. Hint: There is a far simpler way to do this than working from (b) in Exercise 1.16.

b) Show that N_1 is a rv.

c) Show that $\mathsf{E}[N_1] = \infty$.

d) Let N_2 be the index of the *second* record-to-date in the sequence. Show that N_2 is a rv. Hint: You need not find the CDF of N_2 here.

e) Contrast your result in (c) to the result from (c) of Exercise 1.16 saying that the expected number of records-to-date is infinite over an an infinite number of trials. Note: this should be a shock to your intuition — there is an infinite expected wait for the first of an infinite sequence of occurrences, each of which must eventually occur.

Exercise 1.18. (Another direction from Exercise 1.16) **a)** For any given $n \geq 2$, find the probability that X_n and X_{n+1} are both records-to-date. Hint: The idea in (b) of 1.16 is helpful here, but the result is not.

b) Is the event that X_n is a record-to-date statistically independent of the event that X_{n+1} is a record-to-date?

c) Find the expected number of adjacent pairs of records-to-date over the sequence $X_1, X_2, \ldots$. Hint: A helpful fact here is that $\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}$.

Exercise 1.19. a) Assume that X is a nonnegative discrete rv taking on values $a_1, a_2, \ldots$, and let $Y = h(X)$ for some nonnegative function h. Let $b_i = h(a_i)$, $i \geq 1$ be the i^{th} value taken on by Y. Show that $\mathsf{E}[Y] = \sum_i b_i \mathsf{p}_Y(b_i) = \sum_i h(a_i)\mathsf{p}_X(a_i)$. Find an example where $\mathsf{E}[X]$ exists but $\mathsf{E}[Y] = \infty$.

b) Let X be a nonnegative continuous rv with density $\mathsf{f}_X(x)$ and let $h(x)$ be differentiable, nonnegative, and nondecreasing in x. Let $A(\delta) = \sum_{n \geq 1} h(n\delta)[\mathsf{F}(n\delta) - \mathsf{F}(n\delta - \delta)]$, *i.e.*, $A(\delta)$ is a δth order approximation to the Stieltjes integral $\int h(x) d\mathsf{F}(x)$. Show that if $A(1) < \infty$, then $A(2^{-k}) \leq A(2^{-(k-1)}) < \infty$ for $k \geq 1$. Show from this that $\int h(x)\, d\mathsf{F}(x)$ converges to a finite value. Note: this is a very special case, but it can be extended to many cases of interest. It seems better to consider these convergence questions as required rather than consider them in general.

Exercise 1.20. a) Consider a positive, integer-valued rv Y whose CDF is given at integer values by

$$\mathsf{F}_Y(y) = 1 - \frac{2}{(y+1)(y+2)} \qquad \text{for integer } \; y \geq 0.$$

Use (1.31) to show that $\mathsf{E}[Y] = 2$. Hint: Note that $1/[(y+1)(y+2)] = 1/(y+1) - 1/(y+2)$.

b) Find the PMF of Y and use it to check the value of $\mathsf{E}[Y]$.

c) Let X be another positive, integer-valued rv. Assume its conditional PMF is given by

$$\mathsf{p}_{X|Y}(x|y) = \frac{1}{y} \qquad \text{for } 1 \leq x \leq y.$$

Find $\mathsf{E}[X \mid Y = y]$ and use it to show that $\mathsf{E}[X] = 3/2$. Explore finding $\mathsf{p}_X(x)$ until you are convinced that using the conditional expectation to calculate $\mathsf{E}[X]$ is considerably easier than using $\mathsf{p}_X(x)$.

d) Let Z be another integer-valued rv with the conditional PMF

$$\mathsf{p}_{Z|Y}(z|y) = \frac{1}{y^2} \qquad \text{for } 1 \leq z \leq y^2.$$

Find $\mathsf{E}[Z \mid Y = y]$ for each integer $y \geq 1$ and find $\mathsf{E}[Z]$.

Exercise 1.21. a) Show that, for uncorrelated rv's, the expected value of the product is equal to the product of the expected values (by definition, X and Y are uncorrelated if $\mathsf{E}\left[(X - \overline{X})(Y - \overline{Y})\right] = 0$).

b) Show that if X and Y are uncorrelated, then the variance of $X + Y$ is equal to the variance of X plus the variance of Y.

c) Show that if $X_1, \ldots, X_n$ are uncorrelated, then the variance of the sum is equal to the sum of the variances.

d) Show that independent rv's are uncorrelated.

e) Let X, Y be identically distributed ternary valued random variables with the PMF $\mathsf{p}_X(-1) = \mathsf{p}_X(1) = 1/4$; $\mathsf{p}_X(0) = 1/2$. Find a simple joint probability assignment such that X and Y are uncorrelated but dependent.

f) You have seen that the moment generating function of a sum of independent rv's is equal to the product of the individual moment generating functions. Give an example where this is false if the variables are uncorrelated but dependent.

Exercise 1.22. Suppose X has the Poisson PMF, $\mathsf{p}_X(n) = \lambda^n \exp(-\lambda)/n!$ for $n \geq 0$ and Y has the Poisson PMF, $\mathsf{p}_Y(n) = \mu^n \exp(-\mu)/n!$ for $n \geq 0$. Assume that X and Y are independent. Find the distribution of $Z = X + Y$ and find the conditional distribution of Y conditional on $Z = n$.

Exercise 1.23. a) Suppose X, Y and Z are binary rv's, each taking on the value 0 with probability 1/2 and the value 1 with probability 1/2. Find a simple example in which X, Y, Z are statistically *dependent* but are *pairwise* statistically *independent* (i.e., X, Y are statistically independent, X, Z are statistically independent, and Y, Z are statistically

independent). Give $\mathsf{p}_{XYZ}(x,y,z)$ for your example. Hint: In the simplest example, there are four joint values for x,y,z that have probability 1/4 each.

b) Is pairwise statistical independence enough to ensure that

$$\mathsf{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathsf{E}\left[X_i\right].$$

for a set of rv's $X_1,\ldots,X_n$?

Exercise 1.24. Show that $\mathsf{E}[X]$ is the value of α that minimizes $\mathsf{E}\left[(X-\alpha)^2\right]$.

Exercise 1.25. For each of the following random variables, find the endpoints r_- and r_+ of the interval for which the moment generating function $\mathsf{g}(r)$ exists. Determine in each case whether $\mathsf{g}_X(r)$ exists at r_- and r_+. For parts a) and b) you should also find and sketch $\mathsf{g}(r)$. For parts c) and d), $\mathsf{g}(r)$ has no closed form.

a) Let λ,θ, be positive numbers and let X have the density

$$\mathsf{f}_X(x) = \frac{1}{2}\lambda\exp(-\lambda x);\ x\geq 0; \quad \mathsf{f}_X(x) = \frac{1}{2}\theta\exp(\theta x);\ x<0.$$

b) Let Y be a Gaussian random variable with mean m and variance σ^2.

c) Let Z be a nonnegative random variable with density

$$\mathsf{f}_Z(z) = k(1+z)^{-2}\ \exp(-\lambda z); \quad z\geq 0.$$

where $\lambda>0$ and $k=[\int_{z\geq 0}(1+z)^{-2}\ \exp(-\lambda z)dz]^{-1}$. Hint: Do not try to evaluate $\mathsf{g}_Z(r)$. Instead, investigate values of r for which the integral is finite and infinite.

d) For the Z of (c), find the limit of $\gamma'(r)$ as r approaches λ from below. Then replace $(1+z)^2$ with $|1+z|^3$ in the definition of $\mathsf{f}_Z(z)$ and K and show whether the above limit is then finite or not. Hint: no integration is required.

Exercise 1.26. a) Assume that the MGF of the random variable X exists (*i.e.*, is finite) in the interval (r_-,r_+), $r_-<0<r_+$, and assume $r_-<r<r_+$ throughout. For any finite constant c, express the moment generating function of $X-c$, *i.e.*, $\mathsf{g}_{(X-c)}(r)$, in terms of $\mathsf{g}_X(r)$ and show that $\mathsf{g}_{(X-c)}(r)$ exists for all r in (r_-,r_+). Explain why $\mathsf{g}''_{(X-c)}(r)\geq 0$.

b) Show that $\mathsf{g}''_{(X-c)}(r) = [\mathsf{g}''_X(r) - 2c\mathsf{g}'_X(r) + c^2\mathsf{g}_X(r)]e^{-rc}$.

c) Use **a)** and **b)** to show that $\mathsf{g}''_X(r)\mathsf{g}_X(r) - [\mathsf{g}'_X(r)]^2\geq 0$, Let $\gamma_X(r)=\ln\mathsf{g}_X(r)$ and show that $\gamma''_X(r)\geq 0$. Hint: Choose $c=\mathsf{g}'_X(r)/\mathsf{g}_X(r)$.

d) Assume that X is non-deterministic, *i.e.*, that there is no value of α such that $\Pr\{X=\alpha\}=1$. Show that the inequality sign "$\geq$" may be replaced by "$>$" everywhere in a), b) and c).

Exercise 1.27. A computer system has n users, each with a unique name and password. Due to a software error, the n passwords are randomly permuted internally (i.e. each of the $n!$ possible permutations are equally likely. Only those users lucky enough to have had their passwords unchanged in the permutation are able to continue using the system.

a) What is the probability that a particular user, say user 1, is able to continue using the system?

b) What is the expected number of users able to continue using the system? Hint: Let X_i be a rv with the value 1 if user i can use the system and 0 otherwise.

Exercise 1.28. Suppose the rv X is continuous and has the CDF $\mathsf{F}_X(x)$. Consider another rv $Y = \mathsf{F}_X(X)$. That is, for each sample point ω such that $X(\omega) = x$, we have $Y(\omega) = \mathsf{F}_X(x)$. Show that Y is uniformly distributed in the interval 0 to 1.

Exercise 1.29. Let Z be an integer-valued rv with the PMF $\mathsf{p}_Z(n) = 1/k$ for $0 \leq n \leq k-1$. Find the mean, variance, and moment generating function of Z. Hint: An elegant way to do this is to let U be a uniformly distributed continuous rv over $(0, 1]$ that is independent of Z. Then $U + Z$ is uniform over $(0, k]$. Use the known results about U and $U + Z$ to find the mean, variance, and MGF for Z.

Exercise 1.30. (Alternate approach 1 to the Markov inequality) **a)** Let Y be a nonnegative rv and $y > 0$ be some fixed number. Let A be the event that $Y \geq y$. Show that $y\,\mathbb{I}_A \leq Y$ (*i.e.*, that this inequality is satisfied for every $\omega \in \Omega$).

b) Use your result in (a) to prove the Markov inequality.

Exercise 1.31. (Alternate approach 2 to the Markov inequality) **a)** Minimize $\mathsf{E}[Y]$ over all nonnegative rv's such that $\Pr\{Y \geq b\} = \beta$ for some given $b > 0$ and $0 < \beta < 1$. Hint: Use a graphical argument similar to that in Figure 1.7. What is the rv that achieves the minimum. Hint: It is binary.

b) Use (a) to prove the Markov inequality and also point out the distribution that meets the inequality with equality.

Exercise 1.32. (The one-sided Chebyshev inequality) This inequality states that if a zero-mean rv X has a variance σ^2, then it satisfies the inequality

$$\Pr\{X \geq b\} \leq \frac{\sigma^2}{\sigma^2 + b^2} \qquad \text{for every } b > 0, \tag{1.101}$$

with equality for some b only if X is binary and $\Pr\{X = b\} = \sigma^2/(\sigma^2 + b^2)$. We prove this here using the same approach as in Exercise 1.31. Let X be a zero-mean rv that satisfies $\Pr\{X \geq b\} = \beta$ for some $b > 0$ and $0 < \beta < 1$. The variance σ^2 of X can be expressed as

$$\sigma^2 = \int_{-\infty}^{b^-} x^2 \mathsf{f}_X(x)\,dx + \int_b^{\infty} x^2\,\mathsf{f}_X(x)\,dx \tag{1.102}$$

We will first minimize σ^2 over all zero-mean X satisfying $\Pr\{X \geq b\} = \beta$.

a) Show that the second integral in (1.102) satisfies $\int_b^\infty x^2 \mathsf{f}_X(x)\, dx \geq b^2\beta$.

b) Show that the first integral in (1.102) is constrained by

$$\int_{-\infty}^{b^-} \mathsf{f}_X(x)\, dx = 1 - \beta \qquad \text{and} \qquad \int_{-\infty}^{b^-} x\mathsf{f}_X(x)\, dx \leq -b\beta.$$

c) Minimize the first integral in (1.102) subject to the constraints in (b). Hint: If you scale $f_X(x)$ up by $1/(1-\beta)$, it integrates to 1 over $(-\infty, b)$ and the second constraint becomes an expectation. You can then minimize the first integral in (1.102) by inspection.

d) Combine the results in a) and c) to show that $\sigma^2 \geq b^2\beta/(1-\beta)$. Find the minimizing distribution. Hint: It is binary.

e) Use (d) to establish (1.101). Also show (trivially) that if Y has a mean $\overline{Y}$ and variance σ^2, then $\Pr\{Y - \overline{Y} \geq b\} \leq \sigma^2/(\sigma^2 + b^2)$

Exercise 1.33. (Proof of (1.48)) Here we show that if X is a zero-mean rv with a variance σ^2, then the median α satisfies $|\alpha| \leq \sigma$.

a) First show that $|\alpha| \leq \sigma$ for the special case where X is binary with equiprobable values at $\pm\sigma$.

b) For all zero-mean rv's X with variance σ^2 other than the special case in a), show that

$$\Pr\{X \geq \sigma\} < 0.5.$$

Hint: Use the one-sided Chebyshev inequality of Exercise 1.32.

c) Show that $\Pr\{X \geq \alpha\} \geq 0.5$. Other than the special case in a), show that this implies that $\alpha < \sigma$.

d) Other than the special case in a), show that $|\alpha| < \sigma$. Hint: repeat b) and c) for the rv $-X$. You have then shown that $|\alpha| \leq \sigma$ with equality only for the binary case with values $\pm\sigma$. For rv's Y with a non-zero mean, this shows that $|\alpha - \overline{Y}| \leq \sigma$.

Exercise 1.34. We stressed the importance of the mean of a rv X in terms of its association with the sample average via the WLLN. Here we show that there is a form of WLLN for the median and for the entire CDF, say $\mathsf{F}_X(x)$ of X via sufficiently many independent sample values of X.

a) For any given y, let $\mathbb{I}_j(y)$ be the indicator function of the event $\{X_j \leq y\}$ where $X_1, X_2, \ldots, X_j, \ldots$ are IID rv's with the CDF $\mathsf{F}_X(x)$. State the WLLN for the IID rv's $\{\mathbb{I}_1(y), \mathbb{I}_2(y), \ldots\}$.

b) Does the answer to (a) require X to have a mean or variance?

c) Suggest a procedure for evaluating the median of X from the sample values of $X_1, X_2, \ldots$. Assume that X is a continuous rv and that its PDF is positive in an open interval around the median. You need not be precise, but try to think the issue through carefully.

What you have seen here, without stating it precisely or proving it is that the median has a law of large numbers associated with it, saying that the sample median of n IID samples of a rv is close to the true median with high probability.

Exercise 1.35. a) Show that for any $0 < k < n$

$$\binom{n}{k+1} \leq \binom{n}{k} \frac{n-k}{k}.$$

b) Extend (a) to show that, for all $\ell \leq n-k$,

$$\binom{n}{k+\ell} \leq \binom{n}{k} \left[\frac{n-k}{k}\right]^{\ell}.$$

c) Let $\tilde{p} = k/n$ and $\tilde{q} = 1 - \tilde{p}$. Let S_n be the sum of n binary IID rv's with $\mathsf{p}_X(0) = q$ and $\mathsf{p}_X(1) = p$. Show that for all $\ell \leq n-k$,

$$\mathsf{p}_{S_n}(k+\ell) \leq \mathsf{p}_{S_n}(k) \left(\frac{\tilde{q}p}{\tilde{p}q}\right)^{\ell}.$$

d) For $k/n > p$, show that $\Pr\{S_n \geq k\} \leq \frac{\tilde{p}q}{\tilde{p}-p}\, \mathsf{p}_{S_n}(k)$.

e) Now let ℓ be fixed and $k = \lceil n\tilde{p} \rceil$ for fixed $\tilde{p}$ such that $1 > \tilde{p} > p$. Argue that as $n \to \infty$,

$$\mathsf{p}_{S_n}(k+\ell) \sim \mathsf{p}_{S_n}(k) \left(\frac{\tilde{q}p}{\tilde{p}q}\right)^{\ell} \qquad \text{and} \quad \Pr\{S_n \geq k\} \sim \frac{\tilde{p}q}{\tilde{p}-p}\, \mathsf{p}_{S_n}(k),$$

where $a(n) \sim b(n)$ means that $\lim_{n\to\infty} a(n)/b(n) = 1$.

Exercise 1.36. A sequence $\{a_n;\, n \geq 1\}$ of real numbers has the limit 0 if for all $\epsilon > 0$, there is an $m(\epsilon)$ such that $|a_n| \leq \epsilon$ for all $n \geq m(\epsilon)$. Show that the sequences in parts a) and b) below satisfy $\lim_{n\to\infty} a_n = 0$ but the sequence in (c) does not have a limit.

a) $a_n = \frac{1}{\ln(\ln(n+1))}$

b) $a_n = n^{10} \exp(-n)$

c) $a_n = 1$ for $n = 10^{\ell}$ for each positive integer ℓ and $a_n = 0$ otherwise.

d) Show that the definition can be changed (with no change in meaning) by replacing ϵ with either $1/k$ or 2^{-k} for every positive integer k.

Exercise 1.37. Represent the MGF of a rv X by

$$\mathsf{g}_X(r) = \int_{-\infty}^{0} e^{rx} d\mathsf{F}(x) + \int_{0}^{\infty} e^{rx} d\mathsf{F}(x).$$

In each of the following parts, you are welcome to restrict X to be either discrete or continuous.

a) Show that the first integral always exists (*i.e.*, is finite) for $r \geq 0$ and that the second integral always exists for $r \leq 0$.

b) Show that if the second integral exists for a given $r_1 > 0$, then it also exists for all r in the range $0 \leq r \leq r_1$.

c) Show that if the first integral exists for a given $r_2 < 0$, then it also exists for all r in the range $r_2 \leq r \leq 0$.

d) Show that the range of r over which $\mathsf{g}_X(r)$ exists is an interval from some $r_- \leq 0$ to some $r_+ \geq 0$ (the interval might or might not include each endpoint, and either or both end point might be 0, ∞, or any point between).

e) Find an example where $r_+ = 1$ and the MGF does not exist for $r = 1$. Find another example where $r_+ = 1$ and the MGF does exist for $r = 1$. Hint: Consider $\mathsf{f}_X(x) = e^{-x}$ for $x \geq 0$ and figure out how to modify it to $\mathsf{f}_Y(y)$ so that $\int_0^\infty e^y \mathsf{f}_Y(y)\, dy < \infty$ but $\int_0^\infty e^{y+\epsilon y} \mathsf{f}_Y(y) = \infty$ for all $\epsilon > 0$.

Exercise 1.38. Let $\{X_n;\, n \geq 1\}$ be a sequence of independent but not identically distributed rv's. We say that the weak law of large numbers (WLLN) holds for this sequence if for all $\epsilon > 0$

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{S_n}{n} - \frac{\mathsf{E}\,[S_n]}{n}\right| \geq \epsilon\right\} = 0 \quad \text{where } S_n = X_1 + X_2 + \cdots + X_n. \qquad \text{(WL)}.$$

a) Show that the WLLN holds if there is some constant A such that $\sigma^2_{X_n} \leq A$ for all n.

b) Suppose that $\sigma^2_{X_n} \leq A\, n^{1-\alpha}$ for some $\alpha < 1$ and for all n. Show that the WLLN holds in this case.

Exercise 1.39. Let $\{X_i;\, i \geq 1\}$ be IID binary rv's. Let $\Pr\{X_i = 1\} = \delta$, $\Pr\{X_i = 0\} = 1 - \delta$. Let $S_n = X_1 + \cdots + X_n$. Let m be an arbitrary but fixed positive integer. Think! then evaluate the following and explain your answers:

a) $\lim_{n\to\infty} \sum_{i:n\delta - m \leq i \leq n\delta + m} \Pr\{S_n = i\}$

b) $\lim_{n\to\infty} \sum_{i:0 \leq i \leq n\delta + m} \Pr\{S_n = i\}$

c) $\lim_{n\to\infty} \sum_{i:n(\delta - 1/m) \leq i \leq n(\delta + 1/m)} \Pr\{S_n = i\}$.

Exercise 1.40. The WLLN is said to hold for a zero-mean sequence $\{S_n;\, n \geq 1\}$ if

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{S_n}{n}\right| > \epsilon\right\} = 0 \qquad \text{for every } \epsilon > 0.$$

The CLT is said to hold for $\{S_n;\, n \geq 1\}$ if for some $\sigma > 0$ and all $z \in \mathbb{R}$,

$$\lim_{n\to\infty} \Pr\left\{\frac{S_n}{\sigma\sqrt{n}} \leq z\right\} = \Phi(z),$$

where $\Phi(z)$ is the normal CDF. Show that if the CLT holds, then the WLLN holds also. Note 1: If you hate ϵ, δ arguments, you will hate this. Note 2: It will probably ease the pain if you convert the WLLN statement to: For every $\epsilon > 0, \delta > 0$ there exists an $n(\epsilon, \delta)$ such that for every $n \geq n(\epsilon, \delta)$,

$$\Pr\left\{\frac{S_n}{n} < -\epsilon\right\} \leq \delta \qquad \text{and} \quad \Pr\left\{\frac{S_n}{n} > 1-\epsilon\right\} \leq \delta. \tag{i}$$

Exercise 1.41. (Details in the proof of Theorem 1.7.4)

a) Show that if $X_1, X_2, \ldots$ are IID, then the truncated versions $\breve{X}_1, \breve{X}_2, \ldots$ are also IID.

b) Show that each $\breve{X}_i$ has a finite mean $\mathsf{E}\left[\breve{X}\right]$ and finite variance $\sigma^2_{\breve{X}}$. Show that the variance is upper bounded by the second moment around the original mean $\overline{X}$, *i.e.*, show that $\sigma^2_{\breve{X}} \leq \mathsf{E}\left[|\ \breve{X} - \mathsf{E}[X]\ |^2\right]$.

c) Assume that $\breve{X}_i$ is X_i truncated to $\overline{X} \pm b$. Show that $|\breve{X} - \overline{X}| \leq b$ and that $|\breve{X} - \overline{X}| \leq |X - \overline{X}|$. Use this to show that $\sigma^2_{\breve{X}} \leq b\mathsf{E}\left[|X - \overline{X}|\right] \leq 2b\mathsf{E}[|X|]$.

d) Let $\breve{S}_n = \breve{X}_1 + \cdots + \breve{X}_n$ and show that for any $\epsilon > 0$,

$$\Pr\left\{\left|\frac{\breve{S}_n}{n} - \mathsf{E}\left[\breve{X}\right]\right| \geq \frac{\epsilon}{2}\right\} \ \leq \ \frac{8b\mathsf{E}[|X|]}{n\epsilon^2}.$$

e) Use Figure 1.13, along with (1.34), to show that for all sufficiently large b, $\left|\ \mathsf{E}\left[X - \breve{X}\right]\ \right| \leq \epsilon/2$. Use this to show that

$$\Pr\left\{\left|\frac{\breve{S}_n}{n} - \mathsf{E}[X]\right| \geq \epsilon\right\} \leq \frac{8b\mathsf{E}[|X|]}{n\epsilon^2} \qquad \text{for all large enough } b.$$

f) Use the following equation to justify (1.96).

$$\begin{aligned}\Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}[X]\right| > \epsilon\right\} \quad &= \quad \Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}[X]\right| > \epsilon \bigcap S_n = \breve{S}_n\right\} \\ &\quad + \Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}[X]\right| > \epsilon \bigcap S_n \neq \breve{S}_n\right\}.\end{aligned}$$

Exercise 1.42. Let $\{X_i;\ i \geq 1\}$ be IID rv's with mean 0 and infinite variance. Assume that $\mathsf{E}\left[|X_i|^{1+h}\right] = \beta$ for some given h, $0 < h < 1$ and some finite β. Let $S_n = X_1 + \cdots + X_n$.

a) Show that $\Pr\{|X_i| \geq y\} \leq \beta\, y^{-1-h}$

b) Let $\{\breve{X}_i; i \geq 1\}$ be truncated variables $\breve{X}_i = \begin{cases} b & : \ X_i \geq b \\ X_i & : \ -b \leq X_i \leq b \\ -b & : \ X_i \leq -b \end{cases}$

Show that $\mathsf{E}\left[\breve{X}^2\right] \leq \frac{2\beta b^{1-h}}{1-h}$. Hint: For a nonnegative rv Z, $\mathsf{E}\left[X^2\right] = \int_0^\infty 2\,z\,\Pr\{Z \geq z\}\,dz$ (you can establish this, if you wish, by integration by parts).

c) Let $\breve{S}_n = \breve{X}_1 + \ldots + \breve{X}_n$. Show that $\Pr\left\{S_n \neq \breve{S}_n\right\} \leq n\beta b^{-1-h}$

d) Show that $\Pr\left\{\left|\frac{S_n}{n}\right| \geq \epsilon\right\} \leq \beta\left[\frac{2b^{1-h}}{(1-h)n\epsilon^2} + \frac{n}{b^{1+h}}\right]$.

e) Optimize your bound with respect to b. How fast does this optimized bound approach 0 with increasing n?

Exercise 1.43. (**MS convergence $\implies$ convergence in probability**) Assume that $\{Z_n;\, n \geq 1\}$ is a sequence of rv's and α is a number with the property that $\lim_{n\to\infty} \mathsf{E}\left[(Z_n - \alpha)^2\right] = 0$.

a) Let $\epsilon > 0$ be arbitrary and show that for each $n \geq 0$,

$$\Pr\{|Z_n - \alpha| \geq \epsilon\} \leq \frac{\mathsf{E}\left[(Z_n - \alpha)^2\right]}{\epsilon^2}.$$

b) For the ϵ above, let $\delta > 0$ be arbitrary. Show that there is an integer m such that $\mathsf{E}\left[(Z_n - \alpha)^2\right] \leq \epsilon^2\delta$ for all $n \geq m$.

c) Show that this implies convergence in probability.

Exercise 1.44. Let $X_1, X_2 \ldots$ be a sequence of IID rv's each with mean 0 and variance σ^2. Let $S_n = X_1 + \cdots + X_n$ for all n and consider the random variable $S_n/\sigma\sqrt{n} - S_{2n}/\sigma\sqrt{2n}$. Find the limiting CDF for this sequence of $rv's$ as $n \to \infty$. The point of this exercise is to see clearly that *the CDF* of $S_n/\sigma\sqrt{n} - S_{2n}/\sigma\sqrt{2n}$ is converging in n but that the sequence of rv's is not converging in any reasonable sense.

Exercise 1.45. Verify (1.57), *i.e.*, verify that $\lim_{y\to\infty} y\Pr\{Y \geq y\} = 0$ if $Y \geq 0$ and $\overline{Y} < \infty$. Hint: Show that $y\Pr\{Y \geq y\} \leq \int_{z\geq y} z d\mathsf{F}_Y(z)$ and show that $\lim_{y\to\infty} \int_{z\geq y} z d\mathsf{F}_Y(z) = 0$.

Exercise 1.46. Show that $\prod_{m\geq n}(1 - 1/m) = 0$. Hint: Show that

$$\left(1 - \frac{1}{m}\right) = \exp\left(\ln\left(1 - \frac{1}{m}\right)\right) \leq \exp\left(-\frac{1}{m}\right).$$

Exercise 1.47. Consider a discrete rv X with the PMF

$$\begin{aligned} \mathsf{p}_X(-1) &= (1 - 10^{-10})/2, \\ \mathsf{p}_X(1) &= (1 - 10^{-10})/2, \\ \mathsf{p}_X(10^{12}) &= 10^{-10}. \end{aligned}$$

a) Find the mean and variance of X. Assuming that $\{X_m;\, m \geq 1\}$ is an IID sequence with the distribution of X and that $S_n = X_1 + \cdots + X_n$ for each n, find the mean and variance of S_n. (no explanations needed.)

b) Let $n = 10^6$ and describe the event $\{S_n \leq 10^6\}$ in words. Find an exact expression for $\Pr\{S_n \leq 10^6\} = \mathsf{F}_{S_n}(10^6)$.

c) Find a way to use the union bound to get a simple upper bound and approximation of $1 - \mathsf{F}_{S_n}(10^6)$.

d) Sketch the CDF of S_n for $n = 10^6$. You can choose the horizontal axis for your sketch to go from -1 to $+1$ or from -3×10^3 to 3×10^3 or from -10^6 to 10^6 or from 0 to 10^{12}, whichever you think will best describe this CDF.

e) Now let $n = 10^{10}$. Give an exact expression for $\Pr\{S_n \leq 10^{10}\}$ and show that this can be approximated by e^{-1}. Sketch the CDF of S_n for $n = 10^{10}$, using a horizontal axis going from slightly below 0 to slightly more than 2×10^{12}. Hint: First view S_n as conditioned on an appropriate rv.

f) Can you make a qualitative statement about how the CDF of a rv X affects the required size of n before the WLLN and the CLT provide much of an indication about S_n.

Exercise 1.48. Let $\{Y_n;\, n \geq 1\}$ be a sequence of rv's and assume that $\lim_{n\to\infty} \mathsf{E}\,[|Y_n|] = 0$. Show that $\{Y_n;\, n \geq 1\}$ converges to 0 in probability. Hint 1: Look for the easy way. Hint 2: The easy way uses the Markov inequality.

Chapter 1

INTRODUCTION AND REVIEW OF PROBABILITY

1.1 Probability models

Probability theory is a central field of mathematics, widely applicable to scientific, technological, and human situations involving uncertainty. The most obvious applications are to situations, such as games of chance, in which repeated trials of essentially the same procedure lead to differing outcomes. For example, when we flip a coin, roll a die, pick a card from a shuffled deck, or spin a ball onto a roulette wheel, the procedure is the same from one trial to the next, but the outcome (heads (H) or tails (T) in the case of a coin, one to six in the case of a die, etc.) varies from one trial to another in a seemingly random fashion.

For the case of flipping a coin, the outcome of the flip could be predicted from the initial position, velocity, and angular momentum of the coin and from the nature of the surface on which it lands. Thus, in one sense, a coin flip is deterministic rather than random and the same can be said for the other examples above. When these initial conditions are unspecified, however, as when playing these games, the outcome can again be viewed as random in some intuitive sense.

Many scientific experiments are similar to games of chance in the sense that multiple trials of apparently the *same* procedure lead to results that *vary* from one trial to another. In some cases, this variation is due to slight variations in the experimental procedure, in some it is due to noise, and in some, such as in quantum mechanics, the randomness is generally believed to be fundamental. Similar situations occur in many types of systems, especially those in which noise and random delays are important. Some of these systems, rather than being repetitions of a common basic procedure, are systems that evolve over time while still containing a sequence of underlying similar random occurrences.

This intuitive notion of randomness, as described above, is a very special kind of uncertainty. Rather than involving a lack of understanding, it involves a type of uncertainty that can lead to probabilistic models with precise results. As in any scientific field, the models might or might not correspond to reality very well, but when they do correspond to reality, there

is the sense that the situation is completely understood, while still being random.

For example, we all feel that we understand flipping a coin or rolling a die, but still accept randomness in each outcome. The theory of probability was developed particularly to give precise and quantitative understanding to these types of situations. The remainder of this section introduces this relationship between the precise view of probability theory and the intuitive view as used in applications and everyday language.

After this introduction, the following sections review probability theory as a mathematical discipline, with a special emphasis on the laws of large numbers. In the final section of this chapter, we use the theory and the laws of large numbers to obtain a fuller understanding of the relationship between theory and the real world.[1]

Probability theory, as a mathematical discipline, started to evolve in the 17th century and was initially focused on games of chance. The importance of the theory grew rapidly, particularly in the 20th century, and it now plays a central role in risk assessment, statistics, data networks, operations research, information theory, control theory, theoretical computer science, quantum theory, game theory, neurophysiology, and many other fields.

The core concept in probability theory is that of a *probability model*. Given the extent of the theory, both in mathematics and in applications, the simplicity of probability models is surprising. The first component of a probability model is a *sample space*, which is a *set* whose elements are called *outcomes* or *sample points*. Probability models are particularly simple in the special case where the sample space is finite,[2] and we consider only this case in the remainder of this section. The second component of a probability model is a class of *events*, which can be considered for now simply as the class of all subsets of the sample space. The third component is a *probability measure*, which can be regarded for now as the assignment of a nonnegative number to each outcome, with the restriction that these numbers must sum to one over the sample space. The probability of an event is the sum of the probabilities of the outcomes comprising that event.

These probability models play a dual role. In the first, the many known results about various classes of models, and the many known relationships between models, constitute the essence of probability theory. Thus one often studies a model not because of any relationship to the real world, but simply because the model provides a building block or example useful for the theory and thus ultimately useful for other models. In the other role, when probability theory is applied to some game, experiment, or some other situation involving randomness, a probability model is used to represent the experiment (in what follows, we refer to all of these random situations as experiments).

For example, the standard probability model for rolling a die uses $\{1, 2, 3, 4, 5, 6\}$ as the sample space, with each possible outcome having probability $1/6$. An *odd* result, *i.e.*, the subset $\{1, 3, 5\}$, is an example of an event in this sample space, and this event has probability

[1] It would be appealing to show how probability theory evolved from real-world random situations, but probability theory, like most mathematical theories, has evolved from complex interactions between theoretical developments and initially over-simplified models of real situations. The successes and flaws of such models lead to refinements of the models and the theory, which in turn suggest applications to totally different fields.

[2] A number of mathematical issues arise with infinite sample spaces, as discussed in the following section.

1/2. The correspondence between model and actual experiment seems straightforward here. Both have the same set of outcomes and, given the symmetry between faces of the die, the choice of equal probabilities seems natural. On closer inspection, there is the following important difference between the model and the actual rolling of a die.

The above model corresponds to a single roll of a die, with a probability defined for each possible outcome. In a real-world experiment where a single die is rolled, an outcome k from 1 to 6 occurs, but there is no *observable* probability for k.

Our intuitive notion of rolling dice, however, involves an experiment with repeated rolls of a die or rolls of different dice. With n rolls altogether, there are are 6^n possible outcomes, one for each possible n-tuple of individual die outcomes. The standard probability model for this repeated-roll experiment is to assign probability 6^{-n} to each possible n-tuple. In this n-repetition experiment, the real-world relative frequency of k, *i.e.*, the fraction of rolls for which the result is k, can be compared with the sample value of the relative frequency of k in the model for repeated rolls. The sample value of the relative frequency of k in this n-repetition model resembles the probability of k in the single-roll experiment in a way to be explained later. This relationship through relative frequencies in a repeated experiment helps overcome the non-observable nature of probabilities in the real world.

1.1.1 The sample space of a probability model

An *outcome* or *sample point* in a probability model corresponds to a complete result (with all detail specified) of the experiment being modeled. For example, a game of cards is often appropriately modeled by the arrangement of cards within a shuffled 52 card deck, thus giving rise to a set of 52! outcomes (incredibly detailed, but trivially simple in structure), even though the entire deck might not be played in one trial of the game. A poker hand with 4 aces is an *event* rather than an *outcome* in this model, since many arrangements of the cards can give rise to 4 aces in a given hand. The possible outcomes in a probability model (and in the experiment being modeled) are mutually exclusive and collectively constitute the entire sample space (space of possible results). An outcome is often called a *finest grain* result of the model in the sense that an outcome ω contains no subsets other than the empty set ϕ and the singleton subset $\{\omega\}$. Thus events typically give only partial information about the result of the experiment, whereas an outcome fully specifies the result.

In choosing the sample space for a probability model of an experiment, we often omit details that appear irrelevant for the purpose at hand. Thus in modeling the set of outcomes for a coin toss as $\{H, T\}$, we ignore the type of coin, the initial velocity and angular momentum of the toss, etc. We also omit the rare possibility that the coin comes to rest on its edge. Sometimes, conversely, the sample space is enlarged beyond what is relevant in the interest of structural simplicity. An example is the above use of a shuffled deck of 52 cards.

The choice of the sample space in a probability model is similar to the choice of a mathematical model in any branch of science. That is, one simplifies the physical situation by eliminating detail of little apparent relevance. One often does this in an iterative way, using a very simple model to acquire initial understanding, and then successively choosing more detailed models based on the understanding from earlier models.

The mathematical theory of probability views the sample space simply as an abstract set of elements, and from a strictly mathematical point of view, the idea of doing an experiment and getting an outcome is a distraction. For visualizing the correspondence between the theory and applications, however, it is better to view the abstract set of elements as the set of possible outcomes of an idealized experiment in which, when the idealized experiment is performed, one and only one of those outcomes occurs. The two views are mathematically identical, but it will be helpful to refer to the first view as a probability model and the second as an idealized experiment. In applied probability texts and technical articles, these idealized experiments, rather than real-world situations are often the primary topic of discussion.[3]

1.1.2 Assigning probabilities for finite sample spaces

The word *probability* is widely used in everyday language, and most of us attach various intuitive meanings[4] to the word. For example, everyone would agree that something virtually impossible should be assigned a probability close to 0 and something virtually certain should be assigned a probability close to 1. For these special cases, this provides a good rationale for choosing probabilities. The relationship between *virtually* and *close to* are unclear at the moment, but if there is some implied limiting process, we would all agree that, in the limit, certainty and impossibility correspond to probabilities 1 and 0 respectively.

Between virtual impossibility and certainty, if one outcome appears to be closer to certainty than another, its probability should be correspondingly greater. This intuitive notion is imprecise and highly subjective; it provides little rationale for choosing numerical probabilities for different outcomes, and, even worse, little rationale justifying that probability models bear any precise relation to real-world situations.

Symmetry can often provide a better rationale for choosing probabilities. For example, the symmetry between H and T for a coin, or the symmetry between the the six faces of a die, motivates assigning equal probabilities, $1/2$ each for H and T and $1/6$ each for the six faces of a die. This is reasonable and extremely useful, but there is no completely convincing reason for choosing probabilities based on symmetry.

Another approach is to perform the experiment many times and choose the probability of each outcome as the relative frequency of that outcome (*i.e.*, the number of occurrences of that outcome divided by the total number of trials). Experience shows that the relative frequency of an outcome often approaches a limiting value with an increasing number of trials. Associating the probability of an outcome with that limiting relative frequency is certainly close to our intuition and also appears to provide a testable criterion between model and real world. This criterion is discussed in Sections 1.6.1 and 1.6.2 and provides a very concrete way to use probabilities, since it suggests that the randomness in a single

[3] This is not intended as criticism, since we will see that there are good reasons to concentrate initially on such idealized experiments. However, readers should always be aware that modeling errors are the major cause of misleading results in applications of probability, and thus modeling must be seriously considered before using the results.

[4] It is popular to try to define probability by likelihood, but this is unhelpful since the words are essentially synonyms.

trial tends to disappear in the aggregate of many trials. Other approaches to choosing probability models will be discussed later.

1.2 The axioms of probability theory

As the applications of probability theory became increasingly varied and complex during the 20th century, the need arose to put the theory on a firm mathematical footing. This was accomplished by an axiomatization of the theory, successfully carried out by the great Russian mathematician A. N. Kolmogorov [14] in 1932. Before stating and explaining these axioms of probability theory, the following two examples explain why the simple approach of the last section, assigning a probability to each sample point, often fails with infinite sample spaces.

Example 1.2.1. Suppose we want to model the phase of a sine wave, where the phase is viewed as being "uniformly distributed" between 0 and 2π. If this phase is the only quantity of interest, it is reasonable to choose a sample space consisting of the set of real numbers between 0 and 2π. There are uncountably[5] many possible phases between 0 and 2π, and with any reasonable interpretation of uniform distribution, one must conclude that each sample point has probability zero. Thus, the simple approach of the last section leads us to conclude that any event in this space with a finite or countably infinite set of sample points should have probability zero. That simple approach does not help in finding the probability, say, of the interval $(0, \pi)$.

For this example, the appropriate view is that taken in all elementary probability texts, namely to assign a probability density $\frac{1}{2\pi}$ to the phase. The probability of an event can then usually be found by integrating the density over that event. Useful as densities are, however, they do not lead to a general approach over arbitrary sample spaces.[6]

Example 1.2.2. Consider an infinite sequence of coin tosses. The usual probability model is to assign probability 2^{-n} to each possible initial n-tuple of individual outcomes. Then in the limit $n \to \infty$, the probability of any given sequence is 0. Again, expressing the probability of an event involving infinitely many tosses as a sum of individual sample-point probabilities does not work. The obvious approach (which we often adopt for this and similar situations) is to evaluate the probability of any given event as an appropriate limit, as $n \to \infty$, of the outcome from the first n tosses.

We will later find a number of situations, even for this almost trivial example, where working with a finite number of elementary experiments and then going to the limit is very awkward. One example, to be discussed in detail later, is the strong law of large numbers (SLLN). This

[5] A set is uncountably infinite if it is infinite and its members cannot be put into one-to-one correspondence with the positive integers. For example the set of real numbers over some interval such as $(0, 2\pi)$ is uncountably infinite. The Wikipedia article on countable sets provides a friendly introduction to the concepts of countability and uncountability.

[6] It is possible to avoid the consideration of infinite sample spaces here by quantizing the possible phases. This is analogous to avoiding calculus by working only with discrete functions. Both usually result in both artificiality and added complexity.

law looks directly at events consisting of infinite length sequences and is best considered in the context of the axioms to follow.

Although appropriate probability models can be generated for simple examples such as those above, there is a need for a consistent and general approach. In such an approach, rather than assigning probabilities to sample points, which are then used to assign probabilities to events, *probabilities must be associated directly with events.* The axioms to follow establish consistency requirements between the probabilities of different events. The axioms, and the corollaries derived from them, are consistent with one's intuition, and, for finite sample spaces, are consistent with our earlier approach. Dealing with the countable unions of events in the axioms will be unfamiliar to some students, but will soon become both familiar and consistent with intuition.

The strange part of the axioms comes from the fact that defining the class of events as the set of *all* subsets of the sample space is usually inappropriate when the sample space is uncountably infinite. What is needed is a class of events that is large enough that we can almost forget that some very strange subsets are excluded. This is accomplished by having two simple sets of axioms, one defining the class of events,[7] and the other defining the relations between the probabilities assigned to these events. In this theory, all events have probabilities, but those truly weird subsets that are not events do not have probabilities. This will be discussed more after giving the axioms for events.

The axioms for events use the standard notation of set theory. Let Ω be the set of all sample points for a given experiment. The events are subsets of the sample space. The union of n subsets (events) $A_1, A_2, \cdots, A_n$ is denoted by either $\bigcup_{i=1}^{n} A_i$ or $A_1 \bigcup \cdots \bigcup A_n$, and consists of all points in at least one of $A_1, \ldots, A_n$. Similarly, the intersection of these subsets is denoted by either $\bigcap_{i=1}^{n} A_i$ or[8] $A_1 A_2 \cdots A_n$ and consists of all points in all of $A_1, \ldots, A_n$.

A *sequence* of events is a collection of events in one-to-one correspondence with the positive integers, *i.e.*, $A_1, A_2, \ldots$, ad infinitum. A countable union, $\bigcup_{i=1}^{\infty} A_i$ is the set of points in one or more of $A_1, A_2, \ldots$. Similarly, a countable intersection $\bigcap_{i=1}^{\infty} A_i$ is the set of points in all of $A_1, A_2, \ldots$. Finally, the complement A^c of a subset (event) A is the set of points in Ω but not A.

1.2.1 Axioms for events

Given a sample space Ω, the class of subsets of Ω that constitute the set of events satisfies the following axioms:

1. Ω is an event.
2. For every sequence of events $A_1, A_2, \ldots$,, the union $\bigcup_{n=1}^{\infty} A_n$ is an event.
3. For every event A, the complement A^c is an event.

There are a number of important corollaries of these axioms. First, the empty set ϕ is an event. This follows from Axioms 1 and 3, since $\phi = \Omega^c$. The empty set does not correspond

[7] A class of elements satisfying these axioms is called a σ-algebra or, less commonly, a σ-field.

[8] Intersection is also sometimes denoted as $A_1 \bigcap \cdots \bigcap A_n$, but is usually abbreviated as $A_1 A_2 \cdots A_n$.

to our intuition about events, but the theory would be extremely awkward if it were omitted. Second, every finite union of events is an event. This follows by expressing $A_1 \bigcup \cdots \bigcup A_n$ as $\bigcup_{i=1}^{\infty} A_i$ where $A_i = \phi$ for all $i > n$. Third, every finite or countable intersection of events is an event. This follows from deMorgan's law,

$$\left[\bigcup_n A_n\right]^c = \bigcap_n A_n^c.$$

Although we will not make a big fuss about these axioms in the rest of the text, we will be careful to use only complements and countable unions and intersections in our analysis. Thus subsets that are not events will not arise.

Note that the axioms do not say that all subsets of Ω are events. In fact, there are many rather silly ways to define classes of events that obey the axioms. For example, the axioms are satisfied by choosing only the universal set Ω and the empty set ϕ to be events. We shall avoid such trivialities by assuming that for each sample point ω, the singleton subset $\{\omega\}$ is an event. For finite sample spaces, this assumption, plus the axioms above, imply that all subsets are events.

For uncountably infinite sample spaces, such as the sinusoidal phase above, this assumption, plus the axioms above, still leaves considerable freedom in choosing a class of events. As an example, the class of all subsets of Ω satisfies the axioms but surprisingly does not allow the probability axioms to be satisfied in any sensible way. How to choose an appropriate class of events requires an understanding of measure theory which would take us too far afield for our purposes. Thus we neither assume nor develop measure theory here.[9]

From a pragmatic standpoint, we start with the class of events of interest, such as those required to define the random variables needed in the problem. That class is then extended so as to be closed under complementation and countable unions. Measure theory shows that this extension is always possible, and we simply accept that as a known result.

1.2.2 Axioms of probability

Given any sample space Ω and any class of events $\mathcal{E}$ satisfying the axioms of events, a probability rule is a function $\Pr\{\}$ mapping each $A \in \mathcal{E}$ to a (finite[10]) real number in such a way that the following three probability axioms[11] hold:

1. $\Pr\{\Omega\} = 1$.

2. For every event A, $\Pr\{A\} \geq 0$.

[9] There is no doubt that measure theory is useful in probability theory, and serious students of probability should certainly learn measure theory at some point. For application-oriented people, however, it seems advisable to acquire more insight and understanding of probability, at a graduate level, before concentrating on the abstractions and subtleties of measure theory.

[10] The word *finite* is redundant here, since the set of real numbers, by definition, does not include $\pm\infty$. The set of real numbers with $\pm\infty$ appended, is called the set of *extended* real numbers

[11] Sometimes finite additivity, (1.3), is added as an additional axiom. This addition is quite intuitive and avoids the technical and somewhat peculiar proofs given for (1.2) and (1.3).

3. The probability of the union of any sequence $A_1, A_2, \ldots$ of disjoint events is given by

$$\Pr\Big\{\bigcup_{n=1}^{\infty} A_n\Big\} = \sum_{n=1}^{\infty} \Pr\{A_n\}, \tag{1.1}$$

where $\sum_{n=1}^{\infty} \Pr\{A_n\}$ is shorthand for $\lim_{m\to\infty} \sum_{n=1}^{m} \Pr\{A_n\}$.

The axioms imply the following useful corollaries:

$$\Pr\{\phi\} = 0 \tag{1.2}$$

$$\Pr\Big\{\bigcup_{n=1}^{m} A_n\Big\} = \sum_{n=1}^{m} \Pr\{A_n\} \qquad \text{for } A_1, \ldots, A_m \text{ disjoint} \tag{1.3}$$

$$\Pr\{A^c\} = 1 - \Pr\{A\} \qquad \text{for all } A \tag{1.4}$$

$$\Pr\{A\} \le \Pr\{B\} \qquad \text{for all } A \subseteq B \tag{1.5}$$

$$\Pr\{A\} \le 1 \qquad \text{for all } A \tag{1.6}$$

$$\sum_n \Pr\{A_n\} \le 1 \qquad \text{for } A_1, \ldots, \text{ disjoint} \tag{1.7}$$

$$\Pr\Big\{\bigcup_{n=1}^{\infty} A_n\Big\} = \lim_{m\to\infty} \Pr\Big\{\bigcup_{n=1}^{m} A_n\Big\} \tag{1.8}$$

$$\Pr\Big\{\bigcup_{n=1}^{\infty} A_n\Big\} = \lim_{n\to\infty} \Pr\{A_n\} \qquad \text{for } A_1 \subseteq A_2 \subseteq \cdots. \tag{1.9}$$

To verify (1.2), consider a sequence of events, $A_1, A_2, \ldots$, for which $A_n = \phi$ for each n. These events are disjoint since ϕ contains no outcomes, and thus has no outcomes in common with itself or any other event. Also, $\bigcup_n A_n = \phi$ since this union contains no outcomes. Axiom 3 then says that

$$\Pr\{\phi\} = \lim_{m\to\infty} \sum_{n=1}^{m} \Pr\{A_n\} = \lim_{m\to\infty} m\Pr\{\phi\}.$$

Since $\Pr\{\phi\}$ is a real number, this implies that $\Pr\{\phi\} = 0$.

To verify (1.3), apply Axiom 3 to the disjoint sequence $A_1, \ldots, A_m, \phi, \phi, \ldots$.

One might reasonably guess that (1.3), along with Axioms 1 and 2 implies Axiom 3. Exercise 1.3 shows why this guess is incorrect.

To verify (1.4), note that $\Omega = A \bigcup A^c$. Then apply (1.3) to the disjoint sets A and A^c.

To verify (1.5), note that if $A \subseteq B$, then $B = A \bigcup (B - A)$ where $B - A$ is an alternate way to write $B \bigcap A^c$. We see then that A and $B - A$ are disjoint, so from (1.3),

$$\Pr\{B\} = \Pr\Big\{A \bigcup (B - A)\Big\} = \Pr\{A\} + \Pr\{B - A\} \ge \Pr\{A\},$$

where we have used Axiom 2 in the last step.

To verify (1.6) and (1.7), first substitute Ω for B in (1.5) and then substitute $\bigcup_n A_n$ for A.

Finally, (1.8) is established in Exercise 1.4, part (e), and (1.9) is a simple consequence of (1.8).

The axioms specify the probability of any *disjoint* union of events in terms of the individual event probabilities, but what about a finite or countable union of arbitrary events? Exercise 1.4 (b) shows that in this case, (1.3) can be generalized to

$$\Pr\Big\{\bigcup_{n=1}^{m} A_n\Big\} = \sum_{n=1}^{m} \Pr\{B_n\}\,, \tag{1.10}$$

where $B_1 = A_1$ and for each $n > 1$, $B_n = A_n - \bigcup_{j=1}^{n-1} A_j$ is the set of points in A_n but not in any of the sets $A_1, \ldots, A_{n-1}$. The probability of a countable union is then given by (1.8). In order to use this, one must know not only the event probabilities for $A_1, A_2 \ldots$, but also the probabilities of their intersections. The union bound, which is derived in Exercise 1.4 (c), depends only on the individual event probabilities, and gives the following frequently useful upper bound on the union probability.

$$\Pr\Big\{\bigcup_{n} A_n\Big\} \leq \sum_{n} \Pr\{A_n\} \qquad \text{(Union bound)}. \tag{1.11}$$

1.3 Probability review

1.3.1 Conditional probabilities and statistical independence

Definition 1.3.1. *For any two events A and B (with $\Pr\{B\} > 0$), the conditional probability of A, conditional on B, is defined by*

$$\Pr\{A|B\} = \Pr\{AB\}\,/\Pr\{B\}\,. \tag{1.12}$$

One visualizes an experiment that has been partly carried out with B as the result. Then $\Pr\{A|B\}$ can be viewed as the probability of A normalized to a sample space restricted to event B. Within this restricted sample space, we can view B as the sample space (*i.e.*, as the set of outcomes that remain possible upon the occurrence of B) and AB as an event within this sample space. For a fixed event B, we can visualize mapping each event A in the original space to event AB in the restricted space. It is easy to see that the event axioms are still satisfied in this restricted space. Assigning probability $\Pr\{A|B\}$ to each event AB in the restricted space, it is easy to see that the axioms of probability are satisfied when B is regarded as the entire sample space. In other words, everything we know about probability can also be applied to such a restricted probability space.

Definition 1.3.2. *Two events, A and B, are statistically independent (or, more briefly,* independent*) if*

$$\Pr\{AB\} = \Pr\{A\}\Pr\{B\}\,.$$

For $\Pr\{B\} > 0$, this is equivalent to $\Pr\{A|B\} = \Pr\{A\}$. This latter form corresponds to our intuitive view of independence, since it says that the observation of B does not change the probability of A. Such intuitive statements about "observation" and "occurrence" are helpful in reasoning probabilistically, but sometimes cause confusion. For example, Bayes law, in the form $\Pr\{A|B\}\Pr\{B\} = \Pr\{B|A\}\Pr\{A\}$, is an immediate consequence of the

definition of conditional probability in (1.12). However, if we can only interpret $\Pr\{A|B\}$ when B is 'observed' or occurs 'before' A, then we cannot interpret $\Pr\{B|A\}$ and $\Pr\{A|B\}$ together. This caused immense confusion in probabilistic arguments before the axiomatic theory was developed.

The notion of independence is of vital importance in defining, and reasoning about, probability models. We will see many examples where very complex systems become very simple, both in terms of intuition and analysis, when appropriate quantities are modeled as statistically independent. An example will be given in the next subsection where repeated independent experiments are used to understand arguments about relative frequencies.

Often, when the assumption of independence turns out to be oversimplified, it is reasonable to assume conditional independence, where A and B are said to be *conditionally independent* given C if $\Pr\{AB|C\} = \Pr\{A|C\}\Pr\{B|C\}$. Most of the stochastic processes to be studied here are characterized by particular forms of independence or conditional independence.

For more than two events, the definition of statistical independence is a little more complicated.

Definition 1.3.3. *The events $A_1, \ldots, A_n$, $n > 2$ are statistically independent if for each collection S of two or more of the integers 1 to n.*

$$\Pr\Big\{\bigcap\nolimits_{i\in S} A_i\Big\} = \prod\nolimits_{i\in S} \Pr\{A_i\}\,. \tag{1.13}$$

This includes the entire collection $\{1, \ldots, n\}$, so one necessary condition for independence is that

$$\Pr\Big\{\bigcap\nolimits_{i=1}^{n} A_i\Big\} = \prod\nolimits_{i=1}^{n} \Pr\{A_i\}\,. \tag{1.14}$$

It might be surprising that (1.14) does not imply (1.13), but the example in Exercise 1.5 will help clarify this. This definition will become clearer (and simpler) when we see how to view independence of events as a special case of independence of random variables.

1.3.2 Repeated idealized experiments

Much of our intuitive understanding of probability comes from the notion of repeating the same idealized experiment many times (*i.e.*, performing multiple trials of the same experiment). However, the axioms of probability contain no explicit recognition of such repetitions. The appropriate way to handle n repetitions of an idealized experiment is through an extended experiment whose sample points are n-tuples of sample points from the original experiment. Such an extended experiment is viewed as n *trials* of the original experiment. The notion of multiple trials of a given experiment is so common that one sometimes fails to distinguish between the original experiment and an extended experiment with multiple trials of the original experiment.

To be more specific, given an original sample space Ω, the sample space of an n-repetition model is the Cartesian product

$$\Omega^{\times n} = \{(\omega_1, \omega_2, \ldots, \omega_n) \,:\, \omega_i \in \Omega \text{ for each } i, 1 \le i \le n\}, \tag{1.15}$$

i.e., the set of all n-tuples for which each of the n components of the n-tuple is an element of the original sample space Ω. Since each sample point in the n-repetition model is an n-tuple of points from the original Ω, it follows that an event in the n-repetition model is a subset of $\Omega^{\times n}$, *i.e.*, a collection of n-tuples $(\omega_1, \ldots, \omega_n)$, where each ω_i is a sample point from Ω. This class of events in $\Omega^{\times n}$ should include each event of the form $\{(A_1 A_2 \cdots A_n)\}$, where $\{(A_1 A_2 \cdots A_n)\}$ denotes the collection of n-tuples $(\omega_1, \ldots, \omega_n)$ where $\omega_i \in A_i$ for $1 \leq i \leq n$. The set of events (for n-repetitions) mult also be extended to be closed under complementation and countable unions and intersections.

The simplest and most natural way of creating a probability model for this extended sample space and class of events is through the assumption that the n-trials are statistically independent. More precisely, we assume that for each extended event $\{(A_1 A_2 \cdots A_n)\}$ contained in $\Omega^{\times n}$, we have

$$\Pr\{(A_1 A_2 \cdots A_n)\} = \prod_{i=1}^{n} \Pr\{A_i\}, \tag{1.16}$$

where $\Pr\{A_i\}$ is the probability of event A_i in the original model. Note that since Ω can be substituted for any A_i in this formula, the subset condition of (1.13) is automatically satisfied. In other words, for any probability model, there is an extended independent n-repetition model for which the events in each trial are independent of those in the other trials. In what follows, we refer to this as the probability model for n independent identically distributed (IID) trials of a given experiment.

The niceties of how to create this model for n IID arbitrary experiments depend on measure theory, but we simply rely on the existence of such a model and the independence of events in different repetitions. What we have done here is very important conceptually. A probability model for an experiment does not say anything directly about repeated experiments. However, questions about independent repeated experiments can be handled directly within this extended model of n IID repetitions. This can also be extended to a countable number of IID trials.

1.3.3 Random variables

The outcome of a probabilistic experiment often specifies a collection of numerical values such as temperatures, voltages, numbers of arrivals or departures in various time intervals, etc. Each such numerical value varies, depending on the particular outcome of the experiment, and thus can be viewed as a mapping from the set Ω of sample points to the set $\mathbb{R}$ of real numbers (note that $\mathbb{R}$ does not include $\pm\infty$). These mappings from sample points to real numbers are called random variables.

Definition 1.3.4. *A random variable (rv) is essentially a function X from the sample space Ω of a probability model to the set of real numbers $\mathbb{R}$. Three modifications are needed to make this precise. First, X might be undefined or infinite for a subset of Ω that has 0 probability.*[12] *Second, the mapping $X(\omega)$ must have the property that $\{\omega \in \Omega : X(\omega) \leq x\}$*

[12] For example, suppose Ω is the closed interval $[0, 1]$ of real numbers with a uniform probability distribution over $[0, 1]$. If $X(\omega) = 1/\omega$, then the sample point 0 maps to ∞ but X is still regarded as a rv. These subsets of 0 probability are usually ignored, both by engineers and mathematicians. Thus, for example, the set $\{\omega \in \Omega : X(\omega) \leq x\}$ means the set for which $X(\omega)$ is both defined and satisfies $X(\omega) \leq x$.

is an event[13] *for each* $x \in \mathbb{R}$. *Third, every finite set of rv's* $X_1, \ldots, X_n$ *has the property that* $\{\omega : X_1(\omega) \leq x_1, \ldots, X_n(\omega) \leq x_n\}$ *is an event for each* $x_1 \in \mathbb{R}, \ldots, x_n \in \mathbb{R}$.

As with any function, there is often confusion between the function itself, which is called X in the definition above, and the value $X(\omega)$ the function takes on for a sample point ω. This is particularly prevalent with random variables (rv's) since we intuitively associate a rv with its sample value when an experiment is performed. We try to control that confusion here by using X, $X(\omega)$, and x, respectively, to refer to the rv, the sample value taken for a given sample point ω, and a generic sample value.

Definition 1.3.5. *The distribution function*[14] $\mathsf{F}_X(x)$ *of a random variable (rv)* X *is a function,* $\mathbb{R} \to \mathbb{R}$, *defined by* $\mathsf{F}_X(x) = \Pr\{\omega \in \Omega : X(\omega) \leq x\}$. *The argument* ω *is usually omitted for brevity, so* $\mathsf{F}_X(x) = \Pr\{X \leq x\}$.

Note that x is the argument of $\mathsf{F}_X(x)$ and the subscript X denotes the particular rv under consideration. As illustrated in Figure 1.1, the distribution function $\mathsf{F}_X(x)$ is nondecreasing with x and must satisfy the limits $\lim_{x\to-\infty} \mathsf{F}_X(x) = 0$ and $\lim_{x\to\infty} \mathsf{F}_X(x) = 1$. It is not hard to show, from the axioms, that $\mathsf{F}_X(x)$ is continuous from the right (*i.e.*, that for every $x \in \mathbb{R}$, $\lim_{k\to\infty} \mathsf{F}_X(x + 1/k) = \mathsf{F}_X(x)$.

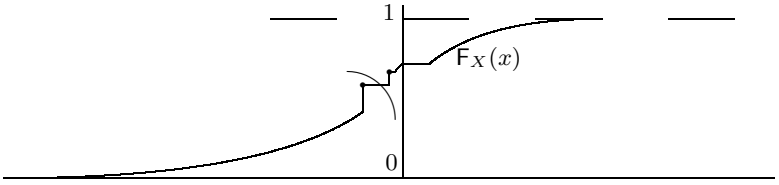


Figure 1.1: Example of a distribution function for a rv that is neither continuous nor discrete. If $\mathsf{F}_X(x)$ has a discontinuity at some x_o, it means that there is a discrete probability at x_o equal to the magnitude of the discontinuity. In this case $\mathsf{F}_X(x_o)$ is given by the height of the upper point at the discontinuity.

Because of the definition of a rv, the set $\{X \leq x\}$ for any rv X and any real number x must be an event, and thus $\Pr\{X \leq x\}$ must be defined for all real x.

The concept of a rv is often extended to complex random variables (rv's) and vector rv's. A *complex random variable* is a mapping from the sample space to the set of finite complex numbers, and a *vector random variable (rv)* is a mapping from the sample space to the finite vectors in some finite dimensional vector space. Another extension is that of defective rvs. X is *defective* if there is an event of *positive* probability for which the mapping is either undefined or defined to be either $+\infty$ or $-\infty$. When we refer to random variables in this text (without any modifier such as complex, vector, or defective), we explicitly restrict attention to the original definition, i.e., a function from Ω to $\mathbb{R}$.

[13] These last two modifications are technical limitations connected with measure theory. They can usually be ignored, since they are satisfied in all but the most bizarre conditions. However, just as it is important to know that not all subsets in a probability space are events, one should know that not all functions from Ω to $\mathbb{R}$ are rv's.

[14] The distribution function is sometimes referred to as the cumulative distribution function.

If X has only a finite or countable number of possible sample values, say $x_1, x_2, \ldots$, the probability $\Pr\{X = x_i\}$ of each sample value x_i is called the probability mass function (PMF) at x_i and denoted by $\mathsf{p}_X(x_i)$; such a random variable is called *discrete*. The distribution function of a discrete rv is a 'staircase function,' staying constant between the possible sample values and having a jump of magnitude $\mathsf{p}_X(x_i)$ at each sample value x_i. Thus the PMF and the distribution function each specify the other for discrete rv's.

If the distribution function $\mathsf{F}_X(x)$ of a rv X has a (finite) derivative at x, the derivative is called the *probability density* (or the density) of X at x and denoted by $\mathsf{f}_X(x)$; for sufficiently small δ, $\delta\mathsf{f}_X(x)$ then approximates the probability that X is mapped to a value between x and $x + \delta$. If the density exists for all x, the rv is said to be *continuous*. More generally, if there is a function $\mathsf{f}_X(x)$ such that, for each $x \in \mathbb{R}$, the distribution function satisfies $\int_{-\infty}^{x} \mathsf{f}_X(y)\, dy$, then the rv is said to be continuous and f_X is the probability density. This generalization allows the density to be discontinuous. In other words, a continuous rv requires a little more than a continuous distribution function and a little less than a continuous density.

Elementary probability courses work primarily with the PMF and the density, since they are convenient for computational exercises. We will often work with the distribution function here. This is partly because it is always defined, partly to avoid saying everything thrice, for discrete, continuous, and other rv's, and partly because the distribution function is often most important in limiting arguments such as steady-state time-average arguments. For distribution functions, density functions, and PMF's, the subscript denoting the rv is often omitted if the rv is clear from the context. The same convention is used for complex rv's and vector rv's.

Appendix A lists the PMF's of a number of widely used discrete rv's and the densities of some equally popular continuous rv's. The mean, variance and moment generating functions of these variables are also listed for ready reference.

1.3.4 Multiple random variables and conditional probabilities

Often we must deal with multiple random variables (rv's) in a single probability experiment. If $X_1, X_2, \ldots, X_n$ are rv's or the components of a vector rv, their joint distribution function is defined by

$$\mathsf{F}_{X_1 \cdots X_n}(x_1, x_2, \ldots, x_n) = \Pr\{\omega \in \Omega : X_1(\omega) \leq x_1, X_2(\omega) \leq x_2, \ldots, X_n(\omega) \leq x_n\}. \quad (1.17)$$

This definition goes a long way toward explaining why we need the notion of a sample space Ω when all we want to talk about is a set of rv's. The distribution function of a rv fully describes the individual behavior of that rv, but Ω and the above mappings are needed to describe how the rv's interact.

For a vector rv $\boldsymbol{X}$ with components $X_1, \ldots, X_n$, or a complex rv X with real and imaginary parts X_1, X_2, the distribution function is also defined by (1.17). Note that $\{X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\}$ is an event and the corresponding probability is nondecreasing in each argument x_i. Also the distribution function of any subset of random variables is obtained

by setting the other arguments to $+\infty$. For example, the distribution of a single rv (called a marginal distribution) is given by

$$\mathsf{F}_{X_i}(x_i) = \mathsf{F}_{X_1 \cdots X_{i-1} X_i X_{i+1} \cdots X_n}(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty).$$

If the rv's are all discrete, there is a joint PMF which specifies and is specified by the joint distribution function. It is given by

$$\mathsf{p}_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \Pr\{X_1 = x_1, \ldots, X_n = x_n\}.$$

Similarly, if the joint distribution function is differentiable everywhere, it specifies and is specified by the joint probability density,

$$\mathsf{f}_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \frac{\partial^n \mathsf{F}(x_1, \ldots, x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n}.$$

.

Two rv's, say X and Y, are *statistically independent* (or, more briefly, *independent*) if

$$\mathsf{F}_{XY}(x, y) = \mathsf{F}_X(x)\mathsf{F}_Y(y) \qquad \text{for each } x \in \mathbb{R}, y \in \mathbb{R}. \tag{1.18}$$

If X and Y are discrete rv's then the definition of independence in (1.18) is equivalent to the corresponding statement for PMF's,

$$\mathsf{p}_{XY}(x_i y_j) = \mathsf{p}_X(x_i)\mathsf{p}_Y(y_j) \qquad \text{for each value } x_i \text{ of } X \text{ and } y_j \text{ of } Y.$$

Since $\{X = x_i\}$ and $\{Y = y_j\}$ are events, the conditional probability of $\{X = x_i\}$ conditional on $\{Y = y_j\}$ (assuming $\mathsf{p}_Y(y_j) > 0$) is given by (1.12) to be

$$\mathsf{p}_{X|Y}(x_i \mid y_j) = \frac{\mathsf{p}_{XY}(x_i, y_j)}{\mathsf{p}_Y(y_j)}.$$

If $\mathsf{p}_{X|Y}(x_i \mid y_j) = \mathsf{p}_X(x_i)$ for all i, j, then it is seen that X and Y are independent. This captures the intuitive notion of independence better than (1.18) for discrete rv's , since it can be viewed as saying that the PMF of X is not affected by the sample value of Y.

If X and Y have a joint density, then (1.18) is equivalent to

$$\mathsf{f}_{XY}(x, y) = \mathsf{f}_X(x)\mathsf{f}_Y(y) \qquad \text{for each } x \in \mathbb{R}, y \in \mathbb{R}.$$

If $\mathsf{f}_Y(y) > 0$, the conditional density can be defined as $\mathsf{f}_{X|Y}(x \mid y) = \frac{\mathsf{f}_{XY}(x,y)}{\mathsf{f}_Y(y)}$. Then statistical independence can be expressed as

$$\mathsf{f}_{X|Y}(x|y) = \mathsf{f}_X(x) \qquad \text{where } \mathsf{f}_Y(y) > 0. \tag{1.19}$$

This captures the intuitive notion of statistical independence for continuous rv's better than (1.18), but it does not quite say that the density of X, conditional on $Y = y$ is the same as the marginal density of X. The event $\{Y = y\}$ has zero probability for a continuous rv, and we cannot condition on events of zero probability. If we look at the derivatives defining these densities, the conditional density looks at the probability that $\{x \leq X \leq x + \delta\}$ given

that $\{y \leq Y \leq y+\epsilon\}$ in the limit $\delta, \epsilon \to 0$. At some level, this is a very technical point and the intuition of conditioning on $\{Y{=}y\}$ works very well. Furthermore, problems are often directly modeled in terms of conditional probability densities, so that viewing a conditional density as a limit is less relevant.

More generally the probability of an arbitrary event A conditional on a given value of a continuous rv Y is given by

$$\Pr\{A \mid Y = y\} = \lim_{\delta \to 0} \frac{\Pr\{A, Y \in [y, y+\delta]\}}{\Pr\{Y \in [y, y+\delta]\}}.$$

We next generalize the above results about two rv's to the case of n rv's $\boldsymbol{X} = X_1, \ldots, X_n$. Statistical independence is then defined by the equation

$$\mathsf{F}_{\boldsymbol{X}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \Pr\{X_i \leq x_i\} = \prod_{i=1}^{n} \mathsf{F}_{X_i}(x_i) \qquad \text{for all values of } x_1, \ldots, x_n. \tag{1.20}$$

In other words, $X_1, \ldots, X_n$ are independent if the events $X_i \leq x_i$ for $1 \leq i \leq n$ are independent for all choices of $x_1, \ldots, x_n$. If the density or PMF exists, (1.20) is equivalent to a product form for the density or mass function. A set of rv's is said to be pairwise independent if each pair of rv's in the set is independent. As shown in Exercise 1.20, pairwise independence does not imply that the entire set is independent.

Independent rv's are very often also identically distributed, *i.e.*, they all have the same distribution function. These cases arise so often that we abbreviate independent identically distributed by IID. For the IID case (1.20) becomes

$$\mathsf{F}_{\boldsymbol{X}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \mathsf{F}_X(x_i). \tag{1.21}$$

1.3.5 Stochastic processes and the Bernoulli process

A stochastic process (or random process[15]) is an infinite collection of rv's, usually indexed by an integer or a real number often interpreted as time.[16] Thus each sample point of the probability model maps to an infinite collection of sample values of rv's. If the index is regarded as time, then each sample point maps to a function of time called a sample path or sample function. These sample paths might vary continuously with time or might vary only at discrete times, and if they vary at discrete times, those times might be deterministic or random.

In many cases, this collection of rv's comprising the stochastic process is the only thing of interest. In this case the sample points of the probability model can be taken to be the

[15]Stochastic and random are synonyms, but *random* has become more popular for rv's and *stochastic* for stochastic processes. The reason for the author's choice is that the common-sense intuition associated with randomness appears more important than mathematical precision in reasoning about rv's, whereas for stochastic processes, common-sense intuition causes confusion much more frequently than with rv's. The less familiar word *stochastic* warns the reader to be more careful.

[16]This definition is deliberately vague, and the choice of whether to call a sequence of rv's a process or a sequence is a matter of custom and choice.

sample paths of the process. Conceptually, then, each event is a collection of sample paths. Often these events are defined in terms of a finite set of rv's.

As an example of sample paths that vary at only discrete times, we might be concerned with the times at which customers arrive at some facility. These 'customers' might be customers entering a store, incoming jobs for a computer system, arriving packets to a communication system, or orders for a merchandising warehouse.

The Bernoulli process is an example of how such customers could be modeled and is perhaps the simplest non-trivial stochastic process. We define this process here and develop a few of its many properties. We will frequently return to it, both to use it as an example and to develop additional properties.

Example 1.3.1. A *Bernoulli process* is a sequence, $Y_1, Y_2, \ldots$, of IID binary random variables.[17] Let $p = \Pr\{Y_i = 1\}$ and $1 - p = \Pr\{Y_i = 0\}$. We usually visualize a Bernoulli process as evolving in discrete time with the event $\{Y_i = 1\}$ representing an arriving customer at time i and $\{Y_i = 0\}$ representing no arrival. Thus at most one arrival occurs at each integer time. We visualize the process as starting at time 0, with the first opportunity for an arrival at time 1.

When viewed as arrivals in time, it is interesting to understand something about the intervals between successive arrivals, and about the aggregate number of arrivals up to any given time (see Figure 1.2). These interarrival times and aggregate numbers of arrivals are rv's that are functions of the underlying sequence $Y_1, Y_2, \ldots$,. The topic of rv's that are defined as functions of other rv's (*i.e.*, whose sample values are functions of the sample values of the other rv's) is taken up in more generality in Section 1.3.7, but the interarrival times and aggregate arrivals for Bernoulli processes are so specialized and simple that it is better to treat them from first principles.

First, consider the first interarrival time, X_1, which is defined as the time of the first arrival. If $Y_1 = 1$, then (and only then) $X_1 = 1$. Thus $\mathsf{p}_{X_1}(1) = p$. Next, $X_1 = 2$ if and only $Y_1 = 0$ and $Y_2 = 1$, so $\mathsf{p}_{X_1}(2) = pq$. Continuing, we see that X_1 has the *geometric* PMF,

$$\mathsf{p}_{X_1}(j) = p(1-p)^{j-1}.$$

Each subsequent interarrival time X_k can be found in this same way.[18] It has the same geometric PMF and is statistically independent of $X_1, \ldots, X_{k-1}$. Thus the sequence of interarrival times is an IID sequence of geometric rv's.

It can be seen from Figure 1.2 that a sample path of interarrival times also determines a sample path of the binary arrival rv's, $\{Y_i;\ i \geq 1\}$. Thus the Bernoulli process can also be characterized as a sequence of IID geometric rv's.

For our present purposes, the most important rv's in a Bernoulli process are the partial sums $S_n = \sum_{i=1}^{n} Y_i$. Each rv S_n is the number of arrivals up to and includng time n, *i.e.*,

[17] We say that a sequence $Y_1, Y_2, \ldots$, of rv's are IID if for each integer n, the rv's $Y_1, \ldots, Y_n$ are IID. There are some subtleties in going to the limit $n \to \infty$, but we can avoid most such subtleties by working with finite n-tuples and going to the limit at the end.

[18] This is one of those maddening arguments that, while intuitively obvious, requires some careful reasoning to be completely convincing. We go through several similar arguments with great care in Chapter 2, and suggest that skeptical readers wait until then to prove this rigorously.

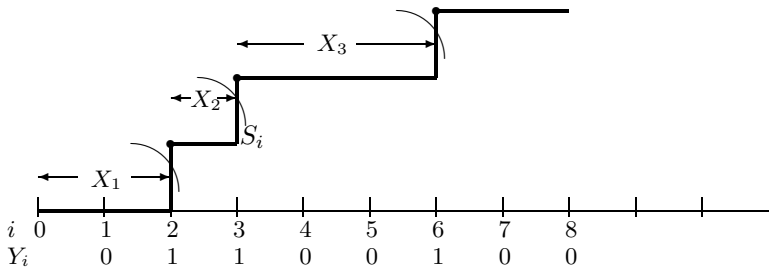


Figure 1.2: Illustration of a sample path for a Bernoulli process: The sample values of the binary rv's Y_i are shown below the time instants. The sample value of the aggregate number of arrivals, $S_n = \sum_{i=1}^{n} Y_i$, is the illustrated step function, and the interarrival intervals are the intervals between steps.

S_n is simply the sum of n binary rv's and thus has the binomial distribution. The PMF $\mathsf{p}_{S_n}(k)$ is the probability that k out of n of the Y_i's have the value 1. There are $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ arrangements of n binary numbers with k 1's, and each has probability $p^k q^{n-k}$. Thus

$$\mathsf{p}_{S_n}(k) = \binom{n}{k} p^k q^{n-k}. \tag{1.22}$$

We will use the binomial PMF extensively as an example in explaining the laws of large numbers later in this chapter, and will often use it in later chapters as an example of a sum of IID rv's. For these examples, we need to know how $\mathsf{p}_{S_n}(k)$ behaves asymptotically as $n \to \infty$ and $k \to \infty$ with k/n essentially constant. The relative frequency k/n will be denoted as $\tilde{p}$. We make a short digression here to state and develop an approximation to the binomial PMF that makes this asymptotic behavior clear.

Lemma 1.3.1. *Let $\mathsf{p}_{S_n}(\tilde{p}n)$ be the PMF of the binomial distribution for an underlying binary PMF $\mathsf{p}_Y(1) = p > 0$, $\mathsf{p}_Y(0) = q > 0$. Then for each integer $\tilde{p}n$, $1 \le \tilde{p}n \le n-1$,*

$$\mathsf{p}_{S_n}(\tilde{p}n)) < \sqrt{\frac{1}{2\pi n\tilde{p}(1-\tilde{p})}}\, \exp\big[n\phi(p,\tilde{p})\big] \Big\} \qquad \textit{where} \tag{1.23}$$

$$\phi(p,\tilde{p}) = \tilde{p}\ln(\frac{p}{\tilde{p}}) + (1-\tilde{p})\ln(\frac{1-p}{1-\tilde{p}}) \;\le\; 0. \tag{1.24}$$

Also, $\phi(p,\tilde{p}) < 0$ for all $\tilde{p} \neq p$. Finally, for any $\epsilon > 0$, there is an $n(\epsilon)$ such that for $n > n(\epsilon)$,

$$\mathsf{p}_{S_n}(\tilde{p}n) > \left(1 - \frac{1}{\sqrt{n}}\right) \sqrt{\frac{1}{2\pi n\tilde{p}(1-\tilde{p})}}\, \exp\big[n\phi(p,\tilde{p})\big] \Big\} \qquad \text{for } \epsilon \le \tilde{p} \le 1-\epsilon \tag{1.25}$$

Discussion: The parameter $\tilde{p} = k/n$ is the relative frequency of 1's in the n-tuple $Y_1, \ldots, Y_n$. For each n, $\tilde{p}$ on the left of (1.23) is restricted so that $\tilde{p}n$ is an integer. The lemma then says that $p_{S_n}(\tilde{p}n)$ is upper bounded by an exponentially decreasing function of n for each $\tilde{p} \neq p$.

If $\tilde{p}$ is bounded away from 0 and 1, the ratio of the upper and lower bounds on $p_{S_n}(\tilde{p}n)$ approaches 1 as $n \to \infty$. A bound that is asymptotically tight in this way is denoted as

$$\mathsf{p}_{S_n}(\tilde{p}n)) \sim \sqrt{\frac{1}{2\pi n\tilde{p}(1-\tilde{p})}}\, \exp\left[n\phi(p,\tilde{p})\right] \qquad \text{for } \epsilon < \tilde{p} < 1-\epsilon \tag{1.26}$$

where the symbol $\sim$ means that the ratio of the left and right side approach 1 as $n \to \infty$

Proof*:[19] The factorial of any positive integer n is bounded by the *Stirling* bounds,[20]

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n < n! < \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{1/12n}. \tag{1.27}$$

The ratio $\sqrt{2\pi n}(n/e)^n/n!$ is monotonically increasing with n toward the limit 1, and the ratio $\sqrt{2\pi n}(n/e)^n \exp(1/12n)/n!$ is monotonically decreasing toward 1. The upper bound is more accurate, but the lower bound is simpler and known as the Stirling approximation.

Since $\sqrt{2\pi n}(n/e)^n/n!$ is increasing in n, we see that $n!/k! < \sqrt{n/k}\, n^n k^{-k} e^{-n+k}$ for $k < n$. Combining this with (1.27) applied to $n-k$,

$$\binom{n}{k} < \sqrt{\frac{n}{2\pi k(n-k)}}\, \frac{n^n}{k^k (n-k)^{n-k}}. \tag{1.28}$$

Using (1.28) in (1.22) to upper bound $\mathsf{p}_{S_n}(k)$,

$$\mathsf{p}_{S_n}(k) < \sqrt{\frac{n}{2\pi k(n-k)}}\, \frac{p^k q^{n-k} n^n}{k^k (n-k)^{n-k}}.$$

Replacing k by $\tilde{p}n$, we get (1.23) where $\phi(p,\tilde{p})$ is given by (1.24). Applying the same argument to the right hand inequality in (1.27),

$$\begin{aligned}\binom{n}{k} &> \sqrt{\frac{n}{2\pi k(n-k)}}\, \frac{n^n}{k^k (n-k)^{n-k}} \exp\left(-\frac{1}{12k} - \frac{1}{12(n-k)}\right) \\ &> \sqrt{\frac{n}{2\pi k(n-k)}}\, \frac{n^n}{k^k (n-k)^{n-k}} \left[1 - \frac{1}{12n\tilde{p}(1-\tilde{p})}\right].\end{aligned} \tag{1.29}$$

For $\epsilon < \tilde{p} < 1-\epsilon$, the term in brackets in (1.29) is lower bounded by $1 - 1/(12n\epsilon(1-\epsilon)$, which is further lower bounded by $1 - 1/\sqrt{n}$ for all sufficiently large n, establishing (1.25).

Finally, to show that $\phi(p,\tilde{p}) \leq 0$, with strict inequality for $\tilde{p} \neq p$, we take the first two derivatives of $\phi(p,\tilde{p})$ with respect to $\tilde{p}$.

$$\frac{\partial \phi(p,\tilde{p})}{\partial \tilde{p}} = \ln\left(\frac{p(1-\tilde{p})}{\tilde{p}(1-p)}\right) \qquad \frac{\partial f^2(p,\tilde{p})}{\partial \tilde{p}^2} = \frac{-1}{\tilde{p}(1-\tilde{p})}.$$

Since the second derivative is negative for $0 < \tilde{p} < 1$, the maximum of $\phi(p,\tilde{p})$ with respect to $\tilde{p}$ is 0, achieved at $\tilde{p} = p$. Thus $\phi(p,\tilde{p}) < 0$ for $\tilde{p} \neq p$. Furthermore, $\phi(p,\tilde{p})$ decreases as $\tilde{p}$ moves in either direction away from p. □

Various aspects of this lemma will be discussed later with respect to each of the laws of large numbers.

[19] Proofs with an asterisk can be omitted without an essential loss of continuity

[20] See Feller [7] for a derivation of these results about the Stirling bounds. Feller also shows that that an improved lower bound to $n!$ is given by $\sqrt{2\pi n}(n/e)^n \exp[\frac{1}{12n} - \frac{1}{360n^3}]$.

We have seen that the Bernoulli process can also be characterized as a sequence of IID geometric interarrival intervals. An interesting generalization of this arises by allowing the interarrival intervals to be arbitrary discrete or continuous nonnegative IID rv's rather than geometric rv's. These processes are known as *renewal processes* and are the topic of Chapter 3. Poisson processes are special cases of renewal processes in which the interarrival intervals have an exponential PDF. These are treated in Chapter 2 and have many connections to Bernoulli processes.

Renewal processes are examples of *discrete stochastic processes*. The distinguishing characteristic of such processes is that interesting things (arrivals, departures, changes of state) occur at discrete instants of time separated by deterministic or random intervals. Discrete stochastic processes are to be distinguished from noise-like stochastic processes in which changes are continuously occurring and the sample paths are continuously varying functions of time. The description of discrete stochastic processes above is not intended to be precise. The various types of stochastic processes developed in subsequent chapters are all discrete in the above sense, however, and we refer to these processes, somewhat loosely, as discrete stochastic processes.

Discrete stochastic processes find wide and diverse applications in operations research, communication, control, computer systems, management science, finance, etc. Paradoxically, we shall spend relatively little of our time discussing these particular applications, and rather develop results and insights about these processes in general. Many examples drawn from the above fields will be discussed, but the examples will be simple, avoiding many of the complications that require a comprehensive understanding of the application area itself.

1.3.6 Expectation

The *expected value* $\mathsf{E}[X]$ of a random variable X is also called the *expectation* or the *mean* and is frequently denoted as $\overline{X}$. Before giving a general definition, we discuss several special cases. First consider nonnegative discrete rv's. The expected value $\mathsf{E}[X]$ is then given by

$$\mathsf{E}[X] = \sum_x x\, \mathsf{p}_X(x). \tag{1.30}$$

If X has a finite number of possible sample values, the above sum must be finite since each sample value must be finite. On the other hand, if X has a countable number of nonnegative sample values, the sum in (1.30) might be either finite or infinite. Example 1.3.2 illustrates a case in which the sum is infinite. The expectation is said to *exist* only if the sum is finite (*i.e.*, if the sum converges to a real number), and in this case $\mathsf{E}[X]$ is given by (1.30). If the sum is infinite, we say that $\mathsf{E}[X]$ does not exist, but also say[21] that $\mathsf{E}[X] = \infty$. In other words, (1.30) can be used in both cases, but $\mathsf{E}[X]$ is said to *exist* only if the sum is finite.

Example 1.3.2. This example will be useful frequently in illustrating rv's that have an infinite expectation. Let N be a positive integer-valued rv with the distribution function

[21] It almost seems metaphysical to say that something has the value infinity when it doesn't exist. However, the word 'exist' here is shorthand for 'exist as a real number,' which makes it quite reasonable to also consider the value in the extended real number system, which includes $\pm\infty$.

$\mathsf{F}_N(n) = n/(n+1)$ for each integer $n \geq 1$. Then N is clearly a positive rv since $\mathsf{F}_N(0) = 0$ and $\lim_{N\to\infty} \mathsf{F}_N(n) = 1$. For each $n \geq 1$, the PMF is given by

$$\mathsf{p}_N(n) = \mathsf{F}_N(n) - \mathsf{F}_N(n-1) = \frac{n}{n+1} - \frac{n-1}{n} = \frac{1}{n(n+1)}. \tag{1.31}$$

Since $\mathsf{p}_N(n)$ is a PMF, we see that $\sum_{n=1}^{\infty} 1/[n(n+1)] = 1$, which is a frequently useful sum. The following equation, however, shows that $\mathsf{E}[N]$ does not exist and has infinite value.

$$\mathsf{E}[N] = \sum_{n=1}^{\infty} n\,\mathsf{p}_N(n) = \sum_{n=1}^{\infty} \frac{n}{n(n+1)} = \sum_{n=1}^{\infty} \frac{1}{n+1} = \infty,$$

where we have used the fact that the harmonic series diverges.

We next derive an alternative expression for the expected value of a nonnegative discrete rv. This new expression is given directly in terms of the distribution function. We then use this new expression as a general definition of expectation which applies to all nonnegative rv's, whether discrete, continuous, or arbitrary. It contains none of the convergence questions that could cause confusion for arbitrary rv's or for continuous rv's with very wild densities.

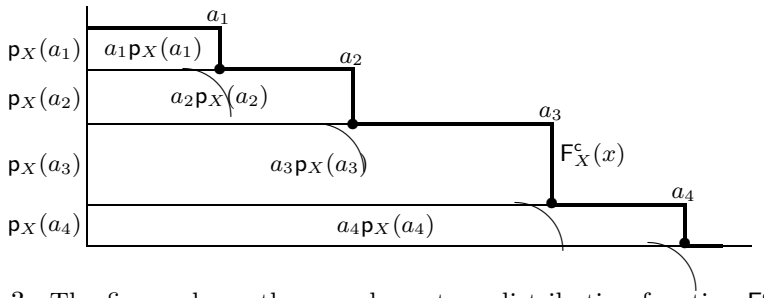


Figure 1.3: The figure shows the complementary distribution function $\mathsf{F}^{\mathsf{c}}_X$ of a nonnegative discrete rv X. For this example, X takes on four possible values, $0 < a_1 < a_2 < a_3 < a_4$. Thus $\mathsf{F}^{\mathsf{c}}_X(x) = \Pr\{X > x\} = 1$ for $x < a_1$. For $a_1 \leq x < a_2$, $\Pr\{X > x\} = 1 - \mathsf{p}_X(a_1)$, and $\Pr\{X > x\}$ has similar drops as x reaches a_2, a_3, and a_4. $\mathsf{E}[X]$, from (1.30), is $\sum_i a_i \mathsf{p}_X(a_i)$, which is the sum of the rectangles in the figure. This is also the area under the curve $\mathsf{F}^{\mathsf{c}}_X(x)$, *i.e.*, $\int_0^\infty \mathsf{F}^{\mathsf{c}}_X(x)\,dx$. It can be seen that this argument applies to any nonnegative rv, thus verifying (1.32).

For a nonnegative discrete rv X, Figure 1.3 illustrates that (1.30) is simply the integral of the complementary distribution function, where the *complementary distribution function* F^{c} of a rv is defined as $\mathsf{F}^{\mathsf{c}}_X(x) = \Pr\{X > x\} = 1 - \mathsf{F}_X(x)$.

$$\mathsf{E}[X] = \int_0^\infty \mathsf{F}^{\mathsf{c}}_X\,dx = \int_0^\infty \Pr\{X > x\}\,dx. \tag{1.32}$$

Although Figure 1.3 only illustrates the equality of (1.30) and (1.32) for one special case, one easily sees that the argument applies to any nonnegative discrete rv, including those with countably many values, by equating the sum of the indicated rectangles with the integral.

For a continuous nonnegative rv X, the conventional definition of expectation is given by

$$\mathsf{E}[X] = \lim_{b\to\infty} \int_0^b x \mathsf{f}_X(x)\, dx. \tag{1.33}$$

Suppose the integral is viewed as a limit of Riemann sums. Each Riemann sum can be viewed as the expectation of a discrete approximation to the continuous rv. The corresponding expectation of the approximation is given by (1.32) using the approximate F_X. Thus (1.32), using the true F_X, yields the expected value of X. This can also be seen using integration by parts. There are no mathematical subtleties in integrating an arbitrary nonnegative nonincreasing function, and this integral must have either a finite or infinite limit. This leads us to the following fundamental definition of expectation for nonnegative rv's:

Definition 1.3.6. *The expectation $\mathsf{E}[X]$ of a nonnegative rv X is defined by (1.32). The expectation is said to exist if and only if the integral is finite. Otherwise the expectation is said to not exist and is also said to be infinite.*

Next consider rv's with both positive and negative sample values. If X has a finite number of positive and negative sample values, say $a_1, a_2, \ldots, a_n$ the expectation $\mathsf{E}[X]$ is given by

$$\begin{aligned} \mathsf{E}[X] &= \sum_i a_i \mathsf{p}_X(a_i) \\ &= \sum_{a_i\le 0} a_i\, \mathsf{p}_X(a_i) + \sum_{a_i>0} a_i\, \mathsf{p}_X(a_i). \end{aligned} \tag{1.34}$$

If X has a countably infinite set of sample values, then (1.34) can still be used if each of the sums in (1.34) converges to a finite value, and otherwise the expectation does not exist (as a real number). It can be seen that each sum in (1.34) converges to a finite value if and only if $\mathsf{E}[|X|]$ exists (*i.e.*, converges to a finite value) for the nonnegative rv $|X|$.

If $\mathsf{E}[X]$ does not exist (as a real number), it still might have the value ∞ if the first sum converges and the second does not, or the value $-\infty$ if the second sum converges and the first does not. If both sums diverge, then $\mathsf{E}[X]$ is undefined, even as an extended real number. In this latter case, the partial sums can be arbitrarily small or large depending on the order in which the terms of (1.34) are summed (see Exercise 1.7).

As illustrated for a finite number of sample values in Figure 1.4, the expression in (1.34) can also be expressed directly in terms of the distribution function and complementary distribution function as

$$\mathsf{E}[X] = -\int_{-\infty}^0 \mathsf{F}_X(x)\, dx + \int_0^\infty \mathsf{F}^{\mathsf{c}}_X(x)\, dx. \tag{1.35}$$

Since $\mathsf{F}^{\mathsf{c}}_X(x) = 1 - \mathsf{F}_X(x)$, this can also be expressed as

$$\mathsf{E}[X] = \int_{-\infty}^\infty \big[u(x) - \mathsf{F}_X(x)\big]\, dx$$

where $u(x)$ is the unit step, $u(x) = 1$ for $x \ge 0$ and $u(x) = 0$ otherwise.

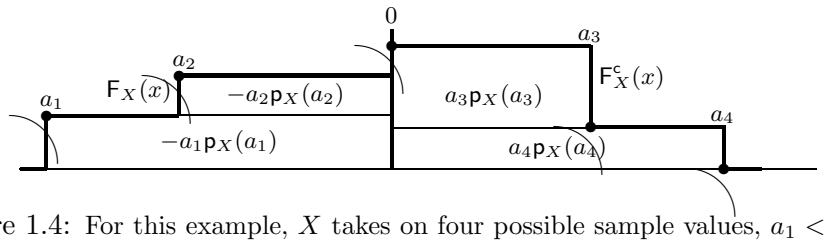


Figure 1.4: For this example, X takes on four possible sample values, $a_1 < a_2 < 0 < a_3 < a_4$. The figure plots $\mathsf{F}_X(x)$ for $x \leq 0$ and $\mathsf{F}^{\mathsf{c}}_X(x)$ for $x > 0$. As in Figure 1.3, $\int_{x\geq 0} \mathsf{F}^{\mathsf{c}}_X(x)\,dx = a_3\mathsf{f}_X(a_3)+a_4\mathsf{f}_X(a_4)$. Similarly, $\int_{x<0} \mathsf{F}_X(x)\,dx = -a_1\mathsf{f}_X(a_1)-a_2\mathsf{f}_X(a_2)$.

The first integral in (1.35) corresponds to the negative sample values and the second to the positive sample values, and $\mathsf{E}[X]$ exists if and only if both integrals are finite (*i.e.*, if $\mathsf{E}[|X|]$ is finite.

For continuous valued rv's with positive and negative sample values, the conventional definition of expectation (assuming that $\mathsf{E}[|X|]$ exists) is given by

$$\mathsf{E}[X] = \int_{-\infty}^{\infty} x\,\mathsf{f}_X(x)\,dx. \tag{1.36}$$

This is equal to (1.35) by the same argument as with nonnegative rv's. Also, as with nonnegative rv's, (1.35) also applies to arbitrary rv's. We thus have the following fundamental definition of expectation:

Definition 1.3.7. *The expectation $\mathsf{E}[X]$ of a rv X exists, with the value given in (1.35), if each of the two terms in (1.35) is finite. The expectation does not exist, but has value ∞ ($-\infty$), if the first term is finite (infinite) and the second infinite (finite). The expectation does not exist and is undefined if both terms are infinite.*

We should not view the general expression in (1.35) for expectation as replacing the need for the conventional expressions in (1.36) and (1.34) for continuous and discrete rv's respectively. We will use all of these expressions frequently, using whichever is most convenient. The main advantages of (1.35) are that it applies equally to all rv's and that it poses no questions about convergence.

Example 1.3.3. The *Cauchy* rv X is the classic example of a rv whose expectation does not exist and is undefined. The probability density is $\mathsf{f}_X(x) = \frac{1}{\pi(1+x^2)}$. Thus $x\mathsf{f}_X(x)$ is proportional to $1/x$ both as $x \to \infty$ and as $x \to -\infty$. It follows that $\int_0^\infty x\mathsf{f}_X(x)\,dx$ and $\int_{-\infty}^0 -x\mathsf{f}_X(x)\,dx$ are both infinite. On the other hand, we see from symmetry that the Cauchy principal value of the integral in (1.36) is given by

$$\lim_{A\to\infty} \int_{-A}^{A} \frac{x}{\pi(1+x)^2}\,dx = 0.$$

There is usually little motivation for considering the upper and lower limits of the integration to have the same magnitude, and the Cauchy principal value usually has little significance for expectations.

1.3.7 Random variables as functions of other random variables

Random variables (rv's) are often defined in terms of each other. For example, if h is a function from $\mathbb{R}$ to $\mathbb{R}$ and X is a rv, then $Y = h(X)$ is the random variable that maps each sample point ω to the composite function $h(X(\omega))$. The distribution function of Y can be found from this, and the expected value of Y can then be evaluated by (1.35).

It is often more convenient to find $\mathsf{E}[Y]$ directly using the distribution function of X. Exercise 1.16 indicates that $\mathsf{E}[Y]$ is given by $\int h(x)\mathsf{f}_X(x)\,dx$ for continuous rv's and by $\sum_x h(x)\mathsf{p}_X(x)$ for discrete rv's. In order to avoid continuing to use separate expressions for continuous and discrete rv's, we express both of these relations by

$$\mathsf{E}[Y] = \int_{-\infty}^{\infty} h(x)\,d\mathsf{F}_X(x), \tag{1.37}$$

This is known as a Stieltjes integral, which can be used as a generalization of both the continuous and discrete cases. For most purposes, we use Stieltjes integrals[22] as a notational shorthand for either $\int h(x)\mathsf{f}_X(x)\,dx$ or $\sum_x h(x)\mathsf{p}_X(x)$.

Knowing that $\mathsf{E}[X]$ exists does not guarantee that $\mathsf{E}[Y]$ exists, but we will treat the question of existence as it arises rather than attempting to establish any general rules.

Particularly important examples of such expected values are the moments $\mathsf{E}[X^n]$ of a rv X and the central moments $\mathsf{E}\left[(X-\overline{X})^n\right]$ of X, where $\overline{X}$ is the mean $\mathsf{E}[X]$. The second central moment is called the *variance*, denoted by σ_X^2 or $\mathsf{VAR}[X]$. It is given by

$$\sigma_X^2 = \mathsf{E}\left[(X-\overline{X})^2\right] = \mathsf{E}\left[X^2\right] - \overline{X}^2. \tag{1.38}$$

The *standard deviation* σ_X of X is the square root of the variance and provides a measure of dispersion of the rv around the mean. Thus the mean is a rough measure of typical values for the outcome of the rv, and σ_X is a measure of the typical difference between X and $\overline{X}$. There are other measures of typical value (such as the median and the mode) and other measures of dispersion, but mean and standard deviation have a number of special properties that make them important. One of these (see Exercise 1.21) is that $\mathsf{E}\left[(X-a)^2\right]$ is minimized over α when $\alpha = \mathsf{E}[X]$.

Next suppose X and Y are rv's and consider the rv[23] $Z = X + Y$. If we assume that X

[22] More specifically, the Riemann-Stieltjes integral, abbreviated here as the Stieltjes integral, is denoted as $\int_a^b h(x)dF_X(x)$. This integral is defined as the limit of a generalized Riemann sum, $\lim_{\delta\to 0}\sum_n h(x_n)[\mathsf{F}(y_n) - \mathsf{F}(y_{n-1})]$ where $\{y_n; n \geq 1\}$ is a sequence of increasing numbers from a to b satisfying $y_n - y_{n-1} \leq \delta$ and $y_{n-1} < x_n \leq y_n$ for all n. The Stieltjes integral exists over finite limits if the limit exists and is independent of the choices of $\{y_n\}$ and $\{x_n\}$ as $\delta \to 0$. It exists over infinite limits if it exists over finite lengths and a limit over the integration limits can be taken. See Rudin [18] for an excellent elementary treatment of Stieltjes integration, and see Exercise 1.12 for some examples.

[23] The question whether a real-valued function of rv's is itself a rv is usually addressed by the use of measure theory, and since we neither use nor develop measure theary in this text, we usually simply assume (within the limits of common sense) that any such function is itself a rv. However, the sum $X + Y$ of rv's is so important throughout this subject that Exercise 1.10 provides a guided derivation of this result for $X + Y$. In the same way, the sum $S_n = X_1 + \cdots + X_n$ of any finite collection of rv's is also a rv.

and Y are independent, then the distribution function of $Z = X + Y$ is given by[24]

$$\mathsf{F}_Z(z) = \int_{-\infty}^{\infty} \mathsf{F}_X(z-y)\, d\mathsf{F}_Y(y) = \int_{-\infty}^{\infty} \mathsf{F}_Y(z-x)\, d\mathsf{F}_X(x). \tag{1.39}$$

If X and Y both have densities, this can be rewritten as

$$\mathsf{f}_Z(z) = \int_{-\infty}^{\infty} \mathsf{f}_X(z-y)\mathsf{f}_Y(y)\, dy = \int_{-\infty}^{\infty} \mathsf{f}_Y(z-x)\mathsf{f}_X(x)\, dx. \tag{1.40}$$

Eq. (1.40) is the familiar convolution equation from linear systems, and we similarly refer to (1.39) as the convolution of distribution functions (although it has a different functional form from (1.40)). If X and Y are nonnegative random variables, then the integrands in (1.39) and (1.40) are non-zero only between 0 and z, so we often use 0 and z as the limits in (1.39) and (1.40).

If $X_1, X_2, \ldots, X_n$ are independent rv's, then the distribution of the rv $S_n = X_1 + X_2 + \cdots + X_n$ can be found by first convolving the distributions of X_1 and X_2 to get the distribution of S_2 and then, for each $i \geq 2$, convolving the distribution of S_i and X_{i+1} to get the distribution of S_{i+1}. The distributions can be convolved in any order to get the same resulting distribution.

Whether or not $X_1, X_2, \ldots, X_n$ are independent, the expected value of $S_n = X_1 + X_2 + \cdots + X_n$ satisfies

$$\mathsf{E}[S_n] = \mathsf{E}[X_1 + X_2 + \cdots + X_n] = \mathsf{E}[X_1] + \mathsf{E}[X_2] + \cdots + \mathsf{E}[X_n]. \tag{1.41}$$

This says that the expected value of a sum is equal to the sum of the expected values, whether or not the rv's are independent (see exercise 1.11). The following example shows how this can be a valuable problem solving aid with an appropriate choice of rv's.

Example 1.3.4. Consider a switch with n input nodes and n output nodes. Suppose each input is randomly connected to a single output in such a way that each output is also connected to a single input. That is, each output is connected to input 1 with probability $1/n$. Given this connection, each of the remaining outputs are connected to input 2 with probability $1/(n-1)$, and so forth.

An input node is said to be *matched* if it is connected to the output of the same number. We want to show that the expected number of matches (for any given n) is 1. Note that the first node is matched with probability $1/n$, and therefore the expectation of a match for node 1 is $1/n$. Whether or not the second input node is matched depends on the choice of output for the first input node, but it can be seen from symmetry that the *marginal distribution* for the output node connected to input 2 is $1/n$ for each output. Thus the expectation of a match for node 2 is also $1/n$. In the same way, the expectation of a match for each input node is $1/n$. From (1.41), the expected total number of matchs is the sum over the expected number for each input, and is thus equal to 1. This exercise would be quite difficult without the use of (1.41).

[24] See Exercise 1.12 for some peculiarties about this definition.

If the rv's $X_1, \ldots, X_n$ are independent, then, as shown in exercises 1.11 and 1.18, the variance of $S_n = X_1 + \cdots + X_n$ is given by

$$\sigma^2_{S_n} = \sum_{i=1}^{n} \sigma^2_{X_i}. \tag{1.42}$$

If $X_1, \ldots, X_n$ are also identically distributed (*i.e.*, $X_1, \ldots, X_n$ are IID) with variance σ^2_X, then $\sigma^2_{S_n} = n\sigma^2_X$. Thus the standard deviation of S_n is $\sigma_{S_n} = \sqrt{n}\sigma_X$. Sums of IID rv's appear everywhere in probability theory and play an especially central role in the laws of large numbers. It is important to remember that the mean of S_n is linear in n but the standard deviation increases only with the square root of n. Figure 1.5 illustrates this behavior.

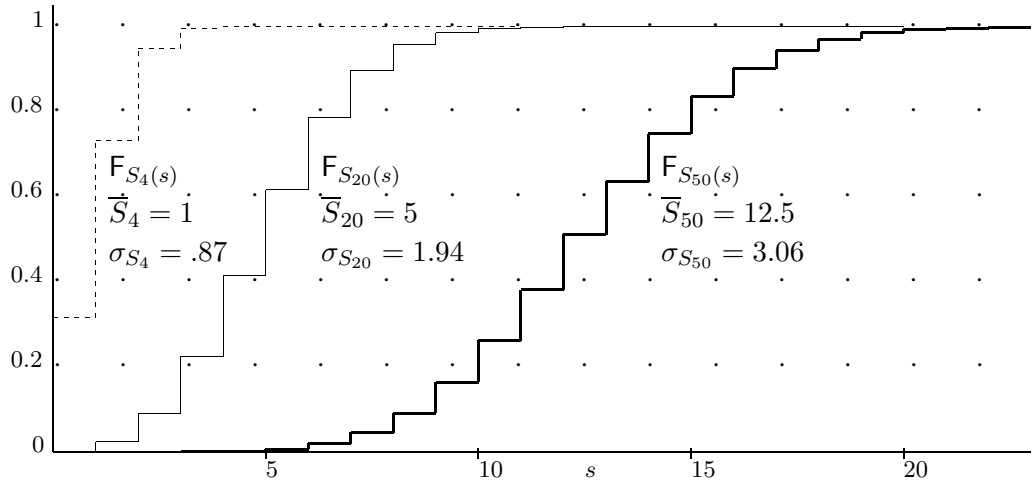


Figure 1.5: The distribution function $\mathsf{F}_{S_n}(s)$ of $S_n = X_1 + \cdots + X_n$ where $X_1, \ldots, X_n$ are typical IID rv's and n takes the values 4, 20, and 50. The particular rv in the figure is binary with $\mathsf{p}_X(1) = 1/4, \mathsf{p}_X(0) = 3/4$. Note that the mean of S_n is proportional to n and the standard deviation to $\sqrt{n}$.

1.3.8 Conditional expectations

Just as the *conditional distribution* of one rv conditioned on a sample value of another rv is important, the *conditional expectation* of one rv based on the sample value of another is equally important. Initially let X be a positive discrete rv and let y be a sample value of another discrete rv Y such that $\mathsf{p}_Y(y) > 0$. Then the conditional expectation of X given $Y = y$ is defined to be

$$\mathsf{E}\,[X \mid Y{=}y] = \sum_x x\, \mathsf{p}_{X|Y}(x \mid y). \tag{1.43}$$

This is simply the ordinary expected value of X using the conditional probabilities in the reduced sample space corresponding to $Y = y$. This value can be finite or infinite as before. More generally, if X can take on positive or negative values, then there is the possibility that

the conditional expectation is undefined. In other words, for discrete rv's, the conditional expectation is exactly the same as the ordinary expectation, except that it is taken using conditional probabilities over the reduced sample space.

More generally yet, let X be an arbitrary rv and let y be a sample value of a discrete rv Y with $\mathsf{p}_Y(y) > 0$. The conditional distribution function of X conditional on $Y = y$ is defined as

$$\mathsf{F}_{X|Y}(x \mid y) = \frac{\Pr\{X \le x, Y = y\}}{\Pr\{Y = y\}}.$$

Since this is an ordinary distribution function in the reduced sample space where $Y = y$, (1.35) expresses the expectation of X conditional on $Y = y$ as

$$\mathsf{E}[X \mid Y = y] = -\int_{-\infty}^{0} \mathsf{F}_{X|Y}(x \mid y)\, dx + \int_{0}^{\infty} \mathsf{F}_{X|Y}(x \mid y)\, dx. \tag{1.44}$$

The forms of conditional expectation in (1.43) and (1.44) are given for individual sample values of Y for which $\mathsf{p}_Y(y) > 0$.

We next show that the conditional expectation of X conditional on a discrete rv Y can also be viewed as a rv. With the possible exception of a set of zero probability, each $\omega \in \Omega$ maps to $\{Y = y\}$ for some y with $\mathsf{p}_Y(y) > 0$ and $\mathsf{E}[X \mid Y = y]$ is defined for that y. Thus we can define $\mathsf{E}[X \mid Y]$ as[25] a rv that is a function of Y, mapping ω to a sample value, say y of Y, and mapping that y to $\mathsf{E}[X \mid Y = y]$. Regarding a conditional expectation as a rv that is a function of the conditioning rv is a powerful tool both in problem solving and in advanced work. For now, we use this to express the unconditional mean of X as

$$\mathsf{E}[X] = \mathsf{E}\big[\mathsf{E}[X \mid Y]\big], \tag{1.45}$$

where the inner expectation is over X for each value of Y and the outer expectation is over the rv $\mathsf{E}[X \mid Y]$, which is a function of Y.

Example 1.3.5. Consider rolling two dice, say a red die and a black die. Let X_1 be the number on the top face of the red die, and X_2 that for the black die. Let $S = X_1 + X_2$. Thus X_1 and X_2 are IID integer rv's, each uniformly distributed from 1 to 6. Conditional on $S = j$, X_1 is uniformly distributed between 1 and $j - 1$ for $j \le 7$ and between $j - 6$ and 6 for $j \ge 7$. For each $j \le 7$, it follows that $\mathsf{E}[X_1 \mid S = j] = j/2$. Similarly, for $j \ge 7$, $\mathsf{E}[X_1 \mid S = j] = j/2$. This can also be seen by the symmetry between X_1 and X_2.

The rv $\mathsf{E}[X_1 \mid S]$ is thus a discrete rv taking on values from 1 to 6 in steps of 1/2 as the sample value of S goes from 2 to 12. The PMF of $\mathsf{E}[X_1 \mid S]$ is given by $\mathsf{p}_{\mathsf{E}[X_1|S]}(j/2) = \mathsf{p}_S(j)$. Using (1.45), we can then calculate $\mathsf{E}[X_1]$ as

$$\mathsf{E}[X_1] = \mathsf{E}\big[\mathsf{E}[X_1 \mid S]\big] = \sum_{j=2}^{12} \frac{j}{2}\, \mathsf{p}_S(j) = \frac{\mathsf{E}[S]}{2} = \frac{7}{2}.$$

This example is not intended to show the value of (1.45) in calculating expectation, since $\mathsf{E}[X_1] = 7/2$ is initially obvious from the uniform integer distribution of X_1. The purpose is simply to illustrate what the rv $\mathsf{E}[X_1 \mid S]$ means.

[25] This assumes that $\mathsf{E}[X \mid Y = y]$ is finite for each y, which is one of the reasons that expectations are said to exist only if they are finite.

To illustrate (1.45) in a more general way, while still assuming X to be discrete, we can write out this expectation by using (1.43) for $\mathsf{E}[X \mid Y=y]$.

$$\begin{aligned}\mathsf{E}[X] &= \mathsf{E}\big[\mathsf{E}[X \mid Y]\big] = \sum_y \mathsf{p}_Y(y)\mathsf{E}[X \mid Y=y] \\ &= \sum_y \mathsf{p}_Y(y) \sum_x x\, \mathsf{p}_{X|Y}(x|y). \end{aligned} \tag{1.46}$$

Operationally, there is nothing very fancy in the example or in (1.45). Combining the sums, (1.46) simply says that $\mathsf{E}[X] = \sum_{y,x} x\, \mathsf{p}_{YX}(y,x)$. As a concept, however, viewing the conditional expectation $\mathsf{E}[X \mid Y]$ as a rv based on the conditioning rv Y is often a useful theoretical tool. This approach is equally useful as a tool in problem solving, since there are many problems where it is easy to find conditional expectations, and then to find the total expectation by averaging over the conditioning variable. For this reason, this result is sometimes called either the total expectation theorem or the iterated expectation theorem. Exercise 1.17 illustrates the advantages of this approach, particularly where it is initially unclear whether or not the expectation is finite. The following cautionary example, however, shows that this approach can sometimes hide convergence questions and give the wrong answer.

Example 1.3.6. Let Y be a geometric rv with the PMF $\mathsf{p}_Y(y) = 2^{-y}$ for integer $y \geq 1$. Let X be an integer rv that, conditional on Y, is binary with equiprobable values $\pm 2^y$ given $Y = y$. We then see that $\mathsf{E}[X \mid Y=y] = 0$ for all y, and thus, (1.46) indicates that $\mathsf{E}[X] = 0$. On the other hand, it is easy to see that $\mathsf{p}_X(2^k) = \mathsf{p}_X(-2^k) = 2^{-k-1}$ for each integer $k \geq 1$. Thus the expectation over positive values of X is ∞ and that over negative values is $-\infty$. In other words, the expected value of X is undefined and (1.46) is incorrect.

The difficulty in the above example cannot occur if X is a nonnegative rv. Then (1.46) is simply a sum of a countable number of nonnegative terms, and thus it either converges to a finite sum independent of the order of summation, or it diverges to ∞, again independent of the order of summation.

If X has both positive and negative components, we can separate it into $X = X^+ + X^-$ where $X^+ = \max(0, X)$ and $X^- = \min(X, 0)$. Then (1.46) applies to X^+ and $-X^-$ separately. If at most one is infinite, then (1.46) applies to X, and otherwise X is undefined. This is summarized in the following theorem:

Theorem 1.3.1 (Total expectation). *Let X and Y be discrete rv's. If X is nonnegative, then $\mathsf{E}[X] = \mathsf{E}\big[\mathsf{E}[X \mid Y]\big] = \sum_y \mathsf{p}_Y(y)\mathsf{E}[X \mid Y=y]$. If X has both positive and negative values, and if at most one of $\mathsf{E}[X^+]$ and $\mathsf{E}[-X^-]$ is infinite, then $\mathsf{E}[X] = \mathsf{E}\big[\mathsf{E}[X \mid Y]\big] = \sum_y \mathsf{p}_Y(y)\mathsf{E}[X \mid Y=y]$.*

We have seen above that if Y is a discrete rv, then the conditional expectation $\mathsf{E}[X|Y=y]$ is little more complicated than the unconditional expectation, and this is true whether X is discrete, continuous, or arbitrary. If X and Y are continuous, we can essentially extend these results to probability densities. In particular, defining $\mathsf{E}[X \mid Y=y]$ as

$$\mathsf{E}[X \mid Y=y] = \int_{-\infty}^{\infty} x\, \mathsf{f}_{X|Y}(x \mid y), \tag{1.47}$$

we have

$$\mathsf{E}[X] = \int_{-\infty}^{\infty} \mathsf{f}_Y(y)\mathsf{E}[X \mid Y{=}y]\, dy = \int_{-\infty}^{\infty} \mathsf{f}_Y(y) \int_{-\infty}^{\infty} x\, \mathsf{f}_{X|Y}(x \mid y)\, dx\, dy. \tag{1.48}$$

We do not state this as a theorem because the details about the integration do not seem necessary for the places where it is useful.

1.3.9 Indicator random variables

For any event A, the *indicator random variable* of A, denoted $\mathbb{I}_A$, is a binary rv that has the value 1 for all $\omega \in A$ and the value 0 otherwise. It then has the PMF $\mathsf{p}_{\mathbb{I}_A}(1) = \Pr\{A\}$ and $\mathsf{p}_{\mathbb{I}_A}(0) = 1 - \Pr\{A\}$. The corresponding distribution function $\mathsf{F}_{\mathbb{I}_A}$ is then illustrated in Figure 1.6. It is easily seen that $\mathsf{E}[\mathbb{I}_A] = \Pr\{A\}$.

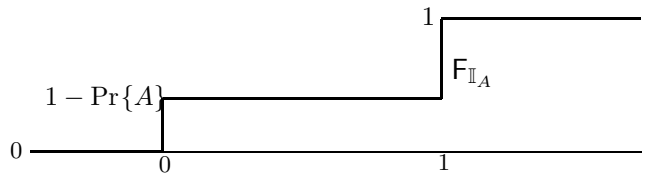


Figure 1.6: The distribution function $\mathsf{F}_{\mathbb{I}_A}$ of an indicator random variable $\mathbb{I}_A$.

Indicator rv's are useful because they allow us to apply the many known results about rv's and particularly binary rv's to events. For example, the laws of large numbers are expressed in terms of sums of rv's, and those results all translate into results about relative frequencies through the use of indicator functions.

1.3.10 Moment generating functions and other transforms

The *moment generating function* (MGF) for a rv X is given by

$$\mathsf{g}_X(r) = \mathsf{E}\left[e^{rX}\right] = \int_{-\infty}^{\infty} e^{rx}\, d\mathsf{F}_X(x). \tag{1.49}$$

where r is a real variable. The integrand is nonnegative, and we can study where the integral exists (*i.e.*, where it is finite) by separating it as follows:

$$\mathsf{g}_X(r) = \int_{0}^{\infty} e^{rx}\, d\mathsf{F}_X(x) + \int_{-\infty}^{0} e^{rx}\, d\mathsf{F}_X(x). \tag{1.50}$$

Both integrals exist for $r = 0$, since the first is $\Pr\{X > 0\}$ and the second is $\Pr\{X \leq 0\}$. The first integral is increasing in r, and thus if it exists for one value of r, it also exists for all smaller values. For example, if X is a nonnegative exponential rv with the density $\mathsf{f}_X(x) = e^{-x}$, then the first integral exists if and only $r < 1$, where it has the value $\frac{1}{1-r}$. As another example, if X satisfies $\Pr\{X > A\} = 0$ for some finite A, then the first integral is at most e^{rA}, which is finite for all real r.

Let $r_+(X)$ be the supremum of values of r for which the first integral exists. The first integral exists for all $r < r_+(X)$, and $0 \leq r_+(X) \leq \infty$. In the same way, let $r_-(X)$ be the infimum of values of r for which the the second integral exists. The second integral exists for all $r > r_-(X)$ and $r_-(X)$ satisfies $0 \geq r_-(X) \geq -\infty$.

Combining the two integrals, the region over which the MGF of an arbitrary rv exists is an interval $I(X)$ from $r_-(X) \leq 0$ to $r_+(X) \geq 0$. Either or both of the end points, $r_-(X)$ and $r_+(X)$, might be included in $I(X)$, and either or both might be either 0 or infinite. We denote these quantities as I, r_-, and r_+ when the rv X is clear from the context. Appendix A gives the interval I for for a number of standard rv's and Exercise 1.22 illustrates $I(X)$ further.

If $\mathsf{g}_X(r)$ exists in an open region of r around 0 (*i.e.*, if $r_- < 0 < r_+$), then derivatives[26] of all orders exist in that region. They are given by

$$\frac{\partial^k \mathsf{g}_X(r)}{\partial r^k} = \int_{-\infty}^{\infty} x^k e^{rx} d\mathsf{F}_X(x) \quad ; \qquad \left.\frac{\partial^k \mathsf{g}_X(r)}{\partial r^k}\right|_{r=0} = \mathsf{E}\left[X^k\right]. \tag{1.51}$$

This shows that finding the moment generating function often provides a convenient way to calculate the moments of a random variable. If any moment fails to exist, however, then the MGF must also fail to exist over each open interval containing 0 (see Exercise 1.31).

Another convenient feature of moment generating functions is their use in treating sums of independent rv's. For example, let $S_n = X_1 + X_2 + \cdots + X_n$. Then

$$\mathsf{g}_{S_n}(r) = \mathsf{E}\left[e^{rS_n}\right] = \mathsf{E}\left[\exp\left(\sum_{i=1}^n rX_i\right)\right] = \mathsf{E}\left[\prod_{i=1}^n \exp(rX_i)\right] = \prod_{i=1}^n \mathsf{g}_{X_i}(r). \tag{1.52}$$

In the last step here, we have used a result of Exercise 1.11, which shows that for independent rv's, the mean of the product is equal to the product of the means. If $X_1, \ldots, X_n$ are also IID, then

$$\mathsf{g}_{S_n}(r) = [\mathsf{g}_X(r)]^n. \tag{1.53}$$

We will use this property frequently in treating sums of IID rv's. Note that this also implies that the region over which the MGF's of S_n and X exist are the same, *i.e.*, $I(S_n) = I(X)$.

The real variable r in the MGF can also be viewed as a complex variable, giving rise to a number of other transforms. A particularly important case is to view r as a pure imaginary variable, say $i\omega$ where $i = \sqrt{-1}$ and ω is real. The MGF is then called the *characteristic function*. Since $|e^{i\omega x}|$ is 1 for all x, $\mathsf{g}_X(i\omega)$ exists for all rv's X and all real ω, and its magnitude is at most one. Note that $\mathsf{g}_X(-i\omega)$ is the Fourier transform of the density of X, so the Fourier transform and characteristic function are the same except for this small notational difference.

The Z-transform is the result of replacing e^r with z in $\mathsf{g}_X(r)$. This is useful primarily for integer valued rv's, but if one transform can be evaluated, the other can be found

[26] This result depends on interchanging the order of differentiation (with respect to r) and integration (with respect to x). This can be shown to be permissible because $\mathsf{g}_X(r)$ exists for r both greater and smaller than 0, which in turn implies, first, that $1 - \mathsf{F}_X(x)$ must approach 0 exponentially as $x \to \infty$ and, second, that $\mathsf{F}_X(x)$ must approach 0 exponentially as $x \to -\infty$.

immediately. Finally, if we use $-s$, viewed as a complex variable, in place of r, we get the two sided Laplace transform of the density of the random variable. Note that for all of these transforms, multiplication in the transform domain corresponds to convolution of the distribution functions or densities, and summation of independent rv's. The simplicity of taking products of transforms is a major reason that transforms are so useful in probability theory.

1.4 Basic inequalities

Inequalities play a particularly fundamental role in probability, partly because many of the models we study are too complex to find exact answers, and partly because many of the most useful theorems establish limiting rather than exact results. In this section, we study three related inequalities, the Markov, Chebyshev, and Chernoff bounds. These are used repeatedly both in the next section and in the remainder of the text.

1.4.1 The Markov inequality

This is the simplest and most basic of these inequalities. It states that if a nonnegative random variable Y has a mean $\mathsf{E}[Y]$, then, for every $y > 0$, $\Pr\{Y \geq y\}$ satisfies[27]

$$\Pr\{Y \geq y\} \leq \frac{\mathsf{E}[Y]}{y} \quad \text{Markov Inequality for nonnegative } Y. \tag{1.54}$$

Figure 1.7 derives this result using the fact (see Figure 1.3) that the mean of a nonnegative rv is the integral of its complementary distribution function, i.e., of the area under the curve $\Pr\{Y > z\}$. Exercise 1.28 gives another simple proof using an indicator random variable.

As an example of this inequality, assume that the average height of a population of people is 1.6 meters. Then the Markov inequality states that at most half of the population have a height exceeding 3.2 meters. We see from this example that the Markov inequality is often very weak. However, for any $y > 0$, we can consider a rv that takes on the value y with probability ϵ and the value 0 with probability $1 - \epsilon$; this rv satisfies the Markov inequality at the point y with equality. Figure 1.7 (as elaborated in Exercise 1.40) also shows that, for any nonnegative rv Y with a finite mean,

$$\lim_{y \to \infty} y \Pr\{Y \geq y\} = 0. \tag{1.55}$$

This will be useful shortly in the proof of Theorem 1.5.3.

[27] The distribution function of any given rv Y is known (at least in principle), and thus one might question why an upper bound is ever preferable to the exact value. One answer is that Y might be given as a function of many other rv's and that the parameters such as the mean in a bound are much easier to find than the distribution function. Another answer is that such inequalities are often used in theorems which state results in terms of simple statistics such as the mean rather than the entire distribution function. This will be evident as we use these bounds.

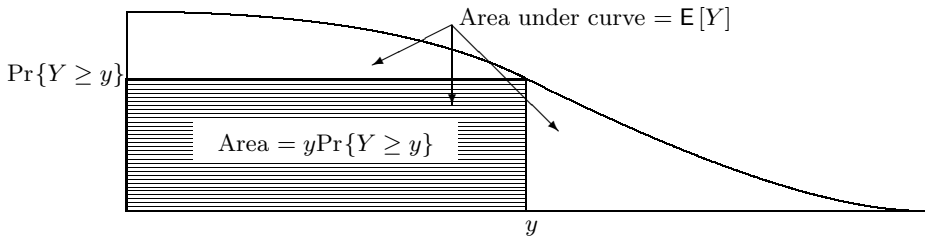


Figure 1.7: Demonstration that $y\Pr\{Y \geq y\} \leq \mathsf{E}[Y]$. By letting $y \to \infty$, it can also be seen that the shaded area becomes a negligible portion of the area $\mathsf{E}[Y]$, so that $\lim_{y\to\infty} y\Pr\{Y > y\} = 0$ if $\mathsf{E}[Y] < \infty$

1.4.2 The Chebyshev inequality

We now use the Markov inequality to establish the well-known Chebyshev inequality. Let Z be an arbitrary rv with finite mean $\mathsf{E}[Z]$ and finite variance σ_Z^2, and define Y as the nonnegative rv $Y = (Z - \mathsf{E}[Z])^2$. Thus $\mathsf{E}[Y] = \sigma_Z^2$. Applying (1.54),

$$\Pr\big\{(Z - \mathsf{E}[Z])^2 \geq y \quad \leq \frac{\sigma_Z^2}{y} \qquad \text{for any } y > 0.$$

Replacing y with ϵ^2 (for any $\epsilon > 0$) and noting that the event $\{(Z - \mathsf{E}[Z])^2 \geq \epsilon^2\}$ is the same as $|Z - \mathsf{E}[Z]| \geq \epsilon$, this becomes

$$\Pr\{|Z - \mathsf{E}[Z]| \geq \epsilon\} \leq \frac{\sigma_Z^2}{\epsilon^2} \qquad \text{(Chebyshev inequality).} \tag{1.56}$$

Note that the Markov inequality bounds just the upper tail of the distribution function and applies only to nonnegative rv's, whereas the Chebyshev inequality bounds both tails of the distribution function. The more important difference, however, is that the Chebyshev bound goes to zero inversely with the square of the distance from the mean, whereas the Markov bound goes to zero inversely with the distance from 0 (and thus asymptotically with distance from the mean).

The Chebyshev inequality is particularly useful when Z is the sample average, $(X_1 + X_2 + \cdots + X_n)/n$, of a set of IID rv's. This will be used shortly in proving the weak law of large numbers.

1.4.3 Chernoff bounds

Chernoff (or exponential) bounds are another variation of the Markov inequality in which the bound on each tail of the distribution function goes to 0 exponentially with distance from the mean. For any given rv Z, let $I(Z)$ be the interval over which the MGF $\mathsf{g}_Z(r) = \mathsf{E}\left[e^{Zr}\right]\}$ exists. Letting $Y = e^{Zr}$ for any $r \in I(Z)$, the Markov inequality (1.54) applied to Y is

$$\Pr\{\exp(rZ) \geq y\} \leq \frac{\mathsf{g}_Z(r)}{y} \qquad \text{for any } y > 0.$$

This takes on a more meaningful form if y is replaced by e^{rb}. Note that $\exp(rZ) \geq \exp(rb)$ is equivalent to $Z \geq b$ for $r > 0$ and to $Z < b$ for $r < 0$. Thus, for any real b, we get the following two bounds, one for $r > 0$ and the other for $r < 0$:

$$\begin{aligned} \Pr\{Z \geq b\} &\leq \mathsf{g}_Z(r)\exp(-rb) \quad ; \qquad \text{(Chernoff bound for } 0 < r \in I(Z) \qquad (1.57) \\ \Pr\{Z \leq b\} &\leq \mathsf{g}_Z(r)\exp(-rb) \quad ; \qquad \text{(Chernoff bound for } 0 > r \in I(Z)). \qquad (1.58) \end{aligned}$$

This provides us with a family of upper bounds on the tails of the distribution function, using values of $r > 0$ for the upper tail and $r < 0$ for the lower tail. For fixed $0 < r \in I(Z)$, this bound on $\Pr\{Z \geq b\}$ decreases exponentially[28] in b at rate r. Similarly, for each $0 > r \in I(Z)$, the bound on $\Pr\{Z \leq b\}$ decreases exponentially at rate r as $b \to -\infty$. We will see shortly that (1.57) is useful only when $b > \mathsf{E}[X]$ and (1.58) is useful only when $b < \mathsf{E}[X]$).

The most important application of these Chernoff bounds is to sums of IID rv's. Let $S_n = X_1 + \cdots + X_n$ where $X_1, \ldots, X_n$ are IID with the MGF $\mathsf{g}_X(r)$. Then $\mathsf{g}_{S_n}(r) = [\mathsf{g}_X(r)]^n$, so (1.57) and (1.58) (with b replaced by na) become

$$\begin{aligned} \Pr\{S_n \geq na\} &\leq [\mathsf{g}_X(r)]^n \exp(-rna) \quad ; \qquad (\text{ for } 0 < r \in I(Z)) \qquad (1.59) \\ \Pr\{S_n \leq na\} &\leq [\mathsf{g}_X(r)]^n \exp(-rna) \quad ; \qquad (\text{ for } 0 > r \in I(Z)). \qquad (1.60) \end{aligned}$$

These equations are easier to understand if we define the *semi-invariant MGF*, $\gamma_X(r)$, as

$$\gamma_X(r) = \ln g_X(r). \tag{1.61}$$

The semi-invariant MGF for a typical rv X is sketched in Figure 1.8. The major features to observe are, first, that $\gamma'_X(0) = \mathsf{E}[X]$ and, second, that $\gamma''_X(r) \geq 0$ for r in the interior of $I(X)$..

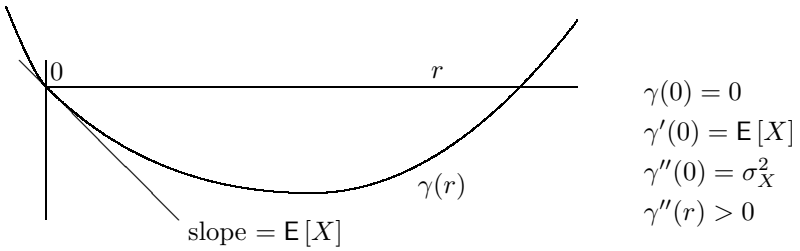


Figure 1.8: Semi-invariant moment-generating function $\gamma(r)$ for a typical rv X assuming $r_- < 0 < r_+$. Since $\gamma(r) = \ln g(r)$, we see that that $\frac{\partial}{\partial r}\gamma(r) = \frac{1}{g(r)}\frac{\partial}{\partial r}g(r)$. Thus $\gamma'(0) = \mathsf{E}[X]$. Also, for r in the interior of $I(X)$, Exercise 1.24 shows that $\gamma''(r) \geq 0$, and in fact, $\gamma''(r)$ is strictly positive except in the uninteresting case where X is deterministic (takes on a single value with probability 1). As indicated in the figure, the straight line of slope $\mathsf{E}[X]$ through the origin is tangent to $\gamma(r)$.

[28] This seems paradoxical, since Z seems to be almost arbitrary. However, since $r \in I(Z)$, we have $\int e^{rb} d\mathsf{F}_Z(b) < \infty$.

In terms of $\gamma_X(r)$, (1.59) and (1.60) become

$$\Pr\{S_n \geq na\} \leq \exp(n[\gamma_X(r) - ra]) \quad ; \quad (\text{ for } 0 < r \in I(X) \tag{1.62}$$

$$\Pr\{S_n \leq na\} \leq \exp(n[\gamma_X(r) - ra]) \quad ; \quad (\text{ for } 0 > r \in I(X). \tag{1.63}$$

These bounds are geometric in n for fixed a and r, so we should ask what value of r provides the tightest bound for any given a. Since $\gamma''_X(r) > 0$, the tightest bound arises either at that r for which $\gamma'(r) = a$ or at one of the end points, r_- or r_+, of $I(X)$. This minimum value is denoted by

$$\mu_X(a) = \inf_r [\gamma_X(r) - ra].$$

Note that $(\gamma_X(r) - ra)|_{r=0} = 0$ and $\frac{\partial}{\partial r}(\gamma_X(r) - ra)|_{r=0} = \mathsf{E}[X] - a$. Thus if $a > \mathsf{E}[X]$, then $\gamma_X(r) - ra$ must be negative for sufficiently small positive r. Similarly, if $a < \mathsf{E}[X]$, then $\gamma_X(r) - ra$ is negative for negative r sufficiently close[29] to 0. In other words,

$$\Pr\{S_n \geq na\} \leq \exp(n\mu_X(a)) \; ; \qquad \text{where } \mu_X(a) < 0 \text{ for } a > \mathsf{E}[X] \tag{1.64}$$

$$\Pr\{S_n \leq na\} \leq \exp(n\mu_X(a)) \; ; \qquad \text{where } \mu_X(a) < 0 \text{ for } a < \mathsf{E}[X]\,. \tag{1.65}$$

This is summarized in the following lemma:

Lemma 1.4.1. *Assume that 0 is in the interior of $I(X)$ and let S_n be the sum of n IID rv's each with the distribution of X. Then $\mu_X(a) = \inf_r[\gamma_X(r) - ra] < 0$ for all $a \neq \mathsf{E}[X]$. Also, $\Pr\{S_n \geq na\} \leq e^{n\mu_X(a)}$ for $a > \mathsf{E}[X]$ and $\Pr\{S_n \leq na\} \leq e^{n\mu_X(a)}$ for $a < \mathsf{E}[X]$.*

Figure 1.9 illustrates the lemma and gives a graphical construction to find[30] $\mu_X(a) = \inf_r[\gamma_X(r) - ra]$.

These Chernoff bounds will be used in the next section to help understand several laws of large numbers. They will also be used extensively in Chapter 7 and are useful for detection, random walks, and information theory.

The following example evaluates these bounds for the case where the IID rv's are binary. We will see that in this case the bounds are exponentially tight in a sense to be described.

Example 1.4.1. Let X be binary with $\mathsf{p}_X(1) = p$ and $\mathsf{p}_X(0) = q = 1 - p$. Then $\mathsf{g}_X(r) = q + pe^r$ for $-\infty < r < \infty$. Also, $\gamma_X(r) = \ln(q + pe^r)$. To be consistent with the expression for the binomial PMF in (1.23), we will find bounds to $\Pr\{S_n \geq \tilde{p}n\}$ and $\Pr\{S_n \leq \tilde{p}n\}$ for $\tilde{p} > p$ and $\tilde{p} < p$ respectively. Thus, according to Lemma 1.4.1, we first evaluate

$$\mu_X(\tilde{p}) = \inf_r [\gamma_X(r) - \tilde{p}r].$$

The minimum occurs at that r for which $\gamma'_X(r) = \tilde{p}$, *i.e.*, at

$$\frac{pe^r}{q + pe^r} = \tilde{p}.$$

[29] In fact, for r sufficiently small, $\gamma(r)$ can be approximated by a second order power series, $\gamma(r) \approx \gamma(0) + r\gamma'(0) + (r^2/2)\gamma''(0) = r\overline{X} + (r^2/2)\sigma_X^2$. It follows that $\mu_X(a) \approx -(a - \overline{X})^2/2\sigma_X^2$ for very small r.

[30] As a special case, the infimum might occur at the edge of the interval of convergence, *i.e.*, at r_- or r_+. As shown in Exercise 1.23, the infimum can be at r_+ (r_-) only if $g_X(r_+)$ ($g_X(r_-)$ exists, and in this case, the graphical techinique in Figure 1.9 still works.

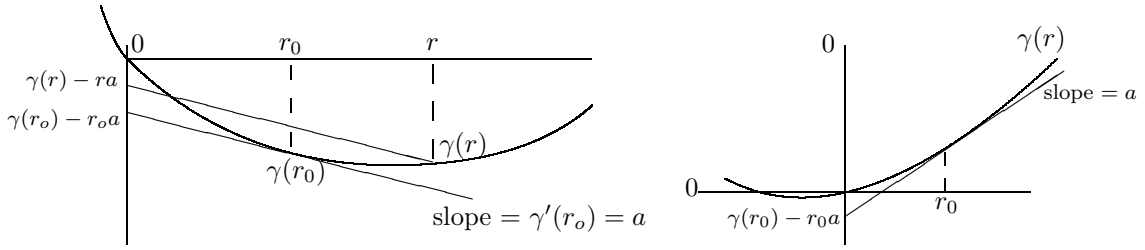


Figure 1.9: Graphical minimization of $\gamma(r) - ar$: For any $r \in I(X)$, $\gamma(r) - ar$ is the vertical axis intercept of a line of slope a through the point $(r, \gamma(r))$. The minimum occurs when the line of slope a is tangent to the curve. The two examples show one case where $\mathsf{E}[X] < 0$ and another where $\mathsf{E}[X] > 0$.

Rearranging terms,

$$e^r = \frac{\tilde{p}q}{p\tilde{q}} \qquad \text{where } \tilde{q} = 1 - \tilde{p}. \tag{1.66}$$

Substituting this minimizing value of r into $\ln(q + pe^r) - r\tilde{p}$ and rearranging terms,

$$\mu_X(\tilde{p}) = \tilde{p} \ln \frac{p}{\tilde{p}} + \tilde{q} \ln \frac{\tilde{q}}{q}. \tag{1.67}$$

Substituting this into (1.64), and (1.65), we get the following Chernoff bounds for binary IID rv's. As shown above, they are exponentially decreasing in n.

$$\Pr\{S_n \geq n\tilde{p}\} \leq \exp\left\{n\left[\tilde{p}\ln\frac{p}{\tilde{p}} + \tilde{q}\ln\frac{q}{\tilde{q}}\right]\right\}; \qquad \text{for } \tilde{p} > p \tag{1.68}$$

$$\Pr\{S_n \leq n\tilde{p}\} \leq \exp\left\{n\left[\tilde{p}\ln\frac{p}{\tilde{p}} + \tilde{q}\ln\frac{q}{\tilde{q}}\right]\right\}; \qquad \text{for } \tilde{p} < p. \tag{1.69}$$

So far, it seems that we have simply developed another upper bound on the tails of the distribution function for the binomial. It will then perhaps be surprising to compare this bound with the asymptotically correct value (repeated below) for the binomial PMF in (1.26).

$$\mathsf{p}_{S_n}(k) \sim \sqrt{\frac{1}{2\pi n\tilde{p}\tilde{q}}}\, \exp n\left[\tilde{p}\ln(p/\tilde{p}) + \tilde{q}\ln(q/\tilde{q})\right] \qquad \text{for } \tilde{p} = \frac{k}{n}. \tag{1.70}$$

For any integer value of $n\tilde{p}$ with $\tilde{p} > p$, we can lower bound $\Pr\{S_n \geq n\tilde{p}\}$ by the single term $\mathsf{p}_{S_n}(n\tilde{p})$. Thus $\Pr\{S_n \geq n\tilde{p}\}$ is both upper and lower bounded by quantities that decrease exponentially with n at the same rate. The lower bound is asymptotic in n and has the coefficient $1/\sqrt{2\pi n\tilde{p}\tilde{q}}$. These differences are essentially negligible for large n compared to the exponential term. We can express this analytically by considering the log of the upper bound in (1.68) and the lower bound in (1.70).

$$\lim_{n\to\infty} \frac{\ln \Pr\{S_n \geq n\tilde{p}\}}{n} = \left[\tilde{p}\ln\frac{p}{\tilde{p}} + \tilde{q}\ln\frac{q}{\tilde{q}}\right] \qquad \text{where } \tilde{p} > p. \tag{1.71}$$

In the same way, for $\tilde{p} < p$,

$$\lim_{n\to\infty} \frac{\ln \Pr\{S_n \le n\tilde{p}\}}{n} = \left[\tilde{p} \ln \frac{p}{\tilde{p}} + \tilde{q} \ln \frac{q}{\tilde{q}}\right] \} \quad \text{where } \tilde{p} < p. \tag{1.72}$$

In other words, these Chernoff bounds are not only upper bounds, but are also exponentially correct in the sense of (1.71) and (1.72). In Chapter 7 we will show that this property is typical for sums of IID rv's . Thus we see that the Chernoff bounds are not 'just bounds,' but rather are bounds that when optimized provide the correct asymptotic exponent for the tails of the distribution of sums of IID rv's. In this sense these bounds are very different from the Markov and Chebyshev bounds.

1.5 The laws of large numbers

The laws of large numbers are a collection of results in probability theory that describe the behavior of the arithmetic average of n rv's for large n. For any n rv's, $X_1, \ldots, X_n$, the *arithmetic average* is the rv $(1/n)\sum_{i=1}^{n} X_i$. Since in any outcome of the experiment, the sample value of this rv is the arithmetic average of the sample values of $X_1, \ldots, X_n$, this random variable is usually called the *sample average*. If $X_1, \ldots, X_n$ are viewed as successive variables in time, this sample average is called the time-average. Under fairly general assumptions, the standard deviation of the sample average goes to 0 with increasing n, and, in various ways depending on the assumptions, the sample average approaches the mean.

These results are central to the study of stochastic processes because they allow us to relate time-averages (i.e., the average over time of individual sample paths) to ensemble-averages (i.e., the mean of the value of the process at a given time). In this section, we develop and discuss two of these results, the weak and the strong law of large numbers for independent identically distributed rv's. The strong law requires considerable patience to understand, but it is a basic and essential result in understanding stochastic processes. We also discuss the central limit theorem, partly because it enhances our understanding of the weak law, and partly because of its importance in its own right.

1.5.1 Weak law of large numbers with a finite variance

Let $X_1, X_2, \ldots, X_n$ be IID rv's with a finite mean $\overline{X}$ and finite variance σ_X^2. Let $S_n = X_1 + \cdots + X_n$, and consider the sample average S_n/n. We saw in (1.42) that $\sigma_{S_n}^2 = n\sigma_X^2$. Thus the variance of S_n/n is

$$\mathsf{VAR}\left[\frac{S_n}{n}\right] = \mathsf{E}\left[\left(\frac{S_n - n\overline{X}}{n}\right)^2\right] = \frac{1}{n^2}\,\mathsf{E}\left[\left(S_n - n\overline{X}\right)^2\right] = \frac{\sigma_X^2}{n}. \tag{1.73}$$

This says that the standard deviation of the sample average S_n/n is $\sigma/\sqrt{n}$, which approaches 0 as n increases. Figure 1.10 illustrates this decrease in the standard deviation of S_n/n with

increasing n. In contrast, recall that Figure 1.5 illustrated how the standard deviation of S_n increases with n. From (1.73), we see that

$$\lim_{n\to\infty} \mathsf{E}\left[\left(\frac{S_n}{n} - \overline{X}\right)^2\right] = 0. \tag{1.74}$$

As a result, we say that S_n/n *converges in mean square* to $\overline{X}$.

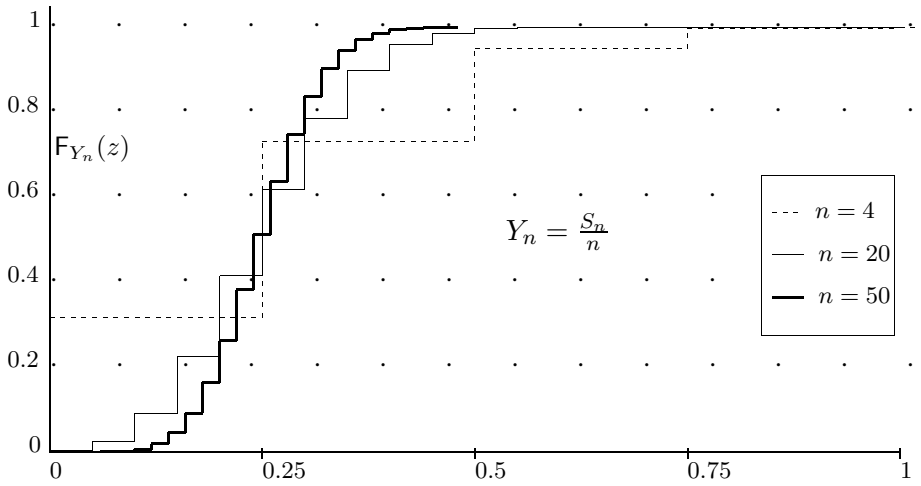


Figure 1.10: The same distribution as Figure 1.5, scaled differently to give the distribution function of the sample average Y_n. It can be visualized that as n increases, the distribution function of Y_n becomes increasingly close to a unit step at the mean, 0.25, of the variables X being summed.

This convergence in mean square says that the sample average, S_n/n, differs from the mean, $\overline{X}$, by a random variable whose standard deviation approaches 0 with increasing n. This convergence in mean square is one sense in which S_n/n approaches $\overline{X}$, but the idea of a sequence of rv's (*i.e.*, a sequence of functions) approaching a constant is clearly much more involved than a sequence of numbers approaching a constant. The laws of large numbers bring out this central idea in a more fundamental, and usually more useful, way. We start the development by applying the Chebyshev inequality (1.56) to the sample average,

$$\Pr\left\{\left|\frac{S_n}{n} - \overline{X}\right| > \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}. \tag{1.75}$$

This is an upper bound on the probability that S_n/n differs by more than ϵ from its mean, $\overline{X}$. This is illustrated in Figure 1.10 which shows the distribution function of S_n/n for various n. The figure suggests that $\lim_{n\to\infty} \mathsf{F}_{S_n/n}(y) = 0$ for all $y < \overline{X}$ and $\lim_{n\to\infty} \mathsf{F}_{S_n/n}(y) = 1$ for all $y > \overline{X}$. This is stated more cleanly in the following weak law of large numbers, abbreviated WLLN

Theorem 1.5.1 (WLLN with finite variance). *For each integer $n \geq 1$, let $S_n = X_1 + \cdots + X_n$ be the sum of n IID rv's with a finite variance. Then the following holds:*

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{S_n}{n} - \overline{X}\right| > \epsilon\right\} = 0 \qquad \textit{for every } \epsilon > 0. \tag{1.76}$$

Proof: For every $\epsilon > 0$, $\Pr\{|S_n/n - \overline{X}| > \epsilon$ is bounded between 0 and $\sigma^2/n\epsilon^2$. Since the upper bound goes to 0 with increasing n, the theorem is proved. □

Discussion: The algebraic proof above is both simple and rigorous. However, the geometric argument in Figure 1.11 probably provides more intuition about how the limit takes place. It is important to understand both.

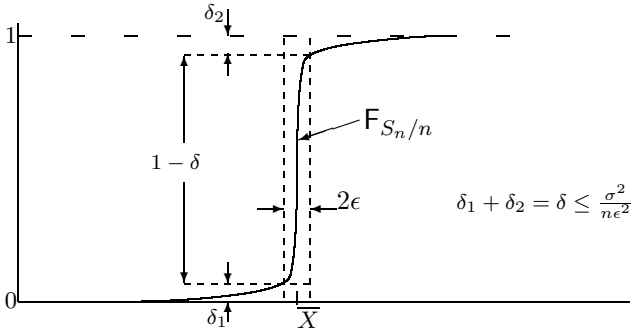


Figure 1.11: Approximation of the distribution function $\mathsf{F}_{S_n/n}$ of a sample average by a step function at the mean: From (1.75), the probability δ that S_n/n differs from $\overline{X}$ by more than ϵ (*i.e.*, $\Pr\{|S_n/n - \overline{X}| \geq \epsilon$) is at most $\sigma^2/n\epsilon^2$. The complementary event, where $|S_n/n - \overline{X}| < \epsilon$, has probability $1 - \delta \geq 1 - \sigma^2/n\epsilon^2$. This means that we can construct a rectangle of width 2ϵ centered on $\overline{X}$ and of height $1 - \delta$ such that $\mathsf{F}_{S_n/n}$ enters the rectangle at the lower left (say at $(\overline{X} - \epsilon, \delta_1)$) and exits at the upper right, say at $(\overline{X} + \epsilon, 1 - \delta_2)$). Now visualize increasing n while holding ϵ fixed. In the limit, $1 - \delta \to 1$ so $\Pr\{|S_n/n - \overline{X}| \geq \epsilon \to 0$. Since this is true for every $\epsilon > 0$ (usually with slower convergence as ϵ gets smaller), $\mathsf{F}_{S_n/n}(y)$ approaches 0 for every $y < \overline{X}$ and approaches 1 for every $y > \overline{X}$, *i.e.*, $\mathsf{F}_{S_n/n}$ approaches a unit step at $\overline{X}$. Note that there are two 'fudge factors' here, ϵ and δ and, since we are approximating an entire distribution function, neither can be omitted, except by directly going to a limit as $n \to \infty$.

We refer to (1.76) as saying that S_n/n converges to $\overline{X}$ in probability. To make sense out of this, we should view $\overline{X}$ as a deterministic random variable, *i.e.*, a rv that takes the value $\overline{X}$ for each sample point of the space. Then (1.76) says that the probability that the absolute difference, $|S_n/n - \overline{X}|$, exceeds any given $\epsilon > 0$ goes to 0 as $n \to \infty$.[31]

One should ask at this point what (1.76) adds to the more specific bound in (1.75). In particular (1.75) provides an upper bound on the rate of convergence for the limit in (1.76). The answer is that (1.76) remains valid when the theorem is generalized. For variables that are not IID or have an infinite variance, (1.75) is no longer necessarily valid. In some situations, as we see later, it is valuable to know that (1.76) holds, even if the rate of convergence is extremely slow or unknown.

One difficulty with the bound in (1.75) is that it is extremely loose in most cases. If S_n/n actually approached $\overline{X}$ this slowly, the weak law of large numbers would often be more a mathematical curiosity than a highly useful result. If we assume that the MGF of X exists

[31] Saying this in words gives one added respect for mathematical notation, and perhaps in this case, it is preferable to simply understand the mathematical statement (1.76).

in an open interval around 0, then (1.75) can be strengthened considerably. Recall from (1.64) and (1.65) that for any $\epsilon > 0$,

$$\Pr\{S_n/n - \overline{X} \geq \epsilon \quad \leq \quad \exp(n\mu_X(\overline{X} + \epsilon)) \tag{1.77}$$

$$\Pr\{S_n/n - \overline{X} \leq -\epsilon \quad \leq \quad \exp(n\mu_X(\overline{X} - \epsilon)), \tag{1.78}$$

where from Lemma 1.4.1, $\mu_X(a) = \inf_r\{\gamma_X(r) - ra\} < 0$ for $a \neq \overline{X}$. Thus, for any $\epsilon > 0$,

$$\Pr\{|S_n/n - \overline{X}| \geq \epsilon \quad \leq \exp[n\mu_X(\overline{X} + \epsilon)] + \exp[n\mu_X(\overline{X} - \epsilon)]. \tag{1.79}$$

The bound here, for any given $\epsilon > 0$, decreases geometrically in n rather than harmonically. In terms of Figure 1.11, the height of the rectangle must approach 1 at least geometrically in n.

1.5.2 Relative frequency

We next show that (1.76) can be applied to the relative frequency of an event as well as to the sample average of a random variable. Suppose that A is some event in a single experiment, and that the experiment is independently repeated n times. Then, in the probability model for the n repetitions, let A_i be the event that A occurs at the ith trial, $1 \leq i \leq n$. The events $A_1, A_2, \ldots, A_n$ are then IID.

If we let $\mathbb{I}_{A_i}$ be the indicator rv for A on the ith trial, then the rv $S_n = \mathbb{I}_{A_1} + \mathbb{I}_{A_2} + \cdots + \mathbb{I}_{A_n}$ is the number of occurrences of A over the n trials. It follows that

$$\text{relative frequency of } A = \frac{S_n}{n} = \frac{\sum_{i=1}^{n} \mathbb{I}_{A_i}}{n}. \tag{1.80}$$

Thus the relative frequency of A is the sample average of the binary rv's $\mathbb{I}_{A_i}$, and everything we know about the sum of IID rv's applies equally to the relative frequency of an event. In fact, everything we know about the sums of IID *binary* rv's applies to relative frequency.

1.5.3 The central limit theorem

The weak law of large numbers says that with high probability, S_n/n is close to $\overline{X}$ for large n, but it establishes this via an upper bound on the tail probabilities rather than an estimate of what $\mathsf{F}_{S_n/n}$ looks like. If we look at the shape of $\mathsf{F}_{S_n/n}$ for various values of n in the example of Figure 1.10, we see that the function $\mathsf{F}_{S_n/n}$ becomes increasingly compressed around $\overline{X}$ as n increases (in fact, this is the essence of what the weak law is saying). If we normalize the random variable S_n/n to 0 mean and unit variance, we get a normalized rv, $Z_n = (S_n/n - \overline{X})\sqrt{n}\,/\sigma$. The distribution function of Z_n is illustrated in Figure 1.12 for the same underlying X as used for S_n/n in Figure 1.10. The curves in the two figures are the same except that each curve has been horizontally scaled by $\sqrt{n}$ in Figure 1.12.

Inspection of Figure 1.12 shows that the normalized distribution functions there seem to be approaching a limiting distribution. The critically important *central limit theorem* states that there is indeed such a limit, and it is the normalized Gaussian distribution function.

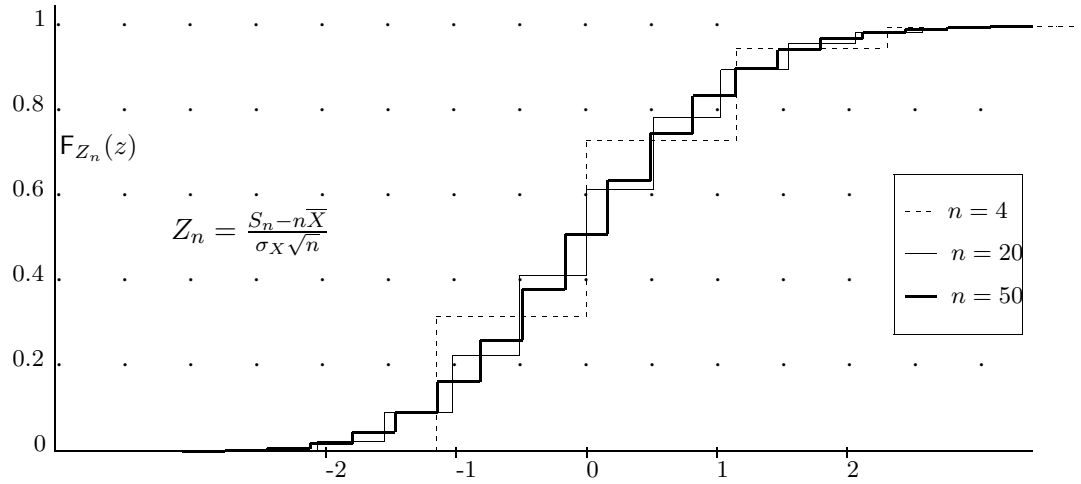


Figure 1.12: The same distribution functions as Figure 1.5 normalized to 0 mean and unit standard deviation, *i.e.*, the distribution functions of $Z_n = (S_n/n - \overline{X})\frac{\sqrt{n}}{\sigma_X}$ for $n = 4$, 20, 50. Note that as n increases, the distribution function of Z_n slowly starts to resemble the normal distribution function.

Theorem 1.5.2 (Central limit theorem (CLT)). *Let $X_1, X_2, \ldots$ be IID rv's with finite mean $\overline{X}$ and finite variance σ^2. Then for every real number z,*

$$\lim_{n\to\infty} \Pr\left\{ \frac{S_n - n\overline{X}}{\sigma\sqrt{n}} \leq z \right\} = \Phi(z), \tag{1.81}$$

where $\Phi(z)$ is the normal distribution function, i.e.*, the Gaussian distribution with mean 0 and variance 1,*

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\frac{y^2}{2})\, dy.$$

Discussion: The rv's $Z_n = (S_n - n\overline{X})/(\sigma\sqrt{n})$ for each $n \geq 1$ on the left side of (1.81) each have mean 0 and variance 1. The central limit theorem (CLT), as expressed in (1.81), says that the sequence of distribution functions, $\mathsf{F}_{Z_1}(z), \mathsf{F}_{Z_2}(z), \ldots$ converges at each value of z to $\Phi(z)$ as $n \to \infty$. In other words, $\lim_{n\to\infty} \mathsf{F}_{Z_n}(z) = \Phi(z)$ for each $z \in \mathbb{R}$. This is called convergence in distribution, since it is the sequence of distribution functions, rather than the sequence of rv's that is converging. The theorem is illustrated by Figure 1.12.

The reason why the word *central* appears in the CLT can be seen by rewriting (1.81) as

$$\lim_{n\to\infty} \Pr\left\{ \frac{S_n}{n} - \overline{X} \leq \frac{\sigma z}{\sqrt{n}} \right\} = \Phi(z), \tag{1.82}$$

Asymptotically, then, we are looking at a sequence of sample averages that differ from the mean by a quantity going to 0 with n as $1/\sqrt{n}$, *i.e.*, by sample averages very close to the mean for large n. This should be contrasted with the optimized Chernoff bound in (1.64)

and (1.65) which look at a sequence of sample averages that differ from the mean by a constant amount for large n. These latter results are exponentially decreasing in n and are known as large deviation results.

The CLT says nothing about speed of convergence to the normal distribution. The Berry-Esseen theorem (see, for example, Feller, [8]) provides some guidance about this for cases in which the third central moment $\mathsf{E}\left[|X-\overline{X}|^3\right]$ exists. This theorem states that

$$\left|\Pr\left\{\frac{S_n - n\overline{X})}{\sigma\sqrt{n}} \le z\right\} - \Phi(z)\right| \;\le\; \frac{C\mathsf{E}\left[|X-\overline{X}|^3\right]}{\sigma^3\sqrt{n}}. \tag{1.83}$$

where C can be upper bounded by 0.766. We will come back shortly to discuss convergence in greater detail.

The CLT helps explain why Gaussian rv's play such a central role in probability theory. In fact, many of the cookbook formulas of elementary statistics are based on the tacit assumption that the underlying variables are Gaussian, and the CLT helps explain why these formulas often give reasonable results.

One should be careful to avoid reading more into the CLT than it says. For example, the normalized sum, $(S_n - n\overline{X})/\sigma\sqrt{n}$ need not have a density that is approximately Gaussian. In fact, if the underlying variables are discrete, the normalized sum is discrete and has no density. The PMF of the normalized sum can have very detailed fine structure; this does not disappear as n increases, but becomes "integrated out" in the distribution function.

The CLT tell us quite a bit about how $\mathsf{F}_{S_n/n}$ converges to a step function at $\overline{X}$. To see this, rewrite (1.81) in the form

$$\lim_{n\to\infty} \Pr\left\{\frac{S_n}{n} - \overline{X} \le \frac{\sigma z}{\sqrt{n}}\right\} = \Phi(z). \tag{1.84}$$

This is illustrated in Figure 1.13 where we have used $\Phi(z)$ as an approximation for the probability on the left.

A proof of the CLT requires mathematical tools that will not be needed subsequently. Thus we give a proof only for the binomial case. Feller ([7] and [8]) gives a thorough and careful exposition and proof of several versions of the CLT including those here.[32]

Proof*[33] of Theorem 1.5.2 (binomial case): We first establish a somewhat simpler result which uses finite limits on both sides of $(S_n - n\overline{X})/\sigma\sqrt{n}$, *i.e.*, we show that for any finite $y < z$,

$$\lim_{n\to\infty} \Pr\left\{y < \frac{S_n - n\overline{X}}{\sigma\sqrt{n}} \le z\right\} = \int_y^z \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right) du. \tag{1.85}$$

[32] Many elementary texts provide 'simple proofs,' using transform techniques, but, among other problems, they usually indicate that the normalized sum has a density that approaches the Gaussian density, which is incorrect for all discrete rv's.

[33] This proof can be omitted without loss of continuity. However, it is an important part of achieving a thorough understanding of the CLT.

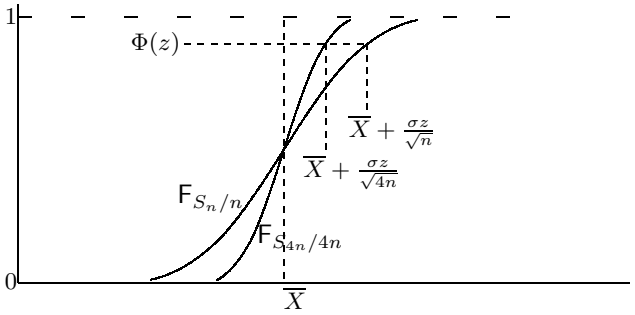


Figure 1.13: Approximation of the distribution function $\mathsf{F}_{S_n/n}$ of a sample average by a Gaussian distribution of the same mean and variance. Whenever n is increased by a factor of 4, the curve is horizontally scaled inward toward $\overline{X}$ by a factor of 2. The CLT says both that these curves are scaled horizontally as $1/\sqrt{n}$ and also that they are better approximated by the Gaussian of the given mean and variance as n increases.

The probability on the left of (1.85) can be rewritten as

$$\begin{aligned}\Pr\left\{y < \frac{S_n - n\overline{X}}{\sigma\sqrt{n}} \le z\right\}\} &= \Pr\left\{\frac{\sigma\, y}{\sqrt{n}} < \frac{S_n}{n} - \overline{X} \le \frac{\sigma\, z}{\sqrt{n}}\right\}\} \\ &= \sum_k \mathsf{p}_{S_n}(k) \quad \text{for} \quad \frac{\sigma\, y}{\sqrt{n}} < \frac{k}{n} - p \le \frac{\sigma\, z}{\sqrt{n}}. \end{aligned} \tag{1.86}$$

Let $\tilde{p} = k/n$, $\tilde{q} = 1 - \tilde{p}$, and $\epsilon(k,n) = \tilde{p} - p$. We abbreviate $\epsilon(k,n)$ as ϵ where k and n are clear from the context. From (1.23), we can express $\mathsf{p}_{S_n}(k)$ as

$$\begin{aligned}\mathsf{p}_{S_n}(k) &\sim \frac{1}{\sqrt{2\pi n\tilde{p}\tilde{q}}} \exp n\left[\tilde{p}\ln(p/\tilde{p}) + \tilde{q}\ln(q/\tilde{q})\right] \\ &= \frac{1}{\sqrt{2\pi n(p+\epsilon)(q-\epsilon)}} \exp\left[-n(p+\epsilon)\ln\left(1+\frac{\epsilon}{p}\right) - n(q-\epsilon)\ln\left(1-\frac{\epsilon}{q}\right)\right]\} \\ &= \frac{1}{\sqrt{2\pi n(p+\epsilon)(q-\epsilon)}} \exp\left[-n\left(\frac{\epsilon^2}{2p} - \frac{\epsilon^3}{6p^2} + \cdots + \frac{\epsilon^2}{2q} + \frac{\epsilon^3}{6q^2}\cdots\right)\right]. \end{aligned} \tag{1.87}$$

where we have used the power series expansion, $\ln(1+u) = u - u^2/2 + u^3/3 \cdots$. From (1.86), $\sigma y/\sqrt{n} < p + \epsilon(k,n) \le \sigma z/\sqrt{n}$, so the omitted terms in (1.87) go to 0 uniformly over the range of k in (1.86). The term $(p+\epsilon)(q-\epsilon)$ also converges (uniformly in k) to pq as $n \to \infty$. Thus

$$\begin{aligned}\mathsf{p}_{S_n}(k) &\sim \frac{1}{\sqrt{2\pi npq}} \exp\left[-n\left(\frac{\epsilon^2}{2p} + \frac{\epsilon^2}{2q}\right)\right]\} \\ &= \frac{1}{\sqrt{2\pi npq}} \exp\left[\frac{-n\epsilon^2}{2pq}\right]\} \\ &= \frac{1}{\sqrt{2\pi npq}} \exp\left[\frac{-u^2(k,n)}{2}\right]\} \qquad \text{where } u(k,n) = \frac{k-np}{\sqrt{pqn}}. \end{aligned} \tag{1.88}$$

Since the ratio of the right to left side in (1.88) approaches 1 uniformly in k over the given range of k,

$$\sum_k \mathsf{p}_{S_n}(k) \sim \sum_k \frac{1}{\sqrt{2\pi npq}} \exp\left[\frac{-u^2(k,n)}{2}\right]. \tag{1.89}$$

Since $u(k,n)$ increases in increments of $1/\sqrt{pqn}$, we can view the sum on the right above as a Riemann sum approximating (1.85). Since the terms of the sum get closer as $n \to \infty$, and since the Riemann integral exists, (1.85) must be satisfied in the limit.

To complete the proof, note from the Chebyshev inequality that

$$\Pr\left\{|S_n - \overline{X}| \geq \frac{\sigma|y|}{\sqrt{n}}\right\} \leq \frac{1}{y^2},$$

and thus

$$\Pr\left\{y < \frac{S_n - n\overline{X}}{\sigma\sqrt{n}} \leq z\right\} \leq \Pr\left\{\frac{S_n - n\overline{X}}{\sigma\sqrt{n}} \leq z\right\} \leq \frac{1}{y^2} + \Pr\left\{y < \frac{S_n - n\overline{X}}{\sigma\sqrt{n}} \leq z\right\}$$

Choosing $y = -n^{1/4}$, we see that $1/y^2 \to 0$ as $n \to \infty$. Also $n\epsilon^3(k,n)$ in (1.87) goes to 0 as $n \to \infty$ for all (k,n) satisfying (1.86) with $y = -n^{1/4}$. Taking the limit as $n \to \infty$ then proves the theorem. □

If we trace through the various approximations in the above proof, we see that the error in $\mathsf{p}_{S_n}(k)$ goes to 0 as $1/n$. This is faster than the $1/\sqrt{n}$ bound in the Berry-Esseen theorem. If we look at Figure 1.3, however, we see that the distribution function of S_n/n contains steps of order $1/\sqrt{n}$. These vertical steps cause the binomial result here to have the same slow $1/\sqrt{n}$ convergence as the general Berry-Esseen bound. It turns out that if we evaluate the distribution function only at the midpoints between these steps, *i.e.*, at $z = (k + 1/2 - np)/\sigma\sqrt{n}$, then the convergence in the distribution function is of order $1/n$.

Since the CLT provides such explicit information about the convergence of S_n/n to $\overline{X}$, it is reasonable to ask why the weak law of large numbers (WLLN) is so important. The first reason is that the WLLN is so simple that it can be used to give clear insights to situations where the CLT could confuse the issue. A second reason is that the CLT requires a variance, where as we see next, the WLLN does not. A third reason is that the WLLN can be extended to many situations in which the variables are not independent and/or not identically distributed.[34] A final reason is that the WLLN provides an upper bound on the tails of $\mathsf{F}_{S_n/n}$, whereas the CLT provides only an approximation.

1.5.4 Weak law with an infinite variance

We now establish the WLLN without assuming a finite variance.

[34]Central limit theorems also hold in many of these more general situations, but they do not hold as widely as the WLLN.

Theorem 1.5.3 (WLLN). *For each integer $n \geq 1$, let $S_n = X_1 + \cdots + X_n$ where $X_1, X_2, \ldots$ are IID rv's satisfying $\mathsf{E}[|X|] < \infty$. Then for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}[X]\right| > \epsilon\right\} = 0. \tag{1.90}$$

Proof:[35] We use a truncation argument; such arguments are used frequently in dealing with rv's that have infinite variance. The underlying idea in these arguments is important, but some less important details are treated in Exercise 1.35. Let b be a positive number (which we later take to be increasing with n), and for each variable X_i, define a new rv $\breve{X}_i$ (see Figure 1.14) by

$$\breve{X}_i = \begin{cases} X_i & \text{for } \mathsf{E}[X] - b \leq X_i \leq \mathsf{E}[X] + b \\ \mathsf{E}[X] + b & \text{for } X_i > \mathsf{E}[X] + b \\ \mathsf{E}[X] - b & \text{for } X_i < \mathsf{E}[X] - b. \end{cases} \tag{1.91}$$

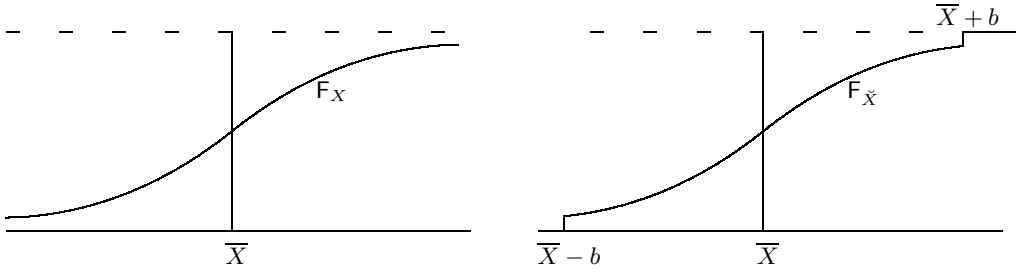


Figure 1.14: The truncated rv $\breve{X}$ for a given rv X has a distribution function which is truncated at $\overline{X} \pm b$.

The truncated variables $\breve{X}_i$ are IID and, because of the truncation, must have a finite second moment. Thus the WLLN applies to the sample average $\breve{S}_n = \breve{X}_1 + \cdots \breve{X}_n$. More particularly, using the Chebshev inequality in the form of (1.75) on $\breve{S}_n/n$, we get

$$\Pr\left\{\left|\frac{\breve{S}_n}{n} - \mathsf{E}\left[\breve{X}\right]\right| > \frac{\epsilon}{2}\right\} \leq \frac{4\sigma^2_{\breve{X}}}{n\epsilon^2} \leq \frac{8b\mathsf{E}[|X|]}{n\epsilon^2},$$

where Exercise 1.35 demonstrates the final inequality. Exercise 1.35 also shows that $\mathsf{E}\left[\breve{X}\right]$ approaches $\mathsf{E}[X]$ as $b \to \infty$ and thus that

$$\Pr\left\{\left|\frac{\breve{S}_n}{n} - \mathsf{E}[X]\right| > \epsilon\right\} \leq \frac{8b\mathsf{E}[|X|]}{n\epsilon^2}, \tag{1.92}$$

for all sufficiently large b. This bound also applies to S_n/n in the case where $S_n = \breve{S}_n$, so we have the following bound (see Exercise 1.35 for further details):

$$\Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}[X]\right| > \epsilon\right\} \leq \Pr\left\{\left|\frac{\breve{S}_n}{n} - \mathsf{E}[X]\right| > \epsilon\right\} + \Pr\left\{S_n \neq \breve{S}_n\right\}. \tag{1.93}$$

[35]The details of this proof can be omitted without loss of continuity. However, the general structure of the truncation argument should be understood.

The original sum S_n is the same as $\breve{S}_n$ unless one of the X_i has an outage, *i.e.*, $|X_i - \overline{X}| > b$. Thus, using the union bound, $\Pr\left\{S_n \neq \breve{S}_n\right\} \leq n\Pr\{|X_i - \overline{X}| > b$. Substituting this and (1.92) into (1.93),

$$\Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}[X]\right| > \epsilon\right\} \leq \frac{8b\mathsf{E}[|X|]}{n\epsilon^2} + \frac{n}{b}\left[b\Pr\{|X - \mathsf{E}[X]| > b\}\right]. \tag{1.94}$$

We now show that for any $\epsilon > 0$ and $\delta > 0$, $\Pr\{|S_n/n - \overline{X}| \geq \epsilon \leq \delta$ for all sufficiently large n. We do this, for given ϵ, δ, by choosing $b(n)$ for each n so that the first term in (1.94) is equal to $\delta/2$. Thus $b(n) = n\delta\epsilon^2/16\mathsf{E}[|X|]$. This means that $n/b(n)$ in the second term is independent of n. Now from (1.55), $\lim_{b\to\infty} b\Pr\{|X - \overline{X}| > b = 0$, so by choosing $b(n)$ sufficiently large (and thus n sufficiently large), the second term in (1.94) is also at most $\delta/2$. □

1.5.5 Convergence of random variables

This section has now developed a number of results about how the sequence of sample averages, $\{S_n/n;\ n \geq 1\}$ for a sequence of IID rv's $\{X_i;\ i \geq 1\}$ approach the mean $\overline{X}$. In the case of the CLT, the limiting distribution around the mean is also specified to be Gaussian. At the outermost intuitive level, *i.e.*, at the level most useful when first looking at some very complicated set of issues, viewing the limit of the sample averages as being essentially equal to the mean is highly appropriate.

At the next intuitive level down, the meaning of the word *essentially* becomes important and thus involves the details of the above laws. All of the results involve how the rv's S_n/n change with n and become better and better approximated by $\overline{X}$. When we talk about a sequence of rv's (namely a sequence of functions on the sample space) being approximated by a rv or numerical constant, we are talking about some kind of *convergence*, but it clearly is not as simple as a sequence of real numbers (such as $1/n$ for example) converging to some given number (0 for example).

The purpose of this section, is to give names and definitions to these various forms of convergence. This will give us increased understanding of the laws of large numbers already developed, but, equally important, It will allow us to develop another law of large numbers called the *strong law of large numbers* (SLLN). Finally, it will put us in a position to use these convergence results later for sequences of rv's other than the sample averages of IID rv's.

We discuss four types of convergence in what follows, convergence in distribution, in probability, in mean square, and with probability 1. For the first three, we first recall the type of large-number result with that type of convergence and then give the general definition.

For convergence with probability 1 (WP1), we first define this type of convergence and then provide some understanding of what it means. This will then be used in Chapter 4 to state and prove the SLLN.

We start with the central limit theorem, which, from (1.81) says

$$\lim_{n\to\infty} \Pr\left\{ \frac{S_n - n\overline{X}}{\sqrt{n}\,\sigma} \le z \right\} = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) dx \qquad \text{for every } z \in \mathbb{R}.$$

This is illustrated in Figure 1.12 and says that the sequence (in n) of distribution functions $\Pr\left\{ \frac{S_n - n\overline{X}}{\sqrt{n}\sigma} \le z \right\}$ converges at every z to the normal distribution function at z. This is an example of *convergence in distribution*.

Definition 1.5.1. *A sequence of random variables, $Z_1, Z_2, \ldots$, converges in distribution to a random variable Z if $\lim_{n\to\infty} \mathsf{F}_{Z_n}(z) = \mathsf{F}_Z(z)$ at each z for which $\mathsf{F}_Z(z)$ is continuous.*

For the CLT example, the rv's that converge in distribution are $\{\frac{S_n - n\overline{X}}{\sqrt{n}\sigma};\ n \ge 1\}$, and they converge in distribution to the normal Gaussian rv.

Convergence in distribution does not say that the rv's themselves converge in any reasonable sense, but only that their distribution functions converge. For example, let $Y_1, Y_2, \ldots$, be IID rv's, each with the distribution function F_Y. For each $n \ge 1$, if we let let $Z_n = Y_n + 1/n$, then it is easy to see that $\{Z_n;\ n \ge 1\}$ converges in distribution to Y. However (assuming Y has variance σ_Y^2 and is independent of each Z_n), we see that $Z_n - Y$ has variance $2\sigma_Y^2$. Thus Z_n does not get close to Y as $n \to \infty$ in any reasonable sense, and $Z_n - Z_m$ does not get small as n and m both get large.[36] As an even more trivial example, the sequence $\{Y_n\ n \ge 1\}$ converges in distribution to Y.

For the CLT, it is the rv's $\frac{S_n - n\overline{X}}{\sqrt{n}\sigma}$ that converge in distribution to the normal. As shown in Exercise 1.38, however, the rv $\frac{S_n - n\overline{X}}{\sqrt{n}\sigma} - \frac{S_{2n} - 2n\overline{X}}{\sqrt{2n}\sigma}$ is not close to 0 in any reasonable sense, even though the two terms have distribution functions that are very close for large n.

For the next type of convergence of rv's, the WLLN, in the form of (1.90), says that

$$\lim_{n\to\infty} \Pr\left\{ \left| \frac{S_n}{n} - \overline{X} \right| > \epsilon \right\} = 0 \qquad \text{for every } \epsilon > 0.$$

This is an example of *convergence in probability,* as defined below:

Definition 1.5.2. *A sequence of random variables $Z_1, Z_2, \ldots$, converges in probability to a rv Z if $\lim_{n\to\infty} \Pr\{|Z_n - Z| > \epsilon\} = 0$ for every $\epsilon > 0$.*

For the WLLN example, Z_n in the definition is the relative frequency S_n/n and Z is the constant rv $\overline{X}$. It is probably simpler and more intuitive in thinking about convergence of rv's to think of the sequence of rv's $\{Y_n = Z_n - Z;\ n \ge 1\}$ as converging to 0 in some sense.[37] As illustrated in Figure 1.10, convergence in probability means that $\{Y_n;\ n \ge 1\}$ converges in distribution to a unit step function at 0.

[36] In fact, saying that a sequence of rv's converges in distribution is unfortunate but standard terminology. It would be just as concise, and far less confusing, to say that a sequence of distribution functions converge rather than saying that a sequence of rv's converge in distribution.

[37] Definition 1.5.2 gives the impression that convergence to a rv Z is more general than convergence to a constant or convergence to 0, but converting the rv's to $Y_n = Z_n - Z$ makes it clear that this added generality is quite superficial.

An equivalent statement, as illustrated in Figure 1.11, is that $\{Y_n;\, n \geq 1\}$ converges in probability to 0 if $\lim_{n\to\infty} \mathsf{F}_{Y_n}(y) = 0$ for all $y < 0$ and $\lim_{n\to\infty} \mathsf{F}_{Y_n}(y) = 1$ for all $y > 0$. This shows that convergence in probability is a special case of convergence in distribution, since with convergence in probability, the sequence F_{Y_n} of distribution functions converges to a unit step at 0. Note that $\lim_{n\to\infty} \mathsf{F}_{Y_n}(y)$ is not specified at $y = 0$. However, the step function is not continuous at 0, so the limit there need not be specified for convergence in distribution.

Convergence in probability says quite a bit more than convergence in distribution. As an important example of this, consider the difference $Y_n - Y_m$ for n and m both large. If $\{Y_n;\, n \geq 1\}$ converges in probability to 0, then Y_n and Y_m are both close to 0 with high probability for large n and m, and thus close to each other. More precisely, $\lim_{m\to\infty,n\to\infty} \Pr\{|Y_n - Y_m| > \epsilon\} = 0$ for every $\epsilon > 0$. If the sequence $\{Y_n;\, n \geq 1\}$ merely converges in distribution to some arbitrary distribution, then, as we saw, $Z_n - Z_m$ can be large with high probability, even when n and m are large. Another example of this is given in Exercise 1.38.

It appears paradoxical that the CLT is more explicit about the convergence of S_n/n to $\overline{X}$ than the weak law, but it corresponds to a weaker type of convergence. The resolution of this paradox is that the sequence of rv's in the CLT is $\{\frac{S_n - n\overline{X}}{\sqrt{n}\sigma};\, n \geq 1\}$. The presence of $\sqrt{n}$ in the denominator of this sequence provides much more detailed information about how S_n/n approaches $\overline{X}$ with increasing n than the limiting unit step of $\mathsf{F}_{S_n/n}$ itself. For example, it is easy to see from the CLT that $\lim_{n\to\infty} \mathsf{F}_{S_n/n}(\overline{X}) = 1/2$, which can not be derived directly from the weak law.

Yet another kind of convergence is *convergence in mean square* (MS). An example of this, for the sample average S_n/n of IID rv's with a variance, is given in (1.74), repeated below:

$$\lim_{n\to\infty} \mathsf{E}\left[\left(\frac{S_n}{n} - \overline{X}\right)^2\right] = 0.$$

The general definition is as follows:

Definition 1.5.3. *A sequence of rv's $Z_1, Z_2, \ldots$, converges in mean square (MS) to a rv Z if $\lim_{n\to\infty} \mathsf{E}\left[(Z_n - Z)^2\right] = 0$.*

Our derivation of the weak law of large numbers (Theorem 1.5.1) was essentially based on the MS convergence of (1.74). Using the same approach, Exercise 1.37 shows in general that convergence in MS implies convergence in probability. Convergence in probability does not imply MS convergence, since as shown in Theorem 1.5.3, the weak law of large numbers holds without the need for a variance.

Figure 1.15 illustrates the relationship between these forms of convergence, *i.e.*, mean square convergence implies convergence in probability, which in turn implies convergence in distribution. The figure also shows convergence with probability 1 (WP1), which is the next form of convergence to be discussed.

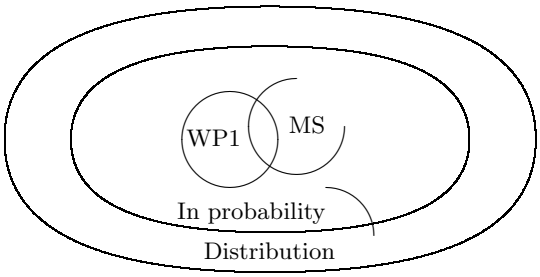


Figure 1.15: Relationship between different kinds of convergence: Convergence in distribution is the most general and is implied by all the others. Convergence in probability is the next most general and is implied by convergence with probability 1 (WP1) and by mean square (MS) convergence, neither of which imply the other.

1.5.6 Convergence with probability 1

We next define and discuss *convergence with probability 1*, abbreviated as convergence WP1. Convergence WP1 is often referred to as convergence a.s. (almost surely) and convergence a.e. (almost everywhere). The strong law of large numbers, which is discussed briefly in this section and further discussed and proven in various forms in Chapters 4 and 7, provides an extremely important example of convergence WP1. The general definition is as follows:

Definition 1.5.4. *Let $Z_1, Z_2, \ldots$, be a sequence of rv's in a sample space Ω and let Z be another rv in Ω. Then $\{Z_n; n \geq 1\}$ is defined to converge to Z with probability 1 (WP1) if*

$$\Pr\Big\{\omega \in \Omega \ : \ \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\Big\} = 1. \tag{1.95}$$

The condition $\Pr\{\omega \in \Omega \ : \ \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\} = 1$ is often stated more compactly as $\Pr\{\lim_n Z_n = Z\} = 1$, and even more compactly as $\lim_n Z_n = Z$ WP1, but the form here is the simplest for initial understanding. As discussed in Chapter 4. the SLLN says that if $\{X_i; i \geq 1\}$ are IID with $\mathsf{E}[|X|] < \infty$, then the sequence of sample averages, $\{S_n/n; n \geq 1\}$ converges WP1 to $\overline{X}$.

In trying to understand (1.95), note that each sample point ω of the underlying sample space Ω maps to a sample value $Z_n(\omega)$ of each rv Z_n, and thus maps to a sample path $\{Z_n(\omega);\ n \geq 1\}$. For any given ω, such a sample path is simply a sequence of real numbers. That sequence of real numbers might converge to $Z(\omega)$ (which is a real number for the given ω), it might converge to something else, or it might not converge at all. Thus there is a set of ω for which the corresponding sample path $\{Z_n(\omega); n \geq 1\}$ converges to $Z(\omega)$, and a second set for which the sample path converges to something else or does not converge at all. Convergence WP1 of the sequence of rv's is thus defined to occur when the first set of sample paths above has probability 1.

For each ω, the sequence $\{Z_n(\omega);\ n \geq 1\}$ is simply a sequence of real numbers, so we briefly review what the limit of such a sequence is. A sequence of real numbers $b_1, b_2, \ldots$ is said to

have a limit b if, for every $\epsilon > 0$, there is an integer m_ϵ such that $|b_n - b| \leq \epsilon$ for all $n \geq m_\epsilon$. An equivalent statement is that $b_1, b_2, \ldots$, has a limit b if, for every integer $k \geq 1$, there is an integer $m(k)$ such that $|b_n - b| \leq 1/k$ for all $n \geq m(k)$.

Figure 1.16 illustrates this definition for those, like the author, whose eyes blur on the second or third 'there exists', 'such that', etc. in a statement. As illustrated, an important aspect of convergence of a sequence $\{b_n;\, n \geq 1\}$ of real numbers is that b_n becomes close to b for large n and stays close for all sufficiently large values of n.

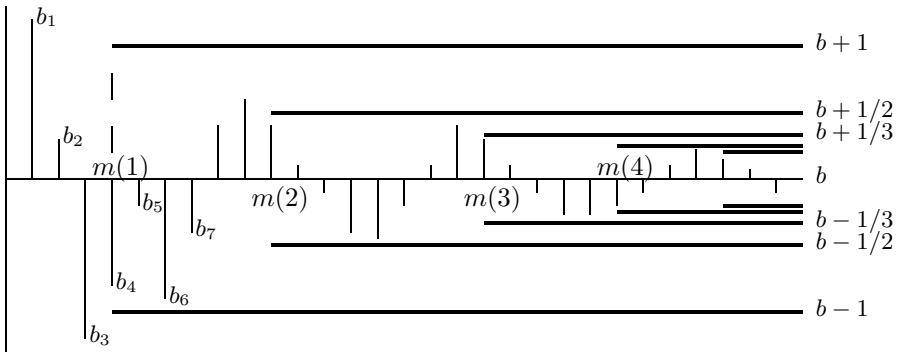


Figure 1.16: Illustration of a sequence of real numbers $b_1, b_2, \ldots$ that converge to a number b. The figure illustrates an integer $m(1)$ such that for all $n \geq m(1)$, b_n lies in the interval $b \pm 1$. Similarly, for each $k \geq 1$, there is an integer $m(k)$ such that b_n lies in $b \pm 1/k$ for all $n \geq m(k)$. Thus $\lim_{n\to\infty} b_n = b$ means that for a sequence of ever tighter constraints, the kth constraint can be met for all sufficiently large n, (*i.e.*, all $n \geq m(k)$). Intuitively, convergence means that the elements $b_1, b_2, \ldots$ get close to b *and stay close*. The sequence of positive integers $m(1), m(2), \ldots$ is nondecreasing, but otherwise arbitrary, depending only on the sequence $\{b_n;\, n \geq 1\}$. For sequences that converge very slowly, the integers $m(1), m(2), \ldots$ are simply correspondingly larger.

Figure 1.17 gives an example of a sequence of real numbers that does not converge. Intuitively, this sequence is close to 0 (and in fact identically equal to 0) for most large n, but it doesn't stay close.

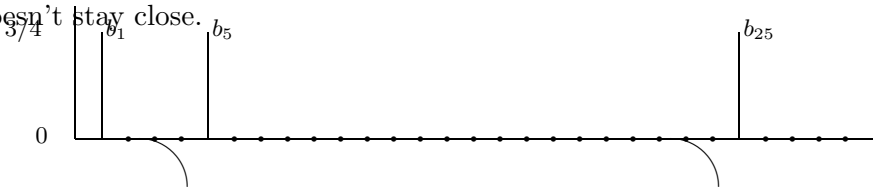


Figure 1.17: Illustration of a non-convergent sequence of real numbers $b_1, b_2, \ldots$. The sequence is defined by $b_n = 3/4$ for $n = 1,\ 5,\ 25, \ldots, 5^j, \ldots$ for all integer $j \geq 0$. For all other n, $b_n = 0$. The terms for which $b_n \neq 0$ become increasingly rare as $n \to \infty$. Note that $b_n \in [1, 1]$ for all n, but there is no $m(2)$ such that $b_n \in [-\frac{1}{2}, \frac{1}{2}]$ for all $n \geq m(2)$. Thus the sequence does not converge.

The following example illustrates how a sequence of rv's can converge in probability but not converge WP1. The example also provides some clues as to why convergence WP1 is important.

Example 1.5.1. Consider a sequence $\{Z_n; n \geq 1\}$ of rv's for which the sample paths constitute the following slight variation of the sequence of real numbers in Figure 1.17. In particular, as illustrated in Figure 1.18, the nonzero term at $n = 5^j$ in Figure 1.17 is replaced by a nonzero term at a randomly chosen n in the interval $[5^j, 5^{j+1})$.

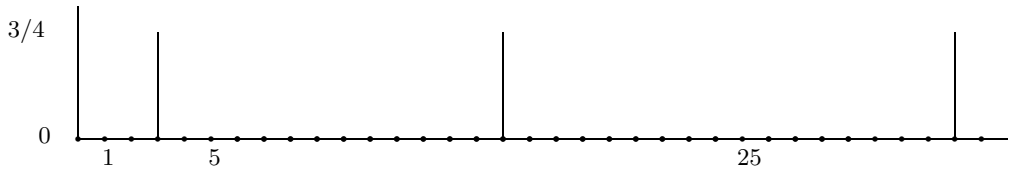


Figure 1.18: Illustration of a sample path of a sequence of rv's $\{Y_n; n \geq 0\}$ where, for each $j \geq 0$, $Y_n = 1$ for an equiprobable choice of $n \in [5^j, 5^{j+1})$ and $Y_n = 0$ otherwise.

Since each sample path contains a single one in each segment $[5^j, 5^{j+1})$, and contains zero's elsewhere, none of the sample paths converge. In other words, $\Pr\{\omega : \lim Z_n(\omega) = 0\} = 0$ rather than 1. On the other hand $\Pr\{Z_n = 0\} = 1 - 5^{-j}$ for $5^j \leq n < 5^{j+1}$, so $\lim_{n\to\infty} \Pr\{Z_n = 0\} = 1$.

Thus this sequence of rv's converges to 0 in probability, but does not converge to 0 WP1. This sequence also converges in mean square and (since it converges in probability) in distribution. Thus we have shown (by example) that convergence WP1 is not implied by any of the other types of convergence we have discussed. We will show in Section 4.2 that convergence WP1 does imply convergence in probability and in distribution but not in mean square (as illustrated in Figure 1.15).

The interesting point in this example is that this sequence of rv's is not bizarre (although it is somewhat specialized to make the analysis simple). Another important point is that this definition of convergence has a long history of being accepted as the 'useful,' 'natural,' and 'correct' way to define convergence for a sequence of real numbers. Thus it is not surprising that convergence WP1 will turn out to be similarly useful for sequences of rv's.

There is a price to be paid in using the concept of convergence WP1. We must then look at the entire sequence of rv's and can no longer analyze finite n-tuples and then go to the limit as $n \to \infty$. This requires a significant additional layer of abstraction, which involves additional mathematical precision and initial loss of intuition. For this reason we put off further discussion of convergence WP1 and the SLLN until Chapter 4 where it is needed.

1.6 Relation of probability models to the real world

Whenever experienced and competent engineers or scientists construct a probability model to represent aspects of some system that either exists or is being designed for some application, they must acquire a deep knowledge of the system and its surrounding circumstances, and concurrently consider various types of probability models used in probabilistic analysis of the same or similar systems. Usually very simple probability models help in understanding

the real-world system, and knowledge about the real-world system helps in understanding what aspects of the system are well-modeled by a given probability model. For a text such as this, there is insufficient space to understand the real-world aspects of each system that might be of interest. We must use the language of various canonical real-world systems for motivation and insight when studying probability models for various classes of systems, but such models must necessarily be chosen more for their tutorial than practical value.

There is a danger, then, that readers will come away with the impression that analysis is more challenging and important than modeling. To the contrary, for work on real-world systems, modeling is almost always more difficult, more challenging, and more important than analysis. The objective here is to provide the necessary knowledge and insight about probabilistic models so that the reader can later combine this with a deep understanding of particular real application areas. This will result in a useful interactive use of models, analysis, and experimentation.

In this section, our purpose is not to learn how to model real-world problems, since, as said above, this requires deep and specialized knowledge of whatever application area is of interest. Rather it is to understand the following conceptual problem that was posed in Section 1.1. Suppose we have a probability model of some real-world experiment involving randomness in the sense expressed there. When the real-world experiment being modeled is performed, there is an outcome, which presumably is one of the outcomes of the probability model, but there is no observable probability.

It appears to be intuitively natural, for experiments that can be carried out repeatedly under essentially the same conditions, to associate the probability of a given event with the relative frequency of that event over many repetitions. We now have the background to understand this approach. We first look at relative frequencies within the probability model, and then within the real world.

1.6.1 Relative frequencies in a probability model

We have seen that for any probability model, an extended probability model exists for n IID idealized experiments of the original model. For any event A in the original model, the indicator function $\mathbb{I}_A$ is a random variable, and the relative frequency of A over n IID experiments is the sample average of n IID rv's each with the distribution of $\mathbb{I}_A$. From the weak law of large numbers, this relative frequency converges in probability to $\mathsf{E}[\mathbb{I}_A] = \Pr\{A\}$. By taking the limit $n \to \infty$, the strong law of large numbers says that the relative frequency of A converges with probability 1 to $\Pr\{A\}$.

In plain English, this says that for large n, the relative frequency of an event (in the n-repetition IID model) is essentially the same as the probability of that event. The word *essentially* is carrying a great deal of hidden baggage. For the weak law, for any $\epsilon, \delta > 0$, the relative frequency is within some ϵ of $\Pr\{A\}$ with a confidence level $1 - \delta$ whenever n is sufficiently large. For the strong law, the ϵ and δ are avoided, but only by looking directly at the limit $n \to \infty$. Despite the hidden baggage, though, relative frequency and probability are related as indicated.

1.6.2 Relative frequencies in the real world

In trying to sort out if and when the laws of large numbers have much to do with real-world experiments, we should ignore the mathematical details for the moment and agree that for large n, the relative frequency of an event A over n IID trials of an idealized experiment is essentially $\Pr\{A\}$. We can certainly visualize a real-world experiment that has the same set of possible outcomes as the idealized experiment and we can visualize evaluating the relative frequency of A over n repetitions with large n. If that real-world relative frequency is essentially equal to $\Pr\{A\}$, and this is true for the various events A of greatest interest, then it is reasonable to hypothesize that the idealized experiment is a reasonable model for the real-world experiment, at least so far as those given events of interest are concerned.

One problem with this comparison of relative frequencies is that we have carefully specified a model for n IID repetitions of the idealized experiment, but have said nothing about how the real-world experiments are repeated. The IID idealized experiments specify that the conditional probability of A at one trial is the same no matter what the results of the other trials are. Intuitively, we would then try to isolate the n real-world trials so they don't affect each other, but this is a little vague. The following examples help explain this problem and several others in comparing idealized and real-world relative frequenices.

Example 1.6.1. Coin tossing: Tossing coins is widely used as a way to choose the first player in various games, and is also sometimes used as a primitive form of gambling. Its importance, however, and the reason for its frequent use, is its simplicity. When tossing a coin, we would argue from the symmetry between the two sides of the coin that each should be equally probable (since any procedure for evaluating the probability of one side should apply equally to the other). Thus since H and T are the only outcomes (the remote possibility of the coin balancing on its edge is omitted from the model), the reasonable and universally accepted model for coin tossing is that H and T each have probability $1/2$.

On the other hand, the two sides of a coin are embossed in different ways, so that the mass is not uniformly distributed. Also the two sides do not behave in quite the same way when bouncing off a surface. Each denomination of each currency behaves slightly differently in this respect. Thus, not only do coins violate symmetry in small ways, but different coins violate it in different ways.

How do we test whether this effect is significant? If we assume for the moment that successive tosses of the coin are well-modeled by the idealized experiment of n IID trials, we can essentially find the probability of H for a particular coin as the relative frequency of H in a sufficiently large number of independent tosses of that coin. This gives us slightly different relative frequencies for different coins, and thus slightly different probability models for different coins.

The assumption of independent tosses is also questionable. Consider building a carefully engineered machine for tossing coins and using it in a vibration-free environment. A standard coin is inserted into the machine in the same way for each toss and we count the number of heads and tails. Since the machine has essentially eliminated the randomness, we would expect all the coins, or almost all the coins, to come up the same way — the more precise the machine, the less independent the results. By inserting the original coin in a random

way, a single trial might have equiprobable results, but successive tosses are certainly not independent. The successive trials would be closer to independent if the tosses were done by a slightly inebriated individual who tossed the coins high in the air.

The point of this example is that there are many different coins and many ways of tossing them, and the idea that one model fits all is reasonable under some conditions and not under others. Rather than retreating into the comfortable world of theory, however, note that we can now find the relative frequency of heads for any given coin and essentially for any given way of tossing that coin.[38]

Example 1.6.2. Binary data: Consider the binary data transmitted over a communication link or stored in a data facility. The data is often a mixture of encoded voice, video, graphics, text, etc., with relatively long runs of each, interspersed with various protocols for retrieving the original non-binary data.

The simplest (and most common) model for this is to assume that each binary digit is 0 or 1 with equal probability and that successive digits are statistically independent. This is the same as the model for coin tossing after the trivial modification of converting $\{H, T\}$ into $\{0, 1\}$. This is also a rather appropriate model for designing a communication or storage facility, since all n-tuples are then equiprobable (in the model) for each n, and thus the facilities need not rely on any special characteristics of the data. On the other hand, if one wants to compress the data, reducing the required number of transmitted or stored bits per incoming bit, then a more elaborate model is needed.

Developing such an improved model would require finding out more about where the data is coming from — a naive application of calculating relative frequencies of n-tuples would probably not be the best choice. On the other hand, there are well-known data compression schemes that in essence track dependencies in the data and use them for compression in a coordinated way. These schemes are called *universal data-compression* schemes since they don't rely on a probability model. At the same time, they are best analyzed by looking at how they perform for various idealized probability models.

The point of this example is that choosing probability models often depends heavily on how the model is to be used. Models more complex than IID binary digits are usually based on what is known about the input processes. Measuring relative frequencies and Associating them with probabilities is the basic underlying conceptual connection between real-world and models, but in practice this is essentially the relationship of last resort. For most of the applications we will study, there is a long history of modeling to build on, with experiments as needed.

Example 1.6.3. Fable: In the year 2008, the financial structure of the USA failed and the world economy was brought to its knees. Much has been written about the role of greed on Wall Street and incompetence in Washington. Another aspect of the collapse, however, was a widespread faith in stochastic models for limiting risk. These models encouraged

[38] We are not suggesting that distinguishing different coins for the sake of coin tossing is an important problem. Rather, we are illustrating that even in such a simple situation, the assumption of identically prepared experiments is questionable and the assumption of independent experiments is questionable. The extension to n repetitions of IID experiments is not necessarily a good model for coin tossing. In other words, one has to question both the original model and the n-repetition model.

people to engage in investments that turned out to be far riskier than the models predicted. These models were created by some of the brightest PhD's from the best universities, but they failed miserably because they modeled everyday events very well, but modeled the rare events and the interconnection of events poorly. They failed badly by not understanding their application, and in particular, by trying to extrapolate typical behavior when their primary goal was to protect against highly atypical situations. The moral of the fable is that brilliant analysis is not helpful when the modeling is poor; as computer engineers say, "garbage in, garbage out."

The examples above show that the problems of modeling a real-world experiment are often connected with the question of creating a model for a set of experiments that are not exactly the same and do not necessarily correspond to the notion of independent repetitions within the model. In other words, the question is not only whether the probability model is reasonable for a single experiment, but also whether the IID repetition model is appropriate for multiple copies of the real-world experiment.

At least we have seen, however, that if a real-world experiment can be performed many times with a physical isolation between performances that is well modeled by the IID repetition model, then the relative frequencies of events in the real-world experiment correspond to relative frequencies in the idealized IID repetition model, which correspond to probabilities in the original model. In other words, under appropriate circumstances, the probabilities in a model become essentially observable over many repetitions.

We will see later that our emphasis on IID repetitions was done for simplicity. There are other models for repetitions of a basic model, such as Markov models, that we study later. These will also lead to relative frequencies approaching probabilities within the repetition model. Thus, for repeated real-world experiments that are well modeled by these repetition models, the real world relative frequencies approximate the probabilities in the model.

1.6.3 Statistical independence of real-world experiments

We have been discussing the use of relative frequencies of an event A in a repeated real-world experiment to test $\Pr\{A\}$ in a probability model of that experiment. This can be done essentially successfully if the repeated trials correpond to IID trials in the idealized experiment. However, the statement about IID trials in the idealized experiment is a statement about probabilities in the extended n-trial model. Thus, just as we tested $\Pr\{A\}$ by repeated real-world trials of a single experiment, we should be able to test $\Pr\{A_1, \ldots, A_n\}$ in the n-repetition model by a much larger number of real-world repetitions of n-tuples rather than single trials.

To be more specific, choose two large integers, m and n, and perform the underlying real-world experiment mn times. Partition the mn trials into m runs of n trials each. For any given n-tuple $A_1, \ldots, A_n$ of successive events, find the relative frequency (over m trials of n tuples) of the n-tuple event $A_1, \ldots, A_n$. This can then be used essentially to test the probability $\Pr\{A_1, \ldots, A_n\}$ in the model for n IID trials. The individual event probabilities can also be tested, so the condition for independence can be tested.

The observant reader will note that there is a tacit assumption above that successive n tuples can be modeled as independent, so it seems that we are simply replacing a big problem with a bigger problem. This is not quite true, since if the trials are dependent with some given probability model for dependent trials, then this test for independence will essentially reject the independence hypothesis for large enough n. In other words, we can not completely verify the correctness of an independence hypothesis for the n-trial model, although in principle we could eventually falsify it if it is false.

Choosing models for real-world experiments is primarily a subject for statistics, and we will not pursue it further except for brief discussions when treating particular application areas. The purpose here has been to treat a fundamental issue in probability theory. As stated before, probabilities are non-observables — they exist in the theory but are not directly measurable in real-world experiments. We have shown that probabilities essentially become observable in the real-world via relative frequencies over repeated trials.

1.6.4 Limitations of relative frequencies

Most real-world applications that are modeled by probability models have such a large sample space that it is impractical to conduct enough trials to choose probabilities from relative frequencies. Even a shuffled deck of 52 cards would require many more than $52! \approx 8 \times 10^{67}$ trials for most of the outcomes to appear even once. Thus relative frequencies can be used to test the probability of given individual events of importance, but are usually impractical for choosing the entire model and even more impractical for choosing a model for repeated trials.

Since relative frequencies give us a concrete interpretation of what probability means, however, we can now rely on other approaches, such as symmetry, for modeling. From symmetry, for example, it is clear that all 52! possible arrangements of a card deck should be equiprobable after shuffling. This leads, for example, to the ability to calculate probabilities of different poker hands, etc., which are such popular exercises in elementary probability classes.

Another valuable modeling procedure is that of constructing a probability model where the possible outcomes are independently chosen n-tuples of outcomes in a simpler model. More generally, most of the random processes to be studied in this text are defined as various ways of combining simpler idealized experiments.

What is really happening as we look at modeling increasingly sophisticated systems and studying increasingly sophisticated models is that we are developing mathematical results for simple idealized models and relating those results to real-world results (such as relating idealized statistically independent trials to real-world independent trials). The association of relative frequencies to probabilities forms the basis for this, but is usually exercised only in the simplest cases.

The way one selects probability models of real-world experiments in practice is to use scientific knowledge and experience, plus simple experiments, to choose a reasonable model. The results from the model (such as the law of large numbers) are then used both to hypothesize results about the real-world experiment and to provisionally reject the model

when further experiments show it to be highly questionable. Although the results about the model are mathematically precise, the corresponding results about the real-world are at best insightful hypotheses whose most important aspects must be validated in practice.

1.6.5 Subjective probability

There are many useful applications of probability theory to situations other than repeated trials of a given experiment. When designing a new system in which randomness (of the type used in probability models) is hypothesized, one would like to analyze the system before actually building it. In such cases, the real-world system does not exist, so indirect means must be used to construct a probability model. Often some sources of randomness, such as noise, can be modeled in the absence of the system. Often similar systems or simulation can be used to help understand the system and help in formulating appropriate probability models. However, the choice of probabilities is to a certain extent subjective.

Another type of situation, of which a canonic example is risk analysis for nuclear reactors, deals with a large number of very unlikely outcomes, each catastrophic in nature. Experimentation clearly cannot be used to establish probabilities, and it is not clear that probabilities have any real meaning here. It can be helpful, however, to choose a probability model on the basis of subjective beliefs which can be used as a basis for reasoning about the problem. When handled well, this can at least make the subjective biases clear, leading to a more rational approach to the problem. When handled poorly, it can hide the arbitrary nature of possibly poor decisions.

We will not discuss the various, often ingenious methods to choose subjective probabilities. The reason is that subjective beliefs should be based on intensive and long term exposure to the particular problem involved; discussing these problems in abstract probability terms weakens this link. We will focus instead on the analysis of idealized models. These can be used to provide insights for subjective models, and more refined and precise results for objective models.

1.7 Summary

This chapter started with an introduction into the correspondence between probability theory and real-world experiments involving randomness. While almost all work in probability theory works with established probability models, it is important to think through what these probabilities mean in the real world, and elementary subjects rarely address these questions seriously.

The next section discussed the axioms of probability theory, along with some insights about why these particular axioms were chosen. This was followed by a review of conditional probabilities, statistical independence, random variables, stochastic processes, and expectations. The emphasis was on understanding the underlying structure of the field rather than reviewing details and problem solving techniques.

This was followed by a fairly extensive treatment of the laws of large numbers. This involved

a fair amount of abstraction, combined with mathematical analysis. The central idea is that the sample average of n IID rv's approaches the mean with increasing n. As a special case, the relative frequency of an event A approaches $\Pr\{A\}$. What the word *approaches* means here is both tricky and vital in understanding probability theory. The strong law of large numbers and convergence WP1 requires mathematical maturity, and is postponed to Chapter 4 where it is first used.

The final section came back to the fundamental problem of understanding the relation between probability theory and randomness in the real-world. It was shown, via the laws of large numbers, that probabilities become essentially observable via relative frequencies calculated over repeated experiments.

There are too many texts on elementary probability to mention here, and most of them serve to give added understanding and background to the material in this chapter. We recommend Bertsekas and Tsitsiklis [2], both for a careful statement of the fundamentals and for a wealth of well-chosen and carefully explained examples.

Texts that cover similar material to that here are [16] and [11]. Kolmogorov [14] is readable for the mathematically mature and is also of historical interest as the translation of the 1933 book that first put probability on a firm mathematical basis. Feller [7] is the classic extended and elegant treatment of elementary material from a mature point of view. Rudin [17] is an excellent text on measure theory for those with advanced mathematical preparation.

1.8 Exercises

Exercise 1.1. Consider a sequence $A_1, A_2, \ldots$ of events each of which have probability zero.

a) Find $\Pr\{\sum_{n=1}^{m} A_n\}$ and find $\lim_{m\to\infty} \Pr\{\sum_{n=1}^{m} A_n\}$. What you have done is to show that the sum of a countably infinite set of numbers each equal to 0 is perfectly well defined as 0.

b) For a sequence of possible phases, $a_1, a_2, \ldots$ between 0 and 2π, and a sequence of singleton events, $A_n = \{a_n\}$, find $\Pr\{\bigcup_n A_n\}$ assuming that the phase is uniformly distributed.

c) Now let A_n be the empty event ϕ for all n. Use (1.1) to show that $\Pr\{\phi\} = 0$.

Exercise 1.2. Let A_1 and A_2 be arbitrary events and show that $\Pr\{A_1 \bigcup A_2\} + \Pr\{A_1 A_2\} = \Pr\{A_1\} + \Pr\{A_2\}$. Explain which parts of the sample space are being double counted on both sides of this equation and which parts are being counted once.

Exercise 1.3. Let $A_1, A_2, \ldots$, be a sequence of disjoint events and assume that $\Pr\{A_n\} = 2^{-n-1}$ for each $n \geq 1$. Assume also that $\Omega = \bigcup_{n=1}^{\infty} A_n$.

a) Show that these assumptions violate the axioms of probability.

b) Show that if (1.3) is substituted for the third of those axioms, then the above assumptions satisfy the axioms.

This shows that the countable additivity of Axiom 3 says something more than the finite additivity of (1.3).

Exercise 1.4. This exercise derives the probability of an arbitrary (non-disjoint) union of events, derives the union bound, and derives some useful limit expressions.

a) For 2 arbitrary events A_1 and A_2, show that

$$A_1 \bigcup A_2 = A_1 \bigcup (A_2 - A_1) \qquad \text{where } A_2 - A_1 = A_2 A_1^c.$$

Show that A_1 and $A_2 - A_1$ are disjoint Hint: This is what Venn diagrams were invented for.

b) For an arbitrary sequence of events, $\{A_n;\, n \geq 1\}$, let $B_1 = A_1$ and for each $n \geq 2$ define $B_n = A_n - \bigcup_{j=1}^{n-1} A_j$. Show that $B_1, B_2, \ldots$, are disjoint events and show that for each $m \geq 2$, $\bigcup_{n=1}^{m} A_n = \bigcup_{n=1}^{n} B_n$. Hint: Use induction.

c) Show that

$$\Pr\Big\{\bigcup_{n=1}^{\infty} A_n\Big\} = \Pr\Big\{\bigcup_{n=1}^{\infty} B_n\Big\} = \sum_{n=1}^{\infty} \Pr\{B_n\}\,.$$

Hint: Use the axioms of probability for the second equality.

d) Show that for each n, $\Pr\{B_n\} \leq \Pr\{A_n\}$. Use this to show that

$$\Pr\Big\{\bigcup_{n=1}^{\infty} A_n\Big\} \leq \sum_{n=1}^{\infty} \Pr\{A_n\}\,.$$

e) Show that $\Pr\{\bigcup_{n=1}^{\infty} A_n\} = \lim_{m\to\infty} \Pr\{\bigcup_{n=1}^{m} A_n\}$. Hint: Combine parts c) and b).. Note that this says that the probability of a limit of unions is equal to the limit of the probabilities. This might well appear to be obvious without a proof, but you will see situations later where similar appearing interchanges cannot be made.

f) Show that $\Pr\{\bigcap_{n=1}^{\infty} A_n\} = \lim_{n\to\infty} \Pr\{\bigcap_{i=1}^{n} A_i\}$. Hint: Remember deMorgan's equalities.

Exercise 1.5. Consider a sample space of 8 equiprobable sample points and let A_1, A_2, A_3 be three events each of probability 1/2 such that $\Pr\{A_1A_2A_3\} = \Pr\{A_1\}\Pr\{A_2\}\Pr\{A_3\}$.

a) Create an example where $\Pr\{A_1A_2\} = \Pr\{A_1A_3\} = \frac{1}{4}$ but $\Pr\{A_2A_3\} = \frac{1}{8}$. Hint: Make a table with a row for each sample point and a column for each event and try different ways of assigning sample points to events (the answer is not unique).

b) Show that, for your example, A_2 and A_3 are not independent. Note that the definition of statistical independence would be very strange if it allowed A_1, A_2, A_3 to be independent while A_2 and A_3 are dependent. This illustrates why the definition of independence requires (1.13) rather than just (1.14).

Exercise 1.6. Suppose X and Y are discrete rv's with the PMF $\mathsf{p}_{XY}(x_i, y_j)$. Show (a picture will help) that this is related to the joint distribution function by

$$\mathsf{p}_{XY}(x_i, y_j) = \lim_{\delta>0,\delta\to 0} [\mathsf{F}(x_i, y_j) - \mathsf{F}(x_i - \delta, y_j) - \mathsf{F}(x_i, y_j - \delta) + \mathsf{F}(x_i - \delta, y_j - \delta)]\,.$$

Exercise 1.7. A variation of Example 1.3.2 is to let M be a random variable that takes on both positive and negative values with the PMF

$$\mathsf{p}_M(m) = \frac{1}{2|m|\,(|m|+1)}.$$

In other words, M is symmetric around 0 and $|M|$ has the same PMF as the nonnegative rv N of Example 1.3.2.

a) Show that $\sum_{m\geq 0} m\mathsf{p}_M(m) = \infty$ and $\sum_{m<0} m\mathsf{p}_M(m) = -\infty$. (Thus show that the expectation of M not only does not exist but is undefined even in the extended real number system.)

b) Suppose that the terms in $\sum_{m=-\infty}^{\infty} m\mathsf{p}_M(m)$ are summed in the order of 2 positive terms for each negative term (*i.e.*, in the order $1, 2, -1, 3, 4, -2, 5, \cdots$). Find the limiting value of the partial sums in this series. Hint: You may find it helpful to know that

$$\lim_{n\to\infty} \left[\sum_{i=1}^{n} \frac{1}{i} - \int_1^n \frac{1}{x}\,dx\right] = \gamma,$$

where γ is the Euler-Mascheroni constant, $\gamma = 0.57721\cdots$.

c) Repeat part b) where, for any given integer $k > 0$, the order of summation is k positive terms for each negative term.

Exercise 1.8. a) For any given rv Y, express $\mathsf{E}[|Y|]$ in terms of $\int_{y<0} \mathsf{F}_Y(y)\,dy$ and $\int_{y\geq 0} \mathsf{F}^{\mathsf{c}}_Y(y)\,dy$.

b) Show that $\mathsf{E}[|Y-\alpha|]$ is minimized by setting α equal to the median of Y (*i.e.*, the value of Y for which $\mathsf{F}_Y(y) = 1/2$). Hint: Use a graphical argument.

Exercise 1.9. Let X be a rv with distribution function $\mathsf{F}_X(x)$. Find the distribution function of the following rv's.

a) The maximum of n IID rv's, each with distribution function $\mathsf{F}_X(x)$.

b) The minimum of n IID rv's, each with distribution $\mathsf{F}_X(x)$.

c) The difference of the rv's defined in a) and b); assume X has a density $\mathsf{f}_X(x)$.

Exercise 1.10. Let X and Y be rv's in some sample space Ω and let $Z = X + Y$, *i.e.*, for each $\omega \in \Omega$, $Z(\omega) = X(\omega) + Y(\omega)$.

a) Show that the set of ω for which $Z(\omega) = \pm\infty$ has probability 0.

b) To show that $Z = X+Y$ is a rv, we must show that for each real number α, the set $\{\omega \in \Omega : X(\omega)+Y(\omega) \leq \alpha\}$ is an event. We proceed indirectly. For an arbitrary positive integer n and an arbitrary integer k, let $B(n,k) = \{\omega : X(\omega) \leq k\alpha/n\} \bigcap \{Y(\omega) \leq (n+1-k)\alpha/n\}$. Let $D(n) = \bigcup_k B(n,k)$ and show that $D(n)$ is an event.

c) On a 2 dimensional sketch for a given α, show the values of $X(\omega)$ and $Y(\omega)$ for which $\omega \in D(n)$. Hint: This set of values should be bounded by a staircase function.

d) Show that

$$\{\omega : X(\omega) + Y(\omega) \leq \alpha\} \;=\; \bigcap\nolimits_n B(n)$$

Explain why this shows that $Z = X + Y$ is a rv.

Exercise 1.11. a) Let $X_1, X_2, \ldots, X_n$ be rv's with expected values $\overline{X}_1, \ldots, \overline{X}_n$. Show that $\mathsf{E}[X_1 + \cdots + X_n] = \overline{X}_1 + \cdots + \overline{X}_n$. You may assume that the rv's have a joint density function, but do not assume that the rv's are independent.

b) Now assume that $X_1, \ldots, X_n$ are statistically independent and show that the expected value of the product is equal to the product of the expected values.

c) Again assuming that $X_1, \ldots, X_n$ are statistically independent, show that the variance of the sum is equal to the sum of the variances.

Exercise 1.12. (Stieltjes integration) **a)** Let $h(x) = u(x)$ and $\mathsf{F}_X(x) = u(x)$ where $u(x)$ is the unit step, *i.e.*, $u(x) = 0$ for $-\infty < x < 0$ and $u(x) = 1$ for $x \geq 0$. Using the definition of the Stieltjes integral in Footnote 22, show that $\int_{-1}^{1} h(x) d\mathsf{F}_X(x)$ does not exist. Hint: Look at the term in the Riemann sum including $x = 0$ and look at the range of choices for $h(x)$ in that interval. Intuitively, it might help initially to view $d\mathsf{F}_X(x)$ as a unit impulse at $x = 0$.

b) Let $h(x) = u(x - a)$ and $\mathsf{F}_X(x) = u(x - b)$ where a and b are in $(-1, +1)$. Show that $\int_{-1}^{1} h(x) d\mathsf{F}_X(x)$ exists if and only if $a \neq b$. Show that the integral has the value 1 for $a < b$ and the value 0 for $a > b$. Argue that this result is still valid in the limit of integration over $(-\infty, \infty)$.

c) Let X and Y be independent discrete rv's, each with a finite set of possible values. Show that $\int_{-\infty}^{\infty} \mathsf{F}_X(z - y) d\mathsf{F}_Y(y)$, defined as a Stieltjes integral, is equal to the distribution of $Z = X + Y$ at each z other than the possible sample values of Z, and is undefined at each sample value of Z. Hint: Express F_X and F_Y as sums of unit steps. Note: This failure of Stieltjes integration is not a serious problem; $\mathsf{F}_Z(z)$ is a step function, and the integral is undefined at its points of discontinuity. We automatically define $\mathsf{F}_Z(z)$ at those step values so that F_Z is a distribution function (*i.e.*, is continuous from the right). This problem does not arise if either X or Y is continuous.

Exercise 1.13. Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of IID continuous rv's with the common probability density function $\mathsf{f}_X(x)$; note that $\Pr\{X{=}\alpha\} = 0$ for all α and that $\Pr\{X_i{=}X_j\} = 0$ for all $i \neq j$. For $n \geq 2$, define X_n as a *record-to-date* of the sequence if $X_n > X_i$ for all $i < n$.

a) Find the probability that X_2 is a record-to-date. Use symmetry to obtain a numerical answer without computation. A one or two line explanation should be adequate.

b) Find the probability that X_n is a record-to-date, as a function of $n \geq 1$. Again use symmetry.

c) Find a simple expression for the expected number of records-to-date that occur over the first m trials for any given integer m. Hint: Use indicator functions. Show that this expected number is infinite in the limit $m \to \infty$.

Exercise 1.14. (Continuation of Exercise 1.13)

a) Let N_1 be the index of the *first* record-to-date in the sequence. Find $\Pr\{N_1 > n\}$ for each $n \geq 2$. Hint: There is a far simpler way to do this than working from part b) in Exercise 1.13.

b) Show that N_1 is a rv.

c) Show that $\mathsf{E}[N_1] = \infty$.

d) Let N_2 be the index of the *second* record-to-date in the sequence. Show that N_2 is a rv. Hint: You need not find the distribution function of N_2 here.

e) Contrast your result in part c) to the result from part c) of Exercise 1.13 saying that the expected number of records-to-date is infinite over an an infinite number of trials. Note:

this should be a shock to your intuition — there is an infinite expected wait for the first of an infinite sequence of occurrences, each of which must eventually occur.

Exercise 1.15. (Another direction from Exercise 1.13) **a)** For any given $n \geq 2$, find the probability that X_n and X_{n+1} are both records-to-date. Hint: The idea in part b) of 1.13 is helpful here, but the result is not.

b) Is the event that X_n is a record-to-date statistically independent of the event that X_{n+1} is a record-to-date?

c) Find the expected number of adjacent pairs of records-to-date over the sequence $X_1, X_2, \ldots$. Hint: A helpful fact here is that $\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}$.

Exercise 1.16. a) Assume that X is a nonnegative discrete rv taking on values $a_1, a_2, \ldots$, and let $Y = h(X)$ for some nonnegative function h. Let $b_i = h(a_i)$, $i \geq 1$ be the i^{th} value taken on by Y. Show that $\mathsf{E}[Y] = \sum_i b_i \mathsf{p}_Y(b_i) = \sum_i h(a_i)\mathsf{p}_X(a_i)$. Find an example where $\mathsf{E}[X]$ exists but $\mathsf{E}[Y] = \infty$.

b) Let X be a nonnegative continuous rv with density $\mathsf{f}_X(x)$ and let $h(x)$ be differentiable, nonnegative, and strictly increasing in x. Let $A(\delta) = \sum_n h(n\delta)[\mathsf{F}(n\delta) - \mathsf{F}(n\delta - \delta)]$, *i.e.*, $A(\delta)$ is the δth order approximation to the Stieltjes integral $\int h(x) d\mathsf{F}(x)$. Show that if $A(1) < \infty$, then $A(2^{-k}) \leq A(2^{k-1}) < \infty$. Show from this that $\int h(x)\, d\mathsf{F}(x)$ converges to a finite value. Note: this is a very special case, but it can be extended to many cases of interest. It seems better to consider these convergence questions as required rather than consider them in general.

Exercise 1.17. a) Consider a positive, integer-valued rv whose distribution function is given at integer values by

$$\mathsf{F}_Y(y) = 1 - \frac{2}{(y+1)(y+2)} \qquad \text{for integer } \ y > 0.$$

Use (1.35) to show that $\mathsf{E}[Y] = 2$. Hint: Note the PMF given in (1.31).

b) Find the PMF of Y and use it to check the value of $\mathsf{E}[Y]$.

c) Let X be another positive, integer-valued rv. Assume its conditional PMF is given by

$$\mathsf{p}_{X|Y}(x|y) = \frac{1}{y} \qquad \text{for } 1 \leq x \leq y.$$

Find $\mathsf{E}[X \mid Y = y]$ and show that $\mathsf{E}[X] = 3/2$. Explore finding $\mathsf{p}_X(x)$ until you are convinced that using the conditional expectation to calculate $\mathsf{E}[X]$ is considerably easier than using $\mathsf{p}_X(x)$.

d) Let Z be another integer-valued rv with the conditional PMF

$$\mathsf{p}_{Z|Y}(z|y) = \frac{1}{y^2} \qquad \text{for } 1 \leq z \leq y^2.$$

Find $\mathsf{E}[Z \mid Y = y]$ for each integer $y \geq 1$ and find $\mathsf{E}[Z]$.

Exercise 1.18. a) Show that, for uncorrelated rv's, the expected value of the product is equal to the product of the expected values (by definition, X and Y are uncorrelated if $\mathsf{E}\left[(X-\overline{X})(Y-\overline{Y})\right] = 0$).

b) Show that if X and Y are uncorrelated, then the variance of $X+Y$ is equal to the variance of X plus the variance of Y.

c) Show that if $X_1, \ldots, X_n$ are uncorrelated, the the variance of the sum is equal to the sum of the variances.

d) Show that independent rv's are uncorrelated.

e) Let X, Y be identically distributed ternary valued random variables with the PMF $\mathsf{p}_X(-1) = \mathsf{p}_X(1) = 1/4$; $\mathsf{p}_X(0) = 1/2$. Find a simple joint probability assignment such that X and Y are uncorrelated but dependent.

f) You have seen that the moment generating function of a sum of independent rv's is equal to the product of the individual moment generating functions. Give an example where this is false if the variables are uncorrelated but dependent.

Exercise 1.19. Suppose X has the Poisson PMF, $\mathsf{p}_X(n) = \lambda^n \exp(-\lambda)/n!$ for $n \geq 0$ and Y has the Poisson PMF, $\mathsf{p}_Y(m) = \mu^n \exp(-\mu)/n!$ for $n \geq 0$. Assume that X and Y are independent. Find the distribution of $Z = X + Y$ and find the conditional distribution of Y conditional on $Z = n$.

Exercise 1.20. a) Suppose X, Y and Z are binary rv's, each taking on the value 0 with probability 1/2 and the value 1 with probability 1/2. Find a simple example in which X, Y, Z are statistically *dependent* but are *pairwise* statistically *independent* (i.e., X, Y are statistically independent, X, Z are statistically independent, and Y, Z are statistically independent). Give $\mathsf{p}_{XYZ}(x,y,z)$ for your example. Hint: In the simplest examle, only 4 of the joint values for x, y, z have positive probabilities.

b) Is pairwise statistical independence enough to ensure that

$$\mathsf{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathsf{E}\left[X_i\right]$$

for a set of rv's $X_1, \ldots, X_n$?

Exercise 1.21. Show that $\mathsf{E}[X]$ is the value of α that minimizes $\mathsf{E}\left[(X-\alpha)^2\right]$.

Exercise 1.22. For each of the following random variables, find the interval (r_-, r_+) over which the moment generating function $g(r)$ exists. Determine in each case whether $g_X(r)$ exists at the end points r_- and r_+. For parts a) and b) you should also find and sketch $g(r)$. For part c), $g(r)$ has no closed form.

a) Let λ, θ, be positive numbers and let X have the density

$$\mathsf{f}_X(x) = \frac{1}{2}\lambda \exp(-\lambda x);\ x \geq 0; \quad \mathsf{f}_X(x) = \frac{1}{2}\theta \exp(\theta x);\ x < 0.$$

b) Let Y be a Gaussian random variable with mean m and variance σ^2.

c) Let Z be a nonnegative random variable with density

$$\mathsf{f}_Z(z) = k(1+z)^{-2}\ \exp(-\lambda z); \quad z \geq 0.$$

where $\lambda > 0$ and $k = [\int_{z\geq 0}(1+z)^2\ \exp(-az)dz]^{-1}$. Hint: Do not try to evaluate $g_Z(r)$. Instead, investigate values of r for which the integral is finite and infinite.

Exercise 1.23. Recall that the MGF of the nonnegative exponential rv with density e^{-x} is $(1-r)^{-1}$ for $r < r_+ = 1$. In other words, $g(r_+)$ does not exist and $\lim_{r\to r_+} g(r) = \infty$, where the limit is over $r < r_+$. In this exercise, you are to assume that X is an arbitrary rv for which $g(r_+)$ does not exist and show that $\lim_{r\to r_+} g(r) = \infty$ where the limit is over $r < r_+$.

a) Explain why

$$\lim_{A\to\infty} \int_0^A e^{xr_+}\, d\mathsf{F}(x) = \infty.$$

b) Show that for any $\epsilon > 0$ and any $A > 0$,

$$g(r_+ - \epsilon) \geq e^{-\epsilon A} \int_0^A e^{xr_+}\, d\mathsf{F}(x).$$

c) Choose $A = 1/\epsilon$ and show that

$$\lim_{\epsilon\to 0} g(r_+ - \epsilon) = \infty.$$

Exercise 1.24. a) Assume that the MGF of the random variable X exists (*i.e.*, is finite) in the interval (r_-, r_+), $r_- < 0 < r_+$, and assume $r_- < r < r_+$ throughout. For any finite constant c, express the moment generating function of $X - c$, *i.e.*, $\mathsf{g}_{(X-c)}(r)$, in terms of $\mathsf{g}_X(r)$ and show that $\mathsf{g}_{(X-c)}(r)$ exists for all r in (r_-, r_+). Explain why $\mathsf{g}''_{(X-c)}(r) \geq 0$.

b) Show that $\mathsf{g}''_{(X-c)}(r) = [\mathsf{g}''_X(r) - 2c\mathsf{g}'_X(r) + c^2\mathsf{g}_X(r)]e^{-rc}$.

c) Use **a)** and **b)** to show that $\mathsf{g}''_X(r)\mathsf{g}_X(r) - [\mathsf{g}'_X(r)]^2 \geq 0$, Let $\gamma_X(r) = \ln \mathsf{g}_X(r)$ and show that $\gamma''_X(r) \geq 0$. Hint: Choose $c = \mathsf{g}'_X(r)/\mathsf{g}_X(r)$.

d) Assume that X is non-deterministic, *i.e.*, that there is no value of α such that $\Pr\{X = \alpha\} = 1$. Show that the inequality sign "$\geq$" may be replaced by "$>$" everywhere in a), b) and c).

Exercise 1.25. A computer system has n users, each with a unique name and password. Due to a software error, the n passwords are randomly permuted internally (i.e. each of the $n!$ possible permutations are equally likely. Only those users lucky enough to have had their passwords unchanged in the permutation are able to continue using the system.

a) What is the probability that a particular user, say user 1, is able to continue using the system?

b) What is the expected number of users able to continue using the system? Hint: Let X_i be a rv with the value 1 if user i can use the system and 0 otherwise.

Exercise 1.26. Suppose the rv X is continuous and has the distribution function $\mathsf{F}_X(x)$. Consider another rv $Y = \mathsf{F}_X(X)$. That is, for each sample point ω such that $X(\omega) = x$, we have $Y(\omega) = \mathsf{F}_X(x)$. Show that Y is uniformly distributed in the interval 0 to 1.

Exercise 1.27. Let Z be an integer valued rv with the PMF $\mathsf{p}_Z(n) = 1/k$ for $0 \leq n \leq k-1$. Find the mean, variance, and moment generating function of Z. Hint: An elegant way to do this is to let U be a uniformly distributed continuous rv over $(0, 1]$ that is independent of Z. Then $U + Z$ is uniform over $(0, k]$. Use the known results about U and $U + Z$ to find the mean, variance, and MGF for Z.

Exercise 1.28. a) Let Y be a nonnegative rv and $y > 0$ be some fixed number. Let A be the event that $Y \geq y$. Show that $y\,\mathbb{I}_A \leq Y$ (*i.e.*, that this inequality is satisfied for every $\omega \in \Omega$).

b) Use your result in part a) to prove the Markov inequality.

Exercise 1.29. a) Show that for any $0 < k < n$

$$\binom{n}{k+1} \leq \binom{n}{k} \frac{n-k}{k}.$$

b) Extend part a) to show that, for all $\ell \leq n - k$,

$$\binom{n}{k+\ell} \leq \binom{n}{k} \left[\frac{n-k}{k}\right]^{\ell}.$$

c) Let $\tilde{p} = k/n$ and $\tilde{q} = 1 - \tilde{p}$. Let S_n be the sum of n binary IID rv's with $\mathsf{p}_X(0) = q$ and $\mathsf{p}_X(1) = p$. Show that for all $\ell \leq n - k$,

$$\mathsf{p}_{S_n}(k+\ell) \leq \mathsf{p}_{S_n}(k) \left(\frac{\tilde{q}p}{\tilde{p}q}\right)^{\ell}.$$

d) For $k/n > p$, show that $\Pr\{S_n \geq kn\} \leq \frac{\tilde{p}q}{\tilde{p}-p}\,\mathsf{p}_{S_n}(k)$.

e) Now let ℓ be fixed and $k = \lceil n\tilde{p} \rceil$ for fixed $\tilde{p}$ such that $1 > \tilde{p} > p$. Argue that as $n \to \infty$,

$$\mathsf{p}_{S_n}(k+\ell) \sim \mathsf{p}_{S_n}(k) \left(\frac{\tilde{q}p}{\tilde{p}q}\right)^{\ell} \qquad \text{and} \quad \Pr\{S_n \geq kn\} \sim \frac{\tilde{p}q}{\tilde{p}-p}\,\mathsf{p}_{S_n}(k).$$

Exercise 1.30. A sequence $\{a_n;\, n \geq 1\}$ of real numbers has the limit 0 if for all $\epsilon > 0$, there is an $m(\epsilon)$ such that $|a_n| \leq \epsilon$ for all $n \geq m(\epsilon)$. Show that the sequences in parts a) and b) below satisfy $\lim_{n\to\infty} a_n = 0$ but the sequence in part c) does not have a limit.

a) $a_n = \frac{1}{\ln(\ln(n+1))}$

b) $a_n = n^{10} \exp(-n)$

c) $a_n = 1$ for $n = 10^{\ell}$ for each positive integer ℓ and $a_n = 0$ otherwise.

d) Show that the definition can be changed (with no change in meaning) by replacing ϵ with either $1/k$ or 2^{-k} for every positive integer k.

Exercise 1.31. Consider the moment generating function of a rv X as consisting of the following two integrals:

$$\mathsf{g}_X(r) = \int_{-\infty}^{0} e^{rx} d\mathsf{F}(x) + \int_{0}^{\infty} e^{rx} d\mathsf{F}(x).$$

In each of the following parts, you are welcome to restrict X to be either discrete or continuous.

a) Show that the first integral always exists (*i.e.*, is finite) for $r \geq 0$ and that the second integral always exists for $r \leq 0$.

b) Show that if the second integral exists for a given $r_1 > 0$, then it also exists for all r in the range $0 \leq r \leq r_1$.

c) Show that if the first integral exists for a given $r_2 < 0$, then it also exists for all r in the range $r_2 \leq r \leq 0$.

d) Show that the range of r over which $\mathsf{g}_X(r)$ exists is an interval from some $r_2 \leq 0$ to some $r_1 \geq 0$ (the interval might or might not include each endpoint, and either or both end point might be 0 or ∞).

e) Find an example where $r_1 = 1$ and the MGF does not exist for $r = 1$. Find another example where $r_1 = 1$ and the MGF does exist for $r = 1$. Hint: Consider $\mathsf{f}_X(x) = e^{-x}$ for $x \geq 0$ and figure out how to modify it to $\mathsf{f}_Y(y)$ so that $\int_0^\infty e^y \mathsf{f}_Y(y)\, dy < \infty$ but $\int_0^\infty e^{y+\epsilon y} \mathsf{f}_Y(y) = \infty$ for all $\epsilon > 0$.

Exercise 1.32. Let $\{X_n;\, n \geq 1\}$ be a sequence of independent but not identically distributed rv's. We say that the weak law of large numbers (WLLN) holds for this sequence if for all $\epsilon > 0$

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{S_n}{n} - \frac{\mathsf{E}\,[S_n]}{n}\right| \geq \epsilon\right\} = 0 \quad \text{where } S_n = X_1 + X_2 + \cdots + X_n. \qquad \text{(WL)}.$$

a) Show that the WLLN holds if there is some constant A such that $\sigma^2_{X_n} \leq A$ for all n.

b) Suppose that $\sigma^2_{X_n} \leq A\, n^{1-\alpha}$ for some $\alpha < 1$ and for all n. Show that the WLLN holds in this case.

Exercise 1.33. Let $\{X_i;\ i \geq 1\}$ be IID binary rv's. Let $\Pr\{X_i = 1\} = \delta$, $\Pr\{X_i = 0\} = 1 - \delta$. Let $S_n = X_1 + \cdots + X_n$. Let m be an arbitrary but fixed positive integer. Think! then evaluate the following and explain your answers:

a) $\lim_{n\to\infty} \sum_{i:n\delta - m \leq i \leq n\delta + m} \Pr\{S_n = i\}$

b) $\lim_{n\to\infty} \sum_{i:0 \leq i \leq n\delta + m} \Pr\{S_n = i\}$

c) $\lim_{n\to\infty} \sum_{i:n(\delta - 1/m) \leq i \leq n(\delta + 1/m)} \Pr\{S_n = i\}$.

Exercise 1.34. Use the Berry-Esseen result, (1.83), to prove the WLLN under the restriction that $\mathsf{E}\left[|X|^3\right]$ exists. Note: This is not intended as a reasonable way to prove the WLLN. Rather, it is to better understand what the convergence result of (1.83) implies. It appears that the CLT, without something extra about convergence, does not establish the WLLN.

Exercise 1.35. (Details in the proof of Theorem 1.5.3)

a) Show that if $X_1, X_2, \ldots$, are IID, then the truncated versions $\breve{X}_1, \breve{X}_2, \ldots$, are also IID.

b) Show that each $\breve{X}_i$ has a finite mean $\mathsf{E}\left[\breve{X}\right]$ and finite variance $\sigma^2_{\breve{X}}$. Show that the variance is upper bounded by the second moment around the original mean $\overline{X}$, *i.e.*, show that $\sigma^2_{\breve{X}} \leq \mathsf{E}\left[\,|\,\breve{X} - \mathsf{E}\left[X\right]\,|^2\right]$.

c) Assume that $\breve{X}_i$ is X_i truncated to $\overline{X} \pm b$. Show that $|\breve{X} - \overline{X}| \leq b$ and that $|\breve{X} - \overline{X}| \leq |X - \overline{X}|$. Use this to show that $\sigma^2_{\breve{X}} \leq b\mathsf{E}\left[|\breve{X} - \overline{X}|\right] \leq 2b\mathsf{E}\left[|X|\right]$.

d) Let $\breve{S}_n = \breve{X}_1 + \cdots + \breve{X}_n$ and show that for any $\epsilon > 0$,

$$\Pr\left\{\left|\frac{\breve{S}_n}{n} - \mathsf{E}\left[\breve{X}\right]\right| \geq \frac{\epsilon}{2}\right\} \leq \frac{8b\mathsf{E}\left[|X|\right]}{n\epsilon^2}.$$

e) Sketch the form of $\mathsf{F}_{\breve{X}-\overline{X}}(x)$ and use this, along with (1.35), to show that for all sufficiently large b, $\left|\,\mathsf{E}\left[\breve{X} - \overline{X}\right]\,\right| \leq \epsilon/2$. Use this to show that

$$\Pr\left\{\left|\frac{\breve{S}_n}{n} - \mathsf{E}\left[X\right]\right| \geq \epsilon\right\} \leq \frac{8b\mathsf{E}\left[|X|\right]}{n\epsilon^2} \qquad \text{for all large enough } b.$$

f) Use the following equation to justify (1.94).

$$\begin{aligned}\Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}\left[X\right]\right| > \epsilon\right\} \;=\; & \Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}\left[X\right]\right| > \epsilon \bigcap S_n = \breve{S}_n\right\} \\ & + \Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}\left[X\right]\right| > \epsilon \bigcap S_n \neq \breve{S}_n\right\}.\end{aligned}$$

Exercise 1.36. Let $\{X_i;\ i \geq 1\}$ be IID rv's with mean 0 and infinite variance. Assume that $\mathsf{E}\left[|X_i|^{1+h}\right] \leq \beta$ for some given h, $0 < h < 1$ and some finite β. Let $S_n = X_1 + \cdots + X_n$.

a) Show that $\Pr\{|X_i| \geq y\} \leq \beta\, y^{-1-h}$

b) Let $\{\breve{X}_i; i \geq 1\}$ be truncated variables $\breve{X}_i = \begin{cases} b & : \ X_i \geq b \\ X_i & : \ -b \leq X_i \leq b \\ -b & : \ X_i \leq -b \end{cases}$

Show that $\mathsf{E}\left[\breve{X}^2\right] \leq \frac{2\beta b^{1-h}}{1-h}$ Hint: For a nonnegative rv Z, $\mathsf{E}\left[X^2\right] \leq \int_0^\infty 2\,z\Pr\{Z \geq z\}\,dz$ (you can establish this, if you wish, by integration by parts).

c) Let $\breve{S}_n = \breve{X}_1 + \ldots + \breve{X}_n$. Show that $\Pr\left\{S_n \neq \breve{S}_n\right\} \leq n\beta b^{-1-h}$

d) Show that $\Pr\left\{\left|\frac{S_n}{n}\right| \geq \epsilon\right. \leq \beta\left[\frac{2b^{1-h}}{(1-h)n\epsilon^2} + \frac{n}{b^{1+h}}\right]$.

e) Optimize your bound with respect to b. How fast does this optimized bound approach 0 with increasing n?

Exercise 1.37. (**MS convergence → convergence in probability**) Assume that $\{Z_n;\, n \geq 1\}$ is a sequence of rv's and α is a number with the property that $\lim_{n\to\infty} \mathsf{E}\left[(Z_n - \alpha)^2\right] = 0$.

a) Let $\epsilon > 0$ be arbitrary and show that for each $n \geq 0$,

$$\Pr\{|Z_n - \alpha| \geq \epsilon\} \leq \frac{\mathsf{E}\left[(Z_n - \alpha)^2\right]}{\epsilon^2}.$$

b) For the ϵ above, let $\delta > 0$ be arbitrary. Show that there is an integer m such that $\mathsf{E}\left[(Z_n - \alpha)^2\right] \leq \epsilon^2\delta$ for all $n \geq m$.

c) Show that this implies convergence in probability.

Exercise 1.38. Let $X_1, X_2 \ldots$, be a sequence of IID rv's each with mean 0 and variance σ^2. Let $S_n = X_1 + \cdots + X_n$ for all n and consider the random variable $S_n/\sigma\sqrt{n} - S_{2n}/\sigma\sqrt{2n}$. Find the limiting distribution function for this sequence of $rv's$ as $n \to \infty$. The point of this exercise is to see clearly that the distribution function of $S_n/\sigma\sqrt{n}$ is converging but that the sequence of rv's is not converging.

Exercise 1.39. A town starts a mosquito control program and the rv Z_n is the number of mosquitos at the end of the nth year ($n = 0, 1, 2, \ldots$). Let X_n be the growth rate of mosquitos in year n; i.e., $Z_n = X_n Z_{n-1}; n \geq 1$. Assume that $\{X_n; n \geq 1\}$ is a sequence of IID rv's with the PMF $\Pr\{X{=}2\} = 1/2$; $\Pr\{X{=}1/2\} = 1/4$; $\Pr\{X{=}1/4\} = 1/4$. Suppose that Z_0, the initial number of mosquitos, is some known constant and assume for simplicity and consistency that Z_n can take on non-integer values.

a) Find $\mathsf{E}[Z_n]$ as a function of n and find $\lim_{n\to\infty} \mathsf{E}[Z_n]$.

b) Let $W_n = \log_2 X_n$. Find $\mathsf{E}[W_n]$ and $\mathsf{E}[\log_2(Z_n/Z_0)]$ as a function of n.

c) There is a constant α such that $\lim_{n\to\infty}(1/n)[\log_2(Z_n/Z_0)] = \alpha$ with probability 1. Find α and explain how this follows from the strong law of large numbers.

d) Using (c), show that $\lim_{n\to\infty} Z_n = \beta$ with probability 1 for some β and evaluate β.

e) Explain carefully how the result in (a) and the result in (d) are possible. What you should learn from this problem is that the expected value of the log of a product of IID rv's might be more significant that the expected value of the product itself.

Exercise 1.40. Use Figure 1.7 to verify (1.55). Hint: Show that $y\Pr\{Y{\geq}y\} \leq \int_{z\geq y} z d\mathsf{F}_Y(z)$ and show that $\lim_{y\to\infty} \int_{z\geq y} z d\mathsf{F}_Y(z) = 0$ if $\mathsf{E}[Y]$ is finite.

Exercise 1.41. Show that $\prod_{m\geq n}(1-1/m)=0$. Hint: Show that

$$\left(1-\frac{1}{m}\right)\Big\} = \exp\left(\ln\left(1-\frac{1}{m}\right)\right)\Big\} \leq \exp\left(-\frac{1}{m}\right).$$

Exercise 1.42. Consider a discrete rv X with the PMF

$$\begin{aligned} \mathsf{p}_X(-1) &= (1-10^{-10})/2, \\ \mathsf{p}_X(1) &= (1-10^{-10})/2, \\ \mathsf{p}_X(10^{12}) &= 10^{-10}. \end{aligned}$$

a) Find the mean and variance of X. Assuming that $\{X_m;\, m\geq 1\}$ is an IID sequence with the distribution of X and that $S_n = X_1+\cdots+X_n$ for each n, find the mean and variance of S_n. (no explanations needed.)

b) Let $n=10^6$ and describe the event $\{S_n \leq 10^6\}$ in words. Find an exact expression for $\Pr\{S_n \leq 10^6 \quad = \mathsf{F}_{S_n}(10^6)$.

c) Find a way to use the union bound to get a simple upper bound and approximation of $1-\mathsf{F}_{S_n}(10^6)$.

d) Sketch the distribution function of S_n for $n=10^6$. You can choose the horizontal axis for your sketch to go from -1 to $+1$ or from -3×10^3 to 3×10^3 or from -10^6 to 10^6 or from 0 to 10^{12}, whichever you think will best describe this distribution function.

e) Now let $n=10^{10}$. Give an exact expression for $\Pr\{S_n \leq 10^{10}$ and show that this can be approximated by e^{-1}. Sketch the distribution function of S_n for $n=10^{10}$, using a horizontal axis going from slightly below 0 to slightly more than 2×10^{12}. Hint: First view S_n as conditioned on an appropriate rv.

d) Can you make a qualitative statement about how the distribution function of a rv X affects the required size of n before the WLLN and the CLT provide much of an indication about S_n.

Chapter 2

POISSON PROCESSES

2.1 Introduction

A Poisson process is a simple and widely used stochastic process for modeling the times at which arrivals enter a system. It is in many ways the continuous-time version of the Bernoulli process. Section 1.4.1 characterized the Bernoulli process by a sequence of IID binary random variables (rv's), $Y_1, Y_2, \ldots$ where $Y_i = 1$ indicates an arrival at increment i and $Y_i = 0$ otherwise. We observed (without any careful proof) that the process could also be characterized by a sequence of geometrically distributed interarrival times.

For the Poisson process, arrivals may occur at arbitrary positive times, and the probability of an arrival at any particular instant is 0. This means that there is no very clean way of describing a Poisson process in terms of the probability of an arrival at any given instant. It is more convenient to define a Poisson process in terms of the sequence of interarrival times, $X_1, X_2, \ldots$, which are defined to be IID. Before doing this, we describe arrival processes in a little more generality.

2.1.1 Arrival processes

An *arrival process* is a sequence of increasing rv's, $0 < S_1 < S_2 < \cdots$, where $S_i < S_{i+1}$ means that $S_{i+1} - S_i$ is a positive rv, *i.e.*, a rv X such that $\mathsf{F}_X(0) = 0$. The rv's $S_1, S_2, \ldots$ are called arrival epochs[1] and represent the successive times at which some random repeating phenomenon occurs. Note that the process starts at time 0 and that multiple arrivals can't occur simultaneously (the phenomenon of bulk arrivals can be handled by the simple extension of associating a positive integer rv to each arrival). We will sometimes permit simultaneous arrivals or arrivals at time 0 as events of zero probability, but these can be ignored. In order to fully specify the process by the sequence $S_1, S_2, \ldots$ of rv's, it is necessary to specify the joint distribution of the subsequences $S_1, \ldots, S_n$ for all $n > 1$.

Although we refer to these processes as arrival processes, they could equally well model departures from a system, or any other sequence of incidents. Although it is quite common,

[1] Epoch is a synonym for time. We use it here to avoid overusing the word time in overlapping ways.

especially in the simulation field, to refer to incidents or arrivals as events, we shall avoid that here. The nth arrival epoch S_n is a rv and $\{S_n \leq t\}$, for example, is an event. This would make it confusing to refer to the nth arrival itself as an event.

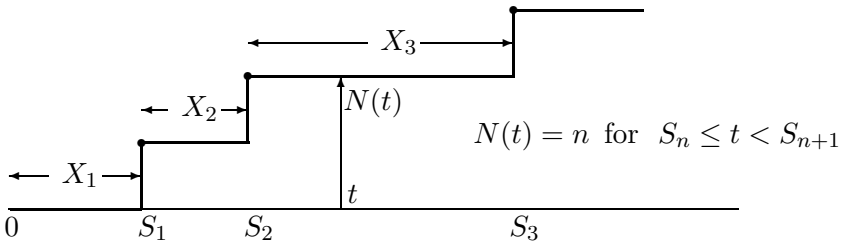


Figure 2.1: A sample function of an arrival process and its arrival epochs $\{S_1, S_2, \dots\}$, its interarrival timess $\{X_1, X_2, \dots\}$, and its counting process $\{N(t);\, t > 0\}$

As illustrated in Figure 2.1, any arrival process $\{S_n;\, n \geq 1\}$ can also be specified by either of two alternative stochastic processes. The first alternative is the sequence of interarrival times, $\{X_i;\, i \geq 1\}$. The X_i here are positive rv's defined in terms of the arrival epochs by $X_1 = S_1$ and $X_i = S_i - S_{i-1}$ for $i > 1$. Similarly, each arrival epoch S_n is specified by $\{X_i;\, i \geq 1\}$ as

$$S_n = \sum_{i=1}^{n} X_i. \tag{2.1}$$

Thus the joint distribution of $X_1, \dots, X_n$ for all $n > 1$ is sufficient (in principle) to specify the arrival process. Since the interarrival times are IID in most cases of interest, it is usually simpler to specify the joint distribution of the X_i than of the S_i.

The second alternative is the counting process $\{N(t);\, t > 0\}$, where for each $t > 0$, the rv $N(t)$ is the aggregate number of arrivals[2] up to and including time t.

The counting process $\{N(t);\, t > 0\}$, illustrated in Figure 2.1, is an uncountably infinite family of rv's where the *counting rv* $N(t)$, for each $t > 0$, is the aggregate number of arrivals in the interval $(0, t]$, *i.e.*, the number of arrivals in the set $\{\tau : 0 < \tau \leq t\}$. By convention, a parenthesis is used to indicate that the corresponding end point of an interval is not contained in the interval and a bracket is used if it is. Thus, as another example, $(a, b) = \{\tau : a < \tau < b\}$. The rv $N(0)$ is defined to be 0 with probability 1, which means that we are considering only arrivals at strictly positive times.

The counting process $\{N(t);\, t > 0\}$ for any arrival process has the property that $N(\tau) \geq N(t)$ for all $\tau \geq t > 0$ (i.e., $N(\tau) - N(t)$ is a nonnegative random variable for $\tau \geq t > 0$).

For any given integer $n \geq 1$ and time $t > 0$, the nth arrival epoch, S_n, and the counting random variable, $N(t)$, are related by

$$\{S_n \leq t\} = \{N(t) \geq n\}. \tag{2.2}$$

[2]Thus, for the Bernoulli process with an increment size of 1, $N(n)$ is the rv denoted as S_n in Section 1.4.1

To see this, note that $\{S_n \leq t\}$ is the event that the nth arrival occurs at some epoch $\tau \leq t$. This event implies that $N(\tau) = n$, and thus that $\{N(t) \geq n\}$. Similarly, $\{N(t) = m\}$ for some $m \geq n$ implies $\{S_m \leq t\}$, and thus that $\{S_n \leq t\}$. This equation is essentially obvious from Figure 2.1, but is one of those peculiar obvious things that is often difficult to see. An alternate form, which is occasionally more transparent, comes from taking the complement of both sides of (2.2), getting

$$\{S_n > t\} = \{N(t) < n\}. \tag{2.3}$$

For example, the event $\{S_1 > t\}$ means that the first arrival occurs after t, which means $\{N(t) < 1\}$ (*i.e.*, $\{N(t) = 0\}$). These relations will be used constantly in going back and forth between arrival epochs and counting rv's. In principle, (2.2) or (2.3) can be used to specify joint CDF of arrival epochs in terms of joint CDF's of counting variables and vice versa, so either characterization can be used to specify an arrival process.

In summary, then, an arrival process can be specified by the joint distributions of the arrival epochs, or of the interarrival times, or of the counting rv's. In principle, specifying any one of these specifies the others also.[3]

2.2 Definition and properties of a Poisson process

The nicest, and in many ways most important, arrival processes are those for which the interarrival times are IID. These arrival processes are called *renewal processes* and form the subject of Chapter 5.

Poisson processes are the nicest, and in many ways most important, class of renewal processes. They are characterized by having exponentially distributed interarrival times. We will soon see why these exponentially distributed interarrival times simplify these processes and make it possible to look at them in so many simple ways. We start by being explicit about the definitions of renewal and Poisson processes.

Definition 2.2.1. *A renewal process is an arrival process for which the sequence of interarrival times is a sequence of positive IID rv's.*

Definition 2.2.2. *A Poisson process is a renewal process in which the interarrival times have an exponential CDF; i.e., for some real $\lambda > 0$, each X_i has the density*[4] $\mathsf{f}_X(x) = \lambda \exp(-\lambda x)$ *for $x \geq 0$.*

The parameter λ is called the rate of the process. We shall see later that for any interval of size t, λt is the expected number of arrivals in that interval, giving some interpretation to denoting λ as a rate.

[3] By definition, a stochastic process is a collection of rv's, so one might ask whether an arrival process (as a stochastic process) is 'really' the arrival-epoch process $\{S_n;\, n \geq 1$ or the interarrival process $X_i;\, i > 0\}$ or the counting process $\{N(t);\, t > 0\}$. The arrival-epoch process comes to grips with the actual arrivals, the interarrival process is often the simplest, and the counting process 'looks' most like a stochastic process in time since $N(t)$ is a rv for each $t \geq 0$. It seems preferable, since the descriptions are so clearly equivalent, to view arrival processes in terms of whichever description is most convenient.

[4] With this density, $\Pr\{X_i > 0\} = 1$, so that we can regard X_i as a positive random variable. Since events of probability zero can be ignored, the density $\lambda \exp(-\lambda x)$ for $x \geq 0$ and zero for $x < 0$ is effectively the same as the density $\lambda \exp(-\lambda x)$ for $x > 0$ and zero for $x \leq 0$.

2.2.1 Memoryless property

What makes the Poisson process unique among renewal processes is the memoryless property of the exponential distribution.

Definition 2.2.3. Memoryless random variables: *A rv X possesses the memoryless property if X is a positive rv (*i.e.*, $\Pr\{X > 0\} = 1$) for which*

$$\Pr\{X > t + x\} = \Pr\{X > x\} \Pr\{X > t\} \qquad \text{for all } x, t \geq 0. \tag{2.4}$$

Note that (2.4) is a statement about the complementary CDF of X. There is no intimation that the *event* $\{X > t+x\}$ in the equation has any particular relation to the events $\{X > t\}$ or $\{X > x\}$.

For an exponential rv X of rate $\lambda > 0$, $\Pr\{X > x\} = e^{-\lambda x}$ for $x \geq 0$. This satisfies (2.4) for all $x \geq 0,\ t \geq 0$, so X is memoryless. Conversely, an arbitrary rv X is memoryless only if it is exponential. To see this, let $h(x) = \ln[\Pr\{X > x\}]$ and observe that since $\Pr\{X > x\}$ is non-increasing in x, $h(x)$ is also. In addition, (2.4) says that $h(t + x) = h(x) + h(t)$ for all $x, t \geq 0$. These two statements (see Exercise 2.6) imply that $h(x)$ must be linear in x, and $\Pr\{X > x\}$ must be exponential in x.

Since a memoryless rv X must be exponential, there must be some $\lambda > 0$ such that $\Pr\{X > t\} = e^{-\lambda t} > 0$ for all $t \geq 0$. Since $\Pr\{X > t\} > 0$, we can rewrite (2.4) as

$$\Pr\{X > t + x \mid X > t\} = \Pr\{X > x\}\,. \tag{2.5}$$

If X is interpreted as the waiting time until some given arrival, then (2.5) states that, given that the arrival has not occured by time t, the distribution of the remaining waiting time (given by x on the left side of (2.5)) is the same as the original waiting time distribution (given on the right side of (2.5)), *i.e.*, the remaining waiting time has no 'memory' of previous waiting.

Example 2.2.1. Suppose X is the waiting time, starting at time 0, for a bus to arrive, and suppose X is memoryless. After waiting from 0 to t, the distribution of the remaining waiting time from t is the same as the original distribution starting from 0. The still waiting customer is, in a sense, no better off at time t than at time 0. On the other hand, if the bus is known to arrive regularly every 16 minutes, then it will certainly arrive within a minute for a person who has already waited 15 minutes. Thus regular arrivals are not memoryless. The opposite situation is also possible. If the bus frequently breaks down, then a 15 minute wait can indicate that the remaining wait is probably very long, so again X is not memoryless. We study these non-memoryless situations when we study renewal processes in Chapter 5.

Although memoryless distributions must be exponential, it can be seen that if the definition of memoryless is restricted to integer times, then the geometric distribution becomes memoryless, and it can be seen as before that this is the only memoryless integer-time distribution. In this respect, the Bernoulli process (which has geometric interarrival times)

is like a discrete-time version of the Poisson process (which has exponential interarrival times).

We now use the memoryless property of exponential rv's to find the distribution of the first arrival in a Poisson process after an arbitrary given time $t > 0$. We not only find this distribution, but also show that this first arrival after t is independent of all arrivals up to and including t. More precisely, we prove the following theorem.

Theorem 2.2.1. *For a Poisson process of rate λ, and any given $t > 0$, the length of the interval from t until the first arrival after t is a nonnegative rv Z with the CDF $1-\exp[-\lambda z]$ for $z \geq 0$. This rv is independent of both $N(t)$ and of the $N(t)$ arrival epochs before time t. It is also independent of the set of rv's $\{N(\tau); \tau \leq t\}$.*

The basic idea behind this theorem is that Z, conditional on the epoch τ of the last arrival before t, is simply the remaining time until the next arrival. Since the interarrival time starting at τ is exponential and thus memoryless, Z is independent of $\tau \leq t$, and of all earlier arrivals. The following proof carries this idea out in detail.

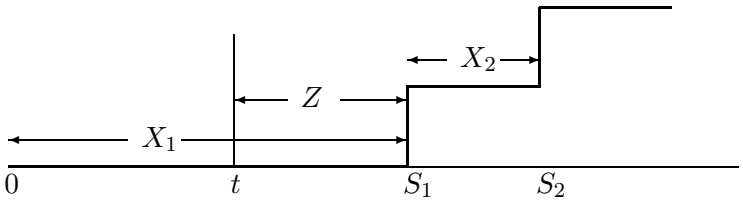


Figure 2.2: For arbitrary fixed $t > 0$, consider the event $\{N(t) = 0\}$. Conditional on this event, Z is the distance from t to S_1; i.e., $Z = X_1 - t$.

Proof: Let Z be the interval from t until the first arrival after t. We first condition on $N(t) = 0$ (see Figure 2.2). Given $N(t) = 0$, we see that $X_1 > t$ and $Z = X_1 - t$. Thus,

$$\begin{aligned} \Pr\{Z > z \mid N(t){=}0\} &= \Pr\{X_1 > z + t \mid N(t){=}0\} \\ &= \Pr\{X_1 > z + t \mid X_1 > t\} \qquad (2.6) \\ &= \Pr\{X_1 > z\} \;=\; e^{-\lambda z}. \qquad (2.7) \end{aligned}$$

In (2.6), we used the fact that $\{N(t) = 0\} = \{X_1 > t\}$, which is clear from Figure 2.1 (and also from (2.3)). In (2.7) we used the memoryless condition in (2.5) and the fact that X_1 is exponential.

Next consider the conditions that $N(t) = n$ (for arbitrary $n > 1$) and $S_n = \tau$ (for arbitrary $\tau \leq t$). The argument here is basically the same as that with $N(t) = 0$, with a few extra details (see Figure 2.3).

Conditional on $N(t) = n$ and $S_n = \tau$, the first arrival after t is the first arrival after the arrival at S_n, *i.e.*, $Z = z$ corresponds to $X_{n+1} = z + (t - \tau)$.

$$\begin{aligned} \Pr\{Z > z \mid N(t){=}n, S_n{=}\tau\} &= \Pr\{X_{n+1} > z{+}t{-}\tau \mid N(t){=}n,\, S_n{=}\tau\} \qquad (2.8) \\ &= \Pr\{X_{n+1} > z{+}t{-}\tau \mid X_{n+1}{>}t{-}\tau,\, S_n{=}\tau\} \qquad (2.9) \\ &= \Pr\{X_{n+1} > z{+}t{-}\tau \mid X_{n+1}{>}t{-}\tau\} \qquad (2.10) \\ &= \Pr\{X_{n+1} > z\} \;=\; e^{-\lambda z}, \qquad (2.11) \end{aligned}$$

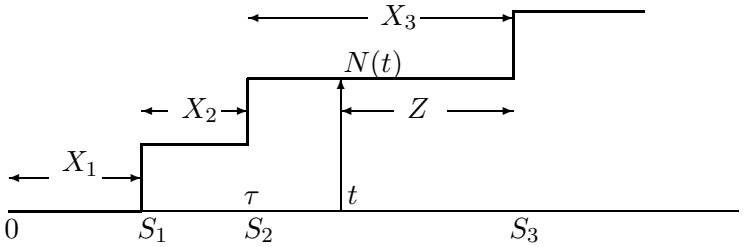


Figure 2.3: Given $N(t) = 2$, and $S_2 = \tau$, X_3 is equal to $Z + (t - \tau)$. Also, the event $\{N(t)=2,\ S_2=\tau\}$ is the same as the event $\{S_2=\tau,\ X_3>t-\tau\}$.

where (2.9) follows because, given $S_n = \tau \le t$, we have $\{N(t) = n\} = \{X_{n+1} > t - \tau\}$ (see Figure 2.3). Eq. (2.10) follows because X_{n+1} is independent of S_n. Eq. (2.11) follows from the memoryless condition in (2.5) and the fact that X_{n+1} is exponential.

The same argument applies if, in (2.8), we condition not only on S_n but also on $S_1, \ldots, S_{n-1}$. Since this is equivalent to conditioning on $N(\tau)$ for all τ in $(0, t]$, we have

$$\Pr\{Z > z \mid \{N(\tau), 0 < \tau \le t\}\} = \exp(-\lambda z). \tag{2.12}$$

□

Next consider the sequence of interarrival times starting from a given time t. The first of these is taken to be the rv Z (now renamed Z_1) of the time from t to the first arrival epoch after t. For $m \ge 2$, let Z_m be the time from the $(m-1)$st arrival after t to the mth arrival epoch after t. Thus, given $N(t) = n$ and $S_n = \tau$, we see that $Z_m = X_{m+n}$ for $m \ge 2$. It follows that, conditional on $N(t) = n$ and $S_n = \tau$, $Z_1, Z_2, \ldots$, are IID exponentially distributed rv's. (see Exercise 2.8). Since this is independent of $N(t)$ and S_n, it follows that that $Z_1, Z_2, \ldots$ are unconditionally IID and also independent of $N(t)$ and S_n. It should also be clear that $Z_1, Z_2, \ldots$ are independent of $\{N(\tau); 0 < \tau \le t\}$.

The above argument shows that the portion of a Poisson process starting at an arbitrary time $t > 0$ is a probabilistic replica of the process starting at 0; that is, the time until the first arrival after t is an exponentially distributed rv with parameter λ, and all subsequent interarrival times are independent of this first arrival epoch and of each other, and all have the same exponential distribution.

Definition 2.2.4. *A counting process $\{N(t);\ t > 0\}$ has the stationary increment property if $N(t') - N(t)$ has the same CDF as $N(t' - t)$ for every $t' > t > 0$.*

Let us define $\widetilde{N}(t, t') = N(t') - N(t)$ as the number of arrivals in the interval $(t, t']$ for any given $t' \ge t$. We have just shown that for a Poisson process, the rv $\widetilde{N}(t, t')$ has the same distribution as $N(t' - t)$, which means that a Poisson process has the stationary increment property. Thus, the distribution of the number of arrivals in an interval depends on the size of the interval but not on its starting point.

Definition 2.2.5. *A counting process $\{N(t);\ t > 0\}$ has the independent increment property if, for every integer $k > 0$, and every k-tuple of times $0 < t_1 < t_2 < \cdots < t_k$, the k-tuple of rv's $N(t_1)$, $\widetilde{N}(t_1, t_2), \ldots, \widetilde{N}(t_{k-1}, t_k)$ are statistically independent.*

For the Poisson process, Theorem 2.2.1 says that for any t, the time Z_1 until the next arrival after t is independent of $N(\tau)$ for all $\tau \leq t$. Letting $t_1 < t_2 < \cdots t_{k-1} < t$, this means that Z_1 is independent of $N(t_1), \widetilde{N}(t_1, t_2), \ldots, \widetilde{N}(t_{k-1}, t)$. We have also seen that the subsequent interarrival times after Z_1, and thus $\widetilde{N}(t, t')$ are independent of $N(t_1), \widetilde{N}(t_1, t_2), \ldots, \widetilde{N}(t_{k-1}, t)$. Renaming t as t_k and t' as t_{k+1}, we see that $\widetilde{N}(t_k, t_{k+1})$ is independent of $N(t_1), \widetilde{N}(t_1, t_2), \ldots, \widetilde{N}(t_{k-1}, t_k)$. Since this is true for all k, the Poisson process has the independent increment property. In summary, we have proved the following:

Theorem 2.2.2. *Poisson processes have both the stationary increment and independent increment properties.*

Note that if we look only at integer times, then the Bernoulli process also has the stationary and independent increment properties.

2.2.2 Probability density of S_n and joint density of $S_1, \ldots, S_n$

Recall from (2.1) that, for a Poisson process, S_n is the sum of n IID rv's, each with the density function $\mathsf{f}_X(x) = \lambda \exp(-\lambda x)$, $x \geq 0$. Also recall that the density of the sum of two independent rv's can be found by convolving their densities, and thus the density of S_2 can be found by convolving $\mathsf{f}_X(x)$ with itself, S_3 by convolving the density of S_2 with $\mathsf{f}_X(x)$, and so forth. The result, for $t \geq 0$, is called the *Erlang density*,[5]

$$\mathsf{f}_{S_n}(t) = \frac{\lambda^n t^{n-1} \exp(-\lambda t)}{(n-1)!}. \tag{2.13}$$

We can understand this density (and other related matters) better by viewing the above mechanical derivation in a slightly different way involving the joint density of $S_1, \ldots, S_n$. For $n = 2$, the joint density of X_1 and S_2 (or equivalently, S_1 and S_2) is given by

$$\mathsf{f}_{X_1 S_2}(x_1, s_2) = \mathsf{f}_{X_1}(x_1)\mathsf{f}_{S_2|X_1}(s_2|x_1) = \lambda e^{-\lambda x_1}\, \lambda e^{-\lambda(s_2 - x_1)} \qquad \text{for } 0 \leq x_1 \leq s_2,$$

where, since $S_2 = X_1 + X_2$, we have used the fact that the density of S_2 conditional on X_1 is just the exponential interarrival density evaluated at $S_2 - X_1$. Thus,

$$\mathsf{f}_{X_1 S_2}(x_1, s_2) = \lambda^2 \exp(-\lambda s_2) \quad \text{for } 0 \leq x_1 \leq s_2. \tag{2.14}$$

This says that the joint density does not contain x_1, except for the constraint $0 \leq x_1 \leq s_2$. Thus, for fixed s_2, the joint density, and thus the conditional density of X_1 given $S_2 = s_2$ is uniform over $0 \leq x_1 \leq s_2$. The integration over x_1 in the convolution equation is then simply multiplication by the interval size s_2, yielding the marginal PDF $\mathsf{f}_{S_2}(s_2) = \lambda^2 s_2 \exp(-\lambda s_2)$, in agreement with (2.13) for $n = 2$.

The following theorem shows that this same curious behavior exhibits itself for the sum of an arbitrary number n of IID exponential rv's.

[5] Another (somewhat rarely used) name for the Erlang density is the *gamma density*.

Theorem 2.2.3. *Let $X_1, X_2, \ldots$ be IID rv's with the density $\mathsf{f}_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. Let $S_n = X_1 + \cdots + X_n$ for each $n \geq 1$. Then for each $n \geq 2$*

$$\mathsf{f}_{S_1 \cdots S_n}(s_1, \ldots, s_n) = \lambda^n \exp(-\lambda s_n) \qquad \text{for } 0 \leq s_1 \leq s_2 \cdots \leq s_n. \tag{2.15}$$

Proof: Replacing X_1 with S_1 in (2.14), we see the theorem holds for $n = 2$. This serves as the basis for the following inductive proof. Assume (2.15) holds for given n. Then

$$\begin{aligned} \mathsf{f}_{S_1 \cdots S_{n+1}}(s_1, \ldots, s_{n+1}) &= \mathsf{f}_{S_1 \cdots S_n}(s_1, \ldots, s_n) \mathsf{f}_{S_{n+1}|S_1 \cdots S_n}(s_{n+1}|s_1, \ldots, s_n) \\ &= \lambda^n \exp(-\lambda s_n)\, \mathsf{f}_{S_{n+1}|S_1 \cdots S_n}(s_{n+1}|s_1, \ldots, s_n), \end{aligned} \tag{2.16}$$

where we used (2.15) for the given n. Now $S_{n+1} = S_n + X_{n+1}$. Since X_{n+1} is independent of $S_1, \ldots, S_n$,

$$\mathsf{f}_{S_{n+1}|S_1 \cdots S_n}(s_{n+1}|s_1, \ldots, s_n) = \lambda \exp\big(-\lambda(s_{n+1} - s_n)\big).$$

Substituting this into (2.16) yields (2.15) □

The interpretation here is the same as with S_2. The joint density does not contain any arrival time other than s_n, except for the ordering constraint $0 \leq s_1 \leq s_2 \leq \cdots \leq s_n$, and thus this joint density is constant over all choices of arrival times satisfying the ordering constraint for a fixed s_n. Mechanically integrating this over s_1, then s_2, etc. we get the Erlang formula (2.13). The Erlang density then is the joint density in (2.15) times the volume $s_n^{n-1}/(n-1)!$ of the region of $s_1, \ldots, s_{n-1}$ satisfying $0 < s_1 < \cdots < s_n$. This will be discussed further later.

Note that (2.15), for all n, specifies the joint distribution for all arrival epochs, and thus fully specifies a Poisson process.

2.2.3 The PMF for $N(t)$

The Poisson counting process, $\{N(t);\, t > 0\}$ consists of a nonnegative integer rv $N(t)$ for each $t > 0$. In this section, we show that the PMF for this rv is the well-known Poisson PMF, as stated in the following theorem. We give two proofs for the theorem, each providing its own type of understanding and each showing the close relationship between $\{N(t) = n\}$ and $\{S_n = t\}$.

Theorem 2.2.4. *For a Poisson process of rate λ, and for any $t > 0$, the PMF for $N(t)$,* i.e.*, the number of arrivals in $(0, t]$, is given by the Poisson PMF,*

$$\mathsf{p}_{N(t)}(n) = \frac{(\lambda t)^n \exp(-\lambda t)}{n!}. \tag{2.17}$$

Proof 1: This proof, for given n and t, is based on two ways of calculating $\Pr\{t < S_{n+1} \leq t + \delta\}$ for some vanishingly small δ. The first way is based on the already known density of S_{n+1} and gives

$$\Pr\{t < S_{n+1} \leq t + \delta\} \;=\; \int_t^{t+\delta} \mathsf{f}_{S_{n+1}}(\tau)\, d\tau \;=\; \mathsf{f}_{S_{n+1}}(t)\,(\delta + \mathrm{o}(\delta)).$$

The term $o(\delta)$ is used to describe a function of δ that goes to 0 faster than δ as $\delta \to 0$. More precisely, a function $g(\delta)$ is said to be of order $o(\delta)$ if $\lim_{\delta \to 0} \frac{g(\delta)}{\delta} = 0$. Thus $\Pr\{t < S_{n+1} \leq t + \delta\} = \mathsf{f}_{S_{n+1}}(t)(\delta + o(\delta))$ is simply a consequence of the fact that S_{n+1} has a continuous probability density in the interval $[t, t + \delta]$.

The second way to calculate $\Pr\{t < S_{n+1} \leq t + \delta\}$ is to first observe that the probability of more than 1 arrival in $(t, t+\delta]$ is $o(\delta)$. Ignoring this possibility, $\{t < S_{n+1} \leq t+\delta\}$ occurs if exactly n arrivals are in the interval $(0, t]$ and one arrival occurs in $(t, t + \delta]$. Because of the independent increment property, this is an event of probability $\mathsf{p}_{N(t)}(n)(\lambda\delta + o(\delta))$. Thus

$$\mathsf{p}_{N(t)}(n)(\lambda\delta + o(\delta)) + o(\delta) = \mathsf{f}_{S_{n+1}}(t)(\delta + o(\delta)).$$

Dividing by δ and taking the limit $\delta \to 0$, we get

$$\lambda \mathsf{p}_{N(t)}(n) = \mathsf{f}_{S_{n+1}}(t).$$

Using the density for f_{S_n} in (2.13), we get (2.17). □

Proof 2: The approach here is to use the fundamental relation that $\{N(t) \geq n\} = \{S_n \leq t\}$. Taking the probabilities of these events,

$$\sum_{i=n}^{\infty} \mathsf{p}_{N(t)}(i) = \int_0^t \mathsf{f}_{S_n}(\tau)\, d\tau \qquad \text{for all } n \geq 1 \text{ and } t > 0.$$

The term on the right above is the CDF of S_n and the term on the left is the complementary CDF of $N(t)$. The complementary CDF and the PMF of $N(t)$ uniquely specify each other, so the theorem is equivalent to showing that

$$\sum\nolimits_{i=n}^{\infty} \frac{(\lambda t)^i \exp(-\lambda t)}{i!} = \int_0^t \mathsf{f}_{S_n}(\tau)\, d\tau. \tag{2.18}$$

If we take the derivative with respect to t of each side of (2.18), we find that almost magically each term except the first on the left cancels out, leaving us with

$$\frac{\lambda^n t^{n-1} \exp(-\lambda t)}{(n-1)!} = \mathsf{f}_{S_n}(t).$$

Thus the derivative with respect to t of each side of (2.18) is equal to the derivative of the other for all $n \geq 1$ and $t > 0$. The two sides of (2.18) are also equal in the limit $t \to \infty$, so it follows that (2.18) is satisfied everywhere, completing the proof. □

2.2.4 Alternate definitions of Poisson processes

Definition 2 of a Poisson process: *A Poisson counting process $\{N(t);\, t > 0\}$ is a counting process that satisfies (2.17) (*i.e.*, has the Poisson PMF) and has the independent and stationary increment properties.*

We have seen that the properties in Definition 2 are satisfied starting with Definition 1 (using IID exponential interarrival times), so Definition 1 implies Definition 2. Exercise

2.4 shows that IID exponential interarrival times are implied by Definition 2, so the two definitions are equivalent.

It may be somewhat surprising at first to realize that a counting process that has the Poisson PMF at each t is not necessarily a Poisson process, and that the independent and stationary increment properties are also necessary. One way to see this is to recall that the Poisson PMF for all t in a counting process is equivalent to the Erlang density for the successive arrival epochs. Specifying the probability density for $S_1, S_2, \ldots$ as Erlang specifies the *marginal* densities of $S_1, S_2, \ldots$ but need not specify the *joint* densities of these rv's. Figure 2.4 illustrates this in terms of the joint density of S_1, S_2, given as

$$\mathsf{f}_{S_1S_2}(s_1, s_2) = \lambda^2 \exp(-\lambda s_2) \qquad \text{for } 0 \le s_1 \le s_2$$

and 0 elsewhere. The figure illustrates how the joint density can be changed without changing the marginals.

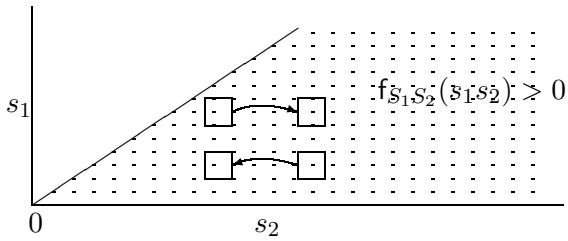


Figure 2.4: The joint density of S_1, S_2 is non-zero in the region shown. It can be changed, while holding the marginals constant, by reducing the joint density by ϵ in the upper left and lower right squares above and increasing it by ϵ in the upper right and lower left squares.

There is a similar effect with the Bernoulli process in that a discrete counting process for which the number of arrivals from 0 to t, for each integer t, is a binomial rv, but the process is not Bernoulli. This is explored in Exercise 2.5.

The next definition of a Poisson process is based on its incremental properties. Consider the number of arrivals in some very small interval $(t, t+\delta]$. Since $\widetilde{N}(t, t+\delta)$ has the same distribution as $N(\delta)$, we can use (2.17) to get

$$\begin{aligned}
\Pr\left\{\widetilde{N}(t,t+\delta)=0\right\} &= e^{-\lambda\delta} \approx 1-\lambda\delta+\mathrm{o}(\delta) \\
\Pr\left\{\widetilde{N}(t,t+\delta)=1\right\} &= \lambda\delta e^{-\lambda\delta} \approx \lambda\delta+\mathrm{o}(\delta) \\
\Pr\left\{\widetilde{N}(t,t+\delta)\ge 2\right\} &\approx \mathrm{o}(\delta). \qquad (2.19)
\end{aligned}$$

Definition 3 of a Poisson process: *A Poisson counting process is a counting process that satisfies (2.19) and has the stationary and independent increment properties.*

We have seen that Definition 1 implies Definition 3. The essence of the argument the other way is that for any interarrival interval X, $\mathsf{F}_X(x+\delta) - \mathsf{F}_X(x)$ is the probability of an arrival

in an appropriate infinitesimal interval of width δ, which by (2.19) is $\lambda\delta + o(\delta)$. Turning this into a differential equation (see Exercise 2.7), we get the desired exponential interarrival intervals. Definition 3 has an intuitive appeal, since it is based on the idea of independent arrivals during arbitrary disjoint intervals. It has the disadvantage that one must do a considerable amount of work to be sure that these conditions are mutually consistent, and probably the easiest way is to start with Definition 1 and derive these properties. Showing that there is a unique process that satisfies the conditions of Definition 3 is even harder, but is not necessary at this point, since all we need is the use of these properties. Section 2.2.5 will illustrate better how to use this definition (or more precisely, how to use (2.19)).

What (2.19) accomplishes in Definition 3, beyond the assumption of independent and stationary increments, is the prevention of bulk arrivals. For example, consider a counting process in which arrivals always occur in pairs, and the intervals between successive pairs are IID and exponentially distributed with parameter λ (see Figure 2.5). For this process, $\Pr\{\widetilde{N}(t, t+\delta)=1\} = 0$, and $\Pr\{\widetilde{N}(t, t+\delta)=2\} = \lambda\delta + o(\delta)$, thus violating (2.19). This process has stationary and independent increments, however, since the process formed by viewing a pair of arrivals as a single incident is a Poisson process.

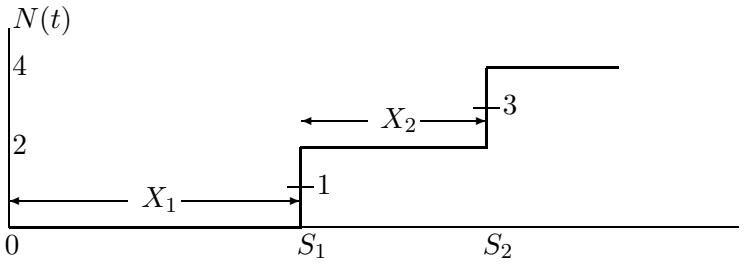


Figure 2.5: A counting process modeling bulk arrivals. X_1 is the time until the first pair of arrivals and X_2 is the interval between the first and second pair of arrivals.

2.2.5 The Poisson process as a limit of shrinking Bernoulli processes

The intuition of Definition 3 can be achieved in a less abstract way by starting with the Bernoulli process, which has the properties of Definition 3 in a discrete-time sense. We then go to an appropriate limit of a sequence of these processes, and find that this sequence of Bernoulli processes converges in some sense to the Poisson process.

Recall that a Bernoulli process is an IID sequence, $Y_1, Y_2, \ldots$ of binary random variables for which $\mathsf{p}_Y(1) = p$ and $\mathsf{p}_Y(0) = 1 - p$. We can visualize $Y_i = 1$ as an *arrival* at time i and $Y_i = 0$ as no arrival, but we can also 'shrink' the time scale of the process so that for some integer $j > 0$, Y_i is an arrival or no arrival at time $i2^{-j}$. We consider a sequence indexed by j of such shrinking Bernoulli processes, and in order to keep the arrival rate constant, we let $p = \lambda 2^{-j}$ for the jth process. Thus for each unit increase in j, the Bernoulli process shrinks by replacing each slot with two slots, each with half the previous arrival probability. The expected number of arrivals per unit time is then λ, matching the Poisson process that

we are approximating.

If we look at this jth process relative to Definition 3 of a Poisson process, we see that for these regularly spaced increments of size $\delta = 2^{-j}$, the probability of one arrival in an increment is $\lambda\delta$ and that of no arrival is $1 - \lambda\delta$, and thus (2.19) is satisfied, and in fact the $o(\delta)$ terms are exactly zero. For arbitrary sized increments, it is clear that disjoint increments have essentially independent arrivals. The increments are not quite stationary, since, for example, an increment of size 2^{-j-1} might contain a time that is a multiple of 2^{-j} or might not, depending on its placement. However, for any fixed increment of size δ, the number of multiples of 2^{-j} (*i.e.*, the number of possible arrival points) is either $\lfloor \delta 2^j \rfloor$ or $1 + \lfloor \delta 2^j \rfloor$. Thus in the limit $j \to \infty$, the increments are both stationary and independent.

For each j, the jth Bernoulli process has an associated Bernoulli counting process $N_j(t) = \sum_{i=1}^{\lfloor t2^j \rfloor} Y_i$. This is the number of arrivals up to time t and is a discrete rv with the binomial PMF. That is, $\mathsf{p}_{N_j(t)}(n) = \binom{\lfloor t2^j \rfloor}{n} p^n (1-p)^{\lfloor t2^j \rfloor - n}$ where $p = \lambda 2^{-j}$. We now show that this PMF approaches the Poisson PMF as j increases.[6]

Theorem 2.2.5 (Poisson's theorem). *Consider the sequence of shrinking Bernoulli processes with arrival probability $\lambda 2^{-j}$ and time-slot size 2^{-j}. Then for every fixed time $t > 0$ and fixed number of arrivals n, the counting PMF $\mathsf{p}_{N_j(t)}(n)$ approaches the Poisson PMF (of the same λ) with increasing j,* i.e.*,*

$$\lim_{j\to\infty} \mathsf{p}_{N_j(t)}(n) = \mathsf{p}_{N(t)}(n). \tag{2.20}$$

Proof: We first rewrite the binomial PMF, for $\lfloor t2^j \rfloor$ variables with $p = \lambda 2^{-j}$ as

$$\begin{aligned}
\lim_{j\to\infty} \mathsf{p}_{N_j(t)}(n) &= \lim_{j\to\infty} \binom{\lfloor t2^j \rfloor}{n} \left(\frac{\lambda 2^{-j}}{1-\lambda 2^{-j}}\right)^n \exp[\lfloor t2^j \rfloor \ln(1-\lambda 2^{-j})] \\
&= \lim_{j\to\infty} \binom{\lfloor t2^j \rfloor}{n} \left(\frac{\lambda 2^{-j}}{1-\lambda 2^{-j}}\right)^n \exp(-\lambda t) && (2.21) \\
&= \lim_{j\to\infty} \frac{\lfloor t2^j \rfloor \cdot \lfloor t2^j - 1 \rfloor \cdots \lfloor t2^j - n + 1 \rfloor}{n!} \left(\frac{\lambda 2^{-j}}{1-\lambda 2^{-j}}\right)^n \exp(-\lambda t) && (2.22) \\
&= \frac{(\lambda t)^n \exp(-\lambda t)}{n!}. && (2.23)
\end{aligned}$$

We used $\ln(1 - \lambda 2^{-j}) = -\lambda 2^{-j} + o(2^{-j})$ in (2.21) and expanded the combinatorial term in (2.22). In (2.23), we recognized that $\lim_{j\to\infty} \lfloor t2^j - i \rfloor \left(\frac{\lambda 2^{-j}}{1-\lambda 2^{-j}}\right) = \lambda t$ for $0 \le i \le n-1$. □

Since the binomial PMF (scaled as above) has the Poisson PMF as a limit for each n, the CDF of $N_j(t)$ also converges to the Poisson CDF for each t. In other words, for each $t > 0$, the counting random variables $N_j(t)$ of the Bernoulli processes converge in distribution to $N(t)$ of the Poisson process.

[6] This limiting result for the binomial distribution is very different from the asymptotic results in Chapter 1 for the binomial. Here the parameter p of the binomial is shrinking with increasing j, whereas there, p is constant while the number of variables is increasing.

This does not say that the Bernoulli *counting processes* converge to the Poisson counting process in any meaningful sense, since the joint distributions are also of concern. The following corollary treats this.

Corollary 2.2.1. *For any finite integer $k > 0$, let $0 < t_1 < t_2 < \cdots < t_k$ be any set of time instants. Then the joint CDF of $N_j(t_1), N_j(t_2), \ldots N_j(t_k)$ approaches the joint CDF of $N(t_1), N(t_2), \ldots N(t_k)$ as $j \to \infty$.*

Proof: We can rewrite the joint PMF for each Bernoulli process as

$$\begin{aligned} \mathsf{p}_{N_j(t_1),\ldots,N_j(t_k)}(n_1,\ldots,n_k) &= \mathsf{p}_{N_j(t_1),\tilde{N}_j(t_1,t_2),\ldots,\tilde{N}_j(t_{k-1},t_k)}(n_1, n_2-n_1, \ldots, n_k-n_{k-1}) \\ &= \mathsf{p}_{N_j(t_1)}(n_1) \prod_{\ell=2}^{k} \mathsf{p}_{\tilde{N}_j(t_\ell,t_{\ell-1})}(n_\ell - n_{\ell-1}) \end{aligned} \tag{2.24}$$

where we have used the independent increment property for the Bernoulli process. For the Poisson process, we similarly have

$$\mathsf{p}_{N(t_1),\ldots,N(t_k)}(n_1,\ldots n_k) = \mathsf{p}_{N(t_1)}(n_1) \prod_{\ell=2}^{k} \mathsf{p}_{\tilde{N}(t_\ell,t_{\ell-1})}(n_\ell - n_{\ell-1}) \tag{2.25}$$

Taking the limit of (2.24) as $j \to \infty$, we recognize from Theorem 2.2.5 that each term of (2.24) goes to the corresponding term in (2.25). For the $\tilde{N}$ rv's , this requires a trivial generalization in Theorem 2.2.5 to deal with the arbitrary starting time. □

We conclude from this that the sequence of Bernoulli processes above converges to the Poisson process in the sense of the corollary. Recall from Section 1.7.5 that there are a number of ways in which a sequence of rv's can converge. As one might imagine, there are many more ways in which a sequence of stochastic processes can converge, and the corollary simply establishes one of these. Note however that showing only that $p_{N_j(t)}(n)$ approaches $p_{N(t)}(n)$ for each t is too weak to be very helpful, because it shows nothing about the time-evolution of the process. On the other hand, we don't even know how to define a joint distribution over an infinite number of epochs, let alone deal with limiting characteristics. Considering arbitrary finite sets of epochs forms a good middle ground.

Both the Poisson process and the Bernoulli process are so easy to analyze that the convergence of shrinking Bernoulli processes to Poisson is rarely the easiest way to establish properties about either. On the other hand, this convergence is a powerful aid to the intuition in understanding each process. In other words, the relation between Bernoulli and Poisson is very useful in suggesting new ways of looking at problems, but is usually not the best way to analyze those problems.

2.3 Combining and splitting Poisson processes

Suppose that $\{N_1(t);\, t > 0\}$ and $\{N_2(t);\, t > 0\}$ are independent Poisson counting processes[7] of rates λ_1 and λ_2 respectively. We want to look at the sum process where $N(t) = N_1(t) +$

[7]Two processes $\{N_1(t);\, t > 0\}$ and $\{N_2(t);\, t > 0\}$ are said to be independent if for all positive integers k and all sets of times $0 < t_1 < t_2 < \cdots < t_k$, the random variables $N_1(t_1), \ldots, N_1(t_k)$ are independent

$N_2(t)$ for all $t \geq 0$. In other words, $\{N(t);\, t > 0\}$ is the process consisting of all arrivals to both process 1 and process 2. We shall show that $\{N(t);\, t > 0\}$ is a Poisson counting process of rate $\lambda = \lambda_1 + \lambda_2$. We show this in three different ways, first using Definition 3 of a Poisson process (since that is most natural for this problem), then using Definition 2, and finally Definition 1. We then draw some conclusions about the way in which each approach is helpful. Since $\{N_1(t);\, t > 0\}$ and $\{N_2(t);\, t > 0\}$ are independent and both possess the stationary and independent increment properties, it follows from the definitions that $\{N(t);\, t > 0\}$ also possesses the stationary and independent increment properties. Using the approximations in (2.19) for the individual processes, we see that

$$\begin{aligned}\Pr\Big\{\widetilde{N}(t,t+\delta)=0\Big\} &= \Pr\Big\{\widetilde{N}_1(t,t+\delta)=0\Big\}\Pr\Big\{\widetilde{N}_2(t,t+\delta)=0\Big\}\\ &= (1-\lambda_1\delta)(1-\lambda_2\delta) \;\approx\; 1-\lambda\delta.\end{aligned}$$

where $\lambda_1\lambda_2\delta^2$ has been dropped. In the same way, $\Pr\Big\{\widetilde{N}(t,\, t{+}\delta)=1\Big\}$ is approximated by $\lambda\delta$ and $\Pr\Big\{\widetilde{N}(t,t+\delta)\geq 2\Big\}$ is approximated by 0, both with errors proportional to δ^2. It follows that $\{N(t);\, t > 0\}$ is a Poisson process.

In the second approach, we have $N(t) = N_1(t) + N_2(t)$. Since $N(t)$, for any given t, is the sum of two independent Poisson rv's, it is also a Poisson rv with mean $\lambda t = \lambda_1 t + \lambda_2 t$. If the reader is not aware that the sum of two independent Poisson rv's is Poisson, it can be derived by discrete convolution of the two PMF's (see Exercise 1.22). More elegantly, one can observe that we have already implicitly shown this fact. That is, if we break an interval I into disjoint subintervals, I_1 and I_2, then the number of arrivals in I (which is Poisson) is the sum of the number of arrivals in I_1 and in I_2 (which are independent Poisson). Finally, since $N(t)$ is Poisson for each t, and since the stationary and independent increment properties are satisfied, $\{N(t);\, t > 0\}$ is a Poisson process.

In the third approach, X_1, the first interarrival interval for the sum process, is the minimum of X_{11}, the first interarrival interval for the first process, and X_{21}, the first interarrival interval for the second process. Thus $X_1 > t$ if and only if both X_{11} and X_{21} exceed t, so

$$\Pr\{X_1 > t\} = \Pr\{X_{11} > t\}\Pr\{X_{21} > t\} = \exp(-\lambda_1 t - \lambda_2 t) = \exp(-\lambda t).$$

Using the memoryless property, each subsequent interarrival interval can be analyzed in the same way.

The first approach above was the most intuitive for this problem, but it required constant care about the order of magnitude of the terms being neglected. The second approach was the simplest analytically (after recognizing that sums of independent Poisson rv's are Poisson), and required no approximations. The third approach was very simple in retrospect, but not very natural for this problem. If we add many independent Poisson processes together, it is clear, by adding them one at a time, that the sum process is again Poisson. What is more interesting is that when many independent counting processes (not necessarily Poisson) are added together, the sum process often tends to be approximately Poisson if

of $N_2(t_1), \ldots, N_2(t_k)$. Here it is enough to extend the independent increment property to independence between increments over the two processes; equivalently, one can require the interarrival intervals for one process to be independent of the interarrivals for the other process.

the individual processes have small rates compared to the sum. To obtain some crude intuition about why this might be expected, note that the interarrival intervals for each process (assuming no bulk arrivals) will tend to be large relative to the mean interarrival interval for the sum process. Thus arrivals that are close together in time will typically come from different processes. The number of arrivals in an interval large relative to the combined mean interarrival interval, but small relative to the individual interarrival intervals, will be the sum of the number of arrivals from the different processes; each of these is 0 with large probability and 1 with small probability, so the sum will be approximately Poisson.

2.3.1 Subdividing a Poisson process

Next we look at how to break $\{N(t);\, t > 0\}$, a Poisson counting process of rate λ, into two processes, $\{N_1(t);\, t > 0\}$ and $\{N_2(t);\, t > 0\}$. Suppose that each arrival in $\{N(t);\, t > 0\}$ is sent to the first process with probability p and to the second process with probability $1 - p$ (see Figure 2.6). Each arrival is switched independently of each other arrival and independently of the arrival epochs. It may be helpful to visualize this as the combination of two independent processes. The first is the Poisson process of rate λ and the second is a Bernoulli process $\{X_n;\, n \geq 1\}$ where $\mathsf{p}_{X_n}(1) = p$ and $\mathsf{p}_{X_n}(2) = 1-p$. The nth arrival of the Poisson process is, with probability p, labeled as a type 1 arrival *i.e.*, labeled as $X_n = 1$. With probability $1 - p$, it is labeled as type 2, *i.e.*, labeled as $X_n = 2$.

We shall show that the resulting processes are each Poisson, with rates $\lambda_1 = \lambda p$ and $\lambda_2 = \lambda(1 - p)$ respectively, and that furthermore the two processes are independent. Note that, conditional on the original process, the two new processes are not independent; in fact one completely determines the other. Thus this independence might be a little surprising.

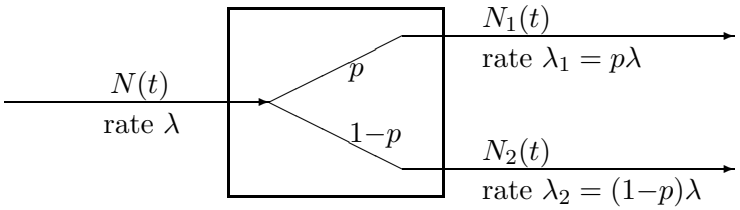


Figure 2.6: Each arrival is independently sent to process 1 with probability p and to process 2 otherwise.

First consider a small increment $(t, t + \delta]$. The original process has an arrival in this incremental interval with probability $\lambda\delta$ (ignoring δ^2 terms as usual), and thus process 1 has an arrival with probability $\lambda\delta p$ and process 2 with probability $\lambda\delta(1 - p)$. Because of the independent increment property of the original process and the independence of the division of each arrival between the two processes, the new processes each have the independent increment property, and from above have the stationary increment property. Thus each process is Poisson. Note now that we cannot verify that the two processes are independent from this small increment model. We would have to show that the number of arrivals for process 1 and 2 are independent over $(t, t + \delta]$. Unfortunately, leaving out the terms of order δ^2, there is at most one arrival to the original process and no possibility of

an arrival to each new process in $(t, t+\delta]$. *If* it is impossible for both processes to have an arrival in the same interval, they cannot be independent. It is possible, of course, for each process to have an arrival in the same interval, but this is a term of order δ^2. Thus, without paying attention to the terms of order δ^2, it is impossible to demonstrate that the processes are independent.

To demonstrate that process 1 and 2 are independent, we first calculate the joint PMF for $N_1(t)$, $N_2(t)$ for arbitrary t. Conditioning on a given number of arrivals $N(t)$ for the original process, we have

$$\Pr\{N_1(t)=m, N_2(t)=k \mid N(t)=m+k\} = \frac{(m+k)!}{m!k!}p^m(1-p)^k. \tag{2.26}$$

Equation (2.26) is simply the binomial distribution, since, given $m+k$ arrivals to the original process, each independently goes to process 1 with probability p. Since the event $\{N_1(t) = m,\ N_2(t) = k\}$ is a subset of the conditioning event above,

$$\Pr\{N_1(t)=m, N_2(t)=k \mid N(t)=m+k\} = \frac{\Pr\{N_1(t)=m, N_2(t)=k\}}{\Pr\{N(t)=m+k\}}.$$

Combining this with (2.26), we have

$$\Pr\{N_1(t)=m, N_2(t)=k\} = \frac{(m+k)!}{m!k!}p^m(1-p)^k\frac{(\lambda t)^{m+k}e^{-\lambda t}}{(m+k)!}. \tag{2.27}$$

Rearranging terms, we get

$$\Pr\{N_1(t)=m, N_2(t)=k\} = \frac{(p\lambda t)^m e^{-\lambda pt}}{m!}\ \frac{[(1-p)\lambda t]^k e^{-\lambda(1-p)t}}{k!}. \tag{2.28}$$

This shows that $N_1(t)$ and $N_2(t)$ are independent. To show that the processes are independent, we must show that for any $k > 1$ and any set of times $0 \le t_1 \le t_2 \le \cdots \le t_k$, the sets $\{N_1(t_i);\ 1 \le i \le k\}$ and $\{N_2(t_j);\ 1 \le j \le k\}$ are independent of each other. It is equivalent to show that the sets $\{\widetilde{N}_1(t_{i-1}, t_i); 1 \le i \le k\}$ and $\{\widetilde{N}_2(t_{j-1}, t_j); 1 \le j \le k\}$ (where t_0 is 0) are independent. The argument above shows this independence for $i = j$, and for $i \ne j$, the independence follows from the independent increment property of $\{N(t);\ t > 0\}$.

2.3.2 Examples using independent Poisson processes

We have observed that if the arrivals of a Poisson process are split into two new arrival processes, with each arrival of the original process independently entering the first of the new processes with some fixed probability p, then each new process is Poisson and independent of the other. The most useful consequence of this is that any two independent Poisson processes can be viewed as being generated from a single process in this way. Thus, if one process has rate λ_1 and the other has rate λ_2, they can be viewed as coming from a process of rate $\lambda_1+\lambda_2$. Each arrival to the combined process is then labeled as a first-process arrival with probability $p = \lambda_1/(\lambda_1 + \lambda_2)$ and as a second-process arrival with probability $1 - p$.

Example 2.3.1. The above point of view is very useful for finding probabilities such as $\Pr\{S_{1k} < S_{2j}\}$ where S_{1k} is the epoch of the kth arrival to the first process and S_{2j} is the epoch of the jth arrival to the second process. The problem can be rephrased in terms of a combined process to ask: out of the first $k+j-1$ arrivals to the combined process, what is the probability that k or more of them are switched to the first process? (Note that if k or more of the first $k+j-1$ go to the first process, at most $j-1$ go to the second, so the kth arrival to the first precedes the jth arrival to the second; similarly if fewer than k of the first $k+j-1$ go to the first process, then the jth arrival to the second process precedes the kth arrival to the first). Since each of these first $k+j-1$ arrivals are switched independently with the same probability p, the answer is

$$\Pr\{S_{1k} < S_{2j}\} = \sum_{i=k}^{k+j-1} \frac{(k+j-1)!}{i!(k+j-1-i)!} p^i (1-p)^{k+j-1-i}. \tag{2.29}$$

Example 2.3.2 (The M/M/1 queue). Queueing theorists use a standard notation of characters separated by slashes to describe common types of queueing systems. The first character describes the arrival process to the queue. M stands for memoryless and means a Poisson arrival process; D stands for deterministic and means that the interarrival interval is fixed and non-random; G stands for general interarrival distribution. We assume that the interarrival intervals are IID (thus making the arrival process a renewal process), but many authors use GI to explicitly indicate IID interarrivals. The second character describes the service process. The same letters are used, with M indicating exponentially distributed service times. The third character gives the number of servers.[8] It is assumed, when this notation is used, that the service times are IID, independent of the arrival epochs, and independent of which server is used.

Consider an M/M/1 queue, *i.e.*, a queueing system with a Poisson arrival system (say of rate λ) and a single server that serves arriving customers in order with a service time distribution $\mathsf{F}(y) = 1 - \exp[-\mu y]$. The service times are independent of each other and of the interarrival intervals. During any period when the server is busy, customers then leave the system according to a Poisson process (process 2) of rate μ. We then see that if j or more customers are waiting at a given time, then (2.29) gives the probability that the kth subsequent arrival comes before the jth departure.

Example 2.3.3 (The sum of geometrically many exponential rv's). Consider waiting for a taxi on a busy street during rush hour. We will model the epochs at which taxis pass our spot on the street as a Poisson process of rate λ. We assume that each taxi is independently empty (*i.e.*, will stop to pick us up) with probability p. The empty taxis then form a Poisson process of rate $p\lambda$ and our waiting time is exponential with parameter λp. Looked at slightly differently, we will be picked by the first taxi with probability p, by the second with probability $(1-p)p$, and in general, by the mth taxi with probability $(1-p)^{m-1}p$. In other words, the number of the taxi that picks us up (numbered in order of arrival) is a geometricly distributed rv, say M, and our waiting time is the sum of the first M interarrival times, *i.e.*, S_M.

[8] Sometimes a fourth character is added which gives the size of the queue plus service facility. Thus, for example, an M/M/m/m queueing system would have Poisson arrivals, m independent servers, each with exponential service time. If an arrival occurs when all servers are busy, then that arrival is dropped.

As illustrated by taxis, waiting times modeled by a geometric sum of exponentials are relatively common. An important example that we shall analyze in Chapter 7 is that of the service time distribution of an M/M/1 queue in steady state (*i.e.*, after the effect of the initially empty queue has disappeared). We will show there that the number M of customers in the system immediately after an arrival is geometric with the PMF

$$\Pr\{M = m\} = (\lambda/\mu)^{m-1}(1 - \lambda/\mu) \qquad \text{for } m \geq 1.$$

Now the time that the new arrival spends in the system is the sum of M service times, each of which is exponential with parameter μ. These service times are independent of each other and of M. Thus the system time (the time in queue and in service) of the new arrival is exponential with parameter $\mu(1 - \lambda/\mu) = \mu - \lambda$.

2.4 Non-homogeneous Poisson processes

The Poisson process, as we defined it, is characterized by a constant arrival rate λ. It is often useful to consider a more general type of process in which the arrival rate varies as a function of time. A *non-homogeneous Poisson process* with time varying arrival rate $\lambda(t)$ is defined[9] as a counting process $\{N(t);\, t > 0\}$ which has the independent increment property and, for all $t \geq 0, \delta > 0$, also satisfies:

$$\begin{aligned} \Pr\left\{\widetilde{N}(t, t+\delta) = 0\right\} &= 1 - \delta\lambda(t) + \mathrm{o}(\delta) \\ \Pr\left\{\widetilde{N}(t, t+\delta) = 1\right\} &= \delta\lambda(t) + \mathrm{o}(\delta) \\ \Pr\left\{\widetilde{N}(t, t+\delta) \geq 2\right\} &= \mathrm{o}(\delta). \end{aligned} \tag{2.30}$$

where $\widetilde{N}(t, t+\delta) = N(t+\delta) - N(t)$. The non-homogeneous Poisson process does not have the stationary increment property.

One common application occurs in optical communication where a non-homogeneous Poisson process is often used to model the stream of photons from an optical modulator; the modulation is accomplished by varying the photon intensity $\lambda(t)$. We shall see another application shortly in the next example. Sometimes a Poisson process, as we defined it earlier, is called a homogeneous Poisson process.

We can use a "shrinking Bernoulli process" again to approximate a non-homogeneous Poisson process. To see how to do this, assume that $\lambda(t)$ is bounded away from zero. We partition the time axis into increments whose lengths δ vary inversely with $\lambda(t)$, thus holding the probability of an arrival in an increment at some fixed value $p = \delta\lambda(t)$. Thus,

[9] We assume that $\lambda(t)$ is right continuous, i.e., that for each t, $\lambda(t)$ is the limit of $\lambda(t+\epsilon)$ as ϵ approaches 0 from above. This allows $\lambda(t)$ to contain discontinuities, as illustrated in Figure 2.7, but follows the convention that the value of the function at the discontinuity is the limiting value from the right. This convention is required in (2.30) to talk about the distribution of arrivals just to the right of time t.

temporarily ignoring the variation of $\lambda(t)$ within an increment,

$$\begin{aligned}
\Pr\left\{\widetilde{N}\left(t, t+\frac{p}{\lambda(t)}\right)=0\right\} &= 1-p+\mathrm{o}(p)\\
\Pr\left\{\widetilde{N}\left(t, t+\frac{p}{\lambda(t)}\right)=1\right\} &= p+\mathrm{o}(p)\\
\Pr\left\{\widetilde{N}\left(t, t+\frac{p}{\lambda(t)}\right)\geq 2\right\} &= \mathrm{o}(p). \qquad (2.31)
\end{aligned}$$

This partition is defined more precisely by defining $m(t)$ as

$$m(t)=\int_0^t \lambda(\tau)d\tau. \qquad (2.32)$$

Then the ith increment ends at that t for which $m(t) = i\ p$.

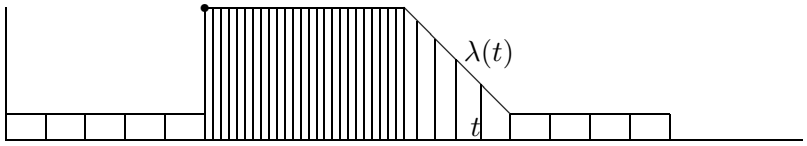


Figure 2.7: Partitioning the time axis into increments each with an expected number of arrivals equal to p. Each rectangle or trapezoid above has the same area, which ensures that the ith partition ends where $m(t) = i\ p$.

As before, let $\{Y_i;\ i \geq 1\}$ be a sequence of IID binary rv's with $\Pr\{Y_i = 1\} = p$ and $\Pr\{Y_i = 0\} = 1 - p$. Consider the counting process $\{N(t);\ t > 0\}$ in which Y_i, for each $i \geq 1$, denotes the number of arrivals in the interval $(t_{i-1}, t_i]$, where t_i satisfies $m(t_i) = i\ p$. Thus, $N(t_i) = Y_1 + Y_2 + \cdots + Y_i$. If p is decreased as 2^{-j}, each increment is successively split into a pair of increments. Thus by the same argument as in (2.23),

$$\Pr\{N(t)=n\}=\frac{[1+\mathrm{o}(p)][m(t)]^n \exp[-m(t)]}{n!}. \qquad (2.33)$$

Similarly, for any interval $(t, \tau]$, taking $\widetilde{m}(t,\tau) = \int_t^\tau \lambda(u)du$, and taking $t = t_k$, $\tau = t_i$ for some k, i, we get

$$\Pr\left\{\widetilde{N}(t,\tau)=n\right\}=\frac{[1+\mathrm{o}(p)][\widetilde{m}(t,\tau)]^n \exp[-\widetilde{m}(t,\tau)]}{n!}. \qquad (2.34)$$

Going to the limit $p \to 0$, the counting process $\{N(t);\ t > 0\}$ above approaches the non-homogeneous Poisson process under consideration, and we have the following theorem:

Theorem 2.4.1. *For a non-homogeneous Poisson process with right-continuous arrival rate $\lambda(t)$ bounded away from zero, the distribution of $\widetilde{N}(t,\tau)$, the number of arrivals in $(t,\tau]$, satisfies*

$$\Pr\left\{\widetilde{N}(t,\tau)=n\right\}=\frac{[\widetilde{m}(t,\tau)]^n \exp[-\widetilde{m}(t,\tau)]}{n!} \qquad \text{where } \widetilde{m}(t,\tau)=\int_t^\tau \lambda(u)\,du. \qquad (2.35)$$

Hence, one can view a non-homogeneous Poisson process as a (homogeneous) Poisson process over a non-linear time scale. That is, let $\{N^*(s); s \geq 0\}$ be a (homogeneous) Poisson process with rate 1. The non-homogeneous Poisson process is then given by $N(t) = N^*(m(t))$ for each t.

Example 2.4.1 (The M/G/∞ Queue). Using the queueing notation explained in Example 2.3.2, an $M/G/\infty$ queue indicates a queue with Poisson arrivals, a general service distribution, and an infinite number of servers. Since the M/G/∞ queue has an infinite number of servers, no arriving customers are ever queued. Each arrival immediately starts to be served by some server, and the service time Y_i of customer i is IID over i with some given CDF $G(y)$; the service time is the interval from start to completion of service and is also independent of arrival epochs. We would like to find the CDF of the number of customers being served at a given epoch τ.

Let $\{N(t); t > 0\}$ be the Poisson counting process, at rate λ, of customer arrivals. Consider the arrival times of those customers that are still in service at some fixed time τ. In some arbitrarily small interval $(t, t+\delta]$, the probability of an arrival is $\delta\lambda + o(\delta)$ and the probability of 2 or more arrivals is negligible (i.e., $o(\delta)$). The probability that a customer arrives in $(t, t+\delta]$ and is still being served at time $\tau > t$ is then $\delta\lambda[1 - G(\tau - t)] + o(\delta)$. Consider a counting process $\{N_1(t);\ 0<t\leq\tau\}$ where $N_1(t)$ is the number of arrivals between 0 and t that are still in service at τ. This counting process has the independent increment property. To see this, note that the overall arrivals in $\{N(t); t > 0\}$ have the independent increment property; also the arrivals in $\{N(t); t > 0\}$ have independent service times, and thus are independently in or not in $\{N_1(t); 0 < t \leq \tau\}$. It follows that $\{N_1(t); 0 < t \leq \tau\}$ is a non-homogeneous Poisson process with rate $\lambda[1 - G(\tau - t)]$ at time $t \leq \tau$. The expected number of arrivals still in service at time τ is then

$$m(\tau) = \lambda \int_{t=0}^{\tau} [1 - G(\tau - t)]\, dt = \lambda \int_{t=0}^{\tau} [1 - G(t)]\, dt. \tag{2.36}$$

and the PMF of the number in service at time τ is given by

$$\Pr\{N_1(\tau) = n\} = \frac{m(\tau)^n \exp(-m(\tau))}{n!}. \tag{2.37}$$

Note that as $\tau \to \infty$, the integral in (2.36) approaches the mean of the service time distribution (i.e., it is the integral of the complementary CDF, $1 - G(t)$, of the service time). This means that in steady state (as $\tau \to \infty$), the distribution of the number in service at τ depends on the service time distribution only through its mean. This example can be used to model situations such as the number of phone calls taking place at a given epoch. This requires arrivals of new calls to be modeled as a Poisson process and the holding time of each call to be modeled as a random variable independent of other holding times and of call arrival times. Finally, as shown in Figure 2.8, we can regard $\{N_1(t);\ 0<t \leq \tau\}$ as a splitting of the arrival process $\{N(t);\ t>0\}$. By the same type of argument as in Section 2.3, the number of customers who have completed service by time τ is independent of the number still in service.

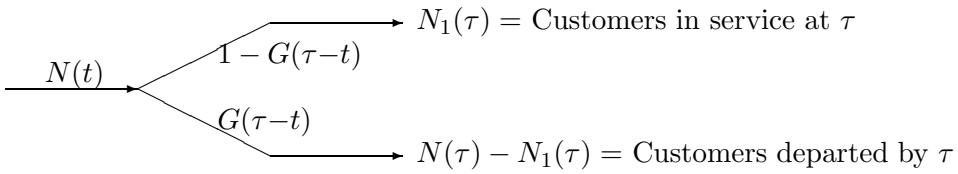


Figure 2.8: Poisson arrivals $\{N(t);\, t > 0\}$ can be considered to be split in a non-homogeneous way. An arrival at t is split with probability $1 - G(\tau - t)$ into a process of customers still in service at τ.

2.5 Conditional arrival densities and order statistics

A diverse range of problems involving Poisson processes are best tackled by conditioning on a given number n of arrivals in the interval $(0, t]$, i.e., on the event $N(t) = n$. Because of the incremental view of the Poisson process as independent and stationary arrivals in each incremental interval of the time axis, we would guess that the arrivals should have some sort of uniform distribution given $N(t) = n$. More precisely, the following theorem shows that the joint density of $\boldsymbol{S}^{(n)} = (S_1, S_2, \ldots, S_n)$ given $N(t) = n$ is uniform over the region $0 < S_1 < S_2 < \cdots < S_n < t$.

Theorem 2.5.1. *Let $\mathsf{f}_{\boldsymbol{S}^{(n)}|N(t)=n}(\boldsymbol{s}^{(n)} \mid n)$ be the joint density of $\boldsymbol{S}^{(n)}$ conditional on $N(t) = n$. This density is constant over the region $0 < s_1 < \cdots < s_n < t$ and has the value*

$$\mathsf{f}_{\boldsymbol{S}^{(n)}|N(t)}(\boldsymbol{s}^{(n)} \mid n) \;=\; \frac{n!}{t^n}. \tag{2.38}$$

Two proofs are given, each illustrative of useful techniques.

Proof 1: Recall that the joint density of the first $n+1$ arrivals $\boldsymbol{S}^{(n+1)} = (S_1, \ldots, S_n, S_{n+1})$ with no conditioning is given in (2.15). We first use Bayes law to calculate the joint density of $\boldsymbol{S}^{(n+1)}$ conditional on $N(t) = n$.

$$\mathsf{f}_{\boldsymbol{S}^{(n+1)}|N(t)}(\boldsymbol{s}^{(n+1)} \mid n)\, \mathsf{p}_{N(t)}(n) = \mathsf{p}_{N(t)|\boldsymbol{S}^{(n+1)}}(n|\boldsymbol{s}^{(n+1)})\mathsf{f}_{\boldsymbol{S}^{(n+1)}}(\boldsymbol{s}^{(n+1)}).$$

Note that $N(t) = n$ if and only if $S_n \le t$ and $S_{n+1} > t$. Thus $\mathsf{p}_{N(t)|\boldsymbol{S}^{(n+1)}}(n|\boldsymbol{s}^{(n+1)})$ is 1 if $S_n \le t$ and $S_{n+1} > t$ and is 0 otherwise. Restricting attention to the case $N(t) = n$, $S_n \le t$ and $S_{n+1} > t$,

$$\begin{aligned}
\mathsf{f}_{\boldsymbol{S}^{(n+1)}|N(t)}(\boldsymbol{s}^{(n+1)} \mid n) &= \frac{\mathsf{f}_{\boldsymbol{S}^{(n+1)}}(\boldsymbol{s}^{(n+1)})}{\mathsf{p}_{N(t)}(n)} \\
&= \frac{\lambda^{n+1}\exp(-\lambda s_{n+1})}{(\lambda t)^n \exp(-\lambda t)\,/n!} \\
&= \frac{n!\lambda \exp[-\lambda(s_{n+1} - t)]}{t^n}.
\end{aligned} \tag{2.39}$$

This is a useful expression, but we are interested in $\boldsymbol{S}^{(n)}$ rather than $\boldsymbol{S}^{(n+1)}$. Thus we break up the left side of (2.39) as follows:

$$\mathsf{f}_{\boldsymbol{S}^{(n+1)}|N(t)}(\boldsymbol{s}^{(n+1)} \mid n) = \mathsf{f}_{\boldsymbol{S}^{(n)}|N(t)}(\boldsymbol{s}^{(n)} \mid n)\; \mathsf{f}_{S_{n+1}|\boldsymbol{S}^{(n)}N(t)}(s_{n+1}|\boldsymbol{s}^{(n)}, n).$$

Conditional on $N(t) = n$, S_{n+1} is the first arrival epoch after t, which by the memoryless property is conditionally independent of $\boldsymbol{S}^{(n)}$, again given $N(t) = n$. Thus that final term is simply $\lambda \exp(-\lambda(s_{n+1} - t))$ for $s_{n+1} > t$. Substituting this into (2.39), the result is (2.38). □

Proof 2: This alternative proof derives (2.38) by looking at arrivals in very small increments of size δ (see Figure 2.9). For a given t and a given set of n times, $0 < s_1 < \cdots < s_n < t$, we calculate the probability that there is a single arrival in each of the intervals $(s_i, s_i + \delta], 1 \leq i \leq n$ and no other arrivals in the interval $(0, t]$. Letting $A(\delta)$ be this event,

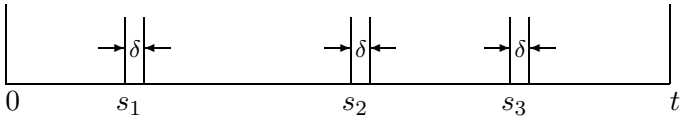


Figure 2.9: Intervals for the event $A(t)$ used to derive the joint arrival density.

$$\Pr\{A(\delta)\} = \mathsf{p}_{N(s_1)}(0)\ \mathsf{p}_{\widetilde{N}(s_1,s_1+\delta)}(1)\ \mathsf{p}_{\widetilde{N}(s_1+\delta,s_2)}(0)\ \mathsf{p}_{\widetilde{N}(s_2,s_2+\delta)}(1) \cdots \mathsf{p}_{\widetilde{N}(s_n+\delta,t)}(0).$$

The sum of the lengths of the above intervals is t. Also $\mathsf{p}_{\widetilde{N}(s_i,s_i+\delta)}(1) = \lambda\delta \exp(-\lambda\delta)$ for each i, so

$$\Pr\{A(\delta)\} = (\lambda\delta)^n \exp(-\lambda t).$$

The event $A(\delta)$ can be characterized as the event that, first, $N(t) = n$ and, second, that the n arrivals occur in $(s_i, s_i+\delta]$ for $1 \leq i \leq n$. Thus we conclude that

$$\mathsf{f}_{\boldsymbol{S}^{(n)}|N(t)}(\boldsymbol{s}^{(n)}|n) = \lim_{\delta \to 0} \frac{\Pr\{A(\delta)\}}{\delta^n \mathsf{p}_{N(t)}(n)},$$

which simplifies to (2.38). □

The joint density of the interarrival intervals, $\boldsymbol{X}^{(n)} = (X_1, \ldots, X_n)$ given $N(t) = n$ can be found directly from Theorem 2.5.1 simply by making the linear transformation $X_1 = S_1$ and $X_i = S_i - S_{i-1}$ for $2 \leq i \leq n$. The density is unchanged, but the constraint region transforms into $\sum_{i=1}^{n} X_i < t$ with $X_i > 0$ for $1 \leq i \leq n$ (see Figure 2.10).

$$\mathsf{f}_{\boldsymbol{X}^{(n)}|N(t)}(\boldsymbol{x}^{(n)} \mid n) = \frac{n!}{t^n} \qquad \text{for } \boldsymbol{X}^{(n)} > 0,\ \sum\nolimits_{i=1}^{n} X_i < t. \tag{2.40}$$

It is also instructive to compare the joint distribution of $\boldsymbol{S}^{(n)}$ conditional on $N(t) = n$ with the joint distribution of n IID uniformly distributed random variables, $\boldsymbol{U}^{(n)} = (U_1, \ldots, U_n)$ on $(0, t]$. For any point $\boldsymbol{U}^{(n)} = \boldsymbol{u}^{(n)}$, this joint density is

$$\mathsf{f}_{\boldsymbol{U}^{(n)}}(\boldsymbol{u}^{(n)}) = \frac{1}{t^n} \qquad \text{for } 0 < u_i \leq t,\ 1 \leq i \leq n.$$

Both $\mathsf{f}_{\boldsymbol{S}^{(n)}}$ and $\mathsf{f}_{\boldsymbol{U}^{(n)}}$ are uniform over the volume of n-space where they are non-zero, but as illustrated in Figure 2.11 for $n = 2$, the volume for the latter is $n!$ times larger than the

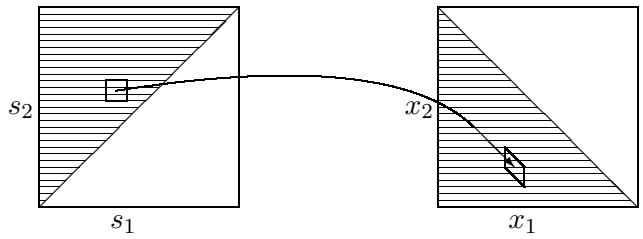


Figure 2.10: Mapping from arrival epochs to interarrival times. Note that incremental cubes in the arrival space map to parallelepipeds of the same volume in the interarrival space.

volume for the former. To explain this more fully, we can define a set of random variables $S_1, \ldots, S_n$, not as arrival epochs in a Poisson process, but rather as the order statistics function of the IID uniform variables $U_1, \ldots, U_n$; that is

$$S_1 = \min(U_1, \ldots, U_n);\ S_2 = 2^{\text{nd}} \text{ smallest } (U_1, \ldots, U_n);\ \text{etc.}$$

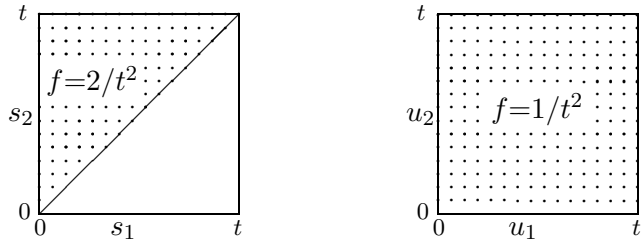


Figure 2.11: Density for the order statistics of an IID 2-dimensional uniform distribution. Note that the square over which $\mathsf{f}_{\boldsymbol{U}^{(2)}}$ is non-zero contains one triangle where $u_2 > u_1$ and another of equal size where $u_1 > u_2$. Each of these maps, by a permutation mapping, to the single triangle where $s_2 > s_1$.

The n-cube is partitioned into $n!$ regions, one where $u_1 < u_2 < \cdots < u_n$. For each permutation $\pi(i)$ of the integers 1 to n, there is another region[10] where $u_{\pi(1)} < u_{\pi(2)} < \cdots < u_{\pi(n)}$. By symmetry, each of these regions has the same volume, which then must be $1/n!$ of the volume t^n of the n-cube.

All of these $n!$ regions map to the same region of ordered values. Thus these order statistics have the same joint probability density function as the arrival epochs $S_1, \ldots, S_n$ conditional on $N(t) = n$. Anything we know (or can discover) about order statistics is valid for Poisson arrival epochs given $N(t) = n$ and vice versa.[11]

[10] As usual, we are ignoring those points where $u_i = u_j$ for some i, j, since the set of such points has 0 probability.

[11] There is certainly also the intuitive notion, given n arrivals in $(0, t]$, and given the stationary and independent increment properties of the Poisson process, that those n arrivals can be viewed as uniformly

Next we want to find the marginal CDF's of the individual S_i, conditional on $N(t) = n$. Starting with S_1, and viewing it as the minimum of the IID uniformly distributed variables $U_1, \ldots, U_n$, we recognize that $S_1 > \tau$ if and only if $U_i > \tau$ for all i, $1 \le i \le n$. Thus,

$$\Pr\{S_1 > \tau \mid N(t){=}n\} = \left[\frac{t-\tau}{t}\right]^n \qquad \text{for } 0 < \tau \le t. \tag{2.41}$$

For S_2 to S_n, the density is slightly simpler in appearance than the CDF. To find $\mathsf{f}_{S_i|N(t)}(s_i \mid n)$, look at n uniformly distributed rv's in $(0,t]$. The probability that one of these lies in the interval $(s_i, s_i + ds_i]$ is $(n\,ds_i)/t$. Out of the remaining $n-1$, the probability that $i-1$ lie in the interval $(0, s_i]$ is given by the binomial distribution with probability of success s_i/t. Thus the desired density is

$$\begin{aligned} \mathsf{f}_{S_i|N(t)}(s_i \mid n)\,dt &= \frac{s_i^{i-1}(t-s_i)^{n-i}(n-1)!}{t^{n-1}(n-i)!(i-1)!} \quad \frac{n\,ds_i}{t} \\ \mathsf{f}_{S_i|N(t)}(s_i \mid n) &= \frac{s_i^{i-1}(t-s_i)^{n-i}n!}{t^n(n-i)!(i-1)!}. \end{aligned} \tag{2.42}$$

It is easy to find the expected value of S_1 conditional on $N(t) = n$ by integrating the complementary CDF in (2.41), getting

$$\mathsf{E}\,[S_1 \mid N(t){=}n] = \frac{t}{n+1}. \tag{2.43}$$

We come back later to find $\mathsf{E}\,[S_i \mid N(t) = n]$ for $2 \le i \le n$. First, we look at the marginal distributions of the interarrival intervals. Recall from (2.40) that

$$\mathsf{f}_{\boldsymbol{X}^{(n)}|N(t)}(\boldsymbol{x}^{(n)} \mid n) = \frac{n!}{t^n} \qquad \text{for } \boldsymbol{X}^{(n)} > 0,\ \textstyle\sum_{i=1}^n X_i < t. \tag{2.44}$$

The joint density is the same for all points in the constraint region, and the constraint does not distinguish between X_1 to X_n. Thus $X_1, \ldots, X_n$ must all have the same marginal distribution, and more generally the marginal distribution of any subset of the X_i can depend only on the size of the subset. We have found the distribution of S_1, which is the same as X_1, and thus

$$\Pr\{X_i > \tau \mid N(t){=}n\} = \left[\frac{t-\tau}{t}\right]^n \qquad \text{for } 1 \le i \le n \text{ and } 0 < \tau \le t. \tag{2.45}$$

$$\mathsf{E}\,[X_i \mid N(t){=}n] = \frac{t}{n+1} \qquad \text{for } 1 \le i \le n. \tag{2.46}$$

From this, we see immediately that for $1 \le i \le n$,

$$\mathsf{E}\,[S_i \mid N(t) = n] = \frac{it}{n+1}. \tag{2.47}$$

distributed. One way to view this is to visualize the Poisson process as the sum of a very large number k of independent processes of rate λ/k each. Then, given $N(t) = n$, with $k >> n$, there is negligible probability of more than one arrival from any one process, and for each of the n processes with arrivals, that arrival is uniformly distributed in $(0,t]$.

One could go on and derive joint distributions of all sorts at this point, but there is one additional type of interval that must be discussed. Define $X^*_{n+1} = t - S_n$ to be the interval from the largest arrival epoch before t to t itself. Rewriting (2.44),

$$\mathsf{f}_{\boldsymbol{X}^{(n)}|N(t)}(\boldsymbol{x}^{(n)} \mid n) = \frac{n!}{t^n} \qquad \text{for } \boldsymbol{X}^{(n)} > 0,\ X^*_{n+1} > 0,\ \sum_{i=1}^{n} X_i + X^*_{n+1} = t.$$

The constraints above are symmetric in $X_1, \ldots, X_n, X^*_{n+1}$, and, within the constraint region, the joint density of $X_1, \ldots, X_n$ (conditional on $N(t) = n$) is uniform. Note that there is no joint density over $X_1, \ldots, X_n, X^*_{n+1}$ conditional on $N(t) = n$, since X^*_{n+1} is then a deterministic function of $X_1, \ldots, X_n$. However, the density over $X_1, \ldots, X_n$ can be replaced by a density over any other n rv's out of $X_1, \ldots, X_n, X^*_{n+1}$ by a linear transformation with unit determinant. Thus X^*_{n+1} has the same marginal distribution as each of the X_i. This gives us a partial check on our work, since the interval $(0, t]$ is divided into $n+1$ intervals of sizes $X_1, X_2, \ldots, X_n, X^*_{n+1}$, and each of these has a mean size $t/(n+1)$. We also see that the joint CDF of any proper subset of $X_1, X_2, \ldots X_n$, X^*_{n+1} depends only on the size of the subset and not the particular rv's in the subset.

One important consequece of this is that we can look at the arrivals in a segment $(0, t)$ of a Poisson process either forward or backward[12] in time and they 'look the same.' Looked at backwards, the interarrival intervals are $X^*_{n+1}, X_n, \ldots, X_2$. These intervals are IID, and X_1 is then determined as $t - X^*_{n+1} - X_n - \cdots - X_2$. We will not make any particular use of this property here, but we will later explore this property of time-reversibility for other types of processes. For Poisson processes, this reversibility is intuitively obvious from the stationary and independent properties. It is less obvious how to express this condition by equations, but that is not really necessary at this point.

2.6 Summary

We started the chapter with three equivalent definitions of a Poisson process—first as a renewal process with exponentially distributed inter-renewal intervals, second as a stationary and independent increment counting process with a Poisson distributed number of arrivals in each interval, and third essentially as a limit of shrinking Bernoulli processes. We saw that each definition provided its own insights into the properties of the process. We emphasized the importance of the memoryless property of the exponential distribution, both as a useful tool in problem solving and as an underlying reason why the Poisson process is so simple.

We next showed that the sum of independent Poisson processes is again a Poisson process. We also showed that if the arrivals in a Poisson process are independently routed to different locations with some fixed probability assignment, then the arrivals at these locations form independent Poisson processes. This ability to view independent Poisson processes either

[12]This must be interpreted carefully, since if we ask about the distribution to the next arrival from some $\tau \in (0, t)$, it is exponential (possibly occurring after t). But if we look for the distribution of the previous arrival before τ, we usually truncate the distribution at 0. Thus the proper interpretation is to look only at the interval $(0, t)$ under the condition of a given number of arrivals in that interval.

independently or as a splitting of a combined process is a powerful technique for finding almost trivial solutions to many problems.

It was next shown that a non-homogeneous Poisson process could be viewed as a (homogeneous) Poisson process on a non-linear time scale. This allows all the properties of (homogeneous) Poisson processes to be applied directly to the non-homogeneous case. The simplest and most useful result from this is (2.35), showing that the number of arrivals in any interval has a Poisson PMF. This result was used to show that the number of customers in service at any given time τ in an M/G/∞ queue has a Poisson PMF with a mean approaching λ times the expected service time in the limit as $\tau \to \infty$.

Finally we looked at the distribution of arrival epochs conditional on n arrivals in the interval $(0, t]$. It was found that these arrival epochs had the same joint distribution as the order statistics of n uniform IID rv's in $(0, t]$. By using symmetry and going back and forth between the uniform variables and the Poisson process arrivals, we found the distribution of the interarrival times, the arrival epochs, and various conditional distributions.

2.7 Exercises

Exercise 2.1. a) Find the Erlang density $\mathsf{f}_{S_n}(t)$ by convolving $\mathsf{f}_X(x) = \lambda \exp(-\lambda x)$ with itself n times.

b) Find the moment generating function of X (or find the Laplace transform of $\mathsf{f}_X(x)$), and use this to find the moment generating function (or Laplace transform) of $S_n = X_1 + X_2 + \cdots + X_n$. Invert your result to find $\mathsf{f}_{S_n}(t)$.

c) Find the Erlang density by starting with (2.15) and then calculating the marginal density for S_n.

Exercise 2.2. a) Find the mean, variance, and moment generating function of $N(t)$, as given by (2.17).

b) Show by discrete convolution that the sum of two independent Poisson rv's is again Poisson.

c) Show by using the properties of the Poisson process that the sum of two independent Poisson rv's must be Poisson.

Exercise 2.3. The purpose of this exercise is to give an alternate derivation of the Poisson distribution for $N(t)$, the number of arrivals in a Poisson process up to time t. Let λ be the rate of the process.

a) Find the conditional probability $\Pr\{N(t) = n \mid S_n = \tau\}$ for all $\tau \leq t$.

b) Using the Erlang density for S_n, use (a) to find $\Pr\{N(t) = n\}$.

Exercise 2.4. Assume that a counting process $\{N(t); t{>}0\}$ has the independent and stationary increment properties and satisfies (2.17) (for all $t > 0$). Let X_1 be the epoch of the first arrival and X_n be the interarrival time between the $n-1^{\text{st}}$ and the nth arrival. Use only these assumptions in doing the following parts of this exercise.

a) Show that $\Pr\{X_1 > x\} = e^{-\lambda x}$.

b) Let S_{n-1} be the epoch of the $n-1^{\text{st}}$ arrival. Show that $\Pr\{X_n > x \mid S_{n-1} = \tau\} = e^{-\lambda x}$.

c) For each $n > 1$, show that $\Pr\{X_n > x\} = e^{-\lambda x}$ and that X_n is independent of S_{n-1}.

d) Argue that X_n is independent of $X_1, X_2, \ldots X_{n-1}$.

Exercise 2.5. The point of this exercise is to show that the sequence of PMF's for a Bernoulli counting process does not specify the process. In other words, knowing that $N(t)$ satisfies the binomial distribution for all t does not mean that the process is Bernoulli. This helps us understand why the second definition of a Poisson process requires stationary and independent increments along with the Poisson distribution for $N(t)$.

a) Let $Y_1, Y_2, Y_3, \ldots$ be a sequence of binary rv's in which each rv is 0 or 1 with equal probability. Find a joint distribution for Y_1, Y_2, Y_3 that satisfies the binomial distribution, $\mathsf{p}_{N(t)}(k) = \binom{t}{k} 2^{-t}$ for $t = 1, 2, 3$ and $0 \leq k \leq t$, but for which Y_1, Y_2, Y_3 are not independent.

One simple solution for this contains four 3-tuples with probability 1/8 each, two 3-tuples with probability 1/4 each, and two 3-tuples with probability 0. Note that by making the subsequent arrivals IID and equiprobable, you have an example where $N(t)$ is binomial for all t but the process is not Bernoulli. Hint: Use the binomial for $t = 3$ to find two 3-tuples that must have probability 1/8. Combine this with the binomial for $t = 2$ to find two other 3-tuples that must have probability 1/8. Finally look at the constraints imposed by the binomial distribution on the remaining four 3-tuples.

b) Generalize (a) to the case where Y_1, Y_2, Y_3 satisfy $\Pr\{Y_i = 1\} = p$ and $\Pr\{Y_i = 0\} = 1 - p$. Assume $p < 1/2$ and find a joint distribution on Y_1, Y_2, Y_3 that satisfies the binomial distribution, but for which the 3-tuple $(0, 1, 1)$ has zero probability.

c) More generally yet, view a joint PMF on binary t-tuples as a nonnegative vector in a 2^t dimensional vector space. Each binomial probability $p_{N(\tau)}(k) = \binom{\tau}{k} q^k (1-q)^{\tau-k}$ constitutes a linear constraint on this vector. For each τ, show that one of these constraints may be replaced by the constraint that the components of the vector sum to 1.

d) Using (c), show that at most $(t + 1)t/2 + 1$ of the binomial constraints are linearly independent. Note that this means that the linear space of vectors satisfying these binomial constraints has dimension at least $2^t - (t + 1)t/2 - 1$. This linear space has dimension 1 for $t = 3$ explaining the results in parts a) and b). It has a rapidly increasing dimension for $t > 3$, suggesting that the binomial constraints are relatively ineffectual for constraining the joint PMF of a joint distribution. More work is required for the case of $t > 3$ because of all the inequality constraints, but it turns out that this large dimensionality remains.

Exercise 2.6. Let $h(x)$ be a positive function of a real variable that satisfies $h(x + t) = h(x) + h(t)$ for all x, t.and let $h(1) = c$.

a) Show that for integer $k > 0$, $h(k) = kc$.

b) Show that for integer $j > 0$, $h(1/j) = c/j$.

c) Show that for all integer k, j, $h(k/j) = ck/j$.

d) The above parts show that $h(x)$ is linear in positive *rational* numbers. For very picky mathematicians, this does not guarantee that $h(x)$ is linear in positive *real* numbers. Show that if $h(x)$ is also monotonic in x, then $h(x)$ is linear in $x > 0$.

Exercise 2.7. Assume that a counting process $\{N(t);\, t{>}0\}$ has the independent and stationary increment properties and, for all $t > 0$, satisfies

$$\begin{aligned}
\Pr\Big\{\widetilde{N}(t, t+\delta) = 0\Big\} &= 1 - \lambda\delta + o(\delta) \\
\Pr\Big\{\widetilde{N}(t, t+\delta) = 1\Big\} &= \lambda\delta + o(\delta) \\
\Pr\Big\{\widetilde{N}(t, t+\delta) > 1\Big\} &= o(\delta).
\end{aligned}$$

a) Let $\mathsf{F}_1^{\mathsf{c}}(\tau) = \Pr\{N(\tau) = 0\}$ and show that $d\mathsf{F}_1^{\mathsf{c}}(\tau)/d\tau = -\lambda \mathsf{F}_1^{\mathsf{c}}(\tau)$.

b) Show that X_1, the time of the first arrival, is exponential with parameter λ.

c) Let $\mathsf{F}_n^{\mathsf{c}}(\tau) = \Pr\left\{\widetilde{N}(t, t+\tau) = 0 \mid S_{n-1} = t\right\}$ and show that $d\mathsf{F}_n^{\mathsf{c}}(\tau)/d\tau = -\lambda \mathsf{F}_n^{\mathsf{c}}(\tau)$.

d) Argue that X_n is exponential with parameter λ and independent of earlier arrival times.

Exercise 2.8. For a Poisson process, let $t > 0$ be arbitrary and let Z_1 be the duration of the interval from t until the next arrival after t. Let Z_m, for each $m > 1$, be the interarrival time from the epoch of the $m-1^{\text{st}}$ arrival after t until the mth arrival.

a) Given that $N(t) = n$, explain why $Z_m = X_{m+n}$ for $m > 1$ and $Z_1 = X_{n+1} - t + S_n$.

b) Conditional on $N(t) = n$ and $S_n = \tau$, show that $Z_1, Z_2, \ldots$ are IID.

c) Show that $Z_1, Z_2, \ldots$ are IID.

Exercise 2.9. Consider a "shrinking Bernoulli" approximation $N_\delta(m\delta) = Y_1 + \cdots + Y_m$ to a Poisson process as described in Subsection 2.2.5.

a) Show that

$$\Pr\{N_\delta(m\delta) = n\} = \binom{m}{n}(\lambda\delta)^n(1-\lambda\delta)^{m-n}.$$

b) Let $t = m\delta$, and let t be fixed for the remainder of the exercise. Explain why

$$\lim_{\delta\to 0} \Pr\{N_\delta(t) = n\} = \lim_{m\to\infty} \binom{m}{n}\left(\frac{\lambda t}{m}\right)^n \left(1 - \frac{\lambda t}{m}\right)^{m-n},$$

where the limit on the left is taken over values of δ that divide t.

c) Derive the following two equalities:

$$\lim_{m\to\infty} \binom{m}{n}\frac{1}{m^n} = \frac{1}{n!}; \qquad \text{and} \qquad \lim_{m\to\infty}\left(1 - \frac{\lambda t}{m}\right)^{m-n} = e^{-\lambda t}.$$

d) Conclude from this that for every t and every n, $\lim_{\delta\to 0} \Pr\{N_\delta(t){=}n\} = \Pr\{N(t){=}n\}$ where $\{N(t);\, t > 0\}$ is a Poisson process of rate λ.

Exercise 2.10. Let $\{N(t);\, t > 0\}$ be a Poisson process of rate λ.

a) Find the joint probability mass function (PMF) of $N(t)$, $N(t+s)$ for $s > 0$.

b) Find $\mathsf{E}\left[N(t) \cdot N(t+s)\right]$ for $s > 0$.

c) Find $\mathsf{E}\left[\widetilde{N}(t_1, t_3) \cdot \widetilde{N}(t_2, t_4)\right]$ where $\widetilde{N}(t, \tau)$ is the number of arrivals in $(t, \tau]$ and $t_1 < t_2 < t_3 < t_4$.

Exercise 2.11. An elementary experiment is independently performed N times where N is a Poisson rv of mean λ. Let $\{a_1, a_2, \ldots, a_K\}$ be the set of sample points of the elementary experiment and let p_k, $1 \leq k \leq K$, denote the probability of a_k.

a) Let N_k denote the number of elementary experiments performed for which the output is a_k. Find the PMF for N_k $(1 \leq k \leq K)$. (Hint: no calculation is necessary.)

b) Find the PMF for $N_1 + N_2$.

c) Find the conditional PMF for N_1 given that $N = n$.

d) Find the conditional PMF for $N_1 + N_2$ given that $N = n$.

e) Find the conditional PMF for N given that $N_1 = n_1$.

Exercise 2.12. Starting from time 0, northbound buses arrive at 77 Mass. Avenue according to a Poisson process of rate λ. Customers arrive according to an independent Poisson process of rate μ. When a bus arrives, all waiting customers instantly enter the bus and subsequent customers wait for the next bus.

a) Find the PMF for the number of customers entering a bus (more specifically, for any given m, find the PMF for the number of customers entering the mth bus).

b) Find the PMF for the number of customers entering the mth bus given that the interarrival interval between bus $m-1$ and bus m is x.

c) Given that a bus arrives at time 10:30 PM, find the PMF for the number of customers entering the next bus.

d) Given that a bus arrives at 10:30 PM and no bus arrives between 10:30 and 11, find the PMF for the number of customers on the next bus.

e) Find the PMF for the number of customers waiting at some given time, say 2:30 PM (assume that the processes started infinitely far in the past). Hint: think of what happens moving backward in time from 2:30 PM.

f) Find the PMF for the number of customers getting on the next bus to arrive after 2:30. Hint: this is different from (a); look carefully at (e).

g) Given that I arrive to wait for a bus at 2:30 PM, find the PMF for the number of customers getting on the next bus.

Exercise 2.13. a) Show that the arrival epochs of a Poisson process satisfy

$$\mathsf{f}_{\boldsymbol{S}^{(n)}|S_{n+1}}(\boldsymbol{s}^{(n)}|s_{n+1}) = n!/s_{n+1}^n.$$

Hint: This is easy if you use only the results of Section 2.2.2.

b) Contrast this with the result of Theorem 2.5.1

Exercise 2.14. Equation (2.42) gives $\mathsf{f}_{S_i|N(t)}(s_i \mid n)$, which is the density of random variable S_i conditional on $N(t) = n$ for $n \geq i$. Multiply this expression by $\Pr\{N(t) = n\}$ and sum over n to find $\mathsf{f}_{S_i}(s_i)$; verify that your answer is indeed the Erlang density.

Exercise 2.15. Consider generalizing the bulk arrival process in Figure 2.5. Assume that the epochs at which arrivals occur form a Poisson process $\{N(t);\, t > 0\}$ of rate λ. At each arrival epoch, S_n, the number of arrivals, Z_n, satisfies $\Pr\{Z_n{=}1\} = p$, $\Pr\{Z_n{=}2)\} = 1 - p$. The variables Z_n are IID.

a) Let $\{N_1(t); t > 0\}$ be the counting process of the epochs at which single arrivals occur. Find the PMF of $N_1(t)$ as a function of t. Similarly, let $\{N_2(t); t \geq 0\}$ be the counting process of the epochs at which double arrivals occur. Find the PMF of $N_2(t)$ as a function of t.

b) Let $\{N_B(t); t \geq 0\}$ be the counting process of the total number of arrivals. Give an expression for the PMF of $N_B(t)$ as a function of t.

Exercise 2.16. a) For a Poisson counting process of rate λ, find the joint probability density of $S_1, S_2, \ldots, S_{n-1}$ conditional on $S_n = t$.

b) Find $\Pr\{X_1 > \tau \mid S_n{=}t\}$.

c) Find $\Pr\{X_i > \tau \mid S_n{=}t\}$ for $1 \leq i \leq n$.

d) Find the density $\mathsf{f}_{S_i|S_n}(s_i|t)$ for $1 \leq i \leq n-1$.

e) Give an explanation for the striking similarity between the condition $N(t) = n - 1$ and the condition $S_n = t$.

Exercise 2.17. a) For a Poisson process of rate λ, find $\Pr\{N(t){=}n \mid S_1{=}\tau\}$ for $t > \tau$ and $n \geq 1$.

b) Using this, find $\mathsf{f}_{S_1}(\tau \mid N(t){=}n)$

c) Check your answer against (2.41).

Exercise 2.18. Consider a counting process in which the rate is a rv Λ with probability density $\mathsf{f}_\Lambda(\lambda) = \alpha e^{-\alpha\lambda}$ for $\lambda > 0$. Conditional on a given sample value λ for the rate, the counting process is a Poisson process of rate λ (i.e., nature first chooses a sample value λ and then generates a sample path of a Poisson process of that rate λ).

a) What is $\Pr\{N(t){=}n \mid \Lambda{=}\lambda\}$, where $N(t)$ is the number of arrivals in the interval $(0, t]$ for some given $t > 0$?

b) Show that $\Pr\{N(t){=}n\}$, the unconditional PMF for $N(t)$, is given by

$$\Pr\{N(t){=}n\} = \frac{\alpha t^n}{(t+\alpha)^{n+1}}.$$

c) Find $\mathsf{f}_\Lambda(\lambda \mid N(t){=}n)$, the density of λ conditional on $N(t){=}n$.

d) Find $\mathsf{E}[\Lambda \mid N(t){=}n]$ and interpret your result for very small t with $n = 0$ and for very large t with n large.

e) Find $\mathsf{E}[\Lambda \mid N(t){=}n, S_1, S_2, \ldots, S_n]$. (Hint: consider the distribution of $S_1, \ldots, S_n$ conditional on $N(t)$ and Λ). Find $\mathsf{E}[\Lambda \mid N(t){=}n, N(\tau){=}m]$ for some $\tau < t$.

Exercise 2.19. a) Use Equation (2.42) to find $\mathsf{E}[S_i \mid N(t){=}n]$. Hint: When you integrate $s_i \mathsf{f}_{S_i}(s_i \mid N(t){=}n)$, compare this integral with $\mathsf{f}_{S_{i+1}}(s_i \mid N(t){=}n+1)$ and use the fact that the latter expression is a probability density.

b) Find the second moment and the variance of S_i conditional on $N(t){=}n$. Hint: Extend the previous hint.

c) Assume that n is odd, and consider $i = (n+1)/2$. What is the relationship between S_i, conditional on $N(t){=}n$, and the sample median of n IID uniform random variables.

d) Give a weak law of large numbers for the above median.

Exercise 2.20. Suppose cars enter a one-way infinite length, infinite lane highway at a Poisson rate λ. The ith car to enter chooses a velocity V_i and travels at this velocity. Assume that the V_i's are independent positive rv's having a common distribution F. Derive the distribution of the number of cars that are located in an interval $(0, a)$ at time t.

Exercise 2.21. Consider an M/G/∞ queue, i.e., a queue with Poisson arrivals of rate λ in which each arrival i, independent of other arrivals, remains in the system for a time X_i, where $\{X_i; i \geq 1\}$ is a set of IID rv's with some given CDF $\mathsf{F}(x)$.

You may assume that the number of arrivals in any interval $(t, t+\epsilon)$ that are still in the system at some later time $\tau \geq t+\epsilon$ is statistically independent of the number of arrivals in that same interval $(t, t+\epsilon)$ that have departed from the system by time τ.

a) Let $N(\tau)$ be the number of customers in the system at time τ. Find the mean, $m(\tau)$, of $N(\tau)$ and find $\Pr\{N(\tau) = n\}$.

b) Let $D(\tau)$ be the number of customers that have departed from the system by time τ. Find the mean, $\mathsf{E}[D(\tau)]$, and find $\Pr\{D(\tau) = d\}$.

c) Find $\Pr\{N(\tau) = n, D(\tau) = d\}$.

d) Let $A(\tau)$ be the total number of arrivals up to time τ. Find $\Pr\{N(\tau) = n \mid A(\tau) = a\}$.

e) Find $\Pr\{D(\tau+\epsilon) - D(\tau) = d\}$.

Exercise 2.22. The voters in a given town arrive at the place of voting according to a Poisson process of rate $\lambda = 100$ voters per hour. The voters independently vote for candidate A and candidate B each with probability 1/2. Assume that the voting starts at time 0 and continues indefinitely.

a) Conditional on 1000 voters arriving during the first 10 hours of voting, find the probability that candidate A receives n of those votes.

b) Again conditional on 1000 voters during the first 10 hours, find the probability that candidate A receives n votes in the first 4 hours of voting.

c) Let T be the epoch of the arrival of the first voter voting for candidate A. Find the density of T.

d) Find the PMF of the number of voters for candidate B who arrive before the first voter for A.

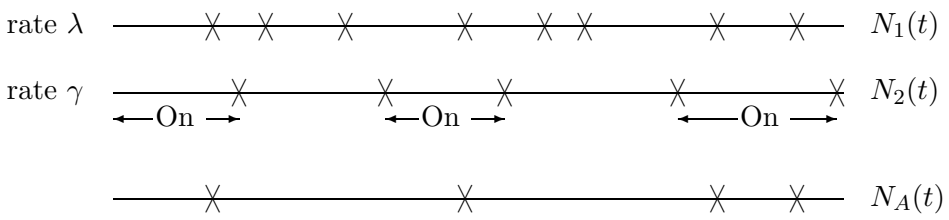
e) Define the nth voter as a *reversal* if the nth voter votes for a different candidate than the $n-1^{st}$. For example, in the sequence of votes $AABAABB$, the third, fourth, and sixth voters are reversals; the third and sixth are A to B reversals and the fourth is a B to A reversal. Let $N(t)$ be the number of reversals up to time t (t in hours). Is $\{N(t);\, t>0\}$ a Poisson process? Explain.

f) Find the expected time (in hours) between reversals.

g) Find the probability density of the time between reversals.

h) Find the density of the time from one A to B reversal to the next A to B reversal.

Exercise 2.23. Let $\{N_1(t);\, t>0\}$ be a Poisson counting process of rate λ. Assume that the arrivals from this process are switched on and off by arrivals from a second independent Poisson process $\{N_2(t); t>0\}$ of rate γ.



Let $\{N_A(t); t\geq 0\}$ be the switched process; that is $N_A(t)$ includes the arrivals from $\{N_1(t);\, t>0\}$ during periods when $N_2(t)$ is even and excludes the arrivals from $\{N_1(t);\, t>0\}$ while $N_2(t)$ is odd.

a) Find the PMF for the number of arrivals of the first process, $\{N_1(t);\, t>0\}$, during the nth period when the switch is on.

b) Given that the first arrival for the second process occurs at epoch τ, find the conditional PMF for the number of arrivals of the first process up to τ.

c) Given that the number of arrivals of the first process, up to the first arrival for the second process, is n, find the density for the epoch of the first arrival from the second process.

d) Find the density of the interarrival time for $\{N_A(t); t\geq 0\}$. Note: This part is quite messy and is done most easily via Laplace transforms.

Exercise 2.24. Let us model the chess tournament between Fisher and Spassky as a stochastic process. Let X_i, for $i\geq 1$, be the duration of the ith game and assume that $\{X_i;\, i\geq 1\}$ is a set of IID exponentially distributed rv's each with density $\mathsf{f}_X(x)=\lambda e^{-\lambda x}$. Suppose that each game (independently of all other games, and independently of the length of the games) is won by Fisher with probability p, by Spassky with probability q, and is a draw with probability $1-p-q$. The first player to win n games is defined to be the winner, but we consider the match up to the point of winning as being embedded in an unending sequence of games.

a) Find the distribution of time, from the beginning of the match, until the completion of the first game that is won (i.e., that is not a draw). Characterize the process of the number $\{N(t);\, t > 0\}$ of games won up to and including time t. Characterize the process of the number $\{N_F(t); t \geq 0\}$ of games won by Fisher and the number $\{N_S(t); t \geq 0\}$ won by Spassky.

b) For the remainder of the problem, assume that the probability of a draw is zero; i.e., that $p + q = 1$. How many of the first $2n - 1$ games must be won by Fisher in order to win the match?

c) What is the probability that Fisher wins the match? Your answer should not involve any integrals. Hint: consider the unending sequence of games and use (b).

d) Let T be the epoch at which the match is completed (i.e., either Fisher or Spassky wins). Find the CDF of T.

e) Find the probability that Fisher wins and that T lies in the interval $(t, t+\delta)$ for arbitrarily small δ.

Exercise 2.25. a) Find the conditional density of S_{i+1}, conditional on $N(t) = n$ and $S_i = s_i$.

b) Use part a) to find the joint density of $S_1, \ldots, S_n$ conditional on $N(t) = n$. Verify that your answer agrees with (2.38).

Exercise 2.26. A two-dimensional Poisson process is a process of randomly occurring special points in the plane such that (i) for any region of area A the number of special points in that region has a Poisson distribution with mean λA, and (ii) the number of special points in nonoverlapping regions is independent. For such a process consider an arbitrary location in the plane and let X denote its distance from its nearest special point (where distance is measured in the usual Euclidean manner). Show that

a) $\Pr\{X > t\} = \exp(-\lambda\pi t^2)$

b) $\mathsf{E}[X] = 1/(2\sqrt{\lambda})$.

Exercise 2.27. This problem is intended to show that one can analyze the long term behavior of queueing problems by using just notions of means and variances, but that such analysis is awkward, justifying understanding the strong law of large numbers. Consider an M/G/1 queue. The arrival process is Poisson with $\lambda = 1$. The expected service time, $\mathsf{E}[Y]$, is 1/2 and the variance of the service time is given to be 1.

a) Consider S_n, the time of the nth arrival, for $n = 10^{12}$. With high probability, S_n will lie within 3 standard derivations of its mean. Find and compare this mean and the 3σ range.

b) Let V_n be the total amount of time during which the server is busy with these n arrivals (i.e., the sum of 10^{12} service times). Find the mean and 3σ range of V_n.

c) Find the mean and 3σ range of I_n, the total amount of time the server is idle up until S_n (take I_n as $S_n - V_n$, thus ignoring any service time after S_n).

d) An idle period starts when the server completes a service and there are no waiting arrivals; it ends on the next arrival. Find the mean and variance of an idle period. Are successive idle periods IID?

e) Combine (c) and (d) to estimate the total number of idle periods up to time S_n. Use this to estimate the total number of busy periods.

f) Combine (e) and (b) to estimate the expected length of a busy period.

Exercise 2.28. The purpose of this problem is to illustrate that for an arrival process with independent but not identically distributed interarrival intervals, $X_1, X_2, \ldots$ the number of arrivals $N(t)$ in the interval $(0, t]$ can be a defective rv. In other words, the 'counting process' is not necessarily a stochastic process according to our definitions. This also suggests that it might be necessary to prove that the counting rv's for a renewal process are actually rv's.

a) Let the CDF of the ith interarrival interval for an arrival process be $\mathsf{F}_{X_i}(x_i) = 1 - \exp(-\alpha^{-i} x_i)$ for some fixed $\alpha \in (0, 1)$. Let $S_n = X_1 + \cdots X_n$ and show that

$$\mathsf{E}[S_n] = \frac{\alpha(1-\alpha^n)}{1-\alpha}.$$

b) Sketch a 'reasonable' sample path for $N(t)$.

c) Use the Markov inequality on $\Pr\{S_n \geq t\}$ to find an upper bound on $\Pr\{N(t) < n\}$ that is smaller than 1 for all n and for large enough t. Use this to show that $N(t)$ is defective for large enough t.

d) (For those looking for a challenge) Show that $N(t)$ is defective for all $t > 0$. Hint: Use the Markov inequality to find an upper bound on $\Pr\{S_m - S_n \leq t/2\}$ for all $m > n$ for any fixed n. Show that, for any $t > 0$, this is bounded below 1 for large enough n. Then show that S_n has a density that is positive for all $t > 0$.

Chapter 3

FINITE-STATE MARKOV CHAINS

3.1 Introduction

The counting processes $\{N(t);\, t > 0\}$ described in Section 2.1.1 have the property that $N(t)$ *changes* at discrete instants of time, but is *defined* for all real $t > 0$. The Markov chains to be discussed in this chapter are stochastic processes *defined* only at integer values of time, $n = 0, 1, \ldots$. At each integer time $n \geq 0$, there is an integer-valued random variable (rv) X_n, called the *state* at time n, and the process is the family of rv's $\{X_n;\, n \geq 0\}$. We refer to these processes as *integer-time processes*. An integer-time process $\{X_n;\, n \geq 0\}$ can also be viewed as a process $\{X(t);\, t \geq 0\}$ defined for all real t by taking $X(t) = X_n$ for $n \leq t < n + 1$, but since changes occur only at integer times, it is usually simpler to view the process only at those integer times.

In general, for Markov chains, the set of possible values for each rv X_n is a countable set $\mathcal{S}$. If $\mathcal{S}$ is countably infinite, it is usually taken to be $\mathcal{S} = \{0, 1, 2, \ldots\}$, whereas if $\mathcal{S}$ is finite, it is usually taken to be $\mathcal{S} = \{1, \ldots, \mathsf{M}\}$. In this chapter (except for Theorems 3.2.2 and 3.2.3), we restrict attention to the case in which $\mathcal{S}$ is finite, *i.e.*, processes whose sample functions are sequences of integers, each between 1 and M. There is no special significance to using integer labels for states, and no compelling reason to include 0 for the countably infinite case and not for the finite case. For the countably infinite case, the most common applications come from queueing theory, where the state often represents the number of waiting customers, which might be zero. For the finite case, we often use vectors and matrices, where positive integer labels simplify the notation. In some examples, it will be more convenient to use more illustrative labels for states.

Definition 3.1.1. *A Markov chain is an integer-time process, $\{X_n, n \geq 0\}$ for which the sample values for each rv X_n, $n \geq 1$, lie in a countable set $\mathcal{S}$ and depend on the past only through the most recent rv X_{n-1}. More specifically, for all positive integers n, and for all $i, j, k, \ldots, m$ in $\mathcal{S}$,*

$$\Pr\{X_n{=}j \mid X_{n-1}{=}i, X_{n-2}{=}k, \ldots, X_0{=}m\} = \Pr\{X_n{=}j \mid X_{n-1}{=}i\}\,. \tag{3.1}$$

Furthermore, $\Pr\{X_n{=}j \mid X_{n-1}{=}i\}$ *depends only on* i *and* j *(not* n*) and is denoted by*

$$\Pr\{X_n{=}j \mid X_{n-1}{=}i\} = P_{ij}. \tag{3.2}$$

The initial state X_0 *has an arbitrary probability distribution. A finite-state Markov chain is a Markov chain in which* $\mathcal{S}$ *is finite.*

Equations such as (3.1) are often easier to read if they are abbreviated as

$$\Pr\{X_n \mid X_{n-1}, X_{n-2}, \ldots, X_0\} = \Pr\{X_n \mid X_{n-1}\}.$$

This abbreviation means that equality holds for all sample values of each of the rv's. *i.e.*, it means the same thing as (3.1).

The rv X_n is called the state of the chain at time n. The possible values for the state at time n, namely $\{1, \ldots, \mathsf{M}\}$ or $\{0, 1, \ldots\}$ are also generally called states, usually without too much confusion. Thus P_{ij} is the probability of going to state j given that the previous state is i; the new state, given the previous state, is independent of all earlier states. The use of the word *state* here conforms to the usual idea of the state of a system — the state at a given time summarizes everything about the past that is relevant to the future.

Definition 3.1.1 is used by some people as the definition of a *homogeneous Markov chain*. For them, Markov chains include more general cases where the transition probabilities can vary with n. Thus they replace (3.1) and (3.2) by

$$\Pr\{X_n{=}j \mid X_{n-1}{=}i, X_{n-2}{=}k, \ldots, X_0{=}m\} = \Pr\{X_n{=}j \mid X_{n-1}{=}i\} = P_{ij}(n). \tag{3.3}$$

We will call a process that obeys (3.3), with a dependence on n, a *non-homogeneous Markov chain*. We will discuss only the homogeneous case, with no dependence on n, and thus restrict the definition to that case. Not much of general interest can be said about non-homogeneous chains.[1]

An initial probability distribution for X_0, combined with the transition probabilities $\{P_{ij}\}$ (or $\{P_{ij}(n)\}$ for the non-homogeneous case), define the probabilities for all events in the Markov chain.

Markov chains can be used to model an enormous variety of physical phenomena and can be used to approximate many other kinds of stochastic processes such as the following example:

Example 3.1.1. Consider an integer process $\{Z_n;\ n \geq 0\}$ where the Z_n are finite integer-valued rv's as in a Markov chain, but each Z_n depends probabilistically on the previous k rv's, $Z_{n-1}, Z_{n-2}, \ldots, Z_{n-k}$. In other words, using abbreviated notation,

$$\Pr\{Z_n \mid Z_{n-1}, Z_{n-2}, \ldots, Z_0\} = \Pr\{Z_n \mid Z_{n-1}, \ldots Z_{n-k}\}. \tag{3.4}$$

[1] On the other hand, we frequently find situations where a small set of rv's, say W, X, Y, Z satisfy the *Markov condition* that $\Pr\{Z \mid Y, X, W\} = \Pr\{Z \mid Y\}$ and $\Pr\{Y \mid X, W\} = \Pr\{Y \mid X\}$ but where the conditional distributions $\Pr\{Z \mid Y\}$ and $\Pr\{Y \mid X\}$ are unrelated. In other words, *Markov chains* imply homogeneity here, whereas the *Markov condition* does not.

We now show how to view the condition on the right side of (3.4), *i.e.*, $(Z_{n-1}, Z_{n-2}, \ldots, Z_{n-k})$ as the state of the process at time $n-1$. We can rewrite (3.4) as

$$\Pr\{Z_n, Z_{n-1}, \ldots, Z_{n-k+1} \mid Z_{n-1}, \ldots, Z_0\} = \Pr\{Z_n, \ldots, Z_{n-k+1} \mid Z_{n-1}, \ldots Z_{n-k}\},$$

since, for each side of the equation, any given set of values for $Z_{n-1}, \ldots, Z_{n-k+1}$ on the right side of the conditioning sign specifies those values on the left side. Thus if we define $X_{n-1} = (Z_{n-1}, \ldots, Z_{n-k})$ for each n, this simplifies to

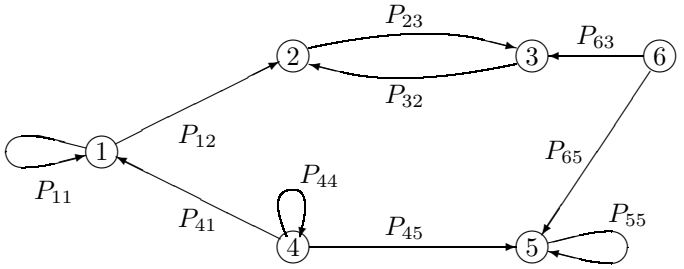
$$\Pr\{X_n \mid X_{n-1}, \ldots, X_{k-1}\} = \Pr\{X_n \mid X_{n-1}\}.$$

We see that by expanding the state space to include k-tuples of the rv's Z_n, we have converted the k dependence over time to a unit dependence over time, *i.e.*, a Markov process is defined using the expanded state space.

Note that in this new Markov chain, the initial state is $X_{k-1} = (Z_{k-1}, \ldots, Z_0)$, so one might want to shift the time axis to start with X_0.

Markov chains are often described by a directed graph (see Figure 3.1 a). In this graphical representation, there is one node for each state and a directed arc for each non-zero transition probability. If$P_{ij} = 0$, then the arc from node i to node j is omitted, so the difference between zero and non-zero transition probabilities stands out clearly in the graph. The classification of states, as discussed in Section 3.2, is determined by the set of transitions with non-zero probabilities, and thus the graphical representation is ideal for that topic.

A finite-state Markov chain is also often described by a matrix $[P]$ (see Figure 3.1 b). If the chain has M states, then $[P]$ is an M by M matrix with elements P_{ij}. The matrix representation is ideally suited for studying algebraic and computational issues.



$$[P] = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{16} \\ P_{21} & P_{22} & \cdots & P_{26} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ P_{61} & P_{62} & \cdots & P_{66} \end{bmatrix}$$

a) Graphical **b)** Matrix

Figure 3.1: Graphical and Matrix Representation of a 6 state Markov Chain; a directed arc from i to j is included in the graph if and only if (iff) $P_{ij} > 0$.

3.2 Classification of states

This section, except where indicated otherwise, applies to Markov chains with both finite and countable state spaces. We start with several definitions.

Definition 3.2.1. *An (n-step) walk is an ordered string of nodes, $(i_0, i_1, \dots i_n)$, $n \geq 1$, in which there is a directed arc from i_{m-1} to i_m for each m, $1 \leq m \leq n$. A path is a walk in which no nodes are repeated. A cycle is a walk in which the first and last nodes are the same and no other node is repeated.*

Note that a walk can start and end on the same node, whereas a path cannot. Also the number of steps in a walk can be arbitrarily large, whereas a path can have at most $\mathsf{M} - 1$ steps and a cycle at most M steps for a finite-state Markov chain with $|\mathcal{S}| = \mathsf{M}$.

Definition 3.2.2. *A state j is* accessible *from i (abbreviated as $i \to j$) if there is a walk in the graph from i to j.*

For example, in Figure 3.1(a), there is a walk from node 1 to node 3 (passing through node 2), so state 3 is accessible from 1. There is no walk from node 5 to 3, so state 3 is not accessible from 5. State 2 is accessible from itself, but state 6 is not accessible from itself. To see the probabilistic meaning of accessibility, suppose that a walk $i_0, i_1, \dots i_n$ exists from node i_0 to i_n. Then, conditional on $X_0 = i_0$, there is a positive probability, $P_{i_0 i_1}$, that $X_1 = i_1$, and consequently (since $P_{i_1 i_2} > 0$), there is a positive probability that $X_2 = i_2$. Continuing this argument, there is a positive probability that $X_n = i_n$, so that $\Pr\{X_n{=}i_n \mid X_0{=}i_0\} > 0$. Similarly, if $\Pr\{X_n{=}i_n \mid X_0{=}i_0\} > 0$, then an n-step walk from i_0 to i_n must exist. Summarizing, $i \to j$ if and only if (iff) $\Pr\{X_n{=}j \mid X_0{=}i\} > 0$ for some $n \geq 1$. We denote $\Pr\{X_n{=}j \mid X_0{=}i\}$ by P_{ij}^n. Thus, for $n \geq 1$, $P_{ij}^n > 0$ if and only if the graph has an n step walk from i to j (perhaps visiting the same node more than once). For the example in Figure 3.1(a), $P_{13}^2 = P_{12}P_{23} > 0$. On the other hand, $P_{53}^n = 0$ for all $n \geq 1$. An important relation that we use often in what follows is that if there is an n-step walk from state i to j and an m-step walk from state j to k, then there is a walk of $m + n$ steps from i to k. Thus

$$P_{ij}^n > 0 \text{ and } \mathrm{P_{jk}^m} > 0 \quad \text{imply} \quad \mathrm{P_{ik}^{n+m}} > 0. \tag{3.5}$$

This also shows that

$$i \to j \text{ and } \mathrm{j} \to \mathrm{k} \quad \text{imply} \quad \mathrm{i} \to \mathrm{k}. \tag{3.6}$$

Definition 3.2.3. *Two distinct states i and j* communicate *(abbreviated $i \leftrightarrow j$) if i is accessible from j and j is accessible from i.*

An important fact about communicating states is that if $i \leftrightarrow j$ and $m \leftrightarrow j$ then $i \leftrightarrow m$. To see this, note that $i \leftrightarrow j$ and $m \leftrightarrow j$ imply that $i \to j$ and $j \to m$, so that $i \to m$. Similarly, $m \to i$, so $i \leftrightarrow m$.

Definition 3.2.4. *A* class *$\mathcal{C}$ of states is a non-empty set of states such that each $i \in \mathcal{C}$ communicates with every other state $j \in \mathcal{C}$ and communicates with no $j \notin \mathcal{C}$.*

For the example of Figure 3.1(a), $\{2, 3\}$ is one class of states, $\{1\}$, $\{4\}$, $\{5\}$, and $\{6\}$ are the other classes. Note that state 6 does not communicate with any other state, and is not even accessible from itself, but the set consisting of $\{6\}$ alone is still a class. The entire set of states in a given Markov chain is partitioned into one or more disjoint classes in this way.

Definition 3.2.5. *For finite-state Markov chains, a* recurrent state *is a state i that is accessible from all states that are accessible from i (i is recurrent if $i \to j$ implies that $j \to i$). A* transient state *is a state that is not recurrent.*

Recurrent and transient states for Markov chains with a countably-infinite state space will be defined in Chapter 5.

According to the definition, a state i in a finite-state Markov chain is recurrent if there is no possibility of going to a state j from which there can be no return. As we shall see later, if a Markov chain ever enters a recurrent state, it returns to that state eventually with probability 1, and thus keeps returning infinitely often (in fact, this property serves as the definition of recurrence for Markov chains without the finite-state restriction). A state i is transient if there is some j that is accessible from i but from which there is no possible return. Each time the system returns to i, there is a possibility of going to j; eventually this possibility will occur with no further returns to i.

Theorem 3.2.1. *For finite-state Markov chains, either all states in a class are transient or all are recurrent.*[2]

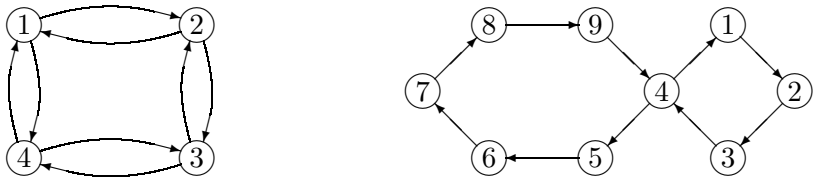
Proof: Assume that state i is transient (i.e., for some j, $i \to j$ but $j \not\to i$) and suppose that i and m are in the same class (i.e., $i \leftrightarrow m$). Then $m \to i$ and $i \to j$, so $m \to j$. Now if $j \to m$, then the walk from j to m could be extended to i; this is a contradiction, and therefore there is no walk from j to m, and m is transient. Since we have just shown that all nodes in a class are transient if any are, it follows that the states in a class are either all recurrent or all transient. □

For the example of Figure 3.1(a), $\{2, 3\}$ and $\{5\}$ are recurrent classes and the other classes are transient. In terms of the graph of a Markov chain, a class is transient if there are any directed arcs going from a node in the class to a node outside the class. Every finite-state Markov chain must have at least one recurrent class of states (see Exercise 3.2), and can have arbitrarily many additional classes of recurrent states and transient states.

States can also be classified according to their periods (see Figure 3.2). For $X_0 = 2$ in Figure 3.2(a), X_n must be 2 or 4 for n even and 1 or 3 for n odd. On the other hand, if X_0 is 1 or 3, then X_n is 2 or 4 for n odd and 1 or 3 for n even. Thus the effect of the starting state never dies out. Figure 3.2(b) illustrates another example in which the memory of the starting state never dies out. The states in both of these Markov chains are said to be periodic with period 2. Another example of periodic states are states 2 and 3 in Figure 3.1(a).

Definition 3.2.6. *The* period *of a state i, denoted $d(i)$, is the greatest common divisor (gcd) of those values of n for which $P_{ii}^n > 0$. If the period is 1, the state is aperiodic, and if the period is 2 or more, the state is periodic.*

[2] As shown in Chapter 5, this theorem is also true for Markov chains with a countably infinite state space, but the proof given here is inadequate. Also recurrent classes with a countably infinite state space are further classified into either *positive-recurrent* or *null-recurrent*, a distinction that does not appear in the finite-state case.



(a) (b)

Figure 3.2: Periodic Markov Chains

For example, in Figure 3.2(a), $P_{11}^n > 0$ for $n = 2, 4, 6, \ldots$. Thus $d(1)$, the period of state 1, is two. Similarly, $d(i) = 2$ for the other states in Figure 3.2(a). For Figure 3.2(b), we have $P_{11}^n > 0$ for $n = 4, 8, 10, 12, \ldots$; thus $d(1) = 2$, and it can be seen that $d(i) = 2$ for all the states. These examples suggest the following theorem.

Theorem 3.2.2. *For any Markov chain (with either a finite or countably infinite number of states), all states in the same class have the same period.*

Proof: Let i and j be any distinct pair of states in a class $\mathcal{C}$. Then $i \leftrightarrow j$ and there is some r such that $P_{ij}^r > 0$ and some s such that $P_{ji}^s > 0$. Since there is a walk of length $r + s$ going from i to j and back to i, $r + s$ must be divisible by $d(i)$. Let t be any integer such that $P_{jj}^t > 0$. Since there is a walk of length $r + t + s$ from i o j, then back to j, and then to i, $r + t + s$ is divisible by $d(i)$, and thus t is divisible by $d(i)$. Since this is true for any t such that $P_{jj}^t > 0$, $d(j)$ is divisible by $d(i)$. Reversing the roles of i and j, $d(i)$ is divisible by $d(j)$, so $d(i) = d(j)$. □

Since the states in a class $\mathcal{C}$ all have the same period and are either all recurrent or all transient, we refer to $\mathcal{C}$ itself as having the period of its states and as being recurrent or transient. Similarly if a Markov chain has a single class of states, we refer to the chain as having the corresponding period.

Theorem 3.2.3. *If a recurrent class $\mathcal{C}$ in a finite-state Markov chain has period d, then the states in $\mathcal{C}$ can be partitioned into d subsets, $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_d$, in such a way that all transitions from $\mathcal{S}_1$ go to $\mathcal{S}_2$, all from $\mathcal{S}_2$ go to $\mathcal{S}_3$, and so forth up to $\mathcal{S}_{m-1}$ to $\mathcal{S}_m$. Finally, all transitions from $\mathcal{S}_m$ go to $\mathcal{S}_1$.*

Proof: See Figure 3.3 for an illustration of the theorem. For a given state in $\mathcal{C}$, say state 1, define the sets $\mathcal{S}_1, \ldots, \mathcal{S}_d$ by

$$\mathcal{S}_m = \{j : P_{1j}^{nd+m} > 0 \text{ for some } n \geq 0\}; \quad 1 \leq m \leq d. \tag{3.7}$$

For each $j \in \mathcal{C}$, we first show that there is one and only one value of m such that $j \in \mathcal{S}_m$. Since $1 \leftrightarrow j$, there is some r for which $P_{1j}^r > 0$ and some s for which $P_{j1}^s > 0$. Thus there is a walk from 1 to 1 (through j) of length $r + s$, so $r + s$ is divisible by d. For the given r,

let m, $1 \leq m \leq d$, satisfy $r = m + nd$, where n is an integer. From (3.7), $j \in \mathcal{S}_m$. Now let r' be any other integer such that $P_{1j}^{r'} > 0$. Then $r' + s$ is also divisible by d, so that $r' - r$ is divisible by d. Thus $r' = m + n'd$ for some integer n' and that same m. Since r' is *any* integer such that $P_{1j}^{r'} > 0$, j is in $\mathcal{S}_m$ for only that one value of m. Since j is arbitrary, this shows that the sets $\mathcal{S}_m$ are disjoint and partition $\mathcal{C}$.

Finally, suppose $j \in \mathcal{S}_m$ and $P_{jk} > 0$. Given a walk of length $r = nd + m$ from state 1 to j, there is a walk of length $nd + m + 1$ from state 1 to k. It follows that if $m < d$, then $k \in \mathcal{S}_{m+1}$ and if $m = d$, then $k \in \mathcal{S}_1$, completing the proof. □

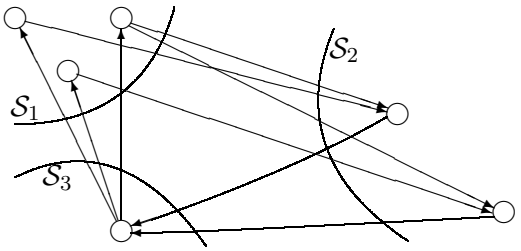


Figure 3.3: Structure of a periodic Markov chain with $d = 3$. Note that transitions only go from one subset $\mathcal{S}_m$ to the next subset $\mathcal{S}_{m+1}$ (or from $\mathcal{S}_d$ to $\mathcal{S}_1$).

We have seen that each class of states (for a finite-state chain) can be classified both in terms of its period and in terms of whether or not it is recurrent. The most important case is that in which a class is both recurrent and aperiodic.

Definition 3.2.7. *For a finite-state Markov chain, an ergodic class of states is a class that is both recurrent and aperiodic*[3]*. A Markov chain consisting entirely of one ergodic class is called an ergodic chain.*

We shall see later that these chains have the desirable property that P_{ij}^n becomes independent of the starting state i as $n \to \infty$. The next theorem establishes the first part of this by showing that $P_{ij}^n > 0$ for all i and j when n is sufficiently large. A guided proof is given in Exercise 3.5.

Theorem 3.2.4. *For an ergodic M state Markov chain, $P_{ij}^m > 0$ for all i, j, and all $m \geq (\mathsf{M}-1)^2 + 1$.*

Figure 3.4 illustrates a situation where the bound $(\mathsf{M}-1)^2 + 1$ is met with equality. Note that there is one cycle of length $\mathsf{M} - 1$ and the single node not on this cycle, node 1, is the unique starting node at which the bound is met with equality.

[3]For Markov chains with a countably infinite state space, ergodic means that the states are positive-recurrent and aperiodic (see Chapter 5, Section 5.1).

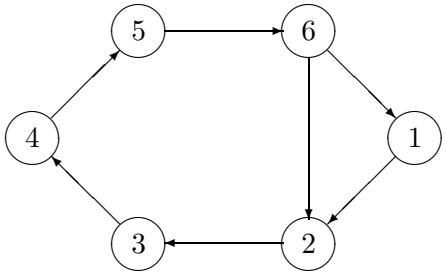


Figure 3.4: An ergodic chain with $\mathsf{M} = 6$ states in which $P_{ij}^m > 0$ for all $m > (\mathsf{M}-1)^2$ and all i, j but $P_{11}^{(\mathsf{M}-1)^2} = 0$ The figure also illustrates that an M state Markov chain must have a cycle with $\mathsf{M}-1$ or fewer nodes. To see this, note that an ergodic chain must have cycles, since each node must have a walk to itself, and subcycles of repeated nodes can be omitted from that walk, converting it into a cycle. Such a cycle might have M nodes, but a chain with only an M node cycle would be periodic. Thus some nodes must be on smaller cycles, such as the cycle of length 5 in the figure.

3.3 The matrix representation

The matrix $[P]$ of transition probabilities of a Markov chain is called a stochastic matrix; that is, a *stochastic matrix* is a square matrix of nonnegative terms in which the elements in each row sum to 1. We first consider the n step transition probabilities P_{ij}^n in terms of [P]. The probability, starting in state i, of going to state j in two steps is the sum over k of the probability of going first to k and then to j. Using the Markov condition in (3.1),

$$P_{ij}^2 = \sum_{k=1}^{\mathsf{M}} P_{ik} P_{kj}.$$

It can be seen that this is just the ij term of the product of the matrix $[P]$ with itself; denoting $[P][P]$ as $[P^2]$, this means that P_{ij}^2 is the (i,j) element of the matrix $[P^2]$. Similarly, P_{ij}^n is the ij element of the nth power of the matrix $[P]$. Since $[P^{m+n}] = [P^m][P^n]$, this means that

$$P_{ij}^{m+n} = \sum_{k=1}^{\mathsf{M}} P_{ik}^m \, P_{kj}^n. \tag{3.8}$$

This is known as the *Chapman-Kolmogorov* equation. An efficient approach to compute $[P^n]$ (and thus P_{ij}^n) for large n, is to multiply $[P]^2$ by $[P]^2$, then $[P]^4$ by $[P]^4$ and so forth. Then $[P], [P^2], [P^4], \ldots$ can be multiplied as needed to get $[P^n]$.

3.3.1 Steady state and $[P^n]$ for large n

The matrix $[P^n]$ (i.e., the matrix of transition probabilities raised to the nth power) is important for a number of reasons. The i, j element of this matrix is $P_{ij}^n = \Pr\{X_n{=}j \mid X_0{=}i\}$. If memory of the past dies out with increasing n, then we would expect the dependence of P_{ij}^n on both n and i to disappear as $n \to \infty$. This means, first, that $[P^n]$ should converge to a limit as $n \to \infty$, and, second, that for each column j, the elements in that column, $P_{1j}^n, P_{2j}^n, \ldots, P_{\mathsf{M}j}^n$ should all tend toward the same value, say π_j, as $n \to \infty$. If this type of convergence occurs, (and we later determine the circumstances under which it occurs), then $P_{ij}^n \to \pi_j$ and each row of the limiting matrix will be $(\pi_1, \ldots, \pi_{\mathsf{M}})$, *i.e.*, each row is the same as each other row.

If we now look at the equation $P_{ij}^{n+1} = \sum_k P_{ik}^n P_{kj}$, and assume the above type of convergence as $n \to \infty$, then the limiting equation becomes $\pi_j = \sum \pi_k P_{kj}$. In vector form, this equation is $\boldsymbol{\pi} = \boldsymbol{\pi}[P]$. We will do this more carefully later, but what it says is that if P_{ij}^n approaches a limit denoted π_j as $n \to \infty$, then $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{\mathsf{M}})$ satisfies $\boldsymbol{\pi} = \boldsymbol{\pi}[P]$. If nothing else, it is easier to solve the linear equations $\boldsymbol{\pi} = \boldsymbol{\pi}[P]$ than to multiply $[P]$ by itself an infinite number of times.

Definition 3.3.1. *A steady-state vector (or a steady-state distribution) for an M state Markov chain with transition matrix $[P]$ is a row vector $\boldsymbol{\pi}$ that satisfies*

$$\boldsymbol{\pi} = \boldsymbol{\pi}[P] \; ; \quad \text{where } \sum_i \pi_i = 1 \text{ and } \pi_i \geq 0 \, , \, 1 \leq i \leq \mathsf{M}. \tag{3.9}$$

If $\boldsymbol{\pi}$ satisfies (3.9), then the last half of the equation says that it must be a probability vector. If $\boldsymbol{\pi}$ is taken as the initial PMF of the chain at time 0, then that PMF is maintained forever. That is, post-multiplyng both sides of (3.9) by $[P]$, we get $\boldsymbol{\pi}[P] = \boldsymbol{\pi}[P^2]$, and iterating this, $\boldsymbol{\pi} = \boldsymbol{\pi}[P^2] = \boldsymbol{\pi}[P^3] = \cdots$ for all n.

It is important to recognize that we have shown that if $[P^n]$ converges to a matrix all of whose rows are $\boldsymbol{\pi}$, then $\boldsymbol{\pi}$ is a steady-state vector, *i.e.*, it satisfies (3.9). However, finding a $\boldsymbol{\pi}$ that satisfies (3.9) does not imply that $[P^n]$ converges as $n \to \infty$. For the example of Figure 3.1, it can be seen that if we choose $\pi_2 = \pi_3 = 1/2$ with $\pi_i = 0$ otherwise, then $\boldsymbol{\pi}$ is a steady-state vector. Reasoning more physically, we see that if the chain is in either state 2 or 3, it simply oscillates between those states for all time. If it starts at time 0 being in states 2 or 3 with equal probability, then it persists forever being in states 2 or 3 with equal probability. Although this choice of $\boldsymbol{\pi}$ satisfies the definition in (3.9) and also is a steady-state distribution in the sense of not changing over time, it is not a very satisfying form of steady state, and almost seems to be concealing the fact that we are dealing with a simple oscillation between states.

This example raises one of a number of questions that should be answered concerning steady-state distributions and the convergence of $[P^n]$:

1. Under what conditions does $\boldsymbol{\pi} = \boldsymbol{\pi}[P]$ have a probability vector solution?
2. Under what conditions does $\boldsymbol{\pi} = \boldsymbol{\pi}[P]$ have a unique probability vector solution?

3. Under what conditions does each row of $[P^n]$ converge to a probability vector solution of $\boldsymbol{\pi} = \boldsymbol{\pi}[P]$?

We first give the answers to these questions for finite-state Markov chains and then derive them. First, (3.9) *always* has a solution (although this is not necessarily true for infinite-state chains). The answers to the second and third questions are simplified if we use the following definition:

Definition 3.3.2. *A unichain is a finite-state Markov chain that contains a single recurrent class plus, perhaps, some transient states. An ergodic unichain is a unichain for which the recurrent class is ergodic.*

A unichain, as we shall see, is the natural generalization of a recurrent chain to allow for some initial transient behavior without disturbing the long term aymptotic behavior of the underlying recurrent chain.

The answer to the second question above is that the solution to (3.9) is unique if and only if [P] is the transition matrix of a *unichain*. If there are c recurrent classes, then (3.9) has c linearly independent solutions, each nonzero only over the elements of the corresponding recurrent class. For the third question, each row of $[P^n]$ converges to the unique solution of (3.9) if and only if [P] is the transition matrix of an *ergodic unichain*. If there are multiple recurrent classes, and each one is aperiodic, then $[P^n]$ still converges, but to a matrix with non-identical rows. If the Markov chain has one or more periodic recurrent classes, then $[P^n]$ does not converge.

We first look at these answers from the standpoint of the transition matrices of finite-state Markov chains, and then proceed in Chapter 5 to look at the more general problem of Markov chains with a countably infinite number of states. There we use renewal theory to answer these same questions (and to discover the differences that occur for infinite-state Markov chains).

The matrix approach is useful computationally and also has the advantage of telling us something about rates of convergence. The approach using renewal theory is very simple (given an understanding of renewal processes), but is more abstract.

In answering the above questions (plus a few more) for finite-state Markov chains, it is simplest to first consider the third question,[4] *i.e.*, the convergence of $[P^n]$. The simplest approach to this, for each column j of $[P^n]$, is to study the difference between the largest and smallest element of that column and how this difference changes with n. The following almost trivial lemma starts this study, and is valid for all finite-state Markov chains.

Lemma 3.3.1. *Let $[P]$ be the transition matrix of a finite-state Markov chain and let $[P^n]$ be the nth power of $[P]$* i.e.*, the matrix of nth order transition probabilities, P_{ij}^n. Then for each state j and each integer $n \geq 1$*

$$\max_i P_{ij}^{n+1} \leq \max_\ell P_{\ell j}^n \qquad \min_i P_{ij}^{n+1} \geq \min_\ell P_{\ell j}^n. \tag{3.10}$$

[4]One might naively try to show that a steady-state vector exists by first noting that each row of P sums to 1. The column vector $\boldsymbol{e} = (1, 1, \ldots, 1)^\mathsf{T}$ then satisfies the eigenvector equation $\boldsymbol{e} = [P]\boldsymbol{e}$. Thus there must also be a left eigenvector satisfying $\boldsymbol{\pi}[P] = \boldsymbol{\pi}$. The problem here is showing that $\boldsymbol{\pi}$ is real and non-negative.

Discussion The lemma says that for each column j, the largest of the elementsis non-increasing with n and the smallest of the elements is non-decreasing with n. The elements in a column that form the maximum and minimum can change with n, but the range covered by those elements is nested in n, either shrinking or staying the same as $n \to \infty$.

Proof: For each i, j, n, we use the Chapman-Kolmogorov equation, (3.8), followed by the fact that $P_{kj}^n \leq \max_\ell P_{\ell j}^n$, to see that

$$P_{ij}^{n+1} = \sum_k P_{ik} P_{kj}^n \leq \sum_k P_{ik} \max_\ell P_{\ell j}^n = \max_\ell P_{\ell j}^n. \tag{3.11}$$

Since this holds for all states i, and thus for the maximizing i, the first half of (3.10) follows. The second half of (3.10) is the same, with minima replacing maxima, *i.e.*,

$$P_{ij}^{n+1} = \sum_k P_{ik} P_{kj}^n \geq \sum_k P_{ik} \min_\ell P_{\ell j}^n = \min_\ell P_{\ell j}^n. \tag{3.12}$$

□

For some Markov chains, the maximizing elements in each column decrease with n and reach a limit equal to the increasing sequence of minimizing elements. For these chains, $[P^n]$ converges to a matrix where each column is constant, *i.e.*, the type of convergence discussed above. For others, the maximizing elements converge at some value strictly above the limit of the minimizing elements, Then $[P^n]$ does not converge to a matrix where each column is constant, and might not converge at all since the location of the maximizing and minimizing elements in each column can vary with n.

The following three subsections establish the above kind of convergence (and a number of subsidiary results) for three cases of increasing complexity. The first assumes that $P_{ij} > 0$ for all i, j. This is denoted as $[P] > 0$ and is not of great interest in its own right, but provides a needed step for the other cases. The second case is where the Markov chain is ergodic, and the third is where the Markov chain is an ergodic unichain.

3.3.2 Steady state assuming $[P] > 0$

Lemma 3.3.2. *Let the transition matrix of a finite-state Markov chain satisfy* $[P] > 0$ *(*i.e.*, $P_{ij} > 0$ for all i, j), and let $\alpha = \min_{i,j} P_{ij}$. Then for all states j and all $n \geq 1$:*

$$\max_i P_{ij}^{n+1} - \min_i P_{ij}^{n+1} \leq \left(\max_\ell P_{\ell j}^n - \min_\ell P_{\ell j}^n \right) (1 - 2\alpha). \tag{3.13}$$

$$\left(\max_\ell P_{\ell j}^n - \min_\ell P_{\ell j}^n \right) \leq (1 - 2\alpha)^n. \tag{3.14}$$

$$\lim_{n\to\infty} \max_\ell P_{\ell j}^n = \lim_{n\to\infty} \min_\ell P_{\ell j}^n > 0. \tag{3.15}$$

Discussion: Since $P_{ij} > 0$ for all i, j, we must have $\alpha > 0$. Thus the theorem says that for each j, the elements P_{ij}^n in column j of $[P^n]$ approach equality over both i and n as

$n \to \infty$, *i.e.*, the state at time n becomes independent of the state at time 0 as $n \to \infty$. The approach is exponential in n.

Proof: We first slightly tighten the inequality in (3.11). For a given j and n, let $\ell_{\min}$ be a value of ℓ that minimizes $P^n_{\ell j}$. Then

$$\begin{aligned} P^{n+1}_{ij} &= \sum_k P_{ik} P^n_{kj} \\ &\leq \sum_{k \neq \ell_{\min}} P_{ik} \max_\ell P^n_{\ell j} \;+\; P_{i\ell_{\min}} \min_\ell P^n_{\ell j} \\ &= \max_\ell P^n_{\ell j} - P_{i\ell_{\min}} \left(\max_\ell P^n_{\ell j} - \min_\ell P^n_{\ell j} \right) \\ &\leq \max_\ell P^n_{\ell j} - \alpha \left(\max_\ell P^n_{\ell j} - \min_\ell P^n_{\ell j} \right), \end{aligned}$$

where in the third step, we added and subtracted $P_{i\ell_{\min}} \min_\ell P^n_{\ell j}$ to the right hand side, and in the fourth step, we used $\alpha \leq P_{i\ell_{\min}}$ in conjuction with the fact that the term in parentheses must be nonnegative.

Repeating the same argument with the roles of max and min reversed,

$$P^{n+1}_{ij} \geq \min_\ell P^n_{\ell j} + \alpha \left(\max_\ell P^n_{\ell j} - \min_\ell P^n_{\ell j} \right).$$

The upper bound above applies to $\max_i P^{n+1}_{ij}$ and the lower bound to $\min_i P^{n+1}_{ij}$. Thus, subtracting the lower bound from the upper bound, we get (3.13).

Finally, note that

$$\min_\ell P_{\ell j} \geq \alpha > 0 \qquad \max_\ell P_{\ell j} \leq 1 - \alpha.$$

Thus $\max_\ell P_{\ell j} - \min_\ell P_{\ell j} \leq 1 - 2\alpha$. Using this as the base for iterating (3.13) over n, we get (3.14). This, in conjuction with (3.10), shows not only that the limits in (3.10) exist and are positive and equal, but that the limits are approached exponentially in n. □

3.3.3 Ergodic Markov chains

Lemma 3.3.2 extends quite easily to arbitrary ergodic finite-state Markov chains. The key to this comes from Theorem 3.2.4, which shows that if $[P]$ is the matrix for an M state ergodic Markov chain, then the matrix $[P^h]$ is positive for any $h \geq (\mathsf{M}-1)^2 - 1$. Thus, choosing $h = (\mathsf{M}-1)^2 - 1$, we can apply Lemma 3.3.2 to $[P^h] > 0$. For each integer $\nu \geq 1$,

$$\max_i P^{h(\nu+1)}_{ij} - \min_i P^{h(\nu+1)}_{ij} \leq \left(\max_m P^{h\nu}_{mj} - \min_m P^{h\nu}_{mj} \right) (1 - 2\beta) \tag{3.16}$$

$$\left(\max_m P^{h\nu}_{mj} - \min_m P^{h\nu}_{mj} \right) \leq (1 - 2\beta)^\nu$$

$$\lim_{\nu \to \infty} \max_m P^{h\nu}_{mj} = \lim_{\nu \to \infty} \min_m P^{h\nu}_{mj} > 0, \tag{3.17}$$

where $\beta = \min_{i,j} P_{ij}^h$. Lemma 3.3.1 states that $\max_i P_{ij}^{n+1}$ is nondecreasing in n, so that the limit on the left in (3.17) can be replaced with a limit in n. Similarly, the limit on the right can be replaced with a limit on n, getting

$$\left(\max_m P_{mj}^n - \min_m P_{mj}^n\right) \leq (1-2\beta)^{\lfloor n/h \rfloor} \tag{3.18}$$

$$\lim_{n\to\infty} \max_m P_{mj}^n = \lim_{n\to\infty} \min_m P_{mj}^n > 0. \tag{3.19}$$

Now define $\boldsymbol{\pi} > 0$ by

$$\pi_j = \lim_{n\to\infty} \max_m P_{mj}^n = \lim_{n\to\infty} \min_m P_{mj}^n > 0. \tag{3.20}$$

Since π_j lies between the minimum and maximum P_{ij}^n for each n,

$$\left|P_{ij}^n - \pi_j\right| \leq (1-2\beta)^{\lfloor n/h \rfloor}. \tag{3.21}$$

In the limit, then,

$$\lim_{n\to\infty} P_{ij}^n = \pi_j \quad \text{for each } i,j. \tag{3.22}$$

This says that the matrix $[P^n]$ has a limit as $n \to \infty$ and the i,j term of that matrix is π_j for all i,j. In other words, each row of this limiting matrix is the same and is the vector $\boldsymbol{\pi}$. This is represented most compactly by

$$\lim_{n\to\infty} [P^n] = \boldsymbol{e}\boldsymbol{\pi} \qquad \text{where } \boldsymbol{e} = (1,1,\ldots,1)^{\mathsf{T}}. \tag{3.23}$$

The following theorem[5] summarizes these results and adds one small additional result.

Theorem 3.3.1. *Let $[P]$ be the matrix of an ergodic finite-state Markov chain. Then there is a unique steady-state vector $\boldsymbol{\pi}$, that vector is positive and satisfies (3.22) and (3.23). The convergence in n is exponential, satisfying (3.18).*

Proof: We need to show that $\boldsymbol{\pi}$ as defined in (3.20) is the unique steady-state vector. Let $\boldsymbol{\mu}$ be any steady state vector, *i.e.*, any probability vector solution to $\boldsymbol{\mu}[P] = \boldsymbol{\mu}$. Then $\boldsymbol{\mu}$ must satisfy $\boldsymbol{\mu} = \boldsymbol{\mu}[P^n]$ for all $n > 1$. Going to the limit,

$$\boldsymbol{\mu} = \boldsymbol{\mu} \lim_{n\to\infty} [P^n] = \boldsymbol{\mu}\boldsymbol{e}\boldsymbol{\pi} = \boldsymbol{\pi}.$$

Thus $\boldsymbol{\pi}$ is a steady state vector and is unique □

3.3.4 Ergodic Unichains

Understanding how P_{ij}^n approaches a limit as $n \to \infty$ for ergodic unichains is a straightforward extension of the results in Section 3.3.3, but the details require a little care. Let $\mathcal{T}$ denote the set of transient states (which might contain several transient classes), and

$$[P] = \left[\begin{array}{c|c} [P_{\mathcal{T}}] & [P_{\mathcal{TR}}] \\ \hline [0] & [P_{\mathcal{R}}] \end{array}\right] \qquad \text{where} \qquad [P_{\mathcal{T}}] = \begin{bmatrix} P_{11} & \cdots & P_{1t} \\ \cdots & \cdots & \cdots \\ P_{t1} & \cdots & P_{tt} \end{bmatrix}$$

$$[P_{\mathcal{TR}}] = \begin{bmatrix} P_{1,t+1} & \cdots & P_{1,t+r} \\ \cdots & \cdots & \cdots \\ P_{t,t+1} & \cdots & P_{t,t+r} \end{bmatrix} \qquad [P_{\mathcal{R}}] = \begin{bmatrix} P_{t+1,t+1} & \cdots & P_{t+r,t+1} \\ \cdots & \cdots & \cdots \\ P_{t+r,t+1} & \cdots & P_{t+r,t+r} \end{bmatrix}$$

Figure 3.5: The transition matrix of a unichain. The block of zeroes in the lower left corresponds to the absence of transitions from recurrent to transient states.

assume the states of $\mathcal{T}$ are numbered $1, 2, \ldots, t$. Let $\mathcal{R}$ denote the recurrent class, assumed to be numbered $t{+}1, \ldots, t{+}r$ (see Figure 3.5).

If i and j are both recurrent states, then there is no possibility of leaving the recurrent class in going from i to j. Assuming this class to be ergodic, the transition matrix $[P_{\mathcal{R}}]$ as shown in Figure 3.5 has been analyzed in Section 3.3.3.

If the initial state is a transient state, then eventually the recurrent class is entered, and eventually after that, the distribution approaches steady state within the recurrent class. This suggests (and we next show) that there is a steady-state vector $\boldsymbol{\pi}$ for $[P]$ itself such that $\pi_j = 0$ for $j \in \mathcal{T}$ and π_j is as given in Section 3.3.3 for each $j \in \mathcal{R}$.

Initially we will show that P_{ij}^n converges to 0 for $i, j \in \mathcal{T}$. The exact nature of how and when the recurrent class is entered starting in a transient state is an interesting problem in its own right, and is discussed more later. For now, a crude bound will suffice.

For each transient state, there must a walk to some recurrent state, and since there are only t transient states, there must be some such path of length at most t. Each such path has positive probability, and thus for each $i \in \mathcal{T}$, $\sum_{j\in\mathcal{R}} P_{ij}^t > 0$. It follows that for each $i \in \mathcal{T}$, $\sum_{j\in\mathcal{T}} P_{ij}^t < 1$. Let $\gamma < 1$ be the maximum of these probabilities over $i \in \mathcal{T}$, *i.e.*,

$$\gamma = \max_{i\in\mathcal{T}} \sum_{j\in\mathcal{T}} P_{ij}^t < 1.$$

Lemma 3.3.3. *Let $[P]$ be a unichain with a set $\mathcal{T}$ of t transient states. Then*

$$\max_{\ell\in\mathcal{T}} \sum_{j\in\mathcal{T}} P_{\ell j}^n \leq \gamma^{\lfloor n/t \rfloor}. \tag{3.24}$$

[5]This is essentially the Frobenius theorem for non-negative irreducible matrices, specialized to Markov chains. A non-negative matrix $[P]$ is *irreducible* if its graph (containing an edge from node i to j if $P_{ij} > 0$) is the graph of a recurrent Markov chain. There is no constraint that each row of $[P]$ sums to 1. The proof of the Frobenius theorem requires some fairly intricate analysis and seems to be far more complex than the simple proof here for Markov chains. A proof of the general Frobenius theorem can be found in [10].

Proof: For each integer multiple νt of t and each $i \in \mathcal{T}$,

$$\sum_{j\in\mathcal{T}} P_{ij}^{(\nu+1)t} = \sum_{k\in\mathcal{T}} P_{ik}^{t} \sum_{j\in\mathcal{T}} P_{kj}^{\nu t} \le \sum_{k\in\mathcal{T}} P_{ik}^{t} \max_{\ell\in\mathcal{T}} \sum_{j\in\mathcal{T}} P_{\ell j}^{\nu t} \le \gamma \max_{\ell\in\mathcal{T}} \sum_{j\in\mathcal{T}} P_{\ell j}^{\nu t}.$$

Recognizing that this applies to all $i \in \mathcal{T}$, and thus to the maximum over i, we can iterate this equation, getting

$$\max_{\ell\in\mathcal{T}} \sum_{j\in\mathcal{T}} P_{\ell j}^{\nu t} \le \gamma^{\nu}.$$

Since this maximum is nonincreasing in n, (3.24) follows. □

We now proceed to the case where the initial state is $i \in \mathcal{T}$ and the final state is $j \in \mathcal{R}$. Let $m = \lfloor n/2 \rfloor$. For each $i \in \mathcal{T}$ and $j \in \mathcal{R}$, the Chapman-Kolmogorov equation, says that

$$P_{ij}^{n} = \sum_{k\in\mathcal{T}} P_{ik}^{m} P_{kj}^{n-m} + \sum_{k\in\mathcal{R}} P_{ik}^{m} P_{kj}^{n-m}.$$

Let π_j be the steady-state probability of state $j \in \mathcal{R}$ in the recurrent Markov chain with states $\mathcal{R}$, *i.e.*, $\pi_j = \lim_{n\to\infty} P_{kj}^{n}$. Then for each $i \in \mathcal{T}$,

$$\begin{aligned}
\left|P_{ij}^{n} - \pi_j\right| &= \left|\sum_{k\in\mathcal{T}} P_{ik}^{m}\left(P_{kj}^{n-m} - \pi_j\right) + \sum_{k\in\mathcal{R}} P_{ik}^{m}\left(P_{kj}^{n-m} - \pi_j\right)\right| \\
&\le \sum_{k\in\mathcal{T}} P_{ik}^{m}\left|P_{kj}^{n-m} - \pi_j\right| + \sum_{k\in\mathcal{R}} P_{ik}^{m}\left|P_{kj}^{n-m} - \pi_j\right| \\
&\le \sum_{k\in\mathcal{T}} P_{ik}^{m} + \sum_{k\in\mathcal{R}} P_{ik}^{m}\left|P_{kj}^{n-m} - \pi_j\right| \qquad (3.25) \\
&\le \gamma^{\lfloor m/t \rfloor} + (1-2\beta)^{\lfloor (n-m)/h \rfloor}, \qquad (3.26)
\end{aligned}$$

where the first step upper bounded the absolute value of a sum by the sum of the absolute values. In the last step, (3.24) was used in the first half of (3.25) and (3.21) (with $h = (r-1)^2 - 1$ and $\beta = \min_{i,j\in\mathcal{R}} P_{ij}^{h} > 0$) was used in the second half.

This is summarized in the following theorem.

Theorem 3.3.2. *Let $[P]$ be the matrix of an ergodic finite-state unichain. Then $\lim_{n\to\infty}[P^n] = \boldsymbol{e\pi}$ where $\boldsymbol{e} = (1, 1, \ldots, 1)^{\mathsf{T}}$ and $\boldsymbol{\pi}$ is the steady-state vector of the recurrent class of states, expanded by 0's for each transient state of the unichain. The convergence is exponential in n for all i, j.*

3.3.5 Arbitrary finite-state Markov chains

The asymptotic behavior of $[P^n]$ as $n \to \infty$ for arbitrary finite-state Markov chains can mostly be deduced from the ergodic unichain case by simple extensions and common sense.

First consider the case of $m > 1$ aperiodic classes plus a set of transient states. If the initial state is in the κth of the recurrent classes, say $\mathcal{R}^{\kappa}$ then the chain remains in $\mathcal{R}^{\kappa}$ and there

is a unique finite-state vector $\boldsymbol{\pi}^\kappa$ that is non-zero only in $\mathcal{R}^\kappa$ that can be found by viewing class κ in isolation.

If the initial state i is transient, then, for each $\mathcal{R}^\kappa$, there is a certain probability that $\mathcal{R}^\kappa$ is eventually reached, and once it is reached there is no exit, so the steady state over that recurrent class is approached. The question of finding the probability of entering each recurrent class from a given transient class will be discussed in the next section.

Next consider a recurrent Markov chain that is periodic with period d. The dth order transition probability matrix, $[P^d]$, is then constrained by the fact that $P^d_{ij} = 0$ for all j not in the same periodic subset as i. In other words, $[P^d]$ is the matrix of a chain with d recurrent classes. We will obtain greater facility in working with this in the next section when eigenvalues and eigenvectors are discussed.

3.4 The eigenvalues and eigenvectors of stochastic matrices

For ergodic unichains, the previous section showed that the dependence of a state on the distant past disappears with increasing n, *i.e.*, $P^n_{ij} \to \pi_j$. In this section we look more carefully at the eigenvalues and eigenvectors of $[P]$ to sharpen our understanding of how fast $[P^n]$ converges for ergodic unichains and what happens for other finite-state Markov chains.

Definition 3.4.1. *The row*[6] *vector $\boldsymbol{\pi}$ is a left eigenvector of $[P]$ of eigenvalue λ if $\boldsymbol{\pi} \neq \boldsymbol{0}$ and $\boldsymbol{\pi}[P] = \lambda\boldsymbol{\pi}$,* i.e., $\sum_i \pi_i P_{ij} = \lambda\pi_j$ *for all j. The column vector $\boldsymbol{\nu}$ is a right eigenvector of eigenvalue λ if $\boldsymbol{\nu} \neq \boldsymbol{0}$ and $[P]\boldsymbol{\nu} = \lambda\boldsymbol{\nu}$,* i.e., $\sum_j P_{ij}\nu_j = \lambda\nu_i$ *for all i.*

We showed that a stochastic matrix always has an eigenvalue $\lambda = 1$, and that for an ergodic unichain, there is a unique steady-state vector $\boldsymbol{\pi}$ that is a left eigenvector with $\lambda = 1$ and (within a scale factor) a unique right eigenvector $\boldsymbol{e} = (1, \ldots, 1)^\mathsf{T}$. In this section we look at the other eigenvalues and eigenvectors and also look at Markov chains other than ergodic unichains. We start by limiting the number of states to $\mathsf{M} = 2$. This provides insight without requiring much linear algebra. After that, the general case with arbitrary $\mathsf{M} < \infty$ is analyzed.

3.4.1 Eigenvalues and eigenvectors for $\mathsf{M} = 2$ states

The eigenvalues and eigenvectors can be found by elementary (but slightly tedious) algebra. The left and right eigenvector equations can be written out as

$$\begin{aligned} \pi_1 P_{11} + \pi_2 P_{21} &= \lambda\pi_1 \\ \pi_1 P_{12} + \pi_2 P_{22} &= \lambda\pi_2 \end{aligned} \quad \text{(left)} \qquad \begin{aligned} P_{11}\nu_1 + P_{12}\nu_2 &= \lambda\nu_1 \\ P_{21}\nu_1 + P_{22}\nu_2 &= \lambda\nu_2 \end{aligned} \quad \text{(right).} \tag{3.27}$$

[6] Students of linear algebra usually work primarily with right eigenvectors (and in abstract linear algebra often ignore matrices and concrete M-tuples altogether). Here a more concrete view is desirable because of the direct connection of $[P^n]$ with transition probabilities. Also, although left eigenvectors could be converted to right eigenvectors by taking the transpose of $[P]$, this would be awkward when Markov chains with rewards are considered and both row and column vectors play important roles.

Each set of equations have a non-zero solution if and only if the matrix $[P - \lambda I]$, where $[I]$ is the identity matrix, is singular (i.e., there must be a non-zero $\boldsymbol{\nu}$ for which $[P-\lambda I]\boldsymbol{\nu} = \boldsymbol{0}$). Thus λ must be such that the determinant of $[P - \lambda I]$, namely $(P_{11} - \lambda)(P_{22} - \lambda) - P_{12}P_{21}$, is equal to 0. Solving this quadratic equation in λ, we find that λ has two solutions,

$$\lambda_1 = 1 \qquad \lambda_2 = 1 - P_{12} - P_{21}.$$

Assuming initially that P_{12} and P_{21} are not both 0, the solution for the left and right eigenvectors, $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\nu}^{(1)}$, of λ_1 and $\boldsymbol{\pi}^{(2)}$ and $\boldsymbol{\nu}^{(2)}$ of λ_2, are given by

$$\begin{array}{llll} \pi_1^{(1)} = \frac{P_{21}}{P_{12}+P_{21}} & \pi_2^{(1)} = \frac{P_{12}}{P_{12}+P_{21}} & \nu_1^{(1)} = 1 & \nu_2^{(1)} = 1 \\ \pi_1^{(2)} = 1 & \pi_2^{(2)} = -1 & \nu_1^{(2)} = \frac{P_{12}}{P_{12}+P_{21}} & \nu_2^{(2)} = \frac{-P_{21}}{P_{12}+P_{21}} \end{array} .$$

These solutions contain arbitrary normalization factors. That for $\boldsymbol{\pi}^{(1)} = (\pi_1^{(1)},\ \pi_2^{(1)})$ has been chosen so that $\boldsymbol{\pi}^{(1)}$ is a steady-state vector (*i.e.*, the components sum to 1). The solutions have also been normalized so that $\boldsymbol{\pi}_i\boldsymbol{\nu}_i = 1$ for $i = 1, 2$. Now define

$$[\Lambda] = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \qquad \text{and} \qquad [U] = \begin{bmatrix} \nu_1^{(1)} & \nu_1^{(2)} \\ \nu_2^{(1)} & \nu_2^{(2)} \end{bmatrix},$$

i.e., $[U]$ is a matrix whose columns are the eigenvectors $\boldsymbol{\nu}^{(1)}$ and $\boldsymbol{\nu}^{(2)}$. Then the two right eigenvector equations in (3.27) can be combined compactly as $[P][U] = [U][\Lambda]$. It turns out (for the given normalization of the eigenvectors) that the inverse of $[U]$ is just the matrix whose rows are the left eigenvectors of $[P]$ (this can be verified by noting that $\boldsymbol{\pi}_1\boldsymbol{\nu}_2 = \boldsymbol{\pi}_2\boldsymbol{\nu}_1 = 0$. We then see that $[P] = [U][\Lambda][U]^{-1}$ and consequently $[P^n] = [U][\Lambda]^n[U]^{-1}$. Multiplying this out, we get

$$[P^n] = \begin{bmatrix} \pi_1 + \pi_2\lambda_2^n & \pi_2 - \pi_2\lambda_2^n \\ \pi_1 - \pi_1\lambda_2^n & \pi_2 + \pi_1\lambda_2^n \end{bmatrix}, \tag{3.28}$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2)$ is the steady state vector $\boldsymbol{\pi}^{(1)}$. Recalling that $\lambda_2 = 1 - P_{12} - P_{21}$, we see that $|\lambda_2| \leq 1$. There are 2 trivial cases where $|\lambda_2| = 1$. In the first, $P_{12} = P_{21} = 0$, so that $[P]$ is just the identity matrix. The Markov chain then has 2 recurrent states and stays forever where it starts. In the other trivial case, $P_{12} = P_{21} = 1$. Then $\lambda_2 = -1$ so that $[P^n]$ alternates between the identity matrix for n even and $[P]$ for n odd. In all other cases, $|\lambda_2| < 1$ and $[P^n]$ approaches the steady state matrix $\lim_{n\to\infty}[P^n] = \boldsymbol{e\pi}$.

What we have learned from this is the exact way in which $[P^n]$ approaches $\boldsymbol{e\pi}$. Each term in $[P^n]$ approaches the steady state value exponentially in n as λ_2^n. Thus, in place of the upper bound in (3.21), we have an exact expression, which in this case is simpler than the bound. As we see shortly, this result is representative of the general case, but the simplicity is lost.

3.4.2 Eigenvalues and eigenvectors for $\mathsf{M} > 2$ states

For the general case of a stochastic matrix, we start with the fact that the set of eigenvalues is given by the set of (possibly complex) values of λ that satisfy the determinant equation $\det[P - \lambda I] = 0$. Since $\det[P - \lambda I]$ is a polynomial of degree M in λ, this equation has M roots (*i.e.*, M eigenvalues), not all of which need be distinct.[7]

Case with M distinct eigenvalues: We start with the simplest case in which the M eigenvalues, say $\lambda_1, \ldots, \lambda_{\mathsf{M}}$, are all distinct. The matrix $[P - \lambda_i I]$ is singular for each i, so there must be a right eigenvector $\boldsymbol{\nu}^{(i)}$ and a left eigenvector $\boldsymbol{\pi}^{(i)}$ for each eigenvalue λ_i. The right eigenvectors span M dimensional space and thus the matrix U with columns $(\boldsymbol{\nu}^{(1)}, \ldots, \boldsymbol{\nu}^{(\mathsf{M})})$ is nonsingular. The left eigenvectors, if normalized to satisfy $\boldsymbol{\pi}^{(i)}\boldsymbol{\nu}^{(i)} = 1$ for each i, then turn out to be the rows of $[U^{-1}]$ (see Exercise 3.11). As in the two state case, we can then express $[P^n]$ as

$$[P^n] = [U^{-1})[\Lambda^n][U], \tag{3.29}$$

where Λ is the diagonal matrix with terms $\lambda_1, \ldots, \lambda_{\mathsf{M}}$.

If Λ is broken into the sum of M diagonal matrices,[8] each with only a single nonzero element, then (see Exercise 3.11) $[P^n]$ can be expressed as

$$[P^n] = \sum_{i=1}^{\mathsf{M}} \lambda_i^n \boldsymbol{\nu}^{(i)}\boldsymbol{\pi}^{(i)}. \tag{3.30}$$

Note that this is the same form as (3.28), where in (3.28), the eigenvalue $\lambda_1 = 1$ simply appears as the value 1. We have seen that there is always one eigenvalue that is 1, with an accompanying steady-state vector $\boldsymbol{\pi}$ as a left eigenvector and the unit vector $\boldsymbol{e} = (1, \ldots, 1)^{\mathsf{T}}$ as a right eigenvector. The other eigenvalues and eigenvectors can be complex, but it is almost self evident from the fact that $[P^n]$ is a stochastic matrix that $\left|\lambda_i\right| \leq 1$. A simple guided proof of this is given in Exercise 3.12.

We have seen that $\lim_{n\to\infty}[P^n] = \boldsymbol{e}\boldsymbol{\pi}$ for ergodic unichains. This implies that all terms except $i = 1$ in (3.30) die out with n, which further implies that $\left|\lambda_i\right| < 1$ for all eigenvalues except $\lambda = 1$. In this case, we see that the rate at which $[P^n]$ approaches steady state is given by the second largest eigenvalue in magnitude, *i.e.*, $\max_{i:|\lambda_i|<1} |\lambda_i|$.

If a recurrent chain is periodic with period d, it turns out that there are d eigenvalues of magnitude 1, and these are uniformly spaced around the unit circle in the complex plane. Exercise 3.19 contains a guided proof of this.

Case with repeated eigenvalues and M linearly independent eigenvectors: If some of the M eigenvalues of $[P]$ are not distinct, the question arises as to how many linearly independent left (or right) eigenvectors exist for an eigenvalue λ_i of multiplicity m, *i.e.*, a λ_i that is an mth order root of $\det[P - \lambda I]$. Perhaps the ugliest part of linear algebra is the fact that an eigenvalue of multiplicity m need not have m linearly independent eigenvectors.

[7] Readers with little exposure to linear algebra can either accept the linear algebra results in this section (without a great deal of lost insight) or can find them in Strang [19] or many other linear algebra texts.

[8] If 0 is one of the M eigenvalues, then only $\mathsf{M} - 1$ such matrices are required.

An example of a very simple Markov chain with $\mathsf{M} = 3$ but only two linearly independent eigenvectors is given in Exercise 3.14. These eigenvectors do not span M-space, and thus the expansion in (3.30) cannot be used.

Before looking at this ugly case, we look at the case where the right eigenvectors, say, span the space, *i.e.*, where each distinct eigenvalue has a number of linearly independent eigenvectors equal to its multiplicity. We can again form a matrix $[U]$ whose columns are the M linearly independent right eigenvectors, and again $[U^{-1}]$ is a matrix whose rows are the corresponding left eigenvectors of $[P]$. We then get (3.30) again. Thus, so long as the eigenvectors span the space, the asymptotic expression for the limiting transition probabilities can be found in the same way.

The most important situation where these repeated eigenvalues make a major difference is for Markov chains with $\kappa > 1$ recurrent classes. In this case, κ is the multiplicity of the eigenvalue 1. It is easy to see that there are κ different steady-state vectors. The steady-state vector for recurrent class ℓ, $1 \leq \ell \leq \kappa$, is strictly positive for each state of the ℓth recurrent class and is zero for all other states.

The eigenvalues for $[P]$ in this case can be found by finding the eigenvalues separately for each recurrent class. If class j contains r_j states, then r_j of the eigenvalues (counting repetitions) of $[P]$ are the eigenvalues of the r_j by r_j matrix for the states in that recurrent class. Thus the rate of convergence of $[P^n]$ within that submatrix is determined by the second largest eigenvalue (in magnitude) in that class.

What this means is that this general theory using eigenvalues says exactly what common sense says: if there are κ recurrent classes, look at each one separately, since they have nothing to do with each other. This also lets us see that for any recurrent class that is aperiodic, all the other eigenvalues for that class are strictly less than 1 in magnitude.

The situation is less obvious if there are κ recurrent classes plus a set of t transient states. All but t of the eigenvalues (counting repetitions) are associated with the recurrent classes, and the remaining t eigenvalues are the eigenvalues of the t by t matrix, say $[P_t]$, between the transient states. Each of these t eigenvalues are strictly less than 1 (as seen in Section 3.3.4) and neither these eigenvalues nor their eigenvectors depend on the transition probabilities from the transient to recurrent states. The left eigenvectors for the recurrent classes also do not depend on these transient to recurrent states. The right eigenvector for $\lambda = 1$ for each recurrent class $\mathcal{R}_\ell$ is very interesting however. It's value is 1 for each state in $\mathcal{R}_\ell$, is 0 for each state in the other recurrent classes, and is equal to $\lim_{n\to\infty} \Pr\{X_n \in \mathcal{R}_\ell \mid X_0 = i\}$ for each transient state i (see Exercise 3.13).

The Jordan form case: As mentioned before, there are cases in which one or more eigenvalues of $[P]$ are repeated (as roots of $\det[P - \lambda I]$) but where the number of linearly independent right eigenvectors for a given eigenvalue is less than the multiplicity of that eigenvalue. In this case, there are not enough eigenvectors to span the space, so there is no M by M matrix whose columns are linearly independent eigenvectors. Thus $[P]$ can not be expressed as $[U^{-1}][\Lambda][U]$ where Λ is the diagonal matrix of the eigenvalues, repeated according to their multiplicity.

The Jordan form is the cure for this unfortunate situation. The Jordan form for a given

$[P]$ is the following modification of the diagonal matrix of eigenvalues: we start with the diagonal matrix of eigenvalues, with the repeated eigenvalues as neighboring elements. Then for each missing eigenvector for a given eigenvalue, a 1 is placed immediately to the right and above a neighboring pair of appearances of that eigenvalue, as seen by example[9] below:

$$[J] = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 & 0 \\ 0 & 0 & \lambda_2 & 1 & 0 \\ 0 & 0 & 0 & \lambda_2 & 1 \\ 0 & 0 & 0 & 0 & \lambda_2 \end{bmatrix}.$$

There is a theorem in linear algebra that says that an invertible matrix $[U]$ exists and a Jordan form exists such that $[P] = [U^{-1}][J][U]$. The major value to us of this result is that it makes it relatively easy to calculate $[J]^n$ for large n (see Exercise 3.15). This exercise also shows that for all stochastic matrices, each eigenvalue of magnitude 1 has precisely one associated eigenvector. This is usually expressed by the statement that all the eigenvalues of magnitude 1 are *simple*, meaning that their multiplicity equals their number of linearly independent eigenvectors. Finally the exercise shows that $[P^n]$ for an aperiodic recurrent chain converges as a polynomial[10] in n times λ_s^n where λ_s is the eigenvalue of largest magnitude less than 1.

The most important results of this section on eigenvalues and eigenvectors can be summarized in the following theorem.

Theorem 3.4.1. *The transition matrix of a finite state unichain has a single eigenvalue $\lambda = 1$ with an accompanying left eigenvector $\boldsymbol{\pi}$ satisfying (3.9) and a left eigenvector $\boldsymbol{e} = (1, 1, \ldots, 1)^\intercal$. The other eigenvalues λ_i all satisfy $|\lambda_i| \leq 1$. The inequality is strict unless the unichain is periodic, say with period d, and then there are d eigenvalues of magnitude 1 spaced equally around the unit circle. If the unichain is ergodic, then $[P^n]$ converges to steady state $\boldsymbol{e\pi}$ with an error in each term bounded by a fixed polynomial in n times $|\lambda_s|^n$, where λ_s is the eignevalue of largest magnitude less than 1.*

Arbitrary Markov chains can be split into their recurrent classes, and this theorem can be applied separately to each class.

3.5 Markov chains with rewards

Suppose that each state i in a Markov chain is associated with a reward, r_i. As the Markov chain proceeds from state to state, there is an associated sequence of rewards that are not independent, but are related by the statistics of the Markov chain. The concept of a reward in each state[11] is quite graphic for modeling corporate profits or portfolio performance, and

[9] See Strang [19], for example, for a more complete description of how to construct a Jordan form

[10] This polynomial is equal to 1 if these eigenvalues are simple.

[11] Occasionally it is more natural to associate rewards with transitions rather than states. If r_{ij} denotes a reward associated with a transition from i to j and P_{ij} denotes the corresponding transition probability, then defining $r_i = \sum_j P_{ij} r_{ij}$ essentially simplifies these transition rewards to rewards over the initial state for the transition. These transition rewards are ignored here, since the details add complexity to a topic that is complex enough for a first treatment.

is also useful for studying queueing delay, the time until some given state is entered, and many other phenomena. The reward r_i associated with a state could equally well be veiwed as a cost or any given real-valued function of the state.

In Section 3.6, we study dynamic programming and Markov decision theory. These topics include a "decision maker," "policy maker," or "control" that modify both the transition probabilities and the rewards at each trial of the 'Markov chain.' The decision maker attempts to maximize the expected reward, but is typically faced with compromising between immediate reward and the longer-term reward arising from the choice of transition probabilities that lead to 'high reward' states. This is a much more challenging problem than the current study of Markov chains with rewards, but a thorough understanding of the current problem provides the machinery to understand Markov decision theory also.

The steady-state expected reward per unit time, assuming a single recurrent class of states, is defined to be the *gain*, expressed as $g = \sum_i \pi_i r_i$ where π_i is the steady-state probability of being in state i.

3.5.1 Examples of Markov chains with rewards

The following examples demonstrate that it is important to understand the transient behavior of rewards as well as the long-term averages. This transient behavior will turn out to be even more important when we study Markov decision theory and dynamic programming.

Example 3.5.1 (Expected first-passage time). First-passage times, *i.e.*, the number of steps taken in going from one given state, say i, to another, say 1, are frequently of interest for Markov chains, and here we solve for the expected value of this random variable.

Since the first-passage time is independent of the transitions after the first entry to state 1, we can modify the chain to convert the final state, say state 1, into a trapping state (a *trapping state* i is a state from which there is no exit, *i.e.*, for which $P_{ii} = 1$). That is, we modify P_{11} to 1 and P_{1j} to 0 for all $j \neq 1$. We leave P_{ij} unchanged for all $i \neq 1$ and all j (see Figure 3.6). This modification of the chain will not change the probability of any sequence of states up to the point that state 1 is first entered.

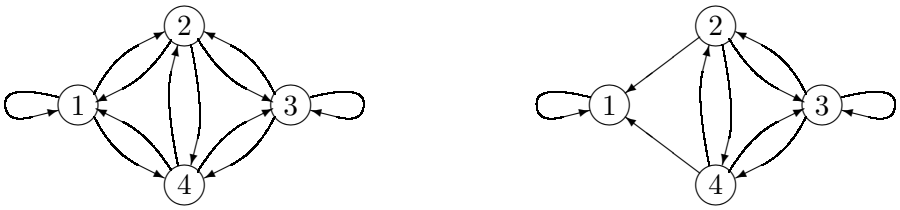


Figure 3.6: The conversion of a recurrent Markov chain with $\mathsf{M} = 4$ into a chain for which state 1 is a trapping state, *i.e.*, the outgoing arcs from node 1 have been removed.

Let v_i be the expected number of steps to first reach state 1 starting in state $i \neq 1$. This number of steps includes the first step plus the expected number of remaining steps to reach state 1 starting from whatever state is entered next (if state 1 is the next state entered, this

remaining number is 0). Thus, for the chain in Figure 3.6, we have the equations

$$\begin{aligned} v_2 &= 1 + P_{23}v_3 + P_{24}v_4. \\ v_3 &= 1 + P_{32}v_2 + P_{33}v_3 + P_{34}v_4. \\ v_4 &= 1 + P_{42}v_2 + P_{43}v_3. \end{aligned}$$

For an arbitrary chain of M states where 1 is a trapping state and all other states are transient, this set of equations becomes

$$v_i = 1 + \sum_{j \neq 1} P_{ij} v_j; \quad i \neq 1. \tag{3.31}$$

If we define $r_i = 1$ for $i \neq 1$ and $r_i = 0$ for $i = 1$, then r_i is a unit reward for not yet entering the trapping state, and v_i is the expected aggregate reward before entering the trapping state. Thus by taking $r_1 = 0$, the reward ceases upon entering the trapping state, and v_i is the expected transient reward, *i.e.*, the expected first-passage time from state i to state 1. Note that in this example, rewards occur only in transient states. Since transient states have zero steady-state probabilities, the steady-state gain per unit time, $g = \sum_i \pi_i r_i$, is 0.

If we define $v_1 = 0$, then (3.31), along with $v_1 = 0$, has the vector form

$$\boldsymbol{v} = \boldsymbol{r} + [P]\boldsymbol{v}; \quad v_1 = 0. \tag{3.32}$$

For a Markov chain with M states, (3.31) is a set of $\mathsf{M} - 1$ equations in the $\mathsf{M} - 1$ variables v_2 to v_{M}. The equation $\boldsymbol{v} = \boldsymbol{r} + [P]\boldsymbol{v}$ is a set of M linear equations, of which the first is the vacuous equation $v_1 = 0 + v_1$, and, with $v_1 = 0$, the last $\mathsf{M} - 1$ correspond to (3.31). It is not hard to show that (3.32) has a unique solution for $\boldsymbol{v}$ under the condition that states 2 to M are all transient and 1 is a trapping state, but we prove this later, in Theorem 3.5.1, under more general circumstances.

Example 3.5.2. Assume that a Markov chain has M states, $\{0, 1, \ldots, \mathsf{M} - 1\}$, and that the state represents the number of customers in an integer-time queueing system. Suppose we wish to find the expected sum of the customer waiting times, starting with i customers in the system at some given time t and ending at the first instant when the system becomes idle. That is, for each of the i customers in the system at time t, the waiting time is counted from t until that customer exits the system. For each new customer entering before the system next becomes idle, the waiting time is counted from entry to exit.

When we discuss Little's theorem in Section 4.5.4, it will be seen that this sum of waiting times is equal to the sum over τ of the state X_τ at time τ, taken from $\tau = t$ to the first subsequent time the system is empty.

As in the previous example, we modify the Markov chain to make state 0 a trapping state and assume the other states are then all transient. We take $r_i = i$ as the "reward" in state i, and v_i as the expected aggregate reward until the trapping state is entered. Using the same reasoning as in the previous example, v_i is equal to the immediate reward $r_i = i$ plus the expected aggregate reward from whatever state is entered next. Thus $v_i = r_i + \sum_{j \geq 1} P_{ij} v_j$. With $v_0 = 0$, this is $\boldsymbol{v} = \boldsymbol{r} + [P]\boldsymbol{v}$. This has a unique solution for $\boldsymbol{v}$, as will be shown later in Theorem 3.5.1. This same analysis is valid for any choice of reward r_i for each transient state i; the reward in the trapping state must be 0 so as to keep the expected aggregate reward finite.

In the above examples, the Markov chain is converted into a trapping state with zero gain, and thus the expected reward is a transient phenomena with no reward after entering the trapping state. We now look at the more general case of a unichain. In this more general case, there can be some gain per unit time, along with some transient expected reward depending on the initial state. We first look at the aggregate gain over a finite number of time units, thus providing a clean way of going to the limit.

Example 3.5.3. The example in Figure 3.7 provides some intuitive appreciation for the general problem. Note that the chain tends to persist in whatever state it is in. Thus if the chain starts in state 2, not only is an immediate reward of 1 achieved, but there is a high probability of additional unit rewards on many successive transitions. Thus the aggregate value of starting in state 2 is considerably more than the immediate reward of 1. On the other hand, we see from symmetry that the gain per unit time, over a long time period, must be one half.

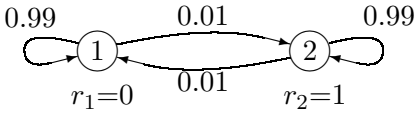


Figure 3.7: Markov chain with rewards and nonzero steady-state gain.

3.5.2 The expected aggregate reward over multiple transitions

Returning to the general case, let X_m be the state at time m and let $R_m = R(X_m)$ be the reward at that m, *i.e.*, if the sample value of X_m is i, then r_i is the sample value of R_m. Conditional on $X_m = i$, the aggregate expected reward $v_i(n)$ over n trials from X_m to X_{m+n-1} is

$$\begin{aligned} v_i(n) &= \mathsf{E}\left[R(X_m) + R(X_{m+1}) + \cdots + R(X_{m+n-1}) \mid X_m = i\right] \\ &= r_i + \sum_j P_{ij} r_j + \cdots + \sum_j P_{ij}^{n-1} r_j. \end{aligned}$$

This expression does not depend on the starting time m because of the homogeneity of the Markov chain. Since it gives the expected reward for each initial state i, it can be combined into the following vector expression $\boldsymbol{v}(n) = (v_1(n), v_2(n), \ldots, v_{\mathsf{M}}(n))^{\mathsf{T}}$,

$$\boldsymbol{v}(n) \;=\; \boldsymbol{r} + [P]\boldsymbol{r} + \cdots + [P^{n-1}]\boldsymbol{r} \;=\; \sum_{h=0}^{n-1} [P^h]\boldsymbol{r}, \tag{3.33}$$

where $\boldsymbol{r} = (r_1, \ldots, r_{\mathsf{M}})^{\mathsf{T}}$ and P^0 is the identity matrix. Now assume that the Markov chain is an ergodic unichain. Then $\lim_{n\to\infty}[P]^n = \boldsymbol{e\pi}$ and $\lim_{n\to\infty}[P]^n\boldsymbol{r} = \boldsymbol{e\pi r} = g\boldsymbol{e}$ where $g = \boldsymbol{\pi r}$ is the steady-state reward per unit time. If $g \neq 0$, then $\boldsymbol{v}(n)$ changes by approximately $g\boldsymbol{e}$ for each unit increase in n, so $\boldsymbol{v}(n)$ does not have a limit as $n \to \infty$. As shown below, however, $\boldsymbol{v}(n) - ng\boldsymbol{e}$ does have a limit, given by

$$\lim_{n\to\infty} (\boldsymbol{v}(n) - ng\boldsymbol{e}) = \lim_{n\to\infty} \sum_{h=0}^{n-1} [P^h - \boldsymbol{e\pi}]\boldsymbol{r}. \tag{3.34}$$

To see that this limit exists, note from (3.26) that $\epsilon > 0$ can be chosen small enough that $P_{ij}^n - \pi_j = o(\exp(-n\epsilon))$ for all states i, j and all $n \geq 1$. Thus $\sum_{h=n}^{\infty}(P_{ij}^h - \pi_j) = o(\exp(-n\epsilon))$ also. This shows that the limits on each side of (3.34) must exist for an ergodic unichain.

The limit in (3.34) is a vector over the states of the Markov chain. This vector gives the asymptotic relative expected advantage of starting the chain in one state relative to another. This is an important quantity in both the next section and the remainder of this one. It is called the relative-gain vector and denoted by $\boldsymbol{w}$,

$$\boldsymbol{w} = \lim_{n\to\infty} \sum_{h=0}^{n-1} [P^h - \boldsymbol{e}\boldsymbol{\pi}]\boldsymbol{r} \tag{3.35}$$

$$= \lim_{n\to\infty} (\boldsymbol{v}(n) - ng\boldsymbol{e}) \,. \tag{3.36}$$

Note from (3.36) that if $g > 0$, then $ng\boldsymbol{e}$ increases linearly with n and and $\boldsymbol{v}(n)$ must asymptotically increase linearly with n. Thus the relative-gain vector $\boldsymbol{w}$ becomes small relative to both $ng\boldsymbol{e}$ and $\boldsymbol{v}(n)$ for large n. As we will see, $\boldsymbol{w}$ is still important, particularly in the next section on Markov decisions.

We can get some feel for $\boldsymbol{w}$ and how $v_i(n) - n\pi_i$ converges to w_i from Example 3.5.3 (as described in Figure 3.7). Since this chain has only two states, $[P^n]$ and $v_i(n)$ can be calculated easily from (3.28). The result is tabulated in Figure 3.8, and it is seen numerically that $\boldsymbol{w} = (-25, +25)^{\mathsf{T}}$. The rather significant advantage of starting in state 2 rather than 1, however, requires hundreds of transitions before the gain is fully apparent.

n	$\boldsymbol{\pi}\boldsymbol{v}(n)$	$v_1(n)$	$v_2(n)$
1	0.5	0	1
2	1	0.01	1.99
4	2	0.0592	3.9408
10	5	0.4268	9.5732
40	20	6.1425	33.8575
100	50	28.3155	71.6845
400	200	175.007	224.9923

Figure 3.8: The expected aggregate reward, as a function of starting state and stage, for the example of figure 3.7. Note that $\boldsymbol{w} = (-25, +25)^{\mathsf{T}}$, but the convergence is quite slow.

This example also shows that it is somewhat inconvenient to calculate $\boldsymbol{w}$ from (3.35), and this inconvenience grows rapidly with the number of states. Fortunately, as shown in the following theorem, $\boldsymbol{w}$ can also be calculated simply by solving a set of linear equations.

Theorem 3.5.1. *Let $[P]$ be the transition matrix for an ergodic unichain. Then the relative-gain vector $\boldsymbol{w}$ given in (3.35) satisfies the following linear vector equation.*

$$\boldsymbol{w} + g\boldsymbol{e} = [P]\boldsymbol{w} + \boldsymbol{r} \qquad \textit{and} \quad \boldsymbol{\pi}\boldsymbol{w} = 0. \tag{3.37}$$

Furthermore (3.37) has a unique solution if $[P]$ is the transition matrix for a unichain (either ergodic or periodic).

Discussion: For an ergodic unichain, the interpretation of $\boldsymbol{w}$ as an asymptotic relative gain comes from (3.35) and (3.36). For a periodic unichain, (3.37) still has a unique solution, but (3.35) no longer converges, so the solution to (3.37) no longer has a clean interpretation as an asymptotic limit of relative gain. This solution is still called a relative-gain vector, and can be interpreted as an asymptotic relative gain over a period, but the important thing is that this equation has a unique solution for arbitrary unichains.

Definition 3.5.1. *The relative-gain vector $\boldsymbol{w}$ of a unichain is the unique vector that satisfies (3.37).*

Proof: Premultiplying both sides of (3.35) by $[P]$,

$$\begin{aligned}
[P]\boldsymbol{w} &= \lim_{n\to\infty} \sum_{h=0}^{n-1} [P^{h+1} - \boldsymbol{e\pi}]\boldsymbol{r} \\
&= \lim_{n\to\infty} \sum_{h=1}^{n} [P^{h} - \boldsymbol{e\pi}]\boldsymbol{r} \\
&= \lim_{n\to\infty} \left(\sum_{h=0}^{n} [P^{h} - \boldsymbol{e\pi}]\boldsymbol{r} \right) - [P^0 - \boldsymbol{e\pi}]\boldsymbol{r} \\
&= \boldsymbol{w} - [P^0]\boldsymbol{r} + \boldsymbol{e\pi r} = \boldsymbol{w} - \boldsymbol{r} + g\boldsymbol{e}.
\end{aligned}$$

Rearranging terms, we get (3.37). For a unichain, the eigenvalue 1 of $[P]$ has multiplicity 1, and the existence and uniqueness of the solution to (3.37) is then a simple result in linear algebra (see Exercise 3.23). □

The above manipulations conceal the intuitive nature of (3.37). To see the intuition, consider the first-passage-time example again. Since all states are transient except state 1, $\pi_1 = 1$. Since $r_1 = 0$, we see that the steady-state gain is $g = 0$. Also, in the more general model of the theorem, $v_i(n)$ is the expected reward over n transitions starting in state i, which for the first-passage-time example is the expected number of transient states visited up to the nth transition. In other words, the quantity v_i in the first-passage-time example is $\lim_{n\to\infty} v_i(n)$. This means that the $\boldsymbol{v}$ in (3.32) is the same as $\boldsymbol{w}$ here, and it is seen that the formulas are the same with g set to 0 in (3.37).

The reason that the derivation of aggregate reward was so simple for first-passage time is that there was no steady-state gain in that example, and thus no need to separate the gain per transition g from the relative gain $\boldsymbol{w}$ between starting states.

One way to apply the intuition of the $g = 0$ case to the general case is as follows: given a reward vector $\boldsymbol{r}$, find the steady-state gain $g = \boldsymbol{\pi r}$, and then define a modified reward vector $\boldsymbol{r'} = \boldsymbol{r} - g\boldsymbol{e}$. Changing the reward vector from $\boldsymbol{r}$ to $\boldsymbol{r'}$ in this way does not change $\boldsymbol{w}$, but the modified limiting aggregate gain, say $\boldsymbol{v'}(n)$ then has a limit, which is in fact $\boldsymbol{w}$. The intuitive derivation used in (3.32) again gives us $\boldsymbol{w} = [P]\boldsymbol{w} + \boldsymbol{r'}$. This is equivalent to (3.37) since $\boldsymbol{r'} = \boldsymbol{r} - g\boldsymbol{e}$.

There are many generalizations of the first-passage-time example in which the reward in each recurrent state of a unichain is 0. Thus reward is accumulated only until a recurrent

state is entered. The following corollary provides a monotonicity result about the relative-gain vector for these circumstances that might seem obvious[12]. Thus we simply state it and give a guided proof in Exercise 3.25.

Corollary 3.5.1. *Let $[P]$ be the transition matrix of a unichain with the recurrent class $\mathcal{R}$. Let $\boldsymbol{r} \geq 0$ be a reward vector for $[P]$ with $r_i = 0$ for $i \in \mathcal{R}$. Then the relative-gain vector $\boldsymbol{w}$ satisfies $\boldsymbol{w} \geq 0$ with $w_i = 0$ for $i \in \mathcal{R}$ and $w_i > 0$ for $r_i > 0$. Furthermore, if $\boldsymbol{r}'$ and $\boldsymbol{r}''$ are different reward vectors for $[P]$ and $\boldsymbol{r}' \geq \boldsymbol{r}''$ with $r'_i = r''_i$ for $i \in \mathcal{R}$, then $\boldsymbol{w}' \geq \boldsymbol{w}''$ with $w'_i = w''_i$ for $i \in \mathcal{R}$ and $w'_i > w''_i$ for $r'_i > r''_i$.*

3.5.3 The expected aggregate reward with an additional final reward

Frequently when a reward is aggregated over n transitions of a Markov chain, it is appropriate to assign some added reward, say u_i, as a function of the final state i. For example, it might be particularly advantageous to end in some particular state. Also, if we wish to view the aggregate reward over $n + \ell$ transitions as the reward over the first n transitions plus that over the following ℓ transitions, we can model the expected reward over the final ℓ transitions as a final reward at the end of the first n transitions. Note that this final expected reward depends only on the state at the end of the first n transitions.

As before, let $R(X_{m+h})$ be the reward at time $m + h$ for $0 \leq h \leq n - 1$ and $U(X_{m+n})$ be the final reward at time $m + n$, where $U(X) = u_i$ for $X = i$. Let $v_i(n, \boldsymbol{u})$ be the expected reward from time m to $m + n$, using the reward $\boldsymbol{r}$ from time m to $m + n - 1$ and using the final reward $\boldsymbol{u}$ at time $m+n$. The expected reward is then the following simple modification of (3.33):

$$\boldsymbol{v}(n, \boldsymbol{u}) \;=\; \boldsymbol{r} + [P]\boldsymbol{r} + \cdots + [P^{n-1}]\boldsymbol{r} + [P^n]\boldsymbol{u} \;=\; \sum_{h=0}^{n-1} [P^h]\boldsymbol{r} \;+\; [P^n]\boldsymbol{u}. \tag{3.38}$$

This simplifies considerably if $\boldsymbol{u}$ is taken to be the relative-gain vector $\boldsymbol{w}$.

Theorem 3.5.2. *Let $[P]$ be the transition matrix of a unichain and let $\boldsymbol{w}$ be the corresponding relative-gain vector. Then for each $n \geq 1$,*

$$\boldsymbol{v}(n, \boldsymbol{w}) = ng\boldsymbol{e} + \boldsymbol{w}. \tag{3.39}$$

Also, for an arbitrary final reward vector $\boldsymbol{u}$,

$$\boldsymbol{v}(n, \boldsymbol{u}) = ng\boldsymbol{e} + \boldsymbol{w} + [P^n](\boldsymbol{u} - \boldsymbol{w}). \tag{3.40}$$

Discussion: An important special case of (3.40) arises from setting the final reward $\boldsymbol{u}$ to 0, thus yielding the following expression for $\boldsymbol{v}(n)$:

$$\boldsymbol{v}(n) = ng\boldsymbol{e} + \boldsymbol{w} - [P^n]\boldsymbol{w}. \tag{3.41}$$

[12] An obvious counterexample if we omit the condition $r_i = 0$ for $i \in \mathcal{R}$ is given by Figure 3.7 where $\boldsymbol{r} = (0, 1)^{\mathsf{T}}$ and $\boldsymbol{w} = (-25, 25)^{\mathsf{T}}$.

For an ergodic unichain, $\lim_{n\to\infty}[P^n] = \boldsymbol{e}\boldsymbol{\pi}$. Since $\boldsymbol{\pi}\boldsymbol{w} = 0$ by definition of $\boldsymbol{w}$, the limit of (3.41) as $n \to \infty$ is

$$\lim_{n\to\infty}(\boldsymbol{v}(n) - ng\boldsymbol{e}) = \boldsymbol{w},$$

which agrees with (3.36). The advantage of (3.41) over (3.36) is that it provides an explicit expression for $\boldsymbol{v}(n)$ for each n and also that it continues to hold for a periodic unichain.

Proof: For $n = 1$, we see from (3.38) that

$$\boldsymbol{v}(1, \boldsymbol{w}) \;=\; \boldsymbol{r} + [P]\boldsymbol{w} \;=\; g\boldsymbol{e} + \boldsymbol{w},$$

so the theorem is satisfied for $n = 1$. For $n > 1$,

$$\begin{aligned}
\boldsymbol{v}(n, \boldsymbol{w}) &= \sum_{h=0}^{n-1}[P^h]\boldsymbol{r} \;+\; [P^n]\boldsymbol{w} \\
&= \sum_{h=0}^{n-2}[P^h]\boldsymbol{r} \;+\; [P^{n-1}]\,(\boldsymbol{r} + [P]\boldsymbol{w}) \\
&= \sum_{h=0}^{n-2}[P^h]\boldsymbol{r} \;+\; [P^{n-1}]\,(g\boldsymbol{e} + \boldsymbol{w}) \\
&= \boldsymbol{v}(n-1, \boldsymbol{w}) + g\boldsymbol{e}.
\end{aligned}$$

Using induction, this implies (3.39).

To establish (3.40), note from (3.38) that

$$\boldsymbol{v}(n, \boldsymbol{u}) - \boldsymbol{v}(n, \boldsymbol{w}) \;=\; [P^n](\boldsymbol{u} - \boldsymbol{w}).$$

Then (3.40) follows by using (3.39) for the value of $\boldsymbol{v}(n, \boldsymbol{w})$. □

3.6 Markov decision theory and dynamic programming

In the previous section, we analyzed the behavior of a Markov chain with rewards. In this section, we consider a much more elaborate structure in which a decision maker can choose among various possible rewards and transition probabilities. In place of the reward r_i and the transition probabilities $\{P_{ij}; 1 \le j \le \mathsf{M}\}$ associated with a given state i, there is a choice between some number K_i of different rewards, say $r_i^{(1)}, r_i^{(2)}, \ldots, r_i^{(K_i)}$ and a corresponding choice between K_i different sets of transition probabilities, say $\{P_{ij}^{(1)}; 1 \le j \le \mathsf{M}\}, \{P_{ij}^{(2)}, 1 \le j \le \mathsf{M}\}, \ldots \{P_{ij}^{(K_i)}; 1 \le j \le \mathsf{M}\}$. At each time m, a decision maker, given $X_m = i$, selects one of the K_i possible choices for state i. Note that if decision k is chosen in state i, then the reward is $r_i^{(k)}$ and the transition probabilities from i are $\{P_{ij}^{(k)}; 1 \le j \le \mathsf{M}\}$; it is not permissable to choose $r_i^{(k)}$ for one k and $\{P_{ij}^{(k)}; 1 \le j \le \mathsf{M}\}$ for another k. We also assume that if decision k is selected at time m, the probability of entering state j at time $m+1$ is $P_{ij}^{(k)}$, independent of earlier states and decisions.

Figure 3.9 shows an example of this situation in which the decision maker can choose between two possible decisions in state 2 ($K_2 = 2$), and has no freedom of choice in state 1 ($K_1 = 1$). This figure illustrates the familiar tradeoff between instant gratification (alternative 2) and long term gratification (alternative 1).



Figure 3.9: A Markov decision problem with two alternatives in state 2.

The set of rules used by the decision maker in selecting an alternative at each time is called a *policy*. We want to consider the expected aggregate reward over n steps of the "Markov chain" as a function of the policy used by the decision maker. If for each state i, the policy uses the same decision, say k_i, at each occurrence of i, then that policy corresponds to a homogeneous Markov chain with transition probabilities $P_{ij}^{(k_i)}$. We denote the matrix of these transition probabilities as $[P^{\boldsymbol{k}}]$, where $\boldsymbol{k} = (k_1, \ldots, k_{\mathsf{M}})$. Such a policy, i.e., mapping each state i into a fixed decision k_i, independent of time and past, is called a stationary policy. The aggregate gain for any such stationary policy was found in the previous section. Since both rewards and transition probabilities depend only on the state and the corresponding decision, and not on time, one feels intuitively that stationary policies make a certain amount of sense over a long period of time. On the other hand, if we look at the example of Figure 3.9, it is clear that decision 2 is the best choice in state 2 at the nth of n trials, but it is less obvious what to do at earlier trials.

In what follows, we first derive the optimal policy for maximizing expected aggregate reward over an arbitrary number n of trials, say at times m to $m + n - 1$. We shall see that the decision at time $m + h$, $0 \leq h < n$, for the optimal policy can in fact depend on h and n (but not m). It turns out to simplify matters considerably if we include a final reward $\{u_i; 1 \leq i \leq \mathsf{M}\}$ at time $m + n$. This final reward $\boldsymbol{u}$ is considered as a fixed vector, to be chosen as appropriate, rather than as part of the choice of policy.

This optimized strategy, as a function of the number of steps n and the final reward $\boldsymbol{u}$, is called an *optimal dynamic policy* for that $\boldsymbol{u}$. This policy is found from the dynamic programming algorithm, which, as we shall see, is conceptually very simple. We then go on to find the relationship between optimal dynamic policies and optimal stationary policies. We shall find that, under fairly general conditions, each has the same long-term gain per trial.

3.6.1 Dynamic programming algorithm

As in our development of Markov chains with rewards, we consider the expected aggregate reward over n time periods, say m to $m + n - 1$, with a final reward at time $m + n$. First consider the optimal decision with $n = 1$. Given $X_m = i$, a decision k is made with immediate reward $r_i^{(k)}$. With probability $P_{ij}^{(k)}$ the next state X_{m+1} is state j and the final

reward is then u_j. The expected aggregate reward over times m and $m+1$, maximized over the decision k, is then

$$v_i^*(1,\boldsymbol{u}) = \max_k\{r_i^{(k)} + \sum_j P_{ij}^{(k)} u_j\}. \tag{3.42}$$

Being explicit about the maximizing decision k', (3.42) becomes

$$v_i^*(1,\boldsymbol{u}) = r_i^{(k')} + \sum_j P_{ij}^{(k')} u_j \qquad \text{for } k' \text{ such that}$$

$$r_i^{(k')} + \sum_j P_{ij}^{(k')} u_j = \max_k\{r_i^{(k)} + \sum_j P_{ij}^{(k)} u_j\}. \tag{3.43}$$

Note that a decision is made only at time m, but that there are two rewards, one at time m and the other, the final reward, at time $m+1$. We use the notation $v_i^*(n,\boldsymbol{u})$ to represent the maximum expected aggregate reward from times m to $m+n$ starting at $X_m = i$. Decisions (with the reward vector $\boldsymbol{r}$) are made at the n times m to $m+n-1$, and this is followed by a final reward vector $\boldsymbol{u}$ (without any decision) at time $m+n$. It often simplifies notation to define the vector of maximal expected aggregate rewards

$$\boldsymbol{v}^*(n,\boldsymbol{u}) = (v_1^*(n,\boldsymbol{u}), v_2^*(n,\boldsymbol{u}), \ldots, v_{\mathsf{M}}^*(1,\boldsymbol{u}))^{\mathsf{T}}.$$

With this notation, (3.42) and (3.43) become

$$\boldsymbol{v}^*(1,\boldsymbol{u}) = \max_{\boldsymbol{k}}\{\boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{u}\} \quad \text{where } \boldsymbol{k} = (k_1,\ldots,k_{\mathsf{M}})^{\mathsf{T}},\ \boldsymbol{r}^{\boldsymbol{k}} = (r_1^{k_1},\ldots,r_{\mathsf{M}}^{k_{\mathsf{M}}})^{\mathsf{T}}. \tag{3.44}$$

$$\boldsymbol{v}^*(1,\boldsymbol{u}) = \boldsymbol{r}^{\boldsymbol{k}'} + [P^{\boldsymbol{k}'}]\boldsymbol{u} \qquad \text{where } \boldsymbol{r}^{\boldsymbol{k}'} + [P^{\boldsymbol{k}'}]\boldsymbol{u} = \max_{\boldsymbol{k}} \boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{u}. \tag{3.45}$$

Now consider $v_i^*(2,\boldsymbol{u})$, *i.e.*, the maximal expected aggregate reward starting at $X_m = i$ with decisions made at times m and $m+1$ and a final reward at time $m+2$. The key to dynamic programming is that an optimal decision at time $m+1$ can be selected based only on the state j at time $m+1$; this decision (given $X_{m+1} = j$) is optimal independent of the decision at time m. That is, whatever decision is made at time m, the maximal expected reward at times $m+1$ and $m+2$, given $X_{m+1} = j$, is $\max_k \left(r_j^{(k)} + \sum_\ell P_{j\ell}^{(k)} u_\ell\right)$. Note that this maximum is $v_j^*(1,\boldsymbol{u})$, as found in (3.42).

Using this optimized decision at time $m+1$, it is seen that if $X_m = i$ and decision k is made at time m, then the sum of expected rewards at times $m+1$ and $m+2$ is $\sum_j P_{ij}^{(k)} v_j^*(1,\boldsymbol{u})$. Adding the expected reward at time m and maximizing over decisions at time m,

$$v_i^*(2,\boldsymbol{u}) = \max_k \left(r_i^{(k)} + \sum\nolimits_j P_{ij}^{(k)} v_j^*(1,\boldsymbol{u})\right). \tag{3.46}$$

In other words, the maximum aggregate gain over times m to $m+2$ (using the final reward $\boldsymbol{u}$ at $m+2$) is the maximum over choices at time m of the sum of the reward at m plus the

maximum aggregate expected reward for $m+1$ and $m+2$. The simple expression of (3.46) results from the fact that the maximization over the choice at time $m+1$ depends on the state at $m+1$ but, given that state, is independent of the policy chosen at time m.

This same argument can be used for all larger numbers of trials. To find the maximum expected aggregate reward from time m to $m+n$, we first find the maximum expected aggregate reward from $m+1$ to $m+n$, conditional on $X_{m+1}=j$ for each state j. This is the same as the maximum expected aggregate reward from time m to $m+n-1$, which is $v_j^*(n-1,\boldsymbol{u})$. This gives us the general expression for $n\geq 2$,

$$v_i^*(n,\boldsymbol{u})=\max_k\left(r_i^{(k)}+\sum\nolimits_j P_{ij}^{(k)}v_j^*(n-1,\boldsymbol{u})\right). \tag{3.47}$$

We can also write this in vector form as

$$\boldsymbol{v}^*(n,\boldsymbol{u})=\max_{\boldsymbol{k}}\left(\boldsymbol{r}^{\boldsymbol{k}}+[P^{\boldsymbol{k}}]\boldsymbol{v}^*(n-1,\boldsymbol{u})\right). \tag{3.48}$$

Here $\boldsymbol{k}$ is a set (or vector) of decisions, $\boldsymbol{k}=(k_1,k_2,\ldots,k_{\mathsf{M}})^{\mathsf{T}}$, where k_i is the decision for state i. $[P^{\boldsymbol{k}}]$ denotes a matrix whose (i,j) element is $P_{ij}^{(k_i)}$, and $\boldsymbol{r}^{\boldsymbol{k}}$ denotes a vector whose ith element is $r_i^{(k_i)}$. The maximization over $\boldsymbol{k}$ in (3.48) is really M separate and independent maximizations, one for each state, i.e., (3.48) is simply a vector form of (3.47). Another frequently useful way to rewrite (3.48) is as follows:

$$\boldsymbol{v}^*(n,\boldsymbol{u})=\boldsymbol{r}^{\boldsymbol{k}'}+[P^{\boldsymbol{k}'}]\boldsymbol{v}^*(n{-}1,\boldsymbol{u})\quad\text{for }\boldsymbol{k}'\text{ such that}$$

$$\boldsymbol{r}^{\boldsymbol{k}'}+[P^{\boldsymbol{k}'}]\boldsymbol{v}^*(n{-}1,\boldsymbol{u})=\max_{\boldsymbol{k}}\left(\boldsymbol{r}^{\boldsymbol{k}}+[P^{\boldsymbol{k}}]\boldsymbol{v}^*(n{-}1,\boldsymbol{u})\right). \tag{3.49}$$

If $\boldsymbol{k}'$ satisfies (3.49), then $\boldsymbol{k}'$ is an optimal decision at an arbitrary time m given, first, that the objective is to maximize the aggregate gain from time m to $m+n$, second, that optimal decisions for this objective are to be made at times $m+1$ to $m+n-1$, and, third, that $\boldsymbol{u}$ is the final reward vector at $m+n$. In the same way, $\boldsymbol{v}^*(n,\boldsymbol{u})$ is the maximum expected reward over this finite sequence of n decisions from m to $m+n-1$ with the final reward $\boldsymbol{u}$ at $m+n$.

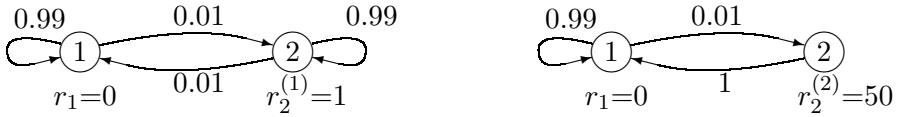
Note that (3.47), (3.48), and (3.49) are valid with no restrictions (such as recurrent or aperiodic states) on the possible transition probabilities $[P^{\boldsymbol{k}}]$. These equations are also valid in principle if the size of the state space is infinite. However, the optimization for each n can then depend on an infinite number of optimizations at $n-1$, which is often infeasible.

The *dynamic programming algorithm* is just the calculation of (3.47), (3.48), or (3.49), performed iteratively for $n=1,2,3,\ldots$. The development of this algorithm, as a systematic tool for solving this class of problems, is due to Bellman [Bel57]. Note that the algorithm is independent of the starting time m; the parameter n, usually referred to as stage n, is the number of decisions over which the aggregate gain is being optimized. This algorithm yields the optimal dynamic policy for any fixed final reward vector $\boldsymbol{u}$ and any given number of trials. Along with the calculation of $\boldsymbol{v}^*(n,\boldsymbol{u})$ for each n, the algorithm also yields the optimal decision at each stage (under the assumption that the optimal policy is to be used for each lower numbered stage, *i.e.*, for each later trial of the process).

The surprising simplicity of the algorithm is due to the Markov property. That is, $v_i^*(n, \boldsymbol{u})$ is the aggregate present and future reward conditional on the present state. Since it is conditioned on the present state, it is independent of the past (i.e., how the process arrived at state i from previous transitions and choices).

Although dynamic programming is computationally straightforward and convenient[13], the asymptotic behavior of $\boldsymbol{v}^*(n, \boldsymbol{u})$ as $n \to \infty$ is not evident from the algorithm. After working out some simple examples, we look at the general question of asymptotic behavior.

Example 3.6.1. Consider Figure 3.9, repeated below, with the final rewards $u_2 = u_1 = 0$.



Since $r_1 = 0$ and $u_1 = u_2 = 0$, the aggregate gain in state 1 at stage 1 is

$$v_1^*(1, \boldsymbol{u}) = r_1 + \sum_j P_{1j} u_j = 0.$$

Similarly, since policy 1 has an immediate reward $r_2^{(1)} = 1$ in state 2, and policy 2 has an immediate reward $r_2^{(2)} = 50$,

$$v_2^*(1, \boldsymbol{u}) = \max\left\{ \left[r_2^{(1)} + \sum_j P_{2j}^{(1)} u_j\right], \quad \left[r_2^{(2)} + \sum_j P_{2j}^{(2)} u_j\right] \right\} = \max\{1,\, 50\} = 50.$$

We can now go on to stage 2, using the results above for $v_j^*(1, \boldsymbol{u})$. From (3.46),

$$\begin{aligned} v_1^*(2, \boldsymbol{u}) &= r_1 + P_{11} v_1^*(1, \boldsymbol{u}) + P_{12} v_2^*(1, \boldsymbol{u}) = P_{12} v_2^*(1, \boldsymbol{u}) = 0.5 \\ v_2^*(2, \boldsymbol{u}) &= \max\left\{ \left[r_2^{(1)} + \sum_j P_{2j}^{(1)} v_j^*(1, \boldsymbol{u})\right], \quad \left[r_2^{(2)} + P_{21}^{(2)} v_1^*(1, \boldsymbol{u})\right] \right\} \\ &= \max\left\{ [1 + P_{22}^{(1)} v_2^*(1, \boldsymbol{u})],\, 50 \right\} = \max\{50.5,\, 50\} = 50.5. \end{aligned}$$

Thus for two trials, decision 1 is optimal in state 2 for the first trial (stage 2), and decision 2 is optimal in state 2 for the second trial (stage 1). What is happening is that the choice of decision 2 at stage 1 has made it very profitable to be in state 2 at stage 1. Thus if the chain is in state 2 at stage 2, it is preferable to choose decision 1 (i.e., the small unit gain) at stage 2 with the corresponding high probability of remaining in state 2 at stage 1. Continuing this computation for larger n, one finds that $v_1^*(n, \boldsymbol{u}) = n/2$ and $v_2^*(n, \boldsymbol{u}) = 50 + n/2$. The optimum dynamic policy (for $\boldsymbol{u} = 0$) is decision 2 for stage 1 (*i.e.*, for the last decision to be made) and decision 1 for all stages $n > 1$ (i.e., for all decisions before the last).

This example also illustrates that the maximization of expected gain is not necessarily what is most desirable in all applications. For example, risk-averse people might well prefer decision 2 at the next to final decision (stage 2). This guarantees a reward of 50, rather than taking a small chance of losing that reward.

[13]Unfortunately, many dynamic programming problems of interest have enormous numbers of states and possible choices of decision (the so called curse of dimensionality), and thus, even though the equations are simple, the computational requirements might be beyond the range of practical feasibility.

Example 3.6.2 (Shortest Path Problems). The problem of finding the shortest paths between nodes in a directed graph arises in many situations, from routing in communication networks to calculating the time to complete complex tasks. The problem is quite similar to the expected first-passage time of example 3.5.1. In that problem, arcs in a directed graph were selected according to a probability distribution, whereas here decisions must be made about which arcs to take. Although this is not a probabilistic problem, the decisions can be posed as choosing a given arc with probability one, thus viewing the problem as a special case of dynamic programming.

Consider finding the shortest path from each node in a directed graph to some particular node, say node 1 (see Figure 3.10). Each arc (except the special arc $(1,1)$) has a positive *link length* associated with it that might reflect physical distance or an arbitrary type of cost. The special arc (1, 1) has 0 link length. The length of a path is the sum of the lengths of the arcs on that path. In terms of dynamic programming, a policy is a choice of arc out of each node (state). Here we want to minimize cost (i.e., path length) rather than maximizing reward, so we simply replace the maximum in the dynamic programming algorithm with a minimum (or, if one wishes, all costs can be replaced with negative rewards).

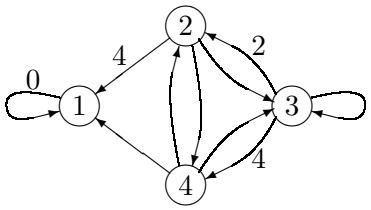


Figure 3.10: A shortest path problem. The arcs are marked with their lengths. Unmarked arcs have unit length.

We start the dynamic programming algorithm with a final cost vector that is 0 for node 1 and infinite for all other nodes. In stage 1, the minimal cost decision for node (state) 2 is arc $(2,1)$ with a cost equal to 4. The minimal cost decision for node 4 is $(4,1)$ with unit cost. The cost from node 3 (at stage 1) is infinite whichever decision is made. The stage 1 costs are then

$$v_1^*(1,\boldsymbol{u}) = 0, \quad v_2^*(1,\boldsymbol{u}) = 4, \quad v_3^*(1,\boldsymbol{u}) = \infty, \quad v_4^*(1,\boldsymbol{u}) = 1.$$

In stage 2, the cost $v_3^*(2,\boldsymbol{u})$, for example, is

$$v_3^*(2,\boldsymbol{u}) = \min\Big[2 + v_2^*(1,\boldsymbol{u}), \quad 4 + v_4^*(1,\boldsymbol{u})\Big] = 5.$$

The set of costs at stage 2 are

$$v_1^*(2,\boldsymbol{u}) = 0, \quad v_2^*(2,\boldsymbol{u}) = 2, \quad v_3^*(2,\boldsymbol{u}) = 5, \quad v_4^*(2,\boldsymbol{u}) = 1.$$

The decision at stage 2 is for node 2 to go to 4, node 3 to 4, and 4 to 1. At stage 3, node 3 switches to node 2, reducing its path length to 4, and nodes 2 and 4 are unchanged. Further iterations yield no change, and the resulting policy is also the optimal stationary policy.

The above results at each stage n can be interpreted as the shortest paths constrained to at most n hops. As n is increased, this constraint is successively relaxed, reaching the true shortest paths in less than M stages.

It can be seen without too much difficulty that these final aggregate costs (path lengths) also result no matter what final cost vector $\boldsymbol{u}$ (with $u_1 = 0$) is used. This is a useful feature for many types of networks where link lengths change very slowly with time and a shortest path algorithm is desired that can track the corresponding changes in the shortest paths.

3.6.2 Optimal stationary policies

In Example 3.6.1, we saw that there was a final transient (at stage 1) in which decision 1 was taken, and in all other stages, decision 2 was taken. Thus, the optimal dynamic policy consisted of a long-term stationary policy, followed by a transient period (for a single stage in this case) over which a different policy was used. It turns out that this final transient can be avoided by choosing an appropriate final reward vector $\boldsymbol{u}$ for the dynamic programming algorithm. If one has very good intuition, one would guess that the appropriate choice of final reward $\boldsymbol{u}$ is the relative-gain vector $\boldsymbol{w}$ associated with the long-term optimal policy.

It seems reasonable to expect this same type of behavior for typical but more complex Markov decision problems. In order to understand this, we start by considering an arbitrary stationary policy $\boldsymbol{k}' = (k'_1, \ldots, k'_{\mathsf{M}})$ and denote the transition matrix of the associated Markov chain as $[P^{\boldsymbol{k}'}]$. We assume that he associated Markov chain is a unichain, or, abbrevating terminology, that $\boldsymbol{k}'$ is a unichain. Let $\boldsymbol{w}'$ be the unique relative-gain vector for $\boldsymbol{k}'$. We then find some necessary conditions for $\boldsymbol{k}'$ to be the optimal dynamic policy at each stage using $\boldsymbol{w}'$ as the final reward vector.

First, from (3.45) $\boldsymbol{k}'$ is an optimal dynamic decision (with the final reward vector $\boldsymbol{w}'$ for $[P^{\boldsymbol{k}'}]$) at stage 1 if

$$\boldsymbol{r}^{\boldsymbol{k}'} + [P^{\boldsymbol{k}'}]\boldsymbol{w}' = \max_{\boldsymbol{k}} \{\boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{w}'\}. \tag{3.50}$$

Note that this is more than a simple statement that $\boldsymbol{k}'$ can be found by maximizing $\boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{w}'$ over $\boldsymbol{k}$. It also involves the fact that $\boldsymbol{w}'$ is the relative-gain vector for $\boldsymbol{k}'$, so there is no immediately obvious way to find a $\boldsymbol{k}'$ that satisfies (3.50), and no a priori assurance that this equation even has a solution. The following theorem, however, says that this is the only condition required to ensure that $\boldsymbol{k}'$ is the optimal dynamic policy at every stage (again using $\boldsymbol{w}'$ as the final reward vector).

Theorem 3.6.1. *Assume that (3.50) is satisfied for some policy $\boldsymbol{k}'$ where the Markov chain for $\boldsymbol{k}'$ is a unichain and $\boldsymbol{w}'$ is the relative-gain vector of $\boldsymbol{k}'$. Then the optimal dynamic policy, using $\boldsymbol{w}'$ as the final reward vector, is the stationary policy $\boldsymbol{k}'$. Furthermore the optimal gain at each stage n is given by*

$$\boldsymbol{v}^*(n, \boldsymbol{w}') = \boldsymbol{w}' + ng'\boldsymbol{e}, \tag{3.51}$$

where $g' = \boldsymbol{\pi}'\boldsymbol{r}^{\boldsymbol{k}'}$ and $\boldsymbol{\pi}'$ is the steady-state vector for $\boldsymbol{k}'$.

Proof: We have seen from (3.45) that $\boldsymbol{k}'$ is an optimal dynamic decision at stage 1. Also, since $\boldsymbol{w}'$ is the relative-gain vector for $\boldsymbol{k}'$, Theorem 3.5.2 asserts that if decision $\boldsymbol{k}'$ is used at each stage, then the aggregate gain satisfies $\boldsymbol{v}(n, \boldsymbol{w}') = ng'\boldsymbol{e} + \boldsymbol{w}'$. Since $\boldsymbol{k}'$ is optimal at stage 1, it follows that (3.51) is satisfied for $n = 1$.

We now use induction on n, with $n = 1$ as a basis, to verify (3.51) and the optimality of this same $\boldsymbol{k}'$ at each stage n. Thus, assume that (3.51) is satisfied for n. Then, from (3.48),

$$\begin{aligned}
\boldsymbol{v}^*(n+1, \boldsymbol{w}') &= \max_{\boldsymbol{k}}\{\boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{v}^*(n, \boldsymbol{w}')\} && (3.52)\\
&= \max_{\boldsymbol{k}}\Big\{\boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\{\boldsymbol{w}' + ng'\boldsymbol{e}\}\Big\} && (3.53)\\
&= ng'\boldsymbol{e} + \max_{\boldsymbol{k}}\{\boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{w}'\} && (3.54)\\
&= ng'\boldsymbol{e} + \boldsymbol{r}^{\boldsymbol{k}'} + [P^{\boldsymbol{k}'}]\boldsymbol{w}'\} && (3.55)\\
&= (n+1)g'\boldsymbol{e} + \boldsymbol{w}'. && (3.56)
\end{aligned}$$

Eqn (3.53) follows from the inductive hypothesis of (3.51), (3.54) follows because $[P^{\boldsymbol{k}}]\boldsymbol{e} = \boldsymbol{e}$ for all $\boldsymbol{k}$, (3.55) follows from (3.50), and (3.56) follows from the definition of $\boldsymbol{w}'$ as the relative-gain vector for $\boldsymbol{k}'$. This verifies (3.51) for $n + 1$. Also, since $\boldsymbol{k}'$ maximizes (3.54), it also maximizes (3.52), showing that $\boldsymbol{k}'$ is the optimal dynamic decision at stage $n + 1$. This completes the inductive step. □

Since our major interest in stationary policies is to help understand the relationship between the optimal dynamic policy and stationary policies, we define an optimal stationary policy as follows:

Definition 3.6.1. *A unichain stationary policy $\boldsymbol{k}'$ is optimal if the optimal dynamic policy with $\boldsymbol{w}'$ as the final reward uses $\boldsymbol{k}'$ at each stage.*

This definition side-steps several important issues. First, we might be interested in dynamic programming for some other final reward vector. Is it possible that dynamic programming performs much better in some sense with a different final reward vector. Is it possible that there is another stationary policy, especially one with a larger gain per stage? We answer these questions later and find that stationary policies that are optimal according to the definition do have maximal gain per stage compared with dynamic policies with arbitrary final reward vectors.

From Theorem 3.6.1, we see that if there is a policy $\boldsymbol{k}'$ which is a unichain with relative-gain vector $\boldsymbol{w}'$, and if that $\boldsymbol{k}'$ is a solution to (3.50), then $\boldsymbol{k}'$ is an optimal stationary policy.

It is easy to imagine Markov decision models for which each policy corresponds to a Markov chain with multiple recurrent classes. There are many special cases of such situations, and their detailed study is inappropriate in an introductory treatment. The essential problem with such models is that it is possible to get into various sets of states from which there is no exit, no matter what decisions are used. These sets might have different gains, so that there is no meaningful overall gain per stage. We avoid these situations by a modeling assumption called *inherent reachability*, which assumes, for each pair (i, j) of states, that there is some decision vector $\boldsymbol{k}$ containing a path from i to j.

The concept of inherent reachability is a little tricky, since it does not say the same $\boldsymbol{k}$ can be used for all pairs of states (*i.e.*, that there is some $\boldsymbol{k}$ for which the Markov chain is recurrent). As shown in Exercise 3.31, however, inherent reachability does imply that for any state j, there is a $\boldsymbol{k}$ for which j is accessible from all other states. As we have seen a number of times, this implies that the Markov chain for $\boldsymbol{k}$ is a unichain in which j is a recurrent state.

Any desired model can be modified to satisfy inherent reachability by creating some new decisions with very large negative rewards; these allow for such paths but very much discourage them. This will allow us to construct optimal unichain policies, but also to use the appearance of these large negative rewards to signal that there was something questionable in the original model.

3.6.3 Policy improvement and the seach for optimal stationary policies

The general idea of policy improvement is to start with an arbitrary unichain stationary policy $\boldsymbol{k'}$ with a relative gain vector $\boldsymbol{w'}$ (as given by (3.37)). We assume inherent reachability throughout this section, so such unichains must exist. We then check whether (3.50), is satisfied, and if so, we know from Theorem 3.6.1 that $\boldsymbol{k'}$ is an optimal stationary policy. If not, we find another stationary policy $\boldsymbol{k}$ that is 'better' than $\boldsymbol{k'}$ in a sense to be described later. Unfortunately, the 'better' policy that we find might not be a unichain, so it will also be necessary to convert this new policy into an equally 'good' unichain policy. This is where the assumption of of inherent reachability is needed. The algorithm then iteratively finds better and better unichain stationary policies, until eventually one of them satisfies (3.50) and is thus optimal.

We now state the policy-improvement algorithm for inherently reachable Markov decision problems. This algorithm is a generalization of Howard's policy-improvement algorithm, [How60].

Policy-improvement Algorithm

1. Choose an arbitrary unichain policy $\boldsymbol{k'}$
2. For policy $\boldsymbol{k'}$, calculate $\boldsymbol{w'}$ and g' from $\boldsymbol{w'} + g'\boldsymbol{e} = \boldsymbol{r}^{\boldsymbol{k'}} + [P^{\boldsymbol{k'}}]\boldsymbol{w'}$ and $\boldsymbol{\pi' w'} = 0$
3. If $\boldsymbol{r}^{\boldsymbol{k'}} + [P^{\boldsymbol{k'}}]\boldsymbol{w'} = \max_{\boldsymbol{k}}\{\boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{w'}\}$, then stop; $\boldsymbol{k'}$ is optimal.
4. Otherwise, choose ℓ and k_ℓ so that $r_\ell^{(k'_\ell)} + \sum_j P_{\ell j}^{(k'_\ell)} w'_j < r_\ell^{(k_\ell)} + \sum_j P_{\ell j}^{(k_\ell)} w'_j$. For $i \neq \ell$, let $k_i = k'_i$.
5. If $\boldsymbol{k} = (k_1, \ldots k_{\mathsf{M}})$ is not a unichain, then let $\mathcal{R}$ be the recurrent class in $\boldsymbol{k}$ that contains state ℓ, and let $\tilde{\boldsymbol{k}}$ be a unichain policy for which $\tilde{k}_i = k_i$ for each $i \in \mathcal{R}$. Alternatively, if $\boldsymbol{k}$ is already a unichain, let $\tilde{\boldsymbol{k}} = \boldsymbol{k}$.
6. Update $\boldsymbol{k'}$ to the value of $\tilde{\boldsymbol{k}}$ and return to step 2.

If the stopping test in step 3 fails, there must be an ℓ and k_ℓ for which $r_\ell^{(k'_\ell)} + \sum_j P_{\ell j}^{(k'_\ell)} w'_j < r_\ell^{(k_\ell)} + \sum_j P_{\ell j}^{(k_\ell)} w'_j$. Thus step 4 can always be executed if the algorithm does not stop in

step 3, and since the decision is changed only for the single state ℓ, the resulting policy $\boldsymbol{k}$ satisfies

$$\boldsymbol{r}^{\boldsymbol{k}'} + [p^{\boldsymbol{k}'}]\boldsymbol{w}' \leq \boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{w}' \qquad \text{with strict inequality for component } \ell. \tag{3.57}$$

The next three lemmas consider the different cases for the state ℓ whose decision is changed in step 4 of the algorithm. Taken together, they show that each iteration of the algorithm either increases the gain per stage or keeps the gain per stage constant while increasing the relative gain vector. After proving these lemmas, we return to show that the algorithm must converge and explain the sense in which the resulting stationary algorithm is optimal.

For each of the lemmas, let $\boldsymbol{k}'$ be the decision vector in step 1 of a given iteration of the policy improvement algorithm and assume that the Markov chain for $\boldsymbol{k}'$ is a unichain. Let g', $\boldsymbol{w}'$, and $\mathcal{R}'$ respectively be the gain per stage, the relative gain vector, and the recurrent set of states for $\boldsymbol{k}'$. Assume that the stopping condition in step 3 is not satisfied and that ℓ denotes the state whose decision is changed. Let k_ℓ be the new decision in step 4 and let $\boldsymbol{k}$ be the new decision vector.

Lemma 3.6.1. *Assume that $\ell \in \mathcal{R}'$. Then the Markov chain for $\boldsymbol{k}$ is a unichain and ℓ is recurrent in $\boldsymbol{k}$. The gain per stage g for $\boldsymbol{k}$ satisfies $g > g'$.*

Proof: The Markov chain for $\boldsymbol{k}$ is the same as that for $\boldsymbol{k}'$ except for the transitions out of state ℓ. Thus every path into ℓ in $\boldsymbol{k}'$ is still a path into ℓ in $\boldsymbol{k}$. Since ℓ is recurrent in the unichain $\boldsymbol{k}'$, it is accessible from all states in $\boldsymbol{k}'$ and thus in $\boldsymbol{k}$. It follows (see Exercise 3.3) that ℓ is recurrent in $\boldsymbol{k}$ and $\boldsymbol{k}$ is a unichain. Since $\boldsymbol{r}^{\boldsymbol{k}'} + [P^{\boldsymbol{k}'}]\boldsymbol{w}' = \boldsymbol{w}' + g'\boldsymbol{e}$ (see (3.37)), we can rewrite (3.57) as

$$\boldsymbol{w}' + g'\boldsymbol{e} \leq \boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{w}' \qquad \text{with strict inequality for component } \ell. \tag{3.58}$$

Premultiplying both sides of (3.58) by the steady-state vector $\boldsymbol{\pi}$ of the Markov chain $\boldsymbol{k}$ and using the fact that ℓ is recurent and thus $\pi_\ell > 0$,

$$\boldsymbol{\pi}\boldsymbol{w}' + g' < \boldsymbol{\pi}\boldsymbol{r}^{\boldsymbol{k}} + \boldsymbol{\pi}[P^{\boldsymbol{k}}]\boldsymbol{w}'.$$

Since $\boldsymbol{\pi}[P^{\boldsymbol{k}}] = \boldsymbol{\pi}$, this simplifies to

$$g' < \boldsymbol{\pi}\boldsymbol{r}^{\boldsymbol{k}}. \tag{3.59}$$

The gain per stage g for $\boldsymbol{k}$ is $\boldsymbol{\pi}\boldsymbol{r}^{\boldsymbol{k}}$, so we have $g' < g$. □

Lemma 3.6.2. *Assume that $\ell \notin \mathcal{R}'$ (*i.e.*, ℓ is transient in $\boldsymbol{k}'$) and that the states of $\mathcal{R}'$ are not accessible from ℓ in $\boldsymbol{k}$. Then $\boldsymbol{k}$ is not a unichain and ℓ is recurrent in $\boldsymbol{k}$. A decision vector $\tilde{\boldsymbol{k}}$ exists that is a unichain for which $\tilde{k}_i = k_i$ for $i \in \mathcal{R}$, and its gain per stage $\tilde{g}$ satisfies $\tilde{g} > g$.*

Proof: Since $\ell \notin \mathcal{R}'$, the transition probabilities from the states of $\mathcal{R}'$ are unchanged in going from $\boldsymbol{k}'$ to $\boldsymbol{k}$. Thus the set of states accessible from $\mathcal{R}'$ remains unchanged, and $\mathcal{R}'$ is a recurrent set of $\boldsymbol{k}$. Since $\mathcal{R}'$ is not accessible from ℓ, there must be another recurrent

set, $\mathcal{R}$, in $\boldsymbol{k}$, and thus $\boldsymbol{k}$ is not a unichain. The states accessible from $\mathcal{R}$ no longer include $\mathcal{R}'$, and since ℓ is the only state whose transition probabilities have changed, all states in $\mathcal{R}$ have paths to ℓ in $\boldsymbol{k}$. It follows that $\ell \in \mathcal{R}$.

Now let $\boldsymbol{\pi}$ be the steady-state vector for $\mathcal{R}$ in the Markov chain for $\boldsymbol{k}$. Since $\pi_\ell > 0$, (3.58) and (3.59) are still valid for this situation. Let $\tilde{\boldsymbol{k}}$ be a decision vector for which $\tilde{k}_i = k_i$ for each $i \in \mathcal{R}$. Using inherent reachability, we can also choose $\tilde{k}_i$ for each $i \notin \mathcal{R}$ so that ℓ is reachable from i (see Exercise 3.31). Thus $\tilde{\boldsymbol{k}}$ is a unichain with the recurrent class $\mathcal{R}$. Since $\tilde{\boldsymbol{k}}$ has the same transition probabilities and rewards in $\mathcal{R}$ as $\boldsymbol{k}$, we see that $\tilde{g} = \pi \boldsymbol{r}^{\boldsymbol{k}}$ and thus $\tilde{g} > g'$. □

The final lemma now includes all cases not in Lemmas 3.6.1 and 3.6.2

Lemma 3.6.3. *Assume that $\ell \notin \mathcal{R}'$ and that $\mathcal{R}'$ is accessible from ℓ in $\boldsymbol{k}$. Then $\boldsymbol{k}$ is a unichain with the same recurrent set $\mathcal{R}'$ as $\boldsymbol{k}'$. The gain per stage g is equal to g' and the relative-gain vector $\boldsymbol{w}$ of $\boldsymbol{k}$ satisfies*

$$\boldsymbol{w}' \leq \boldsymbol{w} \qquad \text{with } w'_\ell < w_\ell \text{ and } w'_i = w_i \text{ for } i \in \mathcal{R}'. \tag{3.60}$$

Proof: Since $\boldsymbol{k}'$ is a unichain, $\boldsymbol{k}'$ contains a path from each state to $\mathcal{R}'$. If such a path does not go through state ℓ, then $\boldsymbol{k}$ also contains that path. If such a path does go through ℓ, then that path can be replaced in $\boldsymbol{k}$ by the same path to ℓ followed by a path in $\boldsymbol{k}$ from ℓ to $\mathcal{R}'$. Thus $\mathcal{R}'$ is accessible from all states in $\boldsymbol{k}$. Since the states accessible from $\mathcal{R}'$ are unchanged from $\boldsymbol{k}'$ to $\boldsymbol{k}$, $\boldsymbol{k}$ is still a unichain with the recurrent set $\mathcal{R}'$ and state ℓ is still transient.

If we write out the defining equation (3.37) for $\boldsymbol{w}'$ component by component, we get

$$w'_i + g' = r_i^{k'_i} + \sum_j P_{ij}^{k'_i} w'_j. \tag{3.61}$$

Consider the set of these equations for which $i \in \mathcal{R}'$. Since $P_{ij}^{k'_i} = 0$ for all transient j in $\boldsymbol{k}'$, these are the same relative-gain equations as for the Markov chain restricted to $\mathcal{R}'$. Therefore $\boldsymbol{w}'$ is uniquely defined for $i \in \mathcal{R}'_i$ by this restricted set of equations. These equations are not changed in going from $\boldsymbol{k}'$ to $\boldsymbol{k}$, so it follows that $w_i = w'_i$ for $i \in \mathcal{R}'$. We have also seen that the steady-state vector $\boldsymbol{\pi}'$ is determined solely by the transition probabilities in the recurrent class, so $\boldsymbol{\pi}'$ is unchanged from $\boldsymbol{k}'$ to $\boldsymbol{k}$, and $g = g'$.

Finally, consider the difference between the relative-gain equations for $\boldsymbol{k}'$ in 3.61 and those for $\boldsymbol{k}$. Since $g' = g$,

$$w_i - w'_i = r_i^{k_i} - r_i^{k'_i} + \sum_j \left(P_{ij}^{k_i} w_j - P_{ij}^{k'_i} w'_j \right). \tag{3.62}$$

For all $i \neq \ell$, this simplifies to

$$w_i - w'_i = \sum_j P_{ij}^{k_i} (w_j - w'_j). \tag{3.63}$$

For $i = \ell$, (3.62) can be rewritten as

$$w_\ell - w'_\ell = \sum_j P_{\ell j}^{k_\ell}(w_j - w'_j) + \left[r_\ell^{k_\ell} - r_\ell^{k'_\ell} + \sum_j \left(P_{\ell j}^{k_\ell} w'_j - P_{\ell j}^{k'_\ell} w'_j \right) \right]. \tag{3.64}$$

The quantity in brackets must be positive because of step 4 of the algorithm, and we denote it as $\hat{r}_\ell - \hat{r}'_\ell$. If we also define $\hat{r}_i = \hat{r}'_i$ for $i \neq \ell$, then we can apply the last part of Corollary 3.5.1 (using $\hat{\boldsymbol{r}}$ and $\hat{\boldsymbol{r}}'$ as reward vectors) to conclude that $\boldsymbol{w} \geq \boldsymbol{w}'$ with $w_\ell > w'_\ell$. □

We now see that each iteration of the algorithm either increases the gain per stage or holds the gain per stage the same and increases the relative-gain vector $\boldsymbol{w}$. Thus the sequence of policies found by the algorithm can never repeat. Since there are a finite number of stationary policies, the algorithm must eventually terminate at step 3. This means that the optimal dynamic policy using the final reward vector $\boldsymbol{w}'$ for the terminating decision vector $\boldsymbol{k}'$ must in fact be the stationary policy $\boldsymbol{k}'$.

The question now arises whether the optimal dynamic policy using some other final reward vector can be substantially better than that using $\boldsymbol{w}'$. The answer is quite simple and is developed in Exercise 3.30. It is shown there that if $\boldsymbol{u}$ and $\boldsymbol{u}'$ are arbitrary final reward vectors used on the dynamic programming algorithm, then $v^*(n, \boldsymbol{u})$ and $v^*(n, \boldsymbol{u}')$ are related by

$$v^*(n, \boldsymbol{u}) \leq v^*(n, \boldsymbol{u}') + \alpha \boldsymbol{e},$$

where $\alpha = \max_i(u_i - u'_i)$. Using $\boldsymbol{w}'$ for $\boldsymbol{u}'$, it is seen that the gain per stage of dynamic programming, with any final reward vector, is at most the gain g' of the stationary policy at the termination of the policy-improvement algorithm.

The above results are summarized in the following theorem.

Theorem 3.6.2. *For any inherently reachable finite-state Markov decision problem, the policy-improvement algorithm terminates with a stationary policy $\boldsymbol{k}'$ that is the same as the solution to the dynamic programming algorithm using $\boldsymbol{w}'$ as the final reward vector. The gain per stage g' of this stationary policy maximizes the gain per stage over all stationary policies and over all final-reward vectors for the dynamic programming algorithm.*

One remaining issue is the question whether the relative-gain vector found by the policy-improvement algorithm is in any sense optimal. The example in Figure 3.11 illustrates two different solutions terminating the policy-improvement algorithm. They each have the same gain (as guaranteed by Theorem 3.6.2) but their relative-gain vectors are not ordered.

In many applications such as variations on the shortest path problem, the interesting issue is what happens before the recurrent class is entered, and there is often only one recurrent class and one set of decisions within that class of interest. The following corollary shows that in this case, the relative-gain vector for the stationary policy that terminates the algorithm is maximal not only among the policies visited by the algorithm but among all policies with the same recurrent class and the same decisions within that class. The proof is almost the same as that of Lemma 3.6.3 and is carried out in Exercise 3.33.

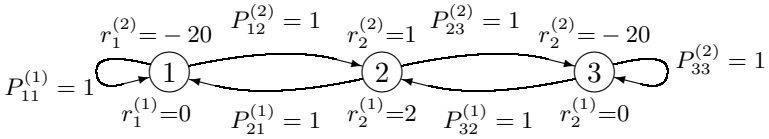


Figure 3.11: A Markov decision problem in which there are two unichain decision vectors (one left-going, and the other right-going). For each, (3.50) is satisfied and the gain per stage is 0. The dynamic programming algorithm (with no final reward) is stationary but has two recurrent classes, one of which is $\{3\}$, using decision 2 and the other of which is $\{1,2\}$, using decision 1 in each state.

Corollary 3.6.1. *Assume the policy improvement algorithm terminates with the recurrent class $\mathcal{R}'$, the decision vector $\boldsymbol{k}'$, and the relative-gain vector $\boldsymbol{w}'$. Then for any stationary policy that has the recurrent class $\mathcal{R}'$ and a decision vector $\boldsymbol{k}$ satisfying $k_i = k_i'$ for all $i \in \mathcal{R}'$, the relative gain vector $\boldsymbol{w}$ satisfies $\boldsymbol{w} \leq \boldsymbol{w}'$.*

3.7 Summary

This chapter has developed the basic results about finite-state Markov chains. It was shown that the states of any finite-state chain can be partitioned into classes, where each class is either transient or recurrent, and each class is periodic or aperiodic. If a recurrent class is periodic of period d, then the states in that class can be partitioned into d subsets where each subset has transitions only into the next subset.

The transition probabilities in the Markov chain can be represented as a matrix $[P]$, and the n-step transition probabilities are given by the matrix product $[P^n]$. If the chain is ergodic, *i.e.*, one aperiodic recurrent class, then the limit of the n-step transition probabilities become independent of the initial state, *i.e.*, $\lim_{n\to\infty} P_{ij}^n = \pi_j$ where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{\mathsf{M}})$ is called the steady-state probability. Thus the limiting value of $[P^n]$ is an M by M matrix whose rows are all the same, *i.e.*, the limiting matrix is the product $\boldsymbol{e\pi}$. The steady state probabilities are uniquely specified by $\sum_j \pi_i P_{ij} = \pi_j$ and $\sum_i \pi_i = 1$. That unique solution must satisfy $\pi_i > 0$ for all i. The same result holds (see Theorem 3.3.2) for aperidodic unichains with the exception that $\pi_i = 0$ for all transient states.

The eigenvalues and eigenvectors of $[P]$ are useful in many ways, but in particular provide precise results about how P_{ij}^n approaches π_j with increasing n. An eigenvalue equal to 1 always exists, and its multiplicity is equal to the number of recurrent classes. For each recurrent class, there is a left eigenvector $\boldsymbol{\pi}$ of eigenvalue 1. It is the steady-state vector for the given recurrent class. If a recurrent class is periodic with period d, then there are d corresponding eigenvalues of magnitude 1 uniformly spaced around the unit circle. The left eigenvector corresponding to each is nonzero only on that periodic recurrent class.

All other eigenvalues of $[P]$ are less than 1 in magnitude. If the eigenvectors of the entire set of eigenvalues span M dimensional space, then $[P^n]$ can be represented by (3.30) which shows explicitly how steady state is approached for aperiodic recurrent classes of states. If the eigenvectors do not span M-space, then (3.30) can be replaced by a Jordan form.

For an arbitrary finite-state Markov chain, if the initial state is transient, then the Markov chain will eventually enter a recurrent state, and the probability that this takes more than n steps approaches zero geometrically in n; Exercise 3.18 shows how to find the probability that each recurrent class is entered. Given an entry into a particular recurrent class, the results about recurrent chains can be used to analyze the behavior within that class.

The results about Markov chains were extended to Markov chains with rewards. The use of reward functions (or cost functions) provides a systematic way to approach a large class of problems ranging from first-passage times to dynamic programming. For unichains, the key result here is Theorem 3.5.2, which provides both an exact expression and an asymptotic expression for the expected aggregate reward over n stages. Markov chains with rewards and multiple recurrent classes are best handled by considering the individual recurrent classes separately.

Finally, the results on Markov chains with rewards were used to understand Markov decision theory. The Bellman dynamic programming algorithm was developed, and the policy improvement algorithm was discussed and analyzed. Theorem 3.6.2 demonstrated the relationship between the optimal dynamic policy and the optimal stationary policy. This section provided only an introduction to dynamic programming and omitted all discussion of discounting (in which future gain is considered worth less than present gain because of interest rates). The development was also restricted to finite-state spaces.

For a review of vectors, matrices, and linear algebra, see any introductory text on linear algebra such as Strang [19]. For further reading on Markov decision theory and dynamic programming, see Bertsekas, [3]. Bellman [1] is of historic interest and quite readable.

3.8 Exercises

Exercise 3.1. Let $[P]$ be the transition matrix for a finite state Markov chain and let state i be recurrent. Prove that i is aperiodic if $P_{ii} > 0$.

Exercise 3.2. Show that every Markov chain with $\mathsf{M} < \infty$ states contains at least one recurrent set of states. Explaining each of the following statements is sufficient.

a) If state i_1 is transient, then there is some other state i_2 such that $i_1 \to i_2$ and $i_2 \not\to i_1$.

b) If the i_2 of part a) is also transient, there is an i_3 such that $i_2 \to i_3$, $i_3 \not\to i_2$, and consequently $i_1 \to i_3$, $i_3 \not\to i_1$.

c) Continuing inductively, if i_k is also transient, there is an i_{k+1} such that $i_j \to i_{k+1}$ and $i_{k+1} \not\to i_j$ for $1 \leq j \leq k$.

d) Show that for some $k \leq \mathsf{M}$, k is not transient, *i.e.*, it is recurrent, so a recurrent class exists.

Exercise 3.3. Consider a finite-state Markov chain in which some given state, say state 1, is accessible from every other state. Show that the chain has at most one recurrent class

$\mathcal{R}$ of states and state $1 \in \mathcal{R}$. (Note that, combined with Exercise 3.2, there is exactly one recurrent class and the chain is then a unichain.)

Exercise 3.4. Show how to generalize the graph in Figure 3.4 to an arbitrary number of states $\mathsf{M} \geq 3$ with one cycle of M nodes and one of $\mathsf{M}-1$ nodes. For $\mathsf{M}=4$, let node 1 be the node not in the cycle of $\mathsf{M}-1$ nodes. List the set of states accessible from node 1 in n steps for each $n \leq 12$ and show that the bound in Theorem 3.2.4 is met with equality. Explain why the same result holds for all larger M.

Exercise 3.5. (Proof of Theorem 3.2.4)

a) Show that an ergodic Markov chain with M states must contain a cycle with $\tau < \mathsf{M}$ states. Hint: Use ergodicity to show that the smallest cycle cannot contain M states.

b) Let ℓ be a fixed state on this cycle of length τ. Let $\mathcal{T}(m)$ be the set of states accessible from ℓ in m steps. Show that for each $m \geq 1$, $\mathcal{T}(m) \subseteq \mathcal{T}(m+\tau)$. Hint: For any given state $j \in \mathcal{T}(m)$, show how to construct a walk of $m+\tau$ steps from ℓ to j from the assumed walk of m steps.

c) Define $\mathcal{T}(0)$ to be the singleton set $\{i\}$ and show that

$$\mathcal{T}(0) \subseteq \mathcal{T}(\tau) \subseteq \mathcal{T}(2\tau) \subseteq \cdots \subseteq \mathcal{T}(n\tau) \subseteq \cdots .$$

d) Show that if one of the inclusions above is satisfied with equality, then all subsequent inclusions are satisifed with equality. Show from this that at most the first $M-1$ inclusions can be satisfied with strict inequality and that $\mathcal{T}(n\tau) = \mathcal{T}((M-1)\tau)$ for all $n \geq M-1$.

e) Show that all states are included in $\mathcal{T}((M-1)\tau)$.

f) Show that $P_{ij}^{(\mathsf{M}-1)^2+1} > 0$ for all i, j.

Exercise 3.6. Consider a Markov chain with one ergodic class of m states, say $\{1, 2, \ldots, m\}$ and $\mathsf{M}-m$ other states that are all transient. Show that $P_{ij}^n > 0$ for all $j \leq m$ and $n \geq (m-1)^2 + 1 + \mathsf{M} - m$.

Exercise 3.7. a) Let τ be the number of states in the smallest cycle of an arbitrary ergodic Markov chain of $\mathsf{M} \geq 3$ states. Show that $P_{ij}^n > 0$ for all $n \geq (\mathsf{M}-2)\tau + \mathsf{M}$. Hint: Look at the proof of Theorem 3.2.4 in Exercise 3.5.

b) For $\tau = 1$, draw the graph of an ergodic Markov chain (generalized for arbitrary $\mathsf{M} \geq 3$) for which there is an i, j for which $P_{ij}^n = 0$ for $n = 2\mathsf{M}-3$. Hint: Look at Figure 3.4.

c) For arbitrary $\tau < \mathsf{M}-1$, draw the graph of an ergodic Markov chain (generalized for arbitrary M) for which there is an i, j for which $P_{ij}^n = 0$ for $n = (\mathsf{M}-2)\tau + \mathsf{M} - 1$.

Exercise 3.8. A transition probability matrix $[P]$ is said to be doubly stochastic if

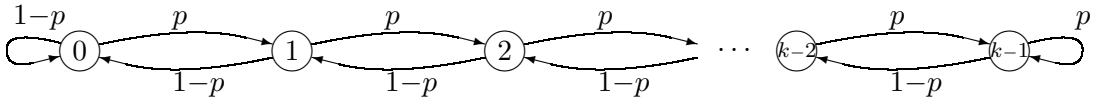
$$\sum_j P_{ij} = 1 \quad \text{for all } i; \quad \sum_i P_{ij} = 1 \quad \text{for all } j.$$

That is, the row sum and the column sum each equal 1. If a doubly stochastic chain has M states and is ergodic (i.e., has a single class of states and is aperiodic), calculate its steady-state probabilities.

Exercise 3.9. **a)** Find the steady-state probabilities $\pi_0, \ldots, \pi_{k-1}$ for the Markov chain below. Express your answer in terms of the ratio $\rho = p/q$. Pay particular attention to the special case $\rho = 1$.

b) Sketch $\pi_0, \ldots, \pi_{k-1}$. Give one sketch for $\rho = 1/2$, one for $\rho = 1$, and one for $\rho = 2$.

c) Find the limit of π_0 as k approaches ∞; give separate answers for $\rho < 1$, $\rho = 1$, and $\rho > 1$. Find limiting values of π_{k-1} for the same cases.



Exercise 3.10. **a)** Find the steady-state probabilities for each of the Markov chains in Figure 3.2. Assume that all clockwise probabilities in the first graph are the same, say p, and assume that $P_{4,5} = P_{4,1}$ in the second graph.

b) Find the matrices $[P]^2$ for the same chains. Draw the graphs for the Markov chains represented by $[P]^2$, i.e., the graph of two step transitions for the original chains. Find the steady-state probabilities for these two step chains. Explain why your steady-state probabilities are not unique.

c) Find $\lim_{n\to\infty}[P^{2n}]$ for each of the chains.

Exercise 3.11. **a)** Assume that $\boldsymbol{\nu}^{(i)}$ is a right eigenvector and $\boldsymbol{\pi}^{(j)}$ is a left eigenvector of an M by M stochastic matrix $[P]$ where $\lambda_i \neq \lambda_j$. Show that $\boldsymbol{\pi}^{(j)}\boldsymbol{\nu}^{(i)} = 0$. Hint: Consider two ways of finding $\boldsymbol{\pi}^{(j)}[P]\boldsymbol{\nu}^{(i)}$.

b) Assume that $[P]$ has M distinct eigenvalues. The right eigenvectors of $[P]$ then span M space (see section 5.2 of Strang, [19]), so the matrix $[U]$ with those eigenvectors as columns is nonsingular. Show that U^{-1} is a matix whose rows are the M left eigenvectors of $[P]$. Hint: use part a).

c) For each i, let $[\Lambda^{(i)}]$ be a diagonal matrix with a single nonzero element, $[\Lambda^{(i)}_{ii}] = \lambda_i$. Assume that $\boldsymbol{\pi}_i\boldsymbol{\nu}_k = 0$. Show that

$$\boldsymbol{\nu}^{(j)}[\Lambda^{(i)}]\boldsymbol{\pi}^{(k)} = \lambda_i\delta_{ik}\delta_{jk},$$

where δ_{ik} is 1 if $i = k$ and 0 otherwise. Hint visualize straightforward vector/matrix multiplication.

d) Verify (3.30).

Exercise 3.12. **a)** Let λ_k be an eigenvalue of a stochastic matrix $[P]$ and let $\boldsymbol{\pi}^{(k)}$ be an eigenvector for λ_k. Show that for each component $\pi^{(k)}_j$ of $\boldsymbol{\pi}^{(k)}$ and each n that

$$\lambda_k^n\,\pi_j^{(k)} = \sum_i \pi_i^{(k)} P_{ij}^n.$$

b) By taking magnitudes of each side, show that

$$\left|\lambda_k\right|^n \left|\pi_j^{(k)}\right| \leq \mathsf{M}.$$

c) Show that $\left|\lambda_k\right| \leq 1$.

Exercise 3.13. Consider a finite state Markov chain with matrix $[P]$ which has κ aperiodic recurrent classes, $\mathcal{R}_1, \ldots, \mathcal{R}_\kappa$ and a set $\mathcal{T}$ of transient states. For any given recurrent class ℓ, consider a vector $\boldsymbol{\nu}$ such that $\nu_i = 1$ for each $i \in \mathcal{R}_\ell$, $\nu_i = \lim_{n\to\infty} \Pr\{X_n \in \mathcal{R}_\ell | X_0 = i\}$ for each $i \in \mathcal{T}$, and $\nu_i = 0$ otherwise. Show that $\boldsymbol{\nu}$ is a right eigenvector of $[P]$ with eigenvalue 1. Hint: Redraw Figure 3.5 for multiple recurrent classes and first show that $\boldsymbol{\nu}$ is an eigenvector of $[P^n]$ in the limit.

Exercise 3.14. Answer the following questions for the following stochastic matrix $[P]$

$$[P] = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}.$$

a) Find $[P]^n$ in closed form for arbitrary $n > 1$.

b) Find all distinct eigenvalues and the multiplicity of each distinct eigenvalue for $[P]$.

c) Find a right eigenvector for each distinct eigenvalue, and show that the eigenvalue of multiplicity 2 does not have 2 linearly independent eigenvectors.

d) Use (c) to show that there is no diagonal matrix $[\Lambda]$ and no invertible matrix $[U]$ for which $[P][U] = [U][\Lambda]$.

e) Rederive the result of part d) using the result of a) rather than c).

Exercise 3.15. a) Let $[J_i]$ be a 3 by 3 block of a Jordan form, *i.e.*,

$$[J_i] = \begin{bmatrix} \lambda_i & 1 & 0 \\ 0 & \lambda_i & 1 \\ 0 & 0 & \lambda_i \end{bmatrix}.$$

Show that the nth power of $[J_i]$ is given by

$$[J_i^n] = \begin{bmatrix} \lambda_i^n & n\lambda_1^{n-1} & \binom{n}{2}\lambda_i^{n-1} \\ 0 & \lambda_i^n & n\lambda_1^{n-1} \\ 0 & 0 & \lambda_i^n \end{bmatrix}.$$

Hint: Perhaps the easiest way is to calculate $[J_i^2]$ and $[J_i^3]$ and then use iteration.

b) Generalize a) to a k by k block of a Jordan form. Note that the nth power of an entire Jordan form is composed of these blocks along the diagonal of the matrix.

c) Let $[P]$ be a stochastic matrix represented by a Jordan form $[J]$ as $[P] = U^{-1}[J][U]$ and consider $[U][P][U^{-1}] = [J]$. Show that any repeated eigenvalue of $[P]$ (*i.e.*, any eigenvalue represented by a Jordan block of 2 by 2 or more) must be strictly less than 1. Hint: Upper bound the elements of $[U][P^n][U^{-1}]$ by taking the magnitude of the elements of $[U]$ and $[U^{-1}]$ and upper bounding each element of a stochastic matrix by 1.

d) Let λ_s be the eigenvalue of largest magnitude less than 1. Assume that the Jordan blocks for λ_s are at most of size k. Show that each ergodic class of $[P]$ converges at least as fast as $n^k \lambda_s^k$.

Exercise 3.16. **a)** Let λ be an eigenvalue of a matrix $[A]$, and let $\boldsymbol{\nu}$ and $\boldsymbol{\pi}$ be right and left eigenvectors respectively of λ, normalized so that $\boldsymbol{\pi\nu} = 1$. Show that

$$[[A] - \lambda \boldsymbol{\nu\pi}]^2 = [A^2] - \lambda^2 \boldsymbol{\nu\pi}.$$

b) Show that $[[A^n] - \lambda^n \boldsymbol{\nu\pi}][[A] - \lambda \boldsymbol{\nu\pi}] = [A^{n+1}] - \lambda^{n+1} \boldsymbol{\nu\pi}$.

c) Use induction to show that $[[A] - \lambda \boldsymbol{\nu\pi}]^n = [A^n] - \lambda^n \boldsymbol{\nu\pi}$.

Exercise 3.17. Let $[P]$ be the transition matrix for an aperiodic Markov unichain with the states numbered as in Figure 3.5.

a) Show that $[P^n]$ can be partitioned as

$$[P^n] = \begin{bmatrix} [P_{\mathcal{T}}{}^n] & [P_x^n] \\ 0 & [P_{\mathcal{R}}^n] \end{bmatrix}.$$

That is, the blocks on the diagonal are simply products of the corresponding blocks of $[P]$, and the upper right block is whatever it turns out to be.

b) Let q_i be the probability that the chain will be in a recurrent state after t transitions, starting from state i, i.e., $q_i = \sum_{t<j\leq t+r} P_{ij}^t$. Show that $q_i > 0$ for all transient i.

c) Let q be the minimum q_i over all transient i and show that $P_{ij}^{nt} \leq (1-q)^n$ for all transient i, j (i.e., show that $[P_{\mathcal{T}}]^n$ approaches the all zero matrix $[0]$ with increasing n).

d) Let $\boldsymbol{\pi} = (\boldsymbol{\pi}_{\mathcal{T}}, \boldsymbol{\pi}_{\mathcal{R}})$ be a left eigenvector of $[P]$ of eigenvalue 1. Show that $\boldsymbol{\pi}_{\mathcal{T}} = \boldsymbol{0}$ and show that $\boldsymbol{\pi}_{\mathcal{R}}$ must be positive and be a left eigenvector of $[P_{\mathcal{R}}]$. Thus show that $\boldsymbol{\pi}$ exists and is unique (within a scale factor).

e) Show that $\boldsymbol{e}$ is the unique right eigenvector of $[P]$ of eigenvalue 1 (within a scale factor).

Exercise 3.18. Generalize Exercise 3.17 to the case of a Markov chain $[P]$ with m recurrent classes and one or more transient classes. In particular,

a) Show that $[P]$ has exactly κ linearly independent left eigenvectors, $\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}, \ldots, \boldsymbol{\pi}^{(\kappa)}$ of eigenvalue 1, and that the mth can be taken as a probability vector that is positive on the mth recurrent class and zero elsewhere.

b) Show that $[P]$ has exactly κ linearly independent right eigenvectors, $\boldsymbol{\nu}^{(1)}, \boldsymbol{\nu}^{(2)}, \ldots, \boldsymbol{\nu}^{(\kappa)}$ of eigenvalue 1, and that the mth can be taken as a vector with $\nu_i^{(m)}$ equal to the probability that recurrent class m will ever be entered starting from state i.

c) Show that

$$\lim_{n\to\infty}[P^n] = \sum_m \boldsymbol{\nu}^{(m)}\boldsymbol{\pi}^{(m)}.$$

Exercise 3.19. Suppose a recurrent Markov chain has period d and let $\mathcal{S}_m$, $1 \le m \le d$, be the mth subset in the sense of Theorem 3.2.3. Assume the states are numbered so that the first s_1 states are the states of $\mathcal{S}_1$, the next s_2 are those of $\mathcal{S}_2$, and so forth. Thus the matrix $[P]$ for the chain has the block form given by

$$[P] = \begin{bmatrix} 0 & [P_1] & \ddots & \ddots & 0 \\ 0 & 0 & [P_2] & \ddots & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & \ddots & \ddots & [P_{d-1}] \\ [P_d] & 0 & \ddots & \ddots & 0 \end{bmatrix},$$

where $[P_m]$ has dimension s_m by s_{m+1} for $1 \le m \le d$, where $d+1$ is interpreted as 1 throughout. In what follows it is usually more convenient to express $[P_m]$ as an M by M matrix $[P'_m]$ whose entries are 0 except for the rows of $\mathcal{S}_m$ and the columns of $\mathcal{S}_{m+1}$, where the entries are equal to those of $[P_m]$. In this view, $[P] = \sum_{m=1}^{d}[P'_m]$.

a) Show that $[P^d]$ has the form

$$[P^d] = \begin{bmatrix} [Q_1] & 0 & \ddots & 0 \\ 0 & [Q_2] & \ddots & \ddots \\ 0 & 0 & \ddots & [Q_d] \end{bmatrix},$$

where $[Q_m] = [P_m][P_{m+1}]\ldots[P_d][P_1]\ldots[P_{m-1}]$. Expressing $[Q_m]$ as an M by M matrix $[Q'_m]$ whose entries are 0 except for the rows and columns of $\mathcal{S}_m$ where the entries are equal to those of $[Q_m]$, this becomes $[P^d] = \sum_{m=1}^{d} Q'_m$.

b) Show that $[Q_m]$ is the matrix of an ergodic Markov chain, so that with the eigenvectors $\hat{\boldsymbol{\pi}}_m, \hat{\boldsymbol{\nu}}_m$ as defined in Exercise 3.18, $\lim_{n\to\infty}[P^{nd}] = \sum_{m=1}^{d} \hat{\boldsymbol{\nu}}^{(m)}\hat{\boldsymbol{\pi}}^{(m)}$.

c) Show that $\hat{\boldsymbol{\pi}}^{(m)}[P'_m] = \hat{\boldsymbol{\pi}}^{(m+1)}$. Note that $\hat{\boldsymbol{\pi}}^{(m)}$ is an M-tuple that is nonzero only on the components of $\mathcal{S}_m$.

d) Let $\phi = \frac{2\boldsymbol{\pi}\sqrt{-1}}{d}$ and let $\boldsymbol{\pi}^{(k)} = \sum_{m=1}^{d} \hat{\boldsymbol{\pi}}^{(m)} e^{mk\phi}$. Show that $\boldsymbol{\pi}^{(k)}$ is a left eigenvector of $[P]$ of eigenvalue $e^{-\phi k}$.

Exercise 3.20. (continuation of Exercise 3.19). **a)** Show that, with the eigenvectors defined in Exercises 3.19,

$$\lim_{n\to\infty}[P^{nd}][P] = \sum_{i=1}^{d} \boldsymbol{\nu}^{(i)}\boldsymbol{\pi}^{(i+1)},$$

where, as before, $d+1$ is taken to be 1.

b) Show that, for $1 \le j < d$,

$$\lim_{n\to\infty}[P^{nd}][P^j] = \sum_{i=1}^{d} \boldsymbol{\nu}^{(i)}\boldsymbol{\pi}^{(i+j)}.$$

c) Show that

$$\lim_{n\to\infty}[P^{nd}]\left\{I + [P] + \ldots, [P^{d-1}]\right\} = \left(\sum_{i=1}^{d} \boldsymbol{\nu}^{(i)}\right)\left(\sum_{i=1}^{d} \boldsymbol{\pi}^{(i+j)}\right).$$

d) Show that

$$\lim_{n\to\infty}\frac{1}{d}\left([P^n] + [P^{n+1}] + \cdots + [P^{n+d-1}]\right) = \boldsymbol{e}\boldsymbol{\pi},$$

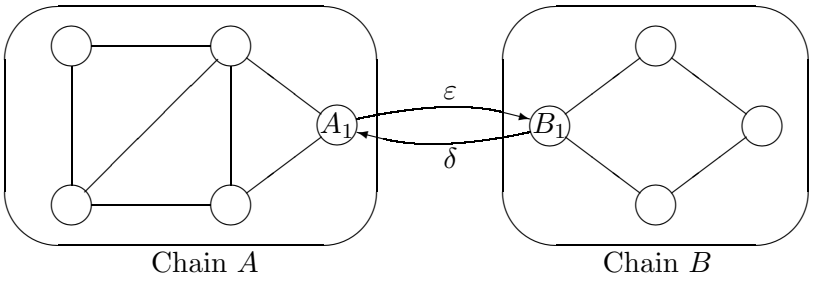
where $\boldsymbol{\pi}$ is the steady-state probability vector for $[P]$. Hint: Show that $\boldsymbol{e} = \sum_m \boldsymbol{\nu}^{(m)}$ and $\boldsymbol{\pi} = (1/d)\sum_m \boldsymbol{\pi}^{(m)}$.

e) Show that the above result is also valid for periodic unichains.

Exercise 3.21. Suppose A and B are each ergodic Markov chains with transition probabilities $\{P_{A_i,A_j}\}$ and $\{P_{B_i,B_j}\}$ respectively. Denote the steady-state probabilities of A and B by $\{\pi_{A_i}\}$ and $\{\pi_{B_i}\}$ respectively. The chains are now connected and modified as shown below. In particular, states A_1 and B_1 are now connected and the new transition probabilities P' for the combined chain are given by

$$\begin{aligned} P'_{A_1,B_1} &= \varepsilon, \quad P'_{A_1,A_j} = (1-\varepsilon)P_{A_1,A_j} \qquad \text{for all } A_j \\ P'_{B_1,A_1} &= \delta, \quad P'_{B_1,B_j} = (1-\delta)P_{B_1,B_j} \qquad \text{for all } B_j. \end{aligned}$$

All other transition probabilities remain the same. Think intuitively of ε and δ as being small, but do not make any approximations in what follows. Give your answers to the following questions as functions of ε, δ, $\{\pi_{A_i}\}$ and $\{\pi_{B_i}\}$.



a) Assume that $\epsilon > 0$, $\delta = 0$ (i.e., that A is a set of transient states in the combined chain). Starting in state A_1, find the conditional expected time to return to A_1 given that the first transition is to some state in chain A.

b) Assume that $\epsilon > 0$, $\delta = 0$. Find $T_{A,B}$, the expected time to first reach state B_1 starting from state A_1. Your answer should be a function of ϵ and the original steady state probabilities $\{\pi_{A_i}\}$ in chain A.

c) Assume $\varepsilon > 0$, $\delta > 0$, find $T_{B,A}$, the expected time to first reach state A_1, starting in state B_1. Your answer should depend only on δ and $\{\pi_{B_i}\}$.

d) Assume $\varepsilon > 0$ and $\delta > 0$. Find $P'(A)$, the steady-state probability that the combined chain is in one of the states $\{A_j\}$ of the original chain A.

e) Assume $\varepsilon > 0$, $\delta = 0$. For each state $A_j \neq A_1$ in A, find v_{A_j}, the expected number of visits to state A_j, starting in state A_1, before reaching state B_1. Your answer should depend only on ε and $\{\pi_{A_i}\}$.

f) Assume $\varepsilon > 0$, $\delta > 0$. For each state A_j in A, find π'_{A_j}, the steady-state probability of being in state A_j in the combined chain. Hint: Be careful in your treatment of state A_1.

Exercise 3.22. Example 3.5.1 showed how to find the expected first passage times to a fixed state, say 1, from all other nodes. It is often desirable to include the expected first recurrence time from state 1 to return to state 1. This can be done by splitting state 1 into 2 states, first an initial state with no transitions coming into it but the original transitions going out, and second, a final trapping state with the original transitions coming in.

a) For the chain on the left side of Figure 3.6, draw the graph for the modified chain with 5 states where state 1 has been split into 2 states.

b) Suppose one has found the expected first-passage-times v_j for states $j = 2$ to 4 (or in general from 2 to M). Find an expression for v_1, the expected first recurrence time for state 1 in terms of $v_2, v_3, \ldots v_{\mathsf{M}}$ and $P_{12}, \ldots, P_{1\mathsf{M}}$.

Exercise 3.23. a) Assume throughout that $[P]$ is the transition matrix of a unichain (and thus the eigenvalue 1 has multiplicity 1). Show that a solution to the equation $[P]\boldsymbol{w} - \boldsymbol{w} = \boldsymbol{r} - g\boldsymbol{e}$ exists if and only if $\boldsymbol{r} - g\boldsymbol{e}$ lies in the column space of $[P - I]$ where I is the identity matrix.

b) Show that this column space is the set of vectors $\boldsymbol{x}$ for which $\boldsymbol{\pi x} = 0$. Then show that $\boldsymbol{r} - g\boldsymbol{e}$ lies in this column space.

c) Show that, with the extra constraint that $\boldsymbol{\pi w} = 0$, the equation $[P]\boldsymbol{w} - \boldsymbol{w} = \boldsymbol{r} - g\boldsymbol{e}$ has a unique solution.

Exercise 3.24. For the Markov chain with rewards in Figure 3.7,

a) Find the solution to (3.5.1) and find the gain g.

b) Modify Figure 3.7 by letting P_{12} be an arbitrary probability. Find g and $\boldsymbol{w}$ again and give an intuitive explanation of why P_{12} effects w_2.

Exercise 3.25. (Proof of Corollary 3.5.1) **a)** Show that the gain per stage g is 0. Hint: Show that $\boldsymbol{r}$ is zero where the steady-state vector $\boldsymbol{\pi}$ is nonzero.

b) Let $[P_{\mathcal{R}}]$ be the transition matrix for the recurrent states and let $\boldsymbol{r}_{\mathcal{R}} = 0$ be the reward vector and $\boldsymbol{w}_{\mathcal{R}}$ the relative-gain vector for $[P_{\mathcal{R}}]$. Show that $\boldsymbol{w}_{\mathcal{R}} = 0$. Hint: Use Theorem 3.5.1.

c) Show that $w_i = 0$ for all $i \in \mathcal{R}$. Hint: Compare the relative-gain equations for $[P]$ to those for $[P_{\mathcal{R}}]$.

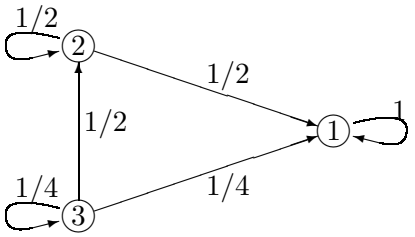
d) Show that for each $n \geq 0$, $[P^n]\boldsymbol{w} = [P^{n+1}]\boldsymbol{w} + [P^n]\boldsymbol{r}$. Hint: Start with the relative-gain equation for $[P]$.

e) Show that $\boldsymbol{w} = [P^{n+1}]\boldsymbol{w} + \sum_{m=0}^{n}[P^m]\boldsymbol{r}$. Hint: Sum the result in b).

f) Show that $\lim_{n\to\infty}[P^{n+1}]\boldsymbol{w} = 0$ and that $\lim_{n\to\infty}\sum_{m=0}^{n}[P^m]\boldsymbol{r}$ is finite, non-negative, and has positive components for $r_i > 0$. Hint: Use lemma 3.3.3.

g) Demonstrate the final result of the corollary by using the previous results on $\boldsymbol{r} = \boldsymbol{r}' - \boldsymbol{r}''$.

Exercise 3.26. Consider the Markov chain below:



a) Suppose the chain is started in state i and goes through n transitions; let $v_i(n)$ be the expected number of transitions (out of the total of n) until the chain enters the trapping state, state 1. Find an expression for $\boldsymbol{v}(n) = (v_1(n), v_2(n), v_3(n))^{\mathsf{T}}$ in terms of $\boldsymbol{v}(n-1)$ (take $v_1(n) = 0$ for all n). (Hint: view the system as a Markov reward system; what is the value of $\boldsymbol{r}$?)

b) Solve numerically for $\lim_{n\to\infty}\boldsymbol{v}(n)$. Interpret the meaning of the elements v_i in the solution of (3.32).

c) Give a direct argument why (3.32) provides the solution directly to the expected time from each state to enter the trapping state.

Exercise 3.27. a) Show that (3.48) can be rewritten in the more compact form

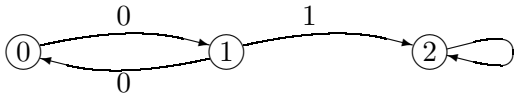
$$\boldsymbol{v}^*(n, \boldsymbol{u}) = \boldsymbol{v}^*(1, \boldsymbol{v}^*(n-1, \boldsymbol{u})).$$

b) Explain why it is also true that

$$\boldsymbol{v}^*(2n, \boldsymbol{u}) = \boldsymbol{v}^*(n, \boldsymbol{v}^*(n, \boldsymbol{u})). \tag{3.65}$$

c) One might guess that (3.65) could be used iteratively, finding $\boldsymbol{v}^*(2^{n+1}, \boldsymbol{u})$ from $\boldsymbol{v}^*(2^n, \boldsymbol{u})$. Explain why this is not possible in any straighttforward way. Hint: Think through explicitly how one might calculate $\boldsymbol{v}^*(n, \boldsymbol{v}^*(n, \boldsymbol{u}))$ from $\boldsymbol{v}^*(n, \boldsymbol{u})$.

Exercise 3.28. Consider a sequence of IID binary rv's $X_1, X_2, \ldots$. Assume that $\Pr\{X_i = 1\} = p_1$, $\Pr\{X_i = 0\} = p_0 = 1 - p_1$. A binary string $(a_1, a_2, \ldots, a_k)$ *occurs* at time n if $X_n = a_k, X_{n-1} = a_{k-1}, \ldots X_{n-k+1} = a_1$. For a given string $(a_1, a_2, \ldots, a_k)$, consider a Markov chain with $k+1$ states $\{0, 1, \ldots, k\}$. State 0 is the initial state, state k is a final trapping state where $(a_1, a_2, \ldots, a_k)$ has already occurred, and each intervening state i, $0 < i < k$, has the property that if the subsequent $k - i$ variables take on the values $a_{i+1}, a_{i+2}, \ldots, a_k$, the Markov chain will move successively from state i to $i+1$ to $i+2$ and so forth to k. For example, if $k = 2$ and $(a_1, a_2) = (0, 1)$, the corresponding chain is given by



a) For the chain above, find the mean first-passage time from state 0 to state 2.

b) For parts b) to d), let $(a_1, a_2, a_3, \ldots, a_k) = (0, 1, 1, \ldots, 1)$, i.e., zero followed by $k-1$ ones. Draw the corresponding Markov chain for $k = 4$.

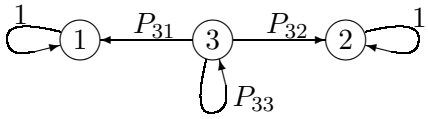
c) Let v_i, $1 \leq i \leq k$ be the expected first-passage time from state i to state k. Note that $v_k = 0$. Show that $v_0 = 1/p_0 + v_1$.

d) For each i, $1 \leq i < k$, show that $v_i = \alpha_i + v_{i+1}$ and $v_0 = \beta_i + v_{i+1}$ where α_i and β_i are each a product of powers of p_0 and p_1. Hint: use induction, or iteration, starting with $i = 1$, and establish both equalities together.

e) Let $k = 3$ and let $(a_1, a_2, a_3) = (1, 0, 1)$. Draw the corresponding Markov chain for this string. Evaluate v_0, the expected first-passage time for the string 1,0,1 to occur.

Exercise 3.29. a) Find $\lim_{n\to\infty}[P^n]$ for the Markov chain below. Hint: Think in terms of the long term transition probabilities. Recall that the edges in the graph for a Markov chain correspond to the positive transition probabilities.

b) Let $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$ denote the first two rows of $\lim_{n\to\infty}[P^n]$ and let $\boldsymbol{\nu}^{(1)}$ and $\boldsymbol{\nu}^{(2)}$ denote the first two columns of $\lim_{n\to\infty}[P^n]$. Show that $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$ are independent left eigenvectors of $[P]$, and that $\boldsymbol{\nu}^{(1)}$ and $\boldsymbol{\nu}^{(2)}$ are independent right eigenvectors of $[P]$. Find the eigenvalue for each eigenvector.



c) Let $\boldsymbol{r}$ be an arbitrary reward vector and consider the equation

$$\boldsymbol{w} + g^{(1)}\boldsymbol{\nu}^{(1)} + g^{(2)}\boldsymbol{\nu}^{(2)} = \boldsymbol{r} + [P]\boldsymbol{w}. \tag{3.66}$$

Determine what values $g^{(1)}$ and $g^{(2)}$ must have in order for (3.66) to have a solution. Argue that with the additional constraints $w_1 = w_2 = 0$, (3.66) has a unique solution for $\boldsymbol{w}$ and find that $\boldsymbol{w}$.

Exercise 3.30. Let $\boldsymbol{u}$ and $\boldsymbol{u}'$ be arbitrary final reward vectors with $\boldsymbol{u} \leq \boldsymbol{u}'$.

a) Let $\boldsymbol{k}$ be an arbitrary stationary policy and prove that $\boldsymbol{v}^{\boldsymbol{k}}(n, \boldsymbol{u}) \leq \boldsymbol{v}^{\boldsymbol{k}}(n, \boldsymbol{u}')$ for each $n \geq 1$.

b) For the optimal dynamic policy, prove that $\boldsymbol{v}^*(n, \boldsymbol{u}) \leq \boldsymbol{v}^*(n, \boldsymbol{u}')$ for each $n \geq 1$. This is known as the monotonicity theorem.

c) Now let $\boldsymbol{u}$ and $\boldsymbol{u}'$ be arbitrary. Let $\alpha = \max_i(u_i - u'_i)$. Show that

$$\boldsymbol{v}^*(n, \boldsymbol{u}) \leq \boldsymbol{v}^*(n, \boldsymbol{u}') + \alpha \boldsymbol{e}.$$

Exercise 3.31. Consider a Markov decision problem with M states in which some state, say state 1, is inherently reachable from each other state.

a) Show that there must be some other state, say state 2, and some decision, k_2, such that $P_{21}^{(k_2)} > 0$.

b) Show that there must be some other state, say state 3, and some decision, k_3, such that either $P_{31}^{(k_3)} > 0$ or $P_{32}^{(k_3)} > 0$.

c) Assume, for some i, and some set of decisions $k_2, \ldots, k_i$ that, for each j, $2 \leq j \leq i$, $P_{jl}^{(k_j)} > 0$ for some $l < j$ (i.e., that each state from 2 to j has a non-zero transition to a lower numbered state). Show that there is some state (other than 1 to i), say $i+1$ and some decision k_{i+1} such that $P_{i+1,l}^{(k_{i+1})} > 0$ for some $l \leq i$.

d) Use parts a), b), and c) to observe that there is a stationary policy $\boldsymbol{k} = k_1, \ldots, k_{\mathsf{M}}$ for which state 1 is accessible from each other state.

Exercise 3.32. George drives his car to the theater, which is at the end of a one-way street. There are parking places along the side of the street and a parking garage that costs \$5 at the theater. Each parking place is independently occupied or unoccupied with probability 1/2. If George parks n parking places away from the theater, it costs him n cents (in time and shoe leather) to walk the rest of the way. George is myopic and can only see the parking place he is currently passing. If George has not already parked by the time he reaches the nth place, he first decides whether or not he will park if the place is unoccupied, and then

observes the place and acts according to his decision. George can never go back and must park in the parking garage if he has not parked before.

a) Model the above problem as a 2 state dynamic programming problem. In the "driving" state, state 2, there are two possible decisions: park if the current place is unoccupied or drive on whether or not the current place is unoccupied.

b) Find $v_i^*(n, \boldsymbol{u})$, the *minimum* expected aggregate cost for n stages (i.e., immediately before observation of the nth parking place) starting in state $i = 1$ or 2; it is sufficient to express $v_i^*(n, \boldsymbol{u})$ in terms of $v_i^*(n-1)$. The final costs, in cents, at stage 0 should be $v_2(0) = 500$, $v_1(0) = 0$.

c) For what values of n is the optimal decision the decision to drive on?

d) What is the probability that George will park in the garage, assuming that he follows the optimal policy?

Exercise 3.33. (Proof of Corollary 3.6.1) **a)** Show that if two stationary policies $\boldsymbol{k}'$ and $\boldsymbol{k}$ have the same recurrent class $\mathcal{R}'$ and if $k_i' = k_i$ for all $i \in \mathcal{R}'$, then $w_i' = w_i$ for all $i \in \mathcal{R}'$. Hint: See the first part of the proof of Lemma 3.6.3.

b) Assume that $\boldsymbol{k}'$ satisfies 3.50 (*i.e.*, that it satisfies the termination condition of the policy improvement algorithm) and that $\boldsymbol{k}$ satisfies the conditions of part a). Show that (3.64) is satisfied for all states ℓ.

c) Show that $\boldsymbol{w} \leq \boldsymbol{w}'$. Hint: Follow the reasoning at the end of the proof of Lemma 3.6.3.

Exercise 3.34. Consider the dynamic programming problem below with two states and two possible policies, denoted $\boldsymbol{k}$ and $\boldsymbol{k}'$. The policies differ only in state 2.



a) Find the steady-state gain per stage, g and g', for stationary policies $\boldsymbol{k}$ and $\boldsymbol{k}'$. Show that $g = g'$.

b) Find the relative-gain vectors, $\boldsymbol{w}$ and $\boldsymbol{w}'$, for stationary policies $\boldsymbol{k}$ and $\boldsymbol{k}'$.

c) Suppose the final reward, at stage 0, is $u_1 = 0$, $u_2 = u$. For what range of u does the dynamic programming algorithm use decision $\boldsymbol{k}$ in state 2 at stage 1?

d) For what range of u does the dynamic programming algorithm use decision $\boldsymbol{k}$ in state 2 at stage 2? at stage n? You should find that (for this example) the dynamic programming algorithm uses the same decision at each stage n as it uses in stage 1.

e) Find the optimal gain $v_2^*(n, \boldsymbol{u})$ and $v_1^*(n, \boldsymbol{u})$ as a function of stage n assuming $u = 10$.

f) Find $\lim_{n\to\infty} v^*(n, \boldsymbol{u})$ and show how it depends on u.

Exercise 3.35. Consider a Markov decision problem in which the stationary policies $\boldsymbol{k}$ and $\boldsymbol{k'}$ each satisfy (3.50) and each correspond to ergodic Markov chains.

a) Show that if $\boldsymbol{r^{k'}} + [P^{\boldsymbol{k'}}]\boldsymbol{w'} \geq \boldsymbol{r^{k}} + [P^{\boldsymbol{k}}]\boldsymbol{w'}$ is not satisfied with equality, then $g' > g$.

b) Show that $\boldsymbol{r^{k'}} + [P^{\boldsymbol{k'}}]\boldsymbol{w'} = \boldsymbol{r^{k}} + [P^{\boldsymbol{k}}]\boldsymbol{w'}$ (Hint: use part a).

c) Find the relationship between the relative gain vector $\boldsymbol{w^k}$ for policy $\boldsymbol{k}$ and the relative-gain vector $\boldsymbol{w'}$ for policy $\boldsymbol{k'}$. (Hint: Show that $\boldsymbol{r^k} + [P^{\boldsymbol{k}}]\boldsymbol{w'} = g\boldsymbol{e} + \boldsymbol{w'}$; what does this say about $\boldsymbol{w}$ and $\boldsymbol{w'}$?)

e) Suppose that policy $\boldsymbol{k}$ uses decision 1 in state 1 and policy $\boldsymbol{k'}$ uses decision 2 in state 1 (i.e., $k_1 = 1$ for policy $\boldsymbol{k}$ and $k_1 = 2$ for policy $\boldsymbol{k'}$). What is the relationship between $r_1^{(k)}, P_{11}^{(k)}, P_{12}^{(k)}, \ldots P_{1J}^{(k)}$ for k equal to 1 and 2?

f) Now suppose that policy $\boldsymbol{k}$ uses decision 1 in each state and policy $\boldsymbol{k'}$ uses decision 2 in each state. Is it possible that $r_i^{(1)} > r_i^{(2)}$ for all i? Explain carefully.

g) Now assume that $r_i^{(1)}$ is the same for all i. Does this change your answer to part f)? Explain.

Exercise 3.36. Consider a Markov decision problem with three states. Assume that each stationary policy corresponds to an ergodic Markov chain. It is known that a particular policy $\boldsymbol{k'} = (k_1, k_2, k_3) = (2, 4, 1)$ is the unique optimal stationary policy (i.e., the gain per stage in steady state is maximized by always using decision 2 in state 1, decision 4 in state 2, and decision 1 in state 3). As usual, $r_i^{(k)}$ denotes the reward in state i under decision k, and $P_{ij}^{(k)}$ denotes the probability of a transition to state j given state i and given the use of decision k in state i. Consider the effect of changing the Markov decision problem in each of the following ways (the changes in each part are to be considered in the absence of the changes in the other parts):

a) $r_1^{(1)}$is replaced by $r_1^{(1)} - 1$.

b) $r_1^{(2)}$is replaced by $r_1^{(2)} + 1$.

c) $r_1^{(k)}$ is replaced by $r_1^{(k)} + 1$ for all state 1 decisions k.

d) for all i, $r_i^{(k_i)}$ is replaced by $r^{(k_i)} + 1$ for the decision k_i of policy $\boldsymbol{k'}$.

For each of the above changes, answer the following questions; *give explanations*:

1) Is the gain per stage, g', increased, decreased, or unchanged by the given change?

2) Is it possible that another policy, $\boldsymbol{k} \neq \boldsymbol{k'}$, is optimal after the given change?

Exercise 3.37. (The Odoni Bound) Let $\boldsymbol{k'}$ be the optimal stationary policy for a Markov decision problem and let g' and $\boldsymbol{\pi'}$ be the corresponding gain and steady-state probability respectively. Let $v_i^*(n, \boldsymbol{u})$ be the optimal dynamic expected reward for starting in state i at stage n with final reward vector $\boldsymbol{u}$.

a) Show that $\min_i[v_i^*(n,\boldsymbol{u}) - v_i^*(n-1,\boldsymbol{u})] \leq g' \leq \max_i[v_i^*(n,\boldsymbol{u}) - v_i^*(n-1,\boldsymbol{u})]$; $n \geq 1$. Hint: Consider premultiplying $\boldsymbol{v}^*(n,\boldsymbol{u}) - \boldsymbol{v}^*(n-1,\boldsymbol{u})$ by $\boldsymbol{\pi}'$ or $\boldsymbol{\pi}$ where $\boldsymbol{k}$ is the optimal dynamic policy at stage n.

b) Show that the lower bound is non-decreasing in n and the upper bound is non-increasing in n and both converge to g' with increasing n.

Exercise 3.38. Consider an integer-time queueing system with a finite buffer of size 2. At the beginning of the n^{th} time interval, the queue contains at most two customers. There is a cost of one unit for each customer in queue (i.e., the cost of delaying that customer). If there is one customer in queue, that customer is served. If there are two customers, an extra server is hired at a cost of 3 units and both customers are served. Thus the total immediate cost for two customers in queue is 5, the cost for one customer is 1, and the cost for 0 customers is 0. At the end of the nth time interval, either 0, 1, or 2 new customers arrive (each with probability 1/3).

a) Assume that the system starts with $0 \leq i \leq 2$ customers in queue at time -1 (i.e., in stage 1) and terminates at time 0 (stage 0) with a final cost $\boldsymbol{u}$ of 5 units for each customer in queue (at the beginning of interval 0). Find the expected aggregate cost $v_i(1,\boldsymbol{u})$ for $0 \leq i \leq 2$.

b) Assume now that the system starts with i customers in queue at time -2 with the same final cost at time 0. Find the expected aggregate cost $v_i(2,\boldsymbol{u})$ for $0 \leq i \leq 2$.

c) For an arbitrary starting time $-n$, find the expected aggregate cost $v_i(n,\boldsymbol{u})$ for $0 \leq i \leq 2$.

d) Find the cost per stage and find the relative cost (gain) vector.

e) Now assume that there is a decision maker who can choose whether or not to hire the extra server when there are two customers in queue. If the extra server is not hired, the 3 unit fee is saved, but only one of the customers is served. If there are two arrivals in this case, assume that one is turned away at a cost of 5 units. Find the minimum dynamic aggregate expected cost $v_i^*(1)$, $0 \leq i \leq$, for stage 1 with the same final cost as before.

f) Find the minimum dynamic aggregate expected cost $v_i^*(n,\boldsymbol{u})$ for stage n, $0 \leq i \leq 2$.

g) Now assume a final cost $\boldsymbol{u}$ of one unit per customer rather than 5, and find the new minimum dynamic aggregate expected cost $v_i^*(n,\boldsymbol{u})$, $0 \leq i \leq 2$.

Chapter 4

RENEWAL PROCESSES

4.1 Introduction

Recall that a renewal process is an arrival process in which the interarrival intervals are positive,[1] independent and identically distributed (IID) random variables (rv's). Renewal processes (since they are arrival processes) can be specified in three standard ways, first, by the joint distributions of the arrival epochs $S_1, S_2, \ldots$, second, by the joint distributions of the interarrival times $X_1, X_2, \ldots$, and third, by the joint distributions of the counting rv's, $N(t)$ for $t > 0$. Recall that $N(t)$ represents the number of arrivals to the system in the interval $(0, t]$.

The simplest characterization is through the interarrival times X_i, since they are IID. Each arrival epoch S_n is simply the sum $X_1 + X_2 + \cdots + X_n$ of n IID rv's. The characterization of greatest interest in this chapter is the renewal counting process, $\{N(t);\ t > 0\}$. Recall from (2.2) and (2.3) that the arrival epochs and the counting rv's are related in each of the following equivalent ways.

$$\{S_n \leq t\} = \{N(t) \geq n\}; \qquad \{S_n > t\} = \{N(t) < n\}. \tag{4.1}$$

The reason for calling these processes *renewal processes* is that the process probabilistically starts over at each arrival epoch, S_n. That is, if the nth arrival occurs at $S_n = \tau$, then, counting from $S_n = \tau$, the j^{th} subsequent arrival epoch is at $S_{n+j} - S_n = X_{n+1} + \cdots + X_{n+j}$. Thus, given $S_n = \tau$, $\{N(\tau + t) - N(\tau); t \geq 0\}$ is a renewal counting process with IID interarrival intervals of the same distribution as the original renewal process. This interpretation of arrivals as renewals will be discussed in more detail later.

The major reason for studying renewal processes is that many complicated processes have randomly occurring instants at which the system returns to a state probabilistically equiva-

[1] Renewal processes are often defined in a slightly more general way, allowing the interarrival intervals X_i to include the possibility $1 > \Pr\{X_i = 0\} > 0$. All of the theorems in this chapter are valid under this more general assumption, as can be verified by complicating the proofs somewhat. Allowing $\Pr\{X_i = 0\} > 0$ allows multiple arrivals at the same instant, which makes it necessary to allow $N(0)$ to take on positive values, and appears to inhibit intuition about renewals. Exercise 4.3 shows how to view these more general renewal processes while using the definition here, thus showing that the added generality is not worth much.

lent to the starting state. These embedded renewal epochs allow us to separate the long term behavior of the process (which can be studied through renewal theory) from the behavior within each renewal period.

Example 4.1.1 (Visits to a given state for a Markov chain). Suppose a recurrent finite-state Markov chain with transition matrix $[P]$ starts in state i at time 0. Then on the first return to state i, say at time n, the Markov chain, from time n on, is a probabilistic replica of the chain starting at time 0. That is, the state at time 1 is j with probability P_{ij}, and, given a return to i at time n, the probability of state j at time $n+1$ is P_{ij}. In the same way, for any $m > 0$,

$$\Pr\{X_1 = j, \ldots, X_m = k \mid X_0 = i\} = \Pr\{X_{n+1} = j, \ldots, X_{n+m} = k \mid X_n = i\}. \tag{4.2}$$

Each subsequent return to state i at a given time n starts a new probabilistic replica of the Markov chain starting in state i at time 0, . Thus the sequence of entry times to state i can be viewed as the arrival epochs of a renewal process.

This example is important, and will form the key to the analysis of Markov chains with a countably infinite set of states in Chapter 5. At the same time, (4.2) does not quite justify viewing successive returns to state i as a renewal process. The problem is that the time of the first entry to state i after time 0 is a random variable rather than a given time n. This will not be a major problem to sort out, but the resolution will be more insightful after developing some basic properties of renewal processes.

Example 4.1.2 (The G/G/m queue:). The customer arrivals to a G/G/m queue form a renewal counting process, $\{N(t);\, t > 0\}$. Each arriving customer waits in the queue until one of m identical servers is free to serve it. The service time required by each customer is a rv, IID over customers, and independent of arrival times and servers. The system is assumed to be empty for $t < 0$, and an arrival, viewed as customer number 0, is assumed at time 0. The subsequent interarrival intervals $X_1, X_2, \ldots$, are IID. Note that $N(t)$ for each $t > 0$ is the number of arrivals in $(0, t]$, so arrival number 0 at $t = 0$ is not counted in $N(t)$.[2]

We define a new counting process, $\{N^r(t);\, t > 0\}$, for which the renewal epochs are those particular arrival epochs in the original process $\{N(t);\ t > 0\}$ at which an arriving customer sees an empty system (i.e., no customer in queue and none in service).[3] We will show in Section 4.5.3 that $\{N^r(t)\ t > 0\}$ is actually a renewal process, but give an intuitive explanation here. Note that customer 0 arrives at time 0 to an empty system, and given a first subsequent arrival to an empty system, at say epoch $S_1^r > 0$, the subsequent customer interarrival intervals are independent of the arrivals in $(0, S_1^r)$ and are identically distributed to those earlier arrivals. The service times after S_1^r are also IID from those earlier. Finally, the conditions that cause queueing starting from the arrival to an empty system at $t = S_1^r$ are the same as those starting from the arrival to an empty system at $t = 0$.

[2]There is always a certain amount of awkwardness in 'starting' a renewal process, and the assumption of an arrival at time 0 which is not counted in $N(t)$ seems strange, but simplifies the notation. The process is defined in terms of the IID inter-renewal intervals $X_1, X_2, \ldots$. The first renewal epoch is at $S_1 = X_1$, and this is the point at which $N(t)$ changes from 0 to 1.

[3]Readers who accept without question that $\{N^r(t)\ t > 0\}$ is a renewal process should be proud of their probabilistic intuition, but should also question exactly how such a conclusion can be proven.

In most situations, we use the words *arrivals* and *renewals* interchangably, but for this type of example, the word *arrival* is used for the counting process $\{N(t);\ t > 0\}$ and the word *renewal* is used for $\{N^r(t);\ t > 0\}$. The reason for being interested in $\{N^r(t);\ t > 0\}$ is that it allows us to analyze very complicated queues such as this in two stages. First, $\{N(t);\ t > 0\}$ lets us analyze the distribution of the inter-renewal intervals X_n^r of $\{N^r(t);\ t > 0\}$. Second, the general renewal results developed in this chapter can be applied to the distribution on X_n^r to understand the overall behavior of the queueing system.

Throughout our study of renewal processes, we use $\overline{X}$ and $\mathsf{E}[X]$ interchangeably to denote the mean inter-renewal interval, and use σ_X^2 or simply σ^2 to denote the variance of the inter-renewal interval. We will usually assume that $\overline{X}$ is finite, but, except where explicitly stated, we need not assume that σ^2 is finite. This means, first, that σ^2 need not be calculated (which is often difficult if renewals are embedded into a more complex process), and second, since modeling errors on the far tails of the inter-renewal distribution typically affect σ^2 more than $\overline{X}$, the results are relatively robust to these kinds of modeling errors.

Much of this chapter will be devoted to understanding the behavior of $N(t)$ and $N(t)/t$ as t becomes large. As might appear to be intuitively obvious, and as is proven in Exercise 4.1, $N(t)$ is a rv (*i.e.*, not defective) for each $t > 0$. Also, as proven in Exercise 4.2, $\mathsf{E}[N(t)] < \infty$ for all $t > 0$. It is then also clear that $N(t)/t$, which is interpreted as the time-average renewal rate over (0,t], is also a rv with finite expectation.

One of the major results about renewal theory, which we establish shortly, concerns the behavior of the rv's $N(t)/t$ as $t \to \infty$. For each sample point $\omega \in \Omega$, $N(t,\omega)/t$ is a nonnegative number for each t and $\{N(t,\omega);\ t > 0\}$ is a sample path of the counting renewal process, taken from $(0,t]$ for each t. Thus $\lim_{t\to\infty} N(t,\omega)/t$, if it exists, is the time-average renewal rate over $(0,\infty)$ for the sample point ω.

The *strong law for renewal processes* states that this limiting time-average renewal rate exists for a set of ω that has probability 1, and that this limiting value is $1/\overline{X}$. We shall often refer to this result by the less precise statement that the time-average renewal rate is $1/\overline{X}$. This result is a direct consequence of the strong law of large numbers (SLLN) for IID rv's. In the next section, we first state and prove the SLLN for IID rv's and then establish the strong law for renewal processes.

Another important theoretical result in this chapter is the elementary renewal theorem, which states that $\mathsf{E}[N(t)/t]$ also approaches $1/\overline{X}$ as $t \to \infty$. Surprisingly, this is more than a trival consequence of the strong law for renewal processes, and we shall develop several widely useful results such as Wald's equality, in establishing this theorem.

The final major theoretical result of the chapter is Blackwell's theorem, which shows that, for appropriate values of δ, the expected number of renewals in an interval $(t, t+\delta]$ approaches $\delta/\overline{X}$ as $t \to \infty$. We shall thus interpret $1/\overline{X}$ as an ensemble-average renewal rate. This rate is the same as the above time-average renewal rate. We shall see the benefits of being able to work with both time-averages and ensemble-averages.

There are a wide range of other results, ranging from standard queueing results to results that are needed in all subsequent chapters.

4.2 The strong law of large numbers and convergence WP1

The concept of a sequence of rv's converging with probability 1 (WP1) was introduced briefly in Section 1.5.6. We discuss this type of convergence more fully here and establish some conditions under which it holds. Next the *strong law of large numbers* (SLLN) is stated for IID rv's (this is essentially the result that the partial sample averages of IID rv's converge to the mean WP1). A proof is given under the added condition that the rv's have a finite fourth moment. Finally, in the following section, we state the strong law for renewal processes and use the SLLN for IID rv's to prove it.

4.2.1 Convergence with probability 1 (WP1)

Recall that a sequence $\{Z_n;\, n \geq 1\}$ of rv's on a sample space Ω is defined to converge WP1 to a rv Z on Ω if

$$\Pr\Big\{\omega \in \Omega : \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\Big\} = 1,$$

i.e., if the set of sample sequences $\{Z_n(\omega);\, n \geq 1\}$ that converge to $Z(\omega)$ has probability 1. This becomes slightly easier to understand if we define $Y_n = Z_n - Z$ for each n. The sequence $\{Y_n;\, n \geq 1\}$ then converges to 0 WP1 if and only if the sequence $\{Z_n;\, n \geq 1\}$ converges to Z WP1. Dealing only with convergence to 0 rather than to an arbitrary rv doesn't cut any steps from the following proofs, but it simplifies the notation and the concepts.

We start with a simple lemma that provides a useful condition under which convergence to 0 WP1 occurs. We shall see later how to use this lemma in an indirect way to prove the SLLN.

Lemma 4.2.1. *Let $\{Y_n; n \geq 1\}$ be a sequence of rv's, each with finite expectation. If $\sum_{n=1}^{\infty} \mathsf{E}\,[|Y_n|] < \infty$, then $\Pr\{\omega : \lim_{n\to\infty} Y_n(\omega) = 0\} = 1$.*

Proof: For any α, $0 < \alpha < \infty$ and any integer $m \geq 1$, the Markov inequality says that

$$\Pr\left\{\sum_{n=1}^{m} |Y_n| > \alpha\right\} \leq \frac{\mathsf{E}\,[\sum_{n=1}^{m} |Y_n|]}{\alpha} = \frac{\sum_{n=1}^{m} \mathsf{E}\,[|Y_n|]}{\alpha}. \tag{4.3}$$

Since $|Y_n|$ is non-negative, $\sum_{n=1}^{m} |Y_n| > \alpha$ implies that $\sum_{n=1}^{m+1} |Y_n| > \alpha$. Thus the left side of (4.3) is nondecreasing in m and we can go to the limit

$$\lim_{m\to\infty} \Pr\left\{\sum_{n=1}^{m} |Y_n| > \alpha\right\} \leq \frac{\sum_{n=1}^{\infty} \mathsf{E}\,[|Y_n|]}{\alpha}.$$

Now let $A_m = \{\omega : \sum_{n=1}^{m} |Y_n(\omega)| > \alpha\}$. As seen above, the sequence $\{A_m;\, m \geq 1\}$ is

nested, $A_1 \subseteq A_2 \cdots$, so from property (1.9) of the axioms of probability,[4]

$$\lim_{m\to\infty} \Pr\left\{\sum_{n=1}^{m} |Y_n| > \alpha\right\} = \Pr\left\{\bigcup_{m=1}^{\infty} A_m\right\}$$
$$= \Pr\left\{\omega : \sum_{n=1}^{\infty} |Y_n(\omega)| > \alpha\right\}, \tag{4.4}$$

where we have used the fact that for any given ω, $\sum_{n=1}^{\infty} |Y_n(\omega)| > \alpha$ if and only if $\sum_{n=1}^{m} |Y_n(\omega)| > \alpha$ for some $m \geq 1$. Combining (4.3) with (4.4),

$$\Pr\left\{\omega : \sum_{n=1}^{\infty} |Y_n(\omega)| > \alpha\right\} \leq \frac{\sum_{n=1}^{\infty} \mathsf{E}\,[|Y_n|]}{\alpha}.$$

Looking at the complementary set and assuming $\alpha > \sum_{n=1}^{\infty} \mathsf{E}\,[|Y_n|]$,

$$\Pr\left\{\omega : \sum_{n=1}^{\infty} |Y_n(\omega)| \leq \alpha\right\} \geq 1 - \frac{\sum_{n=1}^{\infty} \mathsf{E}\,[|Y_n|]}{\alpha}. \tag{4.5}$$

For any ω such that $\sum_{n=1}^{\infty} |Y_n(\omega)| \leq \alpha$, we see that $\{|Y_n(\omega)|;\, n \geq 1\}$ is simply a sequence of non-negative numbers with a finite sum. Thus the individual numbers in that sequence must approach 0, *i.e.*, $\lim_{n\to\infty} |Y_n(\omega)| = 0$ for each such ω. It follows then that

$$\Pr\left\{\omega : \lim_{n\to\infty} |Y_n(\omega)| = 0\right\} \geq \Pr\left\{\omega : \sum_{n=1}^{\infty} |Y_n(\omega)| \leq \alpha\right\}.$$

Combining this with (4.5),

$$\Pr\left\{\omega : \lim_{n\to\infty} |Y_n(\omega)| = 0\right\} \geq 1 - \frac{\sum_{n=1}^{\infty} \mathsf{E}\,[|Y_n|]}{\alpha}.$$

This is true for all α, so $\Pr\{\omega : \lim_{n\to\infty} |Y_n| = 0\} = 1$, and thus $\Pr\{\omega : \lim_{n\to\infty} Y_n = 0\} = 1$. □

It is instructive to recall Example 1.5.1, illustrated in Figure 4.1, where $\{Y_n;\, n \geq 1\}$ converges in probability but does not converge with probability one. Note that $\mathsf{E}\,[Y_n] = 1/(5^{j+1} - 1)$ for $n \in [5^j, 5^{j+1})$. Thus $\lim_{n\to\infty} \mathsf{E}\,[Y_n] = 0$, but $\sum_{n=1}^{\infty} \mathsf{E}\,[Y_n] = \infty$. Thus this sequence does not satisfy the conditions of the lemma. This helps explain how the conditions in the lemma exclude such sequences.

Before proceeding to the SLLN, we want to show that convergence WP1 implies convergence in probability. We give an incomplete argument here with precise versions both in Exercise 4.5 and Exercise 4.6. Exercise 4.6 has the added merit of expressing the set $\{\omega : \lim_n Y_n(\omega) = 0\}$ explicitly in terms of countable unions and intersections of simple events involving finite sets of the Y_n. This representation is valid whether or not the conditions of the lemma are satisfied and shows that this set is indeed an event.

[4]This proof probably appears to be somewhat nitpicking about limits. The reason for this is that the argument is quite abstract and it is difficult to develop the kind of intuition that ordinarily allows us to be somewhat more casual.

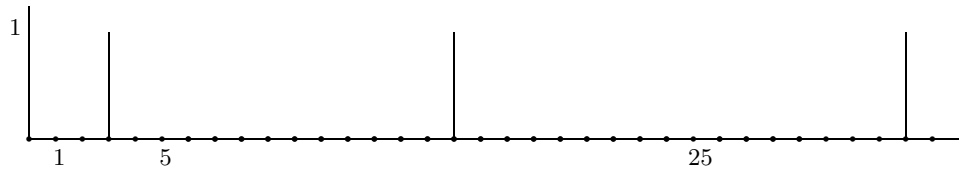


Figure 4.1: Illustration of a sample path of a sequence of rv's $\{Y_n; n \geq 0\}$ where, for each $j \geq 0$, $Y_n = 1$ for an equiprobable choice of $n \in [5^j, 5^{j+1})$ and $Y_n = 0$ otherwise.

Assume that $\{Y_n;\, n \geq 1\}$ is a sequence of rv's such that $\lim_{n\to\infty}(Y_n) = 0$ WP1. Then for any $\epsilon > 0$, each sample sequence $\{Y_n(\omega);\, n \geq 1\}$ that converges to 0 satisfies $|Y_n| \leq \epsilon$ for all sufficiently large n. This means (see Exercise 4.5) that $\lim_{n\to\infty} \Pr\{|Y_n| \leq \epsilon\} = 1$. Since this is true for all $\epsilon > 0$, $\{Y_n;\, n \geq 0\}$ converges in probability to 0.

4.2.2 Strong law of large numbers (SLLN)

We next develop the strong law of large numbers. We do not have the mathematical tools to prove the theorem in its full generality, but will give a fairly insightful proof under the additional assumption that the rv under discussion has a finite 4th moment. The theorem has a remarkably simple and elementary form, considering that it is certainly one of the most important theorems in probability theory. Most of the hard work in understanding the theorem comes from understanding what convergence WP1 means, and that has already been discussed. Given this understanding, the theorem is relatively easy to understand and surprisingly easy to prove (assuming a 4th moment).

Theorem 4.2.1 (Strong Law of Large Numbers (SLLN)). *For each integer $n \geq 1$, let $S_n = X_1 + \cdots + X_n$, where $X_1, X_2, \ldots$ are IID rv's satisfying $\mathsf{E}[|X|] < \infty$. Then*

$$\Pr\left\{\omega \,:\, \lim_{n\to\infty} \frac{S_n(\omega)}{n} = \overline{X}\right\} = 1. \tag{4.6}$$

Proof (for the case where $\overline{X} = 0$ and $\mathsf{E}\left[X^4\right] < \infty$):

Assume that $\overline{X} = 0$ and $\mathsf{E}\left[X^4\right] < \infty$. Denote $\mathsf{E}\left[X^4\right]$ by γ. For any real number x, if $|x| \leq 1$, then $x^2 \leq 1$, and if $|x| > 1$, then $x^2 < x^4$. Thus $x^2 \leq 1 + x^4$ for all x. It follows $\sigma^2 = \mathsf{E}\left[X^2\right] \leq 1 + \mathsf{E}\left[X^4\right]$. Thus σ^2 is finite if $\mathsf{E}\left[X^4\right]$ is.

Now let $S_n = X_1 + \cdots + X_n$ where $X_1, \ldots, X_n$ are IID with the distribution of X.

$$\begin{aligned}
\mathsf{E}\left[S_n^4\right] &= \mathsf{E}\left[(X_1 + \cdots + X_n)(X_1 + \cdots + X_n)(X_1 + \cdots + X_n)(X_1 + \cdots + X_n)\right] \\
&= \mathsf{E}\left[\left(\sum_{i=1}^n X_i\right)\left(\sum_{j=1}^n X_j\right)\left(\sum_{k=1}^n X_k\right)\left(\sum_{\ell=1}^n X_\ell\right)\right] \\
&= \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{\ell=1}^n \mathsf{E}\left[X_i X_j X_k X_\ell\right],
\end{aligned}$$

where we have multiplied out the product of sums to get a sum of n^4 terms.

For each i, $1 \leq i \leq n$, there is a term in this sum with $i = j = k = \ell$. For each such term, $\mathsf{E}[X_i X_j X_k X_\ell] = \mathsf{E}[X^4] = \gamma$. There are n such terms (one for each choice of i, $1 \leq i \leq n$) and they collectively contribute $n\gamma$ to the sum $\mathsf{E}[S_n^4]$. Also, for each $i, k \neq i$, there is a term with $j = i$ and $\ell = k$. For each of these $n(n-1)$ terms, $\mathsf{E}[X_i X_i X_k X_k] = \sigma^4$. There are another $n(n-1)$ terms with $j \neq i$ and $k = i$, $\ell = j$. Each such term contributes σ^4 to the sum. Finally, for each $i \neq j$, there is a term with $\ell = i$ and $k = j$. Collectively all of these terms contribute $3n(n-1)\sigma^4$ to the sum. Each of the remaining terms is 0 since at least one of i, j, k, ℓ is different from all the others, Thus we have

$$\mathsf{E}[S_n^4] = n\gamma + 3n(n-1)\sigma^4.$$

Now consider the sequence of rv's $\{S_n^4/n^4;\ n \geq 1\}$.

$$\sum_{n=1}^{\infty} \mathsf{E}\left[\left|\frac{S_n^4}{n^4}\right|\right] \;=\; \sum_{n=1}^{\infty} \frac{n\gamma + 3n(n-1)\sigma^4}{n^4} \;<\; \infty,$$

where we have used the facts that the series $\sum_{n\geq 1} 1/n^2$ and the series $\sum_{n\geq 1} 1/n^3$ converge.

Using Lemma 4.2.1 applied to $\{S_n^4/n^4;\ n \geq 1\}$, we see that $\lim_{n\to\infty} S_n^4/n^4 = 0$ WP1. For each ω such that $\lim_{n\to\infty} S_n^4(\omega)/n^4 = 0$, the nonnegative fourth root of that sequence of nonnegative numbers also approaches 0. Thus $\lim_{n\to\infty} |S_n/n| = 0$ WP1. □

The above proof assumed that $\mathsf{E}[X] = 0$. It can be extended trivially to the case of an arbitrary finite $\overline{X}$ by replacing X in the proof with $X - \overline{X}$. A proof using the weaker condition that $\sigma_X^2 < \infty$ will be given in Section 7.9.1.

The technique that was used at the end of this proof provides a clue about why the concept of convergence WP1 is so powerful. The technique showed that if one sequence of rv's ($\{S_n^4/n^4;\ n \geq 1\}$) converges to 0 WP1, then another sequence ($|S_n/n|;\ n \geq 1\}$) also converges WP1. We will formalize and generalize this technique in Lemma 4.3.2 as a major step toward establishing the strong law for renewal processes.

4.3 Strong law for renewal processes

To get an intuitive idea why $N(t)/t$ should approach $1/\overline{X}$ for large t, consider Figure 4.2. For any given sample function of $\{N(t);\ t > 0\}$, note that, for any given t, $N(t)/t$ is the slope of a straight line from the origin to the point $(t, N(t))$. As t increases, this slope decreases in the interval between each adjacent pair of arrival epochs and then jumps up at the next arrival epoch. In order to express this as an equation, note that t lies between the $N(t)$th arrival (which occurs at $S_{N(t)}$) and the $(N(t)+1)$th arrival (which occurs at $S_{N(t)+1}$). Thus, for all sample points,

$$\frac{N(t)}{S_{N(t)}} \geq \frac{N(t)}{t} > \frac{N(t)}{S_{N(t)+1}}. \tag{4.7}$$

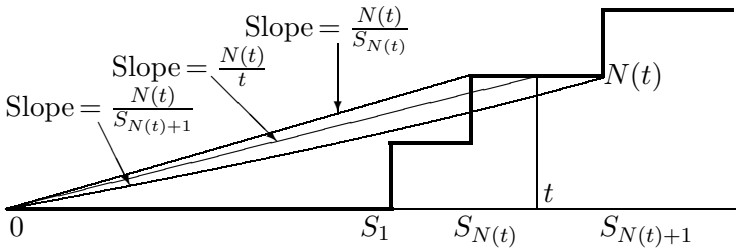


Figure 4.2: Comparison of a sample function of $N(t)/t$ with $\frac{N(t)}{S_{N(t)}}$ and $\frac{N(t)}{S_{N(t)+1}}$ for the same sample point. Note that for the given sample point, $N(t)$ is the number of arrivals up to and including t, and thus $S_{N(t)}$ is the epoch of the last arrival before or at time t. Similarly, $S_{N(t)+1}$ is the epoch of the first arrival strictly after time t.

We want to show intuitively why the slope $N(t)/t$ in the figure approaches $1/\overline{X}$ as $t \to \infty$. As t increases, we would guess that $N(t)$ increases without bound, *i.e.*, that for each arrival, another arrival occurs eventually. Assuming this, the left side of (4.7) increases with increasing t as $1/S_1, 2/S_2, \ldots, n/S_n, \ldots$, where $n = N(t)$. Since S_n/n converges to $\overline{X}$ WP1 from the strong law of large numbers, we might be brave enough or insightful enough to guess that n/S_n converges to $1/\overline{X}$.

We are now ready to state the strong law for renewal processes as a theorem. Before proving the theorem, we formulate the above two guesses as lemmas and prove their validity.

Theorem 4.3.1 (Strong Law for Renewal Processes). *For a renewal process with mean inter-renewal interval $\overline{X} < \infty$, $\lim_{t\to\infty} N(t)/t = 1/\overline{X}$ WP1.*

Lemma 4.3.1. *Let $\{N(t);\, t > 0\}$ be a renewal counting process with inter-renewal rv's $\{X_n;\ n \geq 1\}$. Then (whether or not $\overline{X} < \infty$), $\lim_{t\to\infty} N(t) = \infty$ WP1 and $\lim_{t\to\infty} \mathsf{E}[N(t)] = \infty$.*

Proof of Lemma 4.3.1: Note that for each sample point ω, $N(t,\omega)$ is a nondecreasing real-valued function of t and thus either has a finite limit or an infinite limit. Using (4.1), the probability that this limit is finite with value less than any given n is

$$\lim_{t\to\infty} \Pr\{N(t) < n\} \;=\; \lim_{t\to\infty} \Pr\{S_n > t\} \;=\; 1 - \lim_{t\to\infty} \Pr\{S_n \leq t\}\,.$$

Since the X_i are rv's, the sums S_n are also rv's (*i.e.*, nondefective) for each n (see Section 1.3.7), and thus $\lim_{t\to\infty} \Pr\{S_n \leq t\} = 1$ for each n. Thus $\lim_{t\to\infty} \Pr\{N(t) < n\} = 0$ for each n. This shows that the set of sample points ω for which $\lim_{t\to\infty} N(t(\omega)) < n$ has probability 0 for all n. Thus the set of sample points for which $\lim_{t\to\infty} N(t,\omega)$ is finite has probability 0 and $\lim_{t\to\infty} N(t) = \infty$ WP1.

Next, $\mathsf{E}[N(t)]$ is nondecreasing in t, and thus has either a finite or infinite limit as $t \to \infty$. For each n, $\Pr\{N(t) \geq n\} \geq 1/2$ for large enough t, and therefore $\mathsf{E}[N(t)] \geq n/2$ for such t. Thus $\mathsf{E}[N(t)]$ can have no finite limit as $t \to \infty$, and $\lim_{t\to\infty} \mathsf{E}[N(t)] = \infty$. □

The following lemma is quite a bit more general than the second guess above, but it will be useful elsewhere. This is the formalization of the technique used at the end of the proof of the SLLN.

Lemma 4.3.2. *Let $\{Z_n;\, n \geq 1\}$ be a sequence of rv's such that $\lim_{n\to\infty} Z_n = \alpha$ WP1. Let f be a real valued function of a real variable that is continuous at α. Then*

$$\lim_{n\to\infty} f(Z_n) = f(\alpha) \qquad \text{WP1.} \tag{4.8}$$

Proof of Lemma 4.3.2: First let $z_1, z_2, \ldots$, be a sequence of real numbers such that $\lim_{n\to\infty} z_n = \alpha$. Continuity of f at α means that for every $\epsilon > 0$, there is a $\delta > 0$ such that $|f(z) - f(\alpha)| < \epsilon$ for all z such that $|z - \alpha| < \delta$. Also, since $\lim_{n\to\infty} z_n = \alpha$, we know that for every $\delta > 0$, there is an m such that $|z_n - \alpha| \leq \delta$ for all $n \geq m$. Putting these two statements together, we know that for every $\epsilon > 0$, there is an m such that $|f(z_n) - f(\alpha)| < \epsilon$ for all $n \geq m$. Thus $\lim_{n\to\infty} f(z_n) = f(\alpha)$.

If ω is any sample point such that $\lim_{n\to\infty} Z_n(\omega) = \alpha$, then $\lim_{n\to\infty} f(Z_n(\omega)) = f(\alpha)$. Since this set of sample points has probability 1, (4.8) follows. □

Proof of Theorem 4.3.1, Strong law for renewal processes: Since $\Pr\{X > 0\} = 1$ for a renewal process, we see that $\overline{X} > 0$. Choosing $f(x) = 1/x$, we see that $f(x)$ is continuous at $x = \overline{X}$. It follows from Lemma 4.3.2 that

$$\lim_{n\to\infty} \frac{n}{S_n} = \frac{1}{\overline{X}} \qquad \text{WP1.}$$

From Lemma 4.3.1, we know that $\lim_{t\to\infty} N(t) = \infty$ with probability 1, so, with probability 1, $N(t)$ increases through all the nonnegative integers as t increases from 0 to ∞. Thus

$$\lim_{t\to\infty} \frac{N(t)}{S_{N(t)}} = \lim_{n\to\infty} \frac{n}{S_n} = \frac{1}{\overline{X}} \qquad \text{WP1.}$$

Recall that $N(t)/t$ is sandwiched between $N(t)/S_{N(t)}$ and $N(t)/S_{N(t)+1}$, so we can complete the proof by showing that $\lim_{t\to\infty} N(t)/S_{N(t)+1} = 1/\overline{X}$. To show this,

$$\lim_{t\to\infty} \frac{N(t)}{S_{N(t)+1}} = \lim_{n\to\infty} \frac{n}{S_{n+1}} = \lim_{n\to\infty} \frac{n+1}{S_{n+1}} \frac{n}{n+1} = \frac{1}{\overline{X}} \qquad \text{WP1.}$$

□

We have gone through the proof of this theorem in great detail, since a number of the techniques are probably unfamiliar to many readers. If one reads the proof again, after becoming familiar with the details, the simplicity of the result will be quite striking. The theorem is also true if the mean inter-renewal interval is infinite; this can be seen by a truncation argument (see Exercise 4.8).

As explained in Section 4.2.1, Theorem 4.3.1 also implies the corresponding weak law of large numbers for $N(t)$, *i.e.*, for any $\epsilon > 0$, $\lim_{t\to\infty} \Pr\{|N(t)/t - 1/\overline{X}| \geq \epsilon\} = 0)$. This weak law could also be derived from the weak law of large numbers for S_n (Theorem 1.5.3). We do not pursue that here, since the derivation is tedious and uninstructive. As we will see, it is the strong law that is most useful for renewal processes.

Figure 4.3 helps give some appreciation of what the strong law for $N(t)$ says and doesn't say. The strong law deals with time-averages, $\lim_{t\to\infty} N(t,\omega)/t$, for individual sample points ω;

these are indicated in the figure as horizontal averages, one for each ω. It is also of interest to look at time and ensemble-averages, $\mathsf{E}[N(t)/t]$, shown in the figure as vertical averages. Note that $N(t,\omega)/t$ is the time-average number of renewals from 0 to t, whereas $\mathsf{E}[N(t)/t]$ averages also over the ensemble. Finally, to focus on arrivals in the vicinity of a particular time t, it is of interest to look at the ensemble-average $\mathsf{E}[N(t+\delta)-N(t)]/\delta$.

Given the strong law for $N(t)$, one would hypothesize that $\mathsf{E}[N(t)/t]$ approaches $1/\overline{X}$ as $t \to \infty$. One might also hypothesize that $\lim_{t\to\infty} \mathsf{E}[N(t+\delta)-N(t)]/\delta = 1/\overline{X}$, subject to some minor restrictions on δ. These hypotheses are correct and are discussed in detail in what follows. This equality of time-averages and limiting ensemble-averages for renewal processes carries over to a large number of stochastic processes, and forms the basis of *ergodic theory*. These results are important for both theoretical and practical purposes. It is sometimes easy to find time averages (just like it was easy to find the time-average $N(t,\omega)/t$ from the strong law of large numbers), and it is sometimes easy to find limiting ensemble-averages. Being able to equate the two then allows us to alternate at will between time and ensemble-averages.

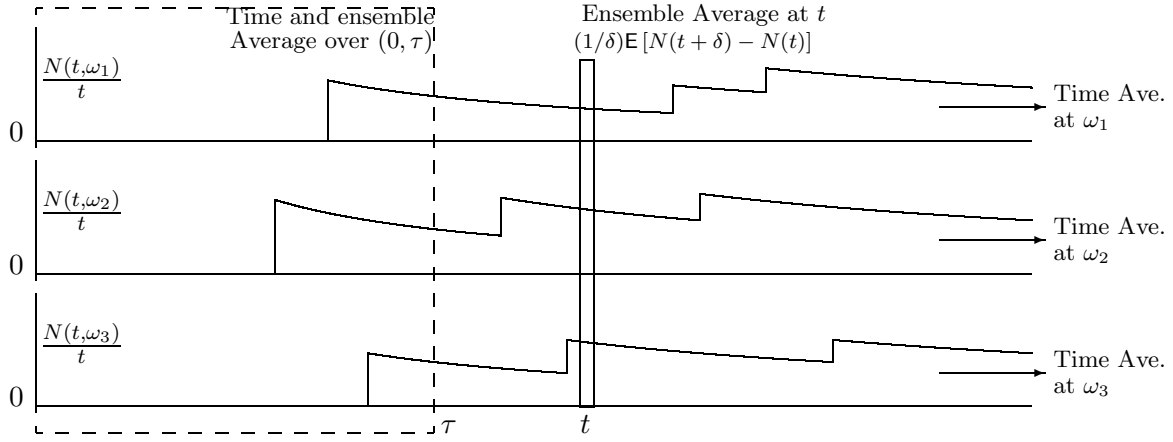


Figure 4.3: The time average at a sample point ω, the time and ensemble average from 0 to a given τ, and the ensemble-average in an interval $(t, t+\delta]$.

Note that in order to equate time-averages and limiting ensemble-averages, quite a few conditions are required. First, the time-average must exist in the limit $t \to \infty$ with probability one and also have a fixed value with probability one; second, the ensemble-average must approach a limit as $t \to \infty$; and third, the limits must be the same. The following example, for a stochastic process very different from a renewal process, shows that equality between time and ensemble averages is not always satisfied for arbitrary processes.

Example 4.3.1. Let $\{X_i; i \geq 1\}$ be a sequence of binary IID random variables, each taking the value 0 with probability 1/2 and 2 with probability 1/2. Let $\{M_n; n \geq 1\}$ be the product process in which $M_n = X_1 X_2 \cdots X_n$. Since $M_n = 2^n$ if X_1 to X_n each take the value 2 (an event of probability 2^{-n}) and $M_n = 0$ otherwise, we see that $\lim_{n\to\infty} M_n = 0$ with probability 1. Also $\mathsf{E}[M_n] = 1$ for all $n \geq 1$. Thus the time-average exists and equals 0 with probability 1 and the ensemble-average exists and equals 1 for all n, but the two are different. The problem is that as n increases, the atypical event in which $M_n = 2^n$ has a

probability approaching 0, but still has a significant effect on the ensemble-average.

Further discussion of ensemble averages is postponed to Section 4.6. Before that, we briefly state and discuss the central limit theorem for counting renewal processes and then introduce the notion of rewards associated with renewal processes.

Theorem 4.3.2 (Central Limit Theorem (CLT) for $N(t)$). *Assume that the inter-renewal intervals for a renewal counting process $\{N(t);\, t>0\}$ have finite standard deviation $\sigma>0$. Then*

$$\lim_{t\to\infty} \Pr\left\{ \frac{N(t)-t/\overline{X}}{\sigma \overline{X}^{-3/2}\sqrt{t}} < \alpha \right\} = \Phi(\alpha). \tag{4.9}$$

where $\Phi(y)=\int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)dx$.

This says that the distribution function of $N(t)$ tends to the Gaussian distribution with mean $t/\overline{X}$ and standard deviation $\sigma\overline{X}^{-3/2}\sqrt{t}$.

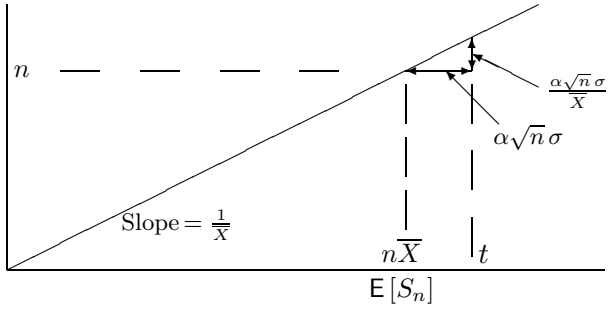


Figure 4.4: Illustration of the central limit theorem (CLT) for renewal processes. A given integer n is shown on the vertical axis, and the corresponding mean, $\mathsf{E}[S_n]=n\overline{X}$ is shown on the horizontal axis. The horizontal line with arrows at height n indicates α standard deviations from $\mathsf{E}[S_n]$, and the vertical line with arrows indicates the distance below $(t/\overline{X})$.

The theorem can be proved by applying Theorem 1.5.2 (the CLT for a sum of IID rv's) to S_n and then using the identity $\{S_n \le t\} = \{N(t)\ge n\}$. The general idea is illustrated in Figure 4.4, but the details are somewhat tedious, and can be found, for example, in [16]. We simply outline the argument here. For any real α, the CLT states that

$$\Pr\big\{S_n \le n\overline{X} + \alpha\sqrt{n}\sigma\big\} \approx \Phi(\alpha),$$

where $\Phi(\alpha)=\int_{-\infty}^{\alpha}\frac{1}{\sqrt{2\pi}}\exp(-x^2/2)\,dx$ and where the approximation becomes exact in the limit $n\to\infty$. Letting

$$t = n\overline{X} + \alpha\sqrt{n}\sigma,$$

and using $\{S_n\le t\}=\{N(t)\ge n\}$,

$$\Pr\{N(t)\ge n\} \approx \Phi(\alpha). \tag{4.10}$$

Since t is monotonic in n for fixed α, we can express n in terms of t, getting

$$n = \frac{t}{\overline{X}} - \frac{\alpha\sigma\sqrt{n}}{\overline{X}} \approx \frac{t}{\overline{X}} - \alpha\sigma t^{1/2}(\overline{X})^{-3/2}.$$

Substituting this into (4.10) establishes the theorem for $-\alpha$, which establishes the theorem since α is arbitrary. The omitted details involve handling the approximations carefully.

4.4 Renewal-reward processes; time-averages

There are many situations in which, along with a renewal counting process $\{N(t);\, t > 0\}$, there is another randomly varying function of time, called a *reward function* $\{R(t); t > 0\}$. $R(t)$ models a rate at which the process is accumulating a reward. We shall illustrate many examples of such processes and see that a "reward" could also be a cost or any randomly varying quantity of interest. The important restriction on these *reward functions* is that $R(t)$ at a given t depends only on the location of t within the inter-renewal interval containing t and perhaps other random variables local to that interval. Before defining this precisely, we start with several examples.

Example 4.4.1. (Time-average residual life) For a renewal counting process $\{N(t), t > 0\}$, let $Y(t)$ be the residual life at time t. The *residual life* is defined as the interval from t until the next renewal epoch, i.e., as $S_{N(t)+1} - t$. For example, if we arrive at a bus stop at time t and buses arrive according to a renewal process, $Y(t)$ is the time we have to wait for a bus to arrive (see Figure 4.5). We interpret $\{Y(t);\, t \geq 0\}$ as a reward function. The time-average of $Y(t)$, over the interval $(0, t]$, is given by[5] $(1/t)\int_0^t Y(\tau)d\tau$. We are interested in the limit of this average as $t \to \infty$ (assuming that it exists in some sense). Figure 4.5 illustrates a sample function of a renewal counting process $\{N(t);\, t > 0\}$ and shows the residual life $Y(t)$ for that sample function. Note that, for a given sample function $\{Y(t) = y(t)\}$, the integral $\int_0^t y(\tau)\, d\tau$ is simply a sum of isosceles right triangles, with part of a final triangle at the end. Thus it can be expressed as

$$\int_0^t y(\tau)d\tau = \frac{1}{2}\sum_{i=1}^{n(t)} x_i^2 + \int_{\tau=s_{n(t)}}^t y(\tau)d\tau,$$

where $\{x_i;\, 0 < i < \infty\}$ is the set of sample values for the inter-renewal intervals.

Since this relationship holds for every sample point, we see that the random variable $\int_0^t Y(\tau)d\tau$ can be expressed in terms of the inter-renewal random variables X_n as

$$\int_{\tau=0}^t Y(\tau)d\tau = \frac{1}{2}\sum_{n=1}^{N(t)} X_n^2 + \int_{\tau=S_{N(t)}}^t Y(\tau)d\tau.$$

[5] $\int_0^t Y(\tau)d\tau$ is a rv just like any other function of a set of rv's. It has a sample value for each sample function of $\{N(t);\, t > 0\}$, and its distribution function could be calculated in a straightforward but tedious way. For arbitrary stochastic processes, integration and differentiation can require great mathematical sophistication, but none of those subtleties occur here.

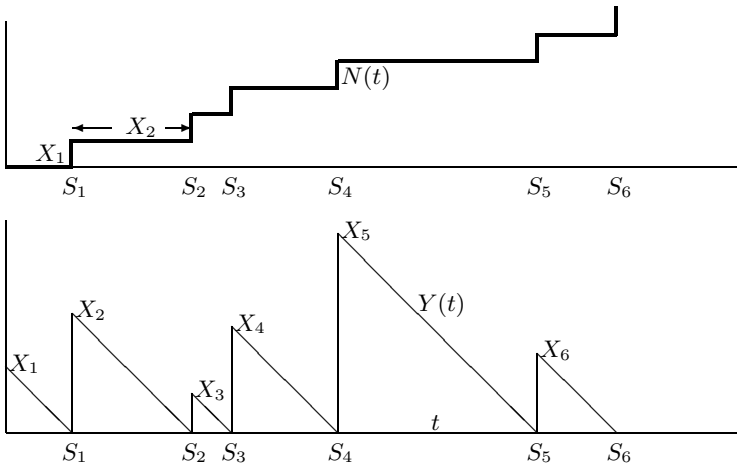


Figure 4.5: Residual life at time t. For any given sample function of the renewal process, the sample function of residual life decreases linearly with a slope of -1 from the beginning to the end of each inter-renewal interval.

Although the final term above can be easily evaluated for a given $S_{N(t)}(t)$, it is more convenient to use the following bound:

$$\frac{1}{2t}\sum_{n=1}^{N(t)} X_n^2 \le \frac{1}{t}\int_{\tau=0}^{t} Y(\tau)d\tau \le \frac{1}{2t}\sum_{n=1}^{N(t)+1} X_n^2. \tag{4.11}$$

The term on the left can now be evaluated in the limit $t \to \infty$ (for all sample functions except a set of probability zero) as follows:

$$\lim_{t\to\infty} \frac{\sum_{n=1}^{N(t)} X_n^2}{2t} = \lim_{t\to\infty} \frac{\sum_{n=1}^{N(t)} X_n^2}{N(t)}\,\frac{N(t)}{2t}. \tag{4.12}$$

Consider each term on the right side of (4.12) separately. For the first term, recall that $\lim_{t\to 0} N(t) = \infty$ with probability 1. Thus as $t \to \infty$, $\sum_{n=1}^{N(t)} X_n^2/N(t)$ goes through the same set of values as $\sum_{n=1}^{k} X_n^2/k$ as $k \to \infty$. Thus, using the SLLN,

$$\lim_{t\to\infty} \frac{\sum_{n=1}^{N(t)} X_n^2}{N(t)} = \lim_{k\to\infty} \frac{\sum_{n=1}^{k} X_n^2}{k} = \mathsf{E}\left[X^2\right] \qquad \text{WP1.}$$

The second term on the right side of (4.12) is simply $N(t)/2t$. By the strong law for renewal processes, $\lim_{t\to\infty} N(t)/2t = 1/(2\mathsf{E}[X])$ WP1. Thus both limits exist WP1 and

$$\lim_{t\to\infty} \frac{\sum_{n=1}^{N(t)} X_n^2}{2t} = \frac{\mathsf{E}\left[X^2\right]}{2\mathsf{E}[X]} \qquad \text{WP1.} \tag{4.13}$$

The right hand term of (4.11) is handled almost the same way:

$$\lim_{t\to\infty} \frac{\sum_{n=1}^{N(t)+1} X_n^2}{2t} = \lim_{t\to\infty} \frac{\sum_{n=1}^{N(t)+1} X_n^2}{N(t)+1}\,\frac{N(t)+1}{N(t)}\,\frac{N(t)}{2t} = \frac{\mathsf{E}\left[X^2\right]}{2\mathsf{E}[X]}. \tag{4.14}$$

Combining these two results, we see that, with probability 1, the time-average residual life is given by

$$\lim_{t\to\infty} \frac{\int_{\tau=0}^{t} Y(\tau)\,d\tau}{t} = \frac{\mathsf{E}\left[X^2\right]}{2\mathsf{E}\left[X\right]}. \tag{4.15}$$

Note that this time-average depends on the second moment of X; this is $\overline{X}^2 + \sigma^2 \geq \overline{X}^2$, so the time-average residual life is at least half the expected inter-renewal interval (which is not surprising). On the other hand, the second moment of X can be arbitrarily large (even infinite) for any given value of $\mathsf{E}\left[X\right]$, so that the time-average residual life can be arbitrarily large relative to $\mathsf{E}\left[X\right]$. This can be explained intuitively by observing that large inter-renewal intervals are weighted more heavily in this time-average than small inter-renewal intervals.

Example 4.4.2. As an example of the effect of improbable but large inter-renewal intervals, let X take on the value ϵ with probability $1-\epsilon$ and value $1/\epsilon$ with probability ϵ. Then, for small ϵ, $\mathsf{E}\left[X\right] \sim 1, \mathsf{E}\left[X^2\right] \sim 1/\epsilon$, and the time average residual life is approximately $1/(2\epsilon)$ (see Figure 4.6).

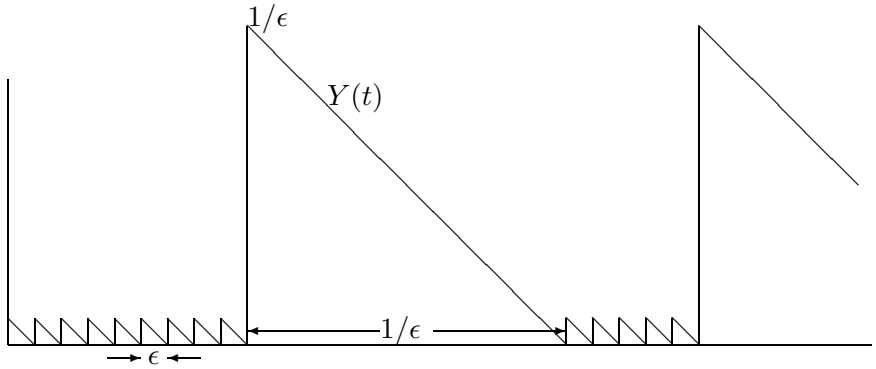


Figure 4.6: Average Residual life is dominated by large interarrival intervals. Each large interval has duration $1/\epsilon$, and the expected aggregate duration between successive large intervals is $1-\epsilon$

Example 4.4.3. (time-average Age) Let $Z(t)$ be the age of a renewal process at time t where *age* is defined as the interval from the most recent arrival before (or at) t until t, i.e., $Z(t) = t - S_{N(t)}$. By convention, if no arrivals have occurred by time t, we take the age to be t (*i.e.*, in this case, $N(t) = 0$ and we take S_0 to be 0).

As seen in Figure 4.16, the age process, for a given sample function of the renewal process, is almost the same as the residual life process—the isosceles right triangles are simply turned around. Thus the same analysis as before can be used to show that the time average of $Z(t)$ is the same as the time-average of the residual life,

$$\lim_{t\to\infty} \frac{\int_{\tau=0}^{t} Z(\tau)\,d\tau}{t} = \frac{\mathsf{E}\left[X^2\right]}{2\mathsf{E}\left[X\right]} \qquad \text{WP1.} \tag{4.16}$$

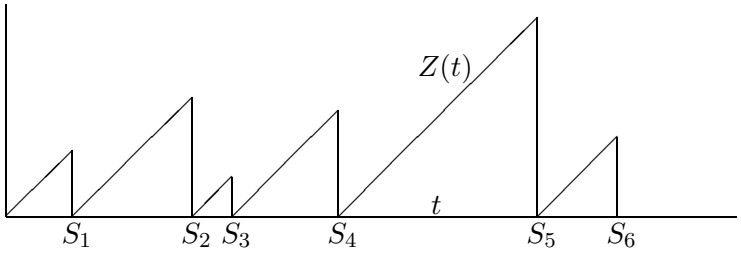


Figure 4.7: Age at time t: For any given sample function of the renewal process, the sample function of age increases linearly with a slope of 1 from the beginning to the end of each inter-renewal interval.

Example 4.4.4. (time-average Duration) Let $\widetilde{X}(t)$ be the duration of the inter-renewal interval containing time t, i.e., $\widetilde{X}(t) = X_{N(t)+1} = S_{N(t)+1} - S_{N(t)}$ (see Figure 4.8). It is clear that $\widetilde{X}(t) = Z(t) + Y(t)$, and thus the time-average of the duration is given by

$$\lim_{t\to\infty} \frac{\int_{\tau=0}^{t} \widetilde{X}(\tau)\, d\tau}{t} = \frac{\mathsf{E}\left[X^2\right]}{\mathsf{E}\left[X\right]} \qquad \text{WP1.} \tag{4.17}$$

Again, long intervals are heavily weighted in this average, so that the time-average duration is at least as large as the mean inter-renewal interval and often much larger.

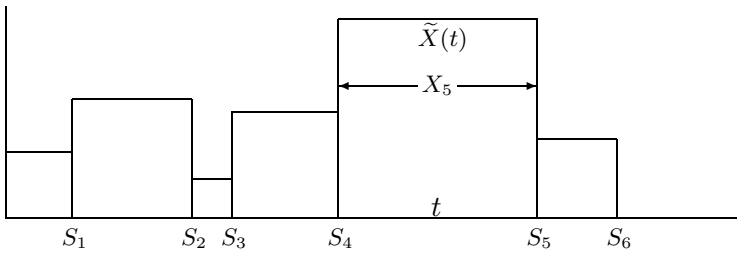


Figure 4.8: Duration $\widetilde{X}(t) = X_{N(t)}$ of the inter-renewal interval containing t.

4.4.1 General renewal-reward processes

In each of these examples, and in many other situations, we have a random function of time (i.e., $Y(t)$, $Z(t)$, or $\widetilde{X}(t)$) whose value at time t depends only on where t is in the current inter-renewal interval (i.e., on the age $Z(t)$ and the duration $\widetilde{X}(t)$ of the current inter-renewal interval). We now investigate the general class of reward functions for which the reward at time t depends at most on the age and the duration at t, *i.e.*, the reward $R(t)$ at time t is given explicitly as a function[6] $\mathcal{R}(Z(t), \widetilde{X}(t))$ of the age and duration at t.

[6]This means that $R(t)$ can be determined at any t from knowing $Z(t)$ and $X(t)$. It does not mean that $R(t)$ must vary as either of those quantities are changed. Thus, for example, $R(t)$ could depend on only one of the two or could even be a constant.

For the three examples above, the function $\mathcal{R}$ is trivial. That is, the residual life, $Y(t)$, is given by $\widetilde{X}(t) - Z(t)$ and the age and duration are given directly.

We now find the time-average value of $R(t)$, namely, $\lim_{t\to\infty} \frac{1}{t} \int_0^t R(\tau)\, d\tau$. As in examples 4.4.1 to 4.4.4 above, we first want to look at the accumulated reward over each inter-renewal period separately. Define R_n as the accumulated reward in the nth renewal interval,

$$\mathrm{R}_n = \int_{S_{n-1}}^{S_n} R(\tau)\, d(\tau) = \int_{S_{n-1}}^{S_n} \mathcal{R}[Z(\tau), \widetilde{X}(\tau)]\, d\tau. \tag{4.18}$$

For residual life (see Example 4.4.1), R_n is the area of the nth isosceles right triangle in Figure 4.5. In general, since $Z(\tau) = \tau - S_{n-1}$,

$$\mathrm{R}_n = \int_{S_{n-1}}^{S_n} \mathcal{R}(\tau - S_{n-1}, X_n)\, d\tau = \int_{z=0}^{X_n} \mathcal{R}(z, X_n)\, dz. \tag{4.19}$$

Note that R_n is a function only of X_n, where the form of the function is determined by $\mathcal{R}(Z, X)$. From this, it is clear that $\{\mathrm{R}_n; n \geq 1\}$ is essentially[7] a set of IID random variables. For residual life, $\mathcal{R}(z, X_n) = X_n - z$, so the integral in (4.19) is $X_n^2/2$, as calculated by inspection before. In general, from (4.19), the expected value of R_n is given by

$$\mathsf{E}\,[\mathrm{R}_n] = \int_{x=0}^{\infty} \int_{z=0}^{x} \mathcal{R}(z, x)\, dz\, d\mathsf{F}_X(x). \tag{4.20}$$

Breaking $\int_0^t R(\tau)\, d\tau$ into the reward over the successive renewal periods, we get

$$\begin{aligned} \int_0^t R(\tau)\, d\tau &= \int_0^{S_1} R(\tau)\, d\tau + \int_{S_1}^{S_2} R(\tau)\, d\tau + \cdots + \int_{S_{N(t)-1}}^{S_{N(t)}} R(\tau)\, d\tau + \int_{S_{N(t)}}^{t} R(\tau) d\tau \\ &= \sum_{n=1}^{N(t)} \mathrm{R}_n + \int_{S_{N(t)}}^{t} R(\tau)\, d\tau. \end{aligned} \tag{4.21}$$

The following theorem now generalizes the results of Examples 4.4.1, 4.4.3, and 4.4.4 to general renewal-reward functions.

Theorem 4.4.1. *Let $\{R(t); t > 0\} \geq 0$ be a nonnegative renewal-reward function for a renewal process with expected inter-renewal time $\mathsf{E}\,[X] = \overline{X} < \infty$. If $\mathsf{E}\,[\mathrm{R}_n] < \infty$, then with probability 1*

$$\lim_{t\to\infty} \frac{1}{t} \int_{\tau=0}^{t} R(\tau)\, d\tau = \frac{\mathsf{E}\,[\mathrm{R}_n]}{\overline{X}}. \tag{4.22}$$

Proof: Using (4.21), the accumulated reward up to time t can be bounded between the accumulated reward up to the renewal before t and that to the next renewal after t,

$$\frac{\sum_{n=1}^{N(t)} \mathrm{R}_n}{t} \leq \frac{\int_{\tau=0}^{t} R(\tau)\, d\tau}{t} \leq \frac{\sum_{n=1}^{N(t)+1} \mathrm{R}_n}{t}. \tag{4.23}$$

[7] One can certainly define functions $\mathcal{R}(Z, X)$ for which the integral in (4.19) is infinite or undefined for some values of X_n, and thus R_n becomes a defective rv. It seems better to handle this type of situation when it arises rather than handling it in general.

The left hand side of (4.23) can now be broken into

$$\frac{\sum_{n=1}^{N(t)} \mathrm{R}_n}{t} = \frac{\sum_{n=1}^{N(t)} \mathrm{R}_n}{N(t)} \; \frac{N(t)}{t}. \tag{4.24}$$

Each R_n is a given function of X_n, so the R_n are IID. As $t \to \infty$, $N(t) \to \infty$, and, thus, as we have seen before, the strong law of large numbers can be used on the first term on the right side of (4.24), getting $\mathsf{E}[\mathrm{R}_n]$ with probability 1. Also the second term approaches $1/\overline{X}$ by the strong law for renewal processes. Since $0 < \overline{X} < \infty$ and $\mathsf{E}[\mathrm{R}_n]$ is finite, the product of the two terms approaches the limit $\mathsf{E}[\mathrm{R}_n]/\overline{X}$. The right-hand inequality of (4.23) is handled in almost the same way,

$$\frac{\sum_{n=1}^{N(t)+1} \mathrm{R}_n}{t} = \frac{\sum_{n=1}^{N(t)+1} \mathrm{R}_n}{N(t)+1} \; \frac{N(t)+1}{N(t)} \; \frac{N(t)}{t}. \tag{4.25}$$

It is seen that the terms on the right side of (4.25) approach limits as before and thus the term on the left approaches $\mathsf{E}[\mathrm{R}_n]/\overline{X}$ with probability 1. Since the upper and lower bound in (4.23) approach the same limit, $(1/t)\int_0^t R(\tau)\,d\tau$ approaches the same limit and the theorem is proved. □

The restriction to nonnegative renewal-reward functions in Theorem 4.4.1 is slightly artificial. The same result holds for non-positive reward functions simply by changing the directions of the inequalities in (4.23). Assuming that $\mathsf{E}[\mathrm{R}_n]$ exists (i.e., that both its positive and negative parts are finite), the same result applies in general by splitting an arbitrary reward function into a positive and negative part. This gives us the corollary:

Corollary 4.4.1. *Let $\{R(t); t > 0\}$ be a renewal-reward function for a renewal process with expected inter-renewal time $\mathsf{E}[X] = \overline{X} < \infty$. If $\mathsf{E}[\mathrm{R}_n]$ exists, then with probability 1*

$$\lim_{t\to\infty} \frac{1}{t} \int_{\tau=0}^{t} R(\tau)\,d\tau = \frac{\mathsf{E}[\mathrm{R}_n]}{\overline{X}}. \tag{4.26}$$

Example 4.4.5. (Distribution of Residual Life) Example 4.4.1 treated the time-average value of the residual life $Y(t)$. Suppose, however, that we would like to find the time-average distribution function of $Y(t)$, i.e., the fraction of time that $Y(t) \le y$ as a function of y. The approach, which applies to a wide variety of applications, is to use an indicator function (for a given value of y) as a reward function. That is, define $R(t)$ to have the value 1 for all t such that $Y(t) \le y$ and to have the value 0 otherwise. Figure 4.9 illustrates this function for a given sample path. Expressing this reward function in terms of $Z(t)$ and $\widetilde{X}(t)$, we have

$$R(t) = \mathcal{R}(Z(t), \widetilde{X}(t)) = \begin{cases} 1; & \widetilde{X}(t) - Z(t) \le y \\ 0; & \text{otherwise} \end{cases}.$$

Note that if an inter-renewal interval is smaller than y (such as the third interval in Figure 4.9), then $R(t)$ has the value one over the entire interval, whereas if the interval is greater

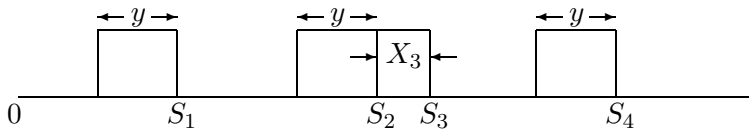


Figure 4.9: Reward function to find the time-average fraction of time that $\{Y(t) \leq y\}$. For the sample function in the figure, $X_1 > y$, $X_2 > y$, and $X_4 > y$, but $X_3 < y$

than y, then $R(t)$ has the value one only over the final y units of the interval. Thus $\mathrm{R}_n = \min[y, X_n]$. Note that the random variable $\min[y, X_n]$ is equal to X_n for $X_n \leq y$, and thus has the same distribution function as X_n in the range 0 to y. Figure 4.10 illustrates this in terms of the complementary distribution function. From the figure, we see that

$$\mathsf{E}[\mathrm{R}_n] = \mathsf{E}[\min(X, y)] = \int_{x=0}^{\infty} \Pr\{\min(X, y) > x\}\, dx = \int_{x=0}^{y} \Pr\{X > x\}\, dx. \tag{4.27}$$

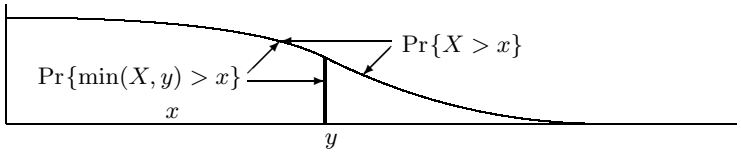


Figure 4.10: R_n for distribution of residual life.

Let $\mathsf{F}_Y(y) = \lim_{t\to\infty}(1/t)\int_0^t R(\tau)\, d\tau$ denote the time-average fraction of time that the residual life is less than or equal to y. From Theorem 4.4.1 and Eq.(4.27), we then have

$$\mathsf{F}_Y(y) = \frac{\mathsf{E}[\mathrm{R}_n]}{\overline{X}} = \frac{1}{\overline{X}} \int_{x=0}^{y} \Pr\{X > x\}\, dx \qquad \text{WP1.} \tag{4.28}$$

As a check, note that this integral is increasing in y and approaches 1 as $y \to \infty$. Note also that the expected value of Y, calculated from (4.28), is given by $\mathsf{E}\left[X^2\right]/2\overline{X}$, in agreement with (4.15).

The same argument can be applied to the time-average distribution of age (see Exercise 4.12). The time-average fraction of time, $\mathsf{F}_Z(z)$, that the age is at most z is given by

$$\mathsf{F}_Z(z) = \frac{1}{\overline{X}} \int_{x=0}^{z} \Pr\{X > x\}\, dx \qquad \text{WP1.} \tag{4.29}$$

In the development so far, the reward function $R(t)$ has been a function solely of the age and duration intervals, and the aggregate reward over the nth inter-renewal interval is a function only of X_n. In more general situations, where the renewal process is embedded in some more complex process, it is often desirable to define $R(t)$ to depend on other aspects of the process as well. The important thing here is for $\{\mathrm{R}_n;\ n \geq 1\}$ to be an IID sequence. How to achieve this, and how it is related to queueing systems, is described in Section 4.5.3. Theorem 4.4.1 clearly remains valid if $\{\mathrm{R}_n;\ n \geq 1\}$ is IID. This more general type

of renewal-reward function will be required and further discussed in Sections 4.5.3 to Rss7 where we discuss Little's theorem and the M/G/1 expected queueing delay, both of which use this more general structure.

Limiting time-averages are sometimes visualized by the following type of experiment. For some given large time t, let T be a uniformly distributed random variable over $(0,t]$; T is independent of the renewal-reward process under consideration. Then $(1/t)\int_0^t R(\tau)\,d\tau$ is the expected value (over T) of $R(T)$ for a given sample path of $\{R(\tau);\ \tau{>}0\}$. Theorem 4.4.1 states that in the limit $t\to\infty$, all sample paths (except a set of probability 0) yield the same expected value over T. This approach of viewing a time-average as a random choice of time is referred to as *random incidence*. Random incidence is awkward mathematically, since the random variable T changes with the overall time t and has no reasonable limit. It also blurs the distinction between time and ensemble-averages, so it will not be used in what follows.

4.5 Random stopping trials

Visualize performing an experiment repeatedly, observing independent successive sample outputs of a given random variable (i.e., observing a sample outcome of $X_1, X_2, \ldots$ where the X_i are IID). The experiment is stopped when enough data has been accumulated for the purposes at hand.

This type of situation occurs frequently in applications. For example, we might be required to make a choice from several hypotheses, and might repeat an experiment until the hypotheses are sufficiently discriminated. If the number of trials is allowed to depend on the outcome, the mean number of trials required to achieve a given error probability is typically a small fraction of the number of trials required when the number is chosen in advance. Another example occurs in tree searches where a path is explored until further extensions of the path appear to be unprofitable.

The first careful study of experimental situations where the number of trials depends on the data was made by the statistician Abraham Wald and led to the field of sequential analysis (see [21]). We study these situations now since one of the major results, Wald's equality, will be useful in studying $\mathsf{E}[N(t)]$ in the next section. Stopping trials are frequently useful in the study of random processes, and in particular will be used in Section 4.7 for the analysis of queues, and again in Chapter 7 as central topics in the study of Random walks and martingales.

An important part of experiments that stop after a random number of trials is the rule for stopping. Such a rule must specify, for each sample path, the trial at which the experiment stops, *i.e.*, the final trial after which no more trials are performed. Thus the rule for stopping should specify a positive, integer valued, random variable J, called the *stopping time*, or *stopping trial*, mapping sample paths to this final trial at which the experiment stops.

We view the sample space as including the set of sample value sequences for the never-ending sequence of random variables $X_1, X_2, \ldots$. That is, even if the experiment is stopped at the end of the second trial, we still visualize the 3rd, 4th, ... random variables as having sample

values as part of the sample function. In other words, we visualize that the experiment continues forever, but that the observer stops watching at the end of the stopping point. From the standpoint of applications, the experiment might or might not continue after the observer stops watching. From a mathematical standpoint, however, it is far preferable to view the experiment as continuing. This avoids confusion and ambiguity about the meaning of IID rv's when the very existence of later variables depends on earlier sample values.

The intuitive notion of stopping a sequential experiment should involve stopping based on the data (*i.e.*, the sample values) gathered up to and including the stopping point. For example, if $X_1, X_2, \ldots$, represent the succesive changes in our fortune when gambling, we might want to stop when our cumulative gain exceeds some fixed value. The stopping trial n then depends on the sample values of $X_1, X_2, \ldots, X_n$. At the same time, we want to exclude from stopping trials those rules that allow the experimenter to peek at subsequent values before making the decision to stop or not.[8] This leads to the following definition.

Definition 4.5.1. *A stopping trial (or stopping time[9]) J for a sequence of rv's $X_1, X_2, \ldots$, is a positive integer-valued rv such that for each $n \geq 1$, the indicator rv $\mathbb{I}_{\{J=n\}}$ is a function of $\{X_1, X_2, \ldots, X_n\}$.*

The last clause of the definition means that any given sample value $x_1, \ldots, x_n$ for $X_1, \ldots, X_n$ uniquely determines whether the corresponding sample value of J is n or not. Note that since the stopping trial J is defined to be a positive integer-valued rv, the events $\{J = n\}$ and $\{J = m\}$ for $m < n$ are disjoint events, so stopping at trial m makes it impossible to also stop at n for a given sample path. Also the union of the events $\{J = n\}$ over $n \geq 1$ has probability 1. Aside from this final restriction, the definition does not depend on the probability measure and depends solely on the set of events $\{J = n\}$ for each n. In many situations, it is useful to relax the definition further to allow J to be a possibly-defective rv. In this case the question of whether stopping occurs with probability 1 can be postponed until after specifying the disjoint events $\{J = n\}$ over $n \geq 1$.

Example 4.5.1. Consider a Bernoulli process $\{X_n; n \geq 1\}$. A very simple stopping trial for this process is to stop at the first occurrence of the string $(1, 0)$. Figure 4.11 illustrates this stopping trial by viewing it as a truncation of the tree of possible binary sequences.

The event $\{J = 2\}$, *i.e.*, the event that stopping occurs at trial 2, is the event $\{X_1{=}1,\ X_2{=}0\}$. Similarly, the event $\{J = 3\}$ is $\{X_1{=}1,\ X_2{=}1,\ X_3{=}0\} \bigcup \{X_1{=}0,\ X_2{=}1,\ X_3{=}0\}$. The disjointness of $\{J = n\}$ and $\{J = m\}$ for $n \neq m$ is represented in the figure by terminating the tree at each stopping node. It can be seen that the tree never dies out completely, and in fact, for each trial n, the number of stopping nodes is $n - 1$. However, the probability that stopping has not occurred by trial n goes to zero exponentially with n, which ensures that J is a random variable.

[8] For example, poker players do not take kindly to a player who attempts to withdraw his bet when someone else wins the hand. Similarly, a statistician gathering data on product failures should not respond to a failure by then recording an earlier trial as a stopping time, thus not recording the failure.

[9] Stopping trials are more often called stopping times or optional stopping times in the literature. In our first major application of a stopping trial, however, the stopping trial is the first trial n at which a renewal epoch S_n exceeds a given time t . Viewing this *trial* as a *time* generates considerable confusion.

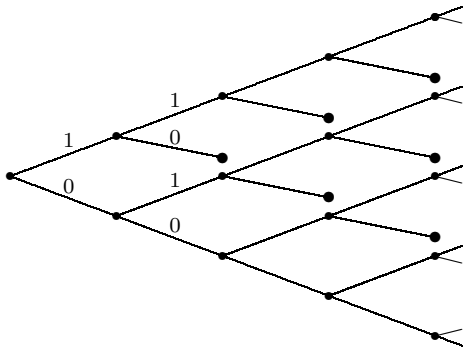


Figure 4.11: A tree representing the set of binary sequences, with a stopping rule viewed as a pruning of the tree. The particular stopping rule here is to stop on the first occurrence of the string $(1, 0)$. The leaves of the tree (*i.e.*, the nodes at which stopping occurs) are marked with large dots and the intermediate nodes (the other nodes) with small dots. Note that each leaf in the tree has a one-to-one correspondence with an initial segment of the tree, so the stopping nodes can be unambiguously viewed either as leaves of the tree or initial segments of the sample sequences.

Representing a stopping rule by a pruned tree can be used for any discrete random sequence, although the tree becomes quite unwieldy in all but trivial cases. Visualizing a stopping rule in terms of a pruned tree is useful conceptually, but stopping rules are usually stated in other terms. For example, we shortly consider a stopping trial for the interarrival intervals of a renewal process as the first n for which the arrival epoch S_n satisfies $S_n > t$ for some given $t > 0$.

4.5.1 Wald's equality

An important question that arises with stopping trials is to evaluate the sum S_J of the random variables up to the stopping trial, *i.e.*, $S_J = \sum_{n=1}^{J} X_n$. Many gambling strategies and investing strategies involve some sort of rule for when to stop, and it is important to understand the rv S_J (which can model the overall gain or loss up to that trial). Wald's equality is very useful in helping to find $\mathsf{E}[S_J]$.

Theorem 4.5.1 (Wald's equality). *Let $\{X_n; n \geq 1\}$ be a sequence of IID rv's, each of mean $\overline{X}$. If J is a stopping trial for $\{X_n; n \geq 1\}$ and if $\mathsf{E}[J] < \infty$, then the sum $S_J = X_1 + X_2 + \cdots + X_J$ at the stopping trial J satisfies*

$$\mathsf{E}[S_J] = \overline{X}\mathsf{E}[J]. \tag{4.30}$$

Proof: Note that X_n is included in $S_J = \sum_{n=1}^{J} X_n$ whenever $n \leq J$, *i.e.*, whenever the indicator function $\mathbb{I}_{\{J \geq n\}} = 1$. Thus

$$S_J = \sum_{n=1}^{\infty} X_n \mathbb{I}_{\{J \geq n\}}\ . \tag{4.31}$$

This includes X_n as part of the sum if stopping has not occurred before trial n. The event $\{J \geq n\}$ is the complement of $\{J < n\} = \{J = 1\} \bigcup \cdots \bigcup \{J = n-1\}$. All of these latter events are determined by $X_1, \ldots, X_{n-1}$ and are thus independent of X_n. It follows that X_n and $\{J < n\}$ are independent and thus X_n and $\{J \geq n\}$ are also independent.[10] Thus

$$\mathsf{E}\left[X_n \mathbb{I}_{\{J \geq n\}}\right] = \overline{X}\mathsf{E}\left[\mathbb{I}_{\{J \geq n\}}\right].$$

We then have

$$\begin{aligned}
\mathsf{E}[S_J] &= \mathsf{E}\left[\sum_{n=1}^{\infty} X_n \mathbb{I}_{\{J \geq n\}}\right] \\
&= \sum_{n=1}^{\infty} \mathsf{E}\left[X_n \mathbb{I}_{\{J \geq n\}}\right] \qquad (4.32) \\
&= \sum_{n=1}^{\infty} \overline{X}\mathsf{E}\left[\mathbb{I}_{\{J \geq n\}}\right] \\
&= \overline{X}\mathsf{E}[J]. \qquad (4.33)
\end{aligned}$$

The interchange of expectation and infinite sum in (4.32) is obviously valid for a finite sum, and is shown in Exercise 4.18 to be valid for an infinite sum if $\mathsf{E}[J] < \infty$. The example below shows that Wald's equality can be invalid when $\mathsf{E}[J] = \infty$. The final step above comes from the observation that $\mathsf{E}\left[\mathbb{I}_{\{J \geq n\}}\right] = \Pr\{J \geq n\}$. Since J is a positive integer rv, $\mathsf{E}[J] = \sum_{n=1}^{\infty} \Pr\{J \geq n\}$. One can also obtain the last step by using $J = \sum_{n=1}^{\infty} \mathbb{I}_{\{J \geq n\}}$ (see Exercise 4.13). □

What this result essentially says in terms of gambling is that strategies for when to stop betting are not really effective as far as the mean is concerned. This sometimes appears obvious and sometimes appears very surprising, depending on the application.

Example 4.5.2 (Stop when you're ahead in coin tossing). We can model a (biased) coin tossing game as a sequence of IID rv's $X_1, X_2, \ldots$ where each X is 1 with probability p and -1 with probability $1-p$. Consider the possibly-defective stopping trial J where J is the first n for which $S_n = X_1 + \cdots + X_n = 1$, *i.e.*, the first trial at which the gambler is ahead.

We first want to see if J is a rv, *i.e.*, if the probability of eventual stopping, say $\theta = \Pr\{J < \infty\}$, is 1. We solve this by a frequently useful trick, but will use other more systematic approaches in Chapters 5 and 7 when we look at this same example as a birth-death Markov chain and then as a simple random walk. Note that $\Pr\{J = 1\} = p$, *i.e.*, $S_1 = 1$ with probability p and stopping occurs at trial 1. With probability $1-p$, $S_1 = -1$. Following $S_1 = -1$, the only way to become one ahead is to first return to $S_n = 0$ for some $n > 1$, and, after the first such return, go on to $S_m = 1$ at some later trial m. The probability of eventually going from -1 to 0 is the same as that of going from 0 to 1, *i.e.*, θ. Also, given a first return to 0 from -1, the probability of reaching 1 from 0 is θ. Thus,

$$\theta = p + (1-p)\theta^2.$$

[10] This can be quite confusing initially, since (as seen in the example of Figure 4.11) X_n is not necessarily independent of the event $\{J = n\}$, nor of $\{J = n+1\}$, etc. In other words, *given that* stopping has not occurred before trial n, then X_n can have a great deal to do with *the trial* at which stopping occurs. However, as shown above, X_n has nothing to do with *whether* $\{J < n\}$ or $\{J \geq n\}$.

This is a quadratic equation in θ with two solutions, $\theta = 1$ and $\theta = p/(1-p)$. For $p > 1/2$, the second solution is impossible since θ is a probability. Thus we conclude that J is a rv. For $p = 1/2$ (and this is the most interesting case), both solutions are the same, $\theta = 1$, and again J is a rv. For $p < 1/2$, the correct solution[11] is $\theta = p/(1-p)$. Thus $\theta < 1$ so J is a defective rv.

For the cases where $p \geq 1/2$, *i.e.*, where J is a rv, we can use the same trick to evaluate $\mathsf{E}[J]$,

$$\mathsf{E}[J] = p + (1-p)(1 + 2\mathsf{E}[J]).$$

The solution to this is

$$\mathsf{E}[J] \;=\; \frac{1}{2(1-p)} \;=\; \frac{1}{2p-1}.$$

We see that $\mathsf{E}[J]$ is finite for $p > 1/2$ and infinite for $p = 1/2$.

For $p > 1/2$, we can check that these results agree with Wald's equality. In particular, since S_J is 1 with probability 1, we also have $\mathsf{E}[S_J] = 1$. Since $\overline{X} = 2p-1$ and $\mathsf{E}[J] = 1/(2p-1)$, Wald's equality is satisfied (which of course it has to be).

For $p = 1/2$, we still have $S_J = 1$ with probability 1 and thus $\mathsf{E}[S_J] = 1$. However $\overline{X} = 0$ so $\overline{X}\mathsf{E}[J]$ has no meaning and Wald's equality breaks down. Thus we see that the restriction $\mathsf{E}[J] < \infty$ in Wald's equality is indeed needed. These results are tabulated below.

	$p > \frac{1}{2}$	$p = \frac{1}{2}$	$p < \frac{1}{2}$
$\Pr\{J < \infty\}$	$\frac{p}{1-p}$	1	1
$\mathsf{E}[J]$	∞	∞	$\frac{1}{2p-1}$

It is surprising that with $p = 1/2$, the gambler can eventually become one ahead with probability 1. This has little practical value, first because the required expected number of trials is infinite, and second (as will be seen later) because the gambler must risk a potentially infinite capital.

4.5.2 Applying Wald's equality to $m(t) = \mathsf{E}[N(t)]$

Let $\{S_n;\, n \geq 1\}$ be the arrival epochs and $\{X_n;\, n \geq 1\}$ the interarrival intervals for a renewal process. For any given $t > 0$, let J be the trial n for which S_n first exceeds t. Note that n is specified by the sample values of $\{X_1, \ldots, X_n\}$ and thus J is a possibly-defective stopping trial for $\{X_n;\, n \geq 1\}$.

Since n is the first trial for which $S_n > t$, we see that $S_{n-1} \leq t$ and $S_n > t$. Thus $N(t)$ is $n-1$ and n is the sample value of $N(t)+1$. Since this is true for all sample sequences, $J = N(t) + 1$. Since $N(t)$ is a non-defective rv, J is also, so J is a stopping trial for $\{X_n;\, n \geq 1\}$.

[11] This will be shown when we view this example as a birth-death Markov chain in Chapter 5.

We can then employ Wald's equality to obtain

$$\mathsf{E}\left[S_{N(t)+1}\right] = \overline{X}\mathsf{E}\left[N(t)+1\right] = \overline{X}[m(t)+1]. \tag{4.34}$$

$$m(t) = \frac{\mathsf{E}\left[S_{N(t)+1}\right]}{\overline{X}} - 1. \tag{4.35}$$

As is often the case with Wald's equality, this provides a relationship between two quantities, $m(t)$ and $\mathsf{E}\left[S_{N(t)+1}\right]$, that are both unknown. This will be used in proving the elementary renewal theorem by upper and lower bounding $\mathsf{E}\left[S_{N(t)+1}\right]$. The lower bound is easy, since $\mathsf{E}\left[S_{N(t)+1}\right] \geq t$, and thus $m(t) \geq t/\overline{X} - 1$. It follows that

$$\frac{m(t)}{t} \geq \frac{1}{\overline{X}} - \frac{1}{t}. \tag{4.36}$$

We derive an upper bound on $\mathsf{E}\left[S_{N(t)+1}\right]$ in the next section. First, however, as a sanity check, consider Figure 4.12 which illustrates (4.35) for the case where each X_n is a deterministic rv where $X_n = \overline{X}$ with probability 1.

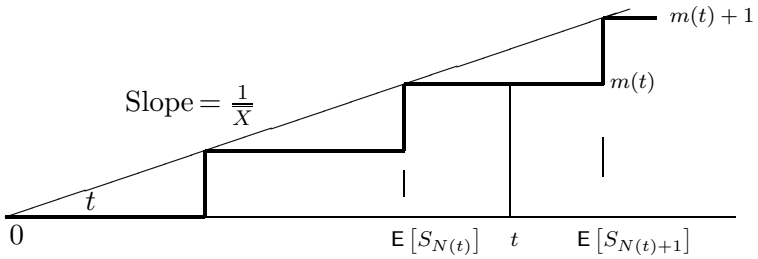


Figure 4.12: Illustration of (4.34) for the special case where X is deterministic. Note that $m(t)$, as a function of t, is then the illustrated staircase function. On each increment of t by $\overline{X}$, $m(t)$ increases by one. Then $m(t)+1$ and $\mathsf{E}\left[S_{N(t)+1}\right]$ are two sides of a right triangle of slope $1/\overline{X}$, yielding (4.34).

It might be puzzling why we used $N(t)+1$ rather than $N(t)$ as a stopping trial for the epochs $\{S_i;\, i \geq 1\}$ in this application of Wald's equality. To understand this, assume for example that $N(t) = n$. When an observer sees the sample values of $S_1, \ldots, S_n$, with $S_n < t$, the observer typically cannot tell (on the basis of $S_1, \ldots, S_n$ alone) whether any other arrivals will occur in the interval $(S_n, t]$. In other words, $N(t) = n$ implies that $S_n \leq t$, but $S_n < t$ does not imply that $N(t) = n$. On the other hand, still assuming $N(t) = n$, an observer seeing $S_1, \ldots, S_{n+1}$ knows that $N(t) = n$.

Any stopping trial (for an arbitrary sequence of rv's $\{S_n;\, n \geq 1\}$) can be viewed as an experiment where an observer views sample values $s_1, s_2, \ldots$, in turn until the stopping rule is satisfied. The stopping rule does not permit either looking ahead or going back to change an earlier decision. Thus the rule: stop at $N(t)+1$ (for the renewal epochs $\{S_n;\, n \geq 1\}$) means stop at the first sample value s_i that exceeds t. Stopping at the final sample value $s_n \leq t$ is not necessarily possible without looking ahead to the following sample value.

4.5.3 Stopping trials, embedded renewals, and G/G/1 queues

The above definition of a stopping trial is quite restrictive in that it refers only to a single sequence of rv's. In many queueing situations, for example, there is both a sequence of interarrival times $\{X_i;\, i \geq 1\}$ and a sequence of service times $\{V_i;\, i \geq 0\}$. Here X_i is the interarrival interval between customer $i-1$ and i, where an initial customer 0 is assumed to arrive at time 0, and X_1 is the arrival time for customer 1. The service time of customer 0 is then V_0 and each V_i, $i > 0$ is the service time of the corresponding ordinary customer. Customer number 0 is not ordinary in the sense that it arrives at the fixed time 0 and is not counted in the arrival counting process $\{N(t);\, t > 0\}$.

Example 4.5.3 (G/G/1 queues:). Consider a G/G/1 queue (the single server case of the G/G/m queue described in Example 4.1.2). We assume that the customers are served in First-Come-First-Served (FCFS) order.[12] Both the interarrival intervals $\{X_i;\, i \geq 1\}$ and the service times $\{V_i;\, i \geq 0\}$ are assumed to be IID and the service times are assumed to be independent of the interarrival intervals. Figure 4.13 illustrates a sample path for these arrivals and departures.

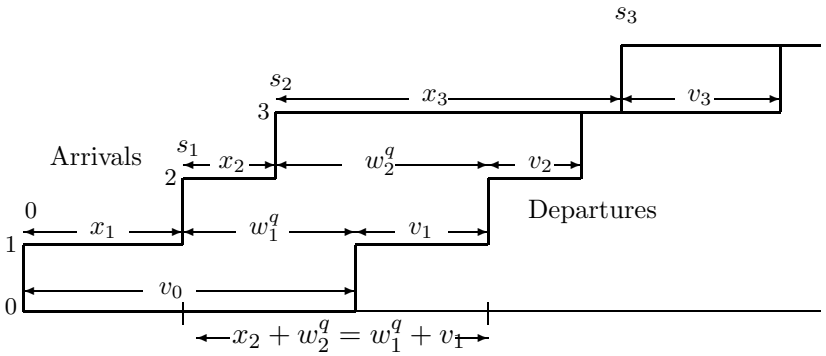


Figure 4.13: Sample path of arrivals and departures from a G/G/1 queue. Customer 0 arrives at time 0 and enters service immediately. Customer 1 arrives at time $s_1 = x_1$. For the case shown above, customer 0 has not yet departed, *i.e.*, $x_1 < v_0$, so customer 1 is queued for the interval $w_1^q = v_0 - x_1$ before entering service. As illustrated, customer 1's system time (queueing time plus service time) is $w_1 = w_1^q + v_1$. Note that the sample path of arrivals in the figure is one plus the sample path of the arrival counting process $\{N(t);\, t > 0\}$, since the counting process, by convention, does not count the initial arrival at time 0.

Customer 2 arrives at $s_2 = x_1 + x_2$. For the case shown above, this is before customer 1 departs at $v_0 + v_1$. Thus, customer 2's wait in queue is $w_2^q = v_0 + v_1 - x_1 - x_2$. As illustrated above, $x_2 + w_2^q$ is also equal to customer 1's system time, so $w_2^q = w_1^q + v_1 - x_2$. Customer 3 arrives when the system is empty, so it enters service immediately with no wait in queue, *i.e.*, $w_3^q = 0$.

The figure illustrates a sample path for which $X_1 < V_0$, so arrival number 1 waits in queue for $W_1^q = V_0 - X_1$. If $X_1 \geq V_0$, on the other hand, then customer one enters service immediately, *i.e.*, customer one 'sees an empty system.' In general, then $W_1^q = \max(V_0 - X_1,\, 0)$. In the same way, as illustrated in the figure, if $W_1^q > 0$, then customer 2

[12] For single server queues, this is sometimes referred to as First-In-First-Out (FIFO) service.

waits for $W_1^q + V_1 - X_2$ if positive and 0 otherwise. This same formula works if $W_1^q = 0$, so $W_2^q = \max(W_1^q + V_1 - X_2,\ 0)$. In general, it can be seen that

$$W_i^q = \max(W_{i-1}^q + V_{i-1} - X_i,\ 0). \tag{4.37}$$

This equation will be analyzed further in Section 7.2 where we are interested in queueing delay and system delay. Here our objectives are simpler, since we only want to show that the subsequence of customer arrivals i for which the event $\{W_i^q = 0\}$ is satisfied form the renewal epochs of a renewal process. To do this, first observe from (4.37) (using induction if desired) that W_i^q is a function of $(X_1, \ldots, X_i)$ and $(V_0, \ldots, V_{i-1})$. Thus, if we let J be the smallest $i > 0$ for which $W_i^q = 0$, then $\mathbb{I}_{J=i}$ is a function of $(X_1, \ldots, X_i)$ and $(V_0, \ldots, V_{i-1})$.

We now interrupt the discussion of G/G/1 queues with the following generalization of the definition of a stopping trial.

Definition 4.5.2 (Generalized stopping trials). *A generalized stopping trial J for a sequence of pairs of rv's $(X_1, V_1), (X_2, V_2) \ldots$, is a positive integer-valued rv such that, for each $n \geq 1$, the indicator rv $\mathbb{I}_{\{J=n\}}$ is a function of $X_1, V_1, X_2, V_2, \ldots, X_n, V_n$.*

Wald's equality can be trivially generalized for these generalized stopping trials.

Theorem 4.5.2 (Generalized Wald's equality). *Let $\{(X_n, V_n);\ n \geq 1\}$ be a sequence of pairs of rv's, where each pair is independent and identically distributed (IID) to all other pairs. Assume that each X_i has finite mean $\overline{X}$. If J is a stopping trial for $\{(X_n, V_n);\ n \geq 1\}$ and if $\mathsf{E}[J] < \infty$, then the sum $S_J = X_1 + X_2 + \cdots + X_J$ satisfies*

$$\mathsf{E}[S_J] = \overline{X}\mathsf{E}[J]. \tag{4.38}$$

The proof of this will be omitted, since it is the same as the proof of Theorem 4.5.1. In fact, the definition of stopping trials could be further generalized by replacing the rv's V_i by vector rv's or by a random number of rv's , and Wald's equality would still hold.[13]

For the example of the G/G/1 queue, we take the sequence of pairs to be $\{(X_1, V_0), (X_2, V_1), \ldots, \}$. Then $\{(X_n, V_{n-1};\ n \geq 1\}$ satisfies the conditions of Theorem 4.5.2 (assuming that $\mathsf{E}[X_i] < \infty$). Let J be the generalized stopping rule specifying the number of the first arrival to find an empty queue. Then the theorem relates $\mathsf{E}[S_J]$, the expected time $t > 0$ until the first arrival to see an empty queue, and $\mathsf{E}[J]$, the expected number of arrivals until seeing an empty queue.

It is important here, as in many applications, to avoid the confusion created by viewing J as a stopping *time*. We have seen that J is the *number* of the first customer to see an empty queue, and S_J is the *time* until that customer arrives.

There is a further possible timing confusion about whether a customer's service time is determined when the customer arrives or when it completes service. This makes no difference, since the ordered sequence of pairs is well-defined and satisfies the appropriate IID condition for using the Wald equality.

[13] In fact, J is sometimes defined to be a stopping rule if $\mathbb{I}_{\{J \geq n\}}$ is independent of $X_n, X_{n+1}, \ldots$ for each n. This makes it easy to prove Wald's equality, but quite hard to see when the definition holds, especially since $\mathbb{I}_{\{J=n\}}$, for example, is typically dependent on X_n (see footnote 7).

As is often the case with Wald's equality, it is not obvious how to compute either quantity in (4.38), but it is nice to know that they are so simply related. It is also interesting to see that, although successive pairs (X_i, V_i) are assumed independent, it is not necessary for X_i and V_i to be independent. This lack of independence does not occur for the G/G/1 (or G/G/m) queue, but can be useful in situations such as packet networks where the interarrival time between two packets at a given node can depend on the service time (the length) of the first packet if both packets are coming from the same node.

Perhaps a more important aspect of viewing the first renewal for the G/G/1 queue as a stopping trial is the ability to show that successive renewals are in fact IID. Let $X_{2,1}, X_{2,2}, \ldots$ be the interarrival times following J, the first arrival to see an empty queue. Conditioning on $J = j$, we have $X_{2,1}{=}X_{j+1},\ X_{2,2}{=}X_{j+2}, \ldots,$. Thus $\{X_{2,k};\ k \geq 1\}$ is an IID sequence with the original interarrival distribution. Similarly $\{(X_{2,k}, V_{2,k});\ k \geq 1\}$ is a sequence of IID pairs with the original distribution. This is valid for all sample values j of the stopping trial J. Thus $\{(X_{2,k}, V_{2,k});\ k \geq 1\}$ is statistically independent of J and $(X_i, V_i);\ 1 \leq i \leq J$.

The argument above can be repeated for subsequent arrivals to an empty system, so we have shown that successive arrivals to an empty system actually form a renewal process.[14]

One can define many different stopping rules for queues, such as the first trial at which a given number of customers are in the queue. Wald's equality can be applied to any such stopping rule, but much more is required for the stopping trial to also form a renewal point. At the first time when n customers are in the system, the subsequent departure times depend partly on the old service times and partly on the new arrival and service times, so the required independence for a renewal point does not exist. Stopping rules are helpful in understanding embedded renewal points, but are by no means equivalent to embedded renewal points.

Finally, nothing in the argument above for the G/G/1 queue made any use of the FCFS service discipline. One can use any service discipline for which the choice of which customer to serve at a given time t is based solely on the arrival and service times of customers in the system by time t. In fact, if the server is never idle when customers are in the system, the renewal epochs will not depend on the service descipline. It is also possible to extend these arguments to the G/G/m queue, although the service discipline can affect the renewal points in this case.

4.5.4 Little's theorem

Little's theorem is an important queueing result stating that the expected number of customers in a queueing system is equal to the product of the arrival rate and the expected time each customer waits in the system. This result is true under very general conditions; we use the G/G/1 queue with FCFS service as a specific example, but the reason for the greater generality will be clear as we proceed. Note that the theorem does not tell us how

[14]Confession by author: For about 15 years, I mistakenly believed that it was obvious that arrivals to an empty system in a G/G/m queue form a renewal process. Thus I can not expect readers to be excited about the above proof. However, it is a nice example of how to use stopping times to see otherwise murky points clearly.

to find either the expected number or expected wait; it only says that if one can be found, the other can also be found.

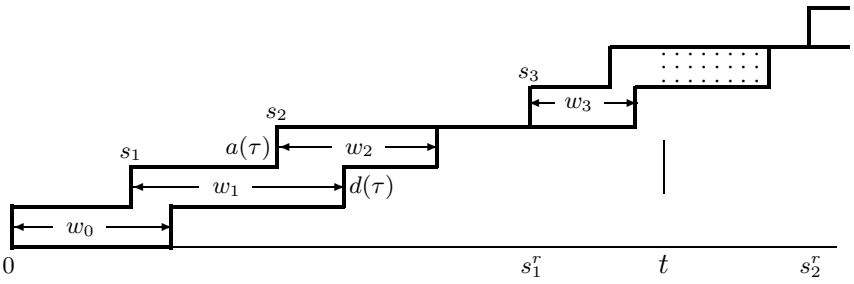


Figure 4.14: Sample path of arrivals, departures, and system waiting times for a G/G/1 queue with FCFS service. The upper step function is the number of customer arrivals, including the customer at time 0 and is denoted $a(\tau)$. Thus $a(\tau)$ is a sample path of $A(\tau) = N(\tau) + 1$, *i.e.*, the arrival counting process incremented by 1 for the initial arrival at $\tau = 0$. The lower step function, $d(\tau)$ is a sample path for $D(\tau)$, which is the number of departures (including customer 0) up to time τ. For each $i \geq 0$, w_i is the sample value of the system waiting time W_i for customer i. Note that $W_i = W_i^q + V_i$.

The figure also shows the sample values s_1^r and s_2^r of the first two arrivals that see an empty system (recall from Section 4.5.3 that the subsequence of arrivals that see an empty system forms a renewal process.)

Figure 4.14 illustrates a sample path for a G/G/1 queue with FCFS service. It illustrates a sample path $a(t)$ for the arrival process $A(t) = N(t) + 1$, *i.e.*, the number of customer arrivals in $[0, t]$, specifically including customer number 0 arriving at $t = 0$. Similarly, it illustrates the departure process $D(t)$, which is the number of departures up to time t, again including customer 0. The difference, $L(t) = A(t) - D(t)$, is then the number in the system at time t.

Recall from Section 4.5.3 that the subsequence of customer arrivals for $t > 0$ that see an empty system form a renewal process. Actually, we showed a little more than that. Not only are the inter-renewal intervals, $X_i^r = S_i^r - S_{i-1}^r$ IID, but the number of customer arrivals in each inter-renewal interval are IID, and the interarrival intervals and service times between inter-renewal intervals are IID. The sample values, s_1^r and s_2^r of the first two renewal epochs are shown in the figure.

The essence of Little's theorem can be seen by observing that $\int_0^{S_1^r} L(\tau)\, d\tau$ in the figure is the area between the upper and lower step functions, integrated out to the first time that the two step functions become equal (*i.e.*, the system becomes empty). For the sample value in the figure, this integral is equal to $w_0 + w_1 + w_2$. In terms of the rv's,

$$\int_0^{S_1^r} L(\tau)\, d\tau = \sum_{i=0}^{N(S_1^r)-1} W_i. \tag{4.39}$$

The same relationship exists in each inter-renewal interval, and in particular we can define

L_n for each $n \geq 1$ as

$$L_n = \int_{S^r_{n-1}}^{S^r_n} L(\tau)\, d\tau \;= \sum_{i=N(S^r_{n-1})}^{N(S^r_n)-1} W_i. \tag{4.40}$$

The interpretation of this is far simpler than the notation. The arrival step function and the departure step function in Figure 4.14 are separated whenever there are customers in the system (the system is busy) and are equal whenever the system is empty. Renewals occur when the system goes from empty to busy, so the nth renewal is at the beginning of the nth busy period. Then L_n is the area of the region between the two step functions over the nth busy period. By simple geometry, this area is also the sum of the customer waiting times over that busy period. Finally, since the interarrival intervals and service times in each busy period are IID with respect to those in each other busy period, the sequence $L_1, L_2, \ldots$, is a sequence of IID rv's.

The function $L(\tau)$ has the same behavior as a renewal reward function, but it is slightly more general, being a function of more than the age and duration of the renewal counting process $\{N^r(t);\, t > 0\}$ at $t = \tau$. However the fact that $\{L_n;\, n \geq 1\}$ is an IID sequence lets us use the same methodology to treat $L(\tau)$ as was used earlier to treat renewal-reward functions. We now state and prove Little's theorem. The proof is almost the same as that of Theorem 4.4.1, so we will not dwell on it.

Theorem 4.5.3 (Little). *For a FCFS G/G/1 queue in which the expected inter-renewal interval is finite, the limiting time-average number of customers in the system is equal, with probability* 1*, to a constant denoted as* $\overline{L}$*. The sample-path-average waiting time per customer is also equal, with probability 1, to a constant denoted as* $\overline{W}$*. Finally* $\overline{L} = \lambda \overline{W}$ *where* λ *is the customer arrival rate,* i.e.*, the reciprocal of the expected interarrival time.*

Proof: Note that for any $t > 0$, $\int_0^t (L(\tau)\, d\tau$ can be expressed as the sum over the busy periods completed before t plus a residual term involving the busy period including t. The residual term can be upper bounded by the integral over that complete busy period. Using this with (4.40), we have

$$\sum_{n=1}^{N^r(t)} \mathrm{L}_n \leq \int_{\tau=0}^{t} L(\tau)\, d\tau \leq \sum_{i=0}^{N(t)} W_i \leq \sum_{n=1}^{N^r(t)+1} \mathrm{L}_n. \tag{4.41}$$

Assuming that the expected inter-renewal interval, $\mathsf{E}[X^r]$, is finite, we can divide both sides of (4.41) by t and go to the limit $t \to \infty$. From the same argument as in Theorem 4.4.1,

$$\lim_{t\to\infty} \frac{\sum_{i=0}^{N(t)} W_i}{t} = \lim_{t\to\infty} \frac{\int_{\tau=0}^{t} L(\tau)\, d\tau}{t} = \frac{\mathsf{E}[\mathrm{L}_n]}{\mathsf{E}[X^r]} \qquad \text{with probability 1.} \tag{4.42}$$

The equality on the right shows that the limiting time average of $L(\tau)$ exists with probability 1 and is equal to $\overline{L} = \mathsf{E}[L_n]\,/\mathsf{E}[X^r]$. The quantity on the left of (4.42) can now be broken up as waiting time per customer multiplied by number of customers per unit time, i.e.,

$$\lim_{t\to\infty} \frac{\sum_{i=0}^{N(t)} W_i}{t} = \lim_{t\to\infty} \frac{\sum_{i=0}^{N(t)} W_i}{N(t)} \lim_{t\to\infty} \frac{N(t)}{t}. \tag{4.43}$$

From (4.42), the limit on the left side of (4.43) exists (and equals $\overline{L}$) with probability 1. The second limit on the right also exists with probability 1 by the strong law for renewal processes, applied to $\{N(t); t > 0\}$. This limit is called the *arrival rate* λ, and is equal to the reciprocal of the mean interarrival interval for $\{N(t)\}$. Since these two limits exist with probability 1, the first limit on the right, which is the sample-path-average waiting time per customer, denoted $\overline{W}$, also exists with probability 1. □

Reviewing this proof and the development of the G/G/1 queue before the theorem, we see that there was a simple idea, expressed by (4.39), combined with a lot of notational complexity due to the fact that we were dealing with both an arrival counting process $\{N(t); t > 0\}$ and an embedded renewal counting process $\{N^r(t); t > 0\}$. The difficult thing, mathematically, was showing that $\{N^r(t); t > 0\}$ is actually a renewal process and showing that the L_n are IID, and this was where we needed to understand stopping rules.

Recall that we assumed earlier that customers departed from the queue in the same order in which they arrived. From Figure 4.15, however, it is clear that FCFS order is not required for the argument. Thus the theorem generalizes to systems with multiple servers and arbitrary service disciplines in which customers do not follow FCFS order. In fact, all that the argument requires is that the system has renewals (which are IID by definition of a renewal) and that the inter-renewal interval is finite with probability 1.

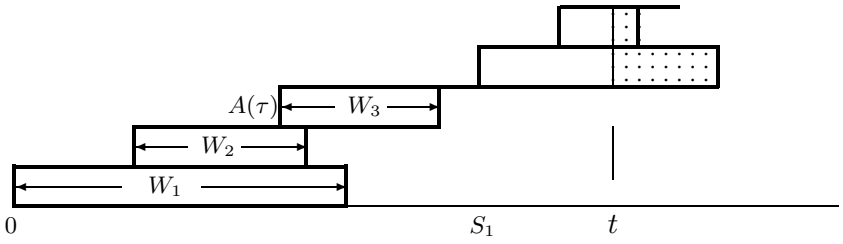


Figure 4.15: Arrivals and departures in a non-FCFS systems. The server, for example, could work simultaneously (at a reduced rate) on all customers in the system, and thus complete service for customers with small service needs before completing earlier arrivals with greater service needs. Note that the jagged right edge of the diagram does not represent number of deparatures, but this is not essential for the argument.

For example, if higher priority is given to customers with small service times, then it is not hard to see that the average number of customers in the system and the average waiting time per customer will be decreased. However, if the server is always busy when there is work to be done, it can be seen that the renewal times are unaffected. Service discipines will be discussed further in Section 5.6.

The same argument as in Little's theorem can be used to relate the average number of customers in a single server queue (not counting service) to the average wait in the queue (not counting service). Renewals still occur on arrivals to an empty system, and the integral of customers in queue over a busy period is still equal to the sum of the queue waiting times. Let $L^q(t)$ be the number in the queue at time t and let $\overline{L}^q = \lim_{t\to\infty}(1/t)\int_0^t L^q(\tau)d\tau$ be the

time-average queue wait. Letting $\overline{W}^q$ be the sample-path-average waiting time in queue,

$$\overline{L}^q = \lambda \overline{W}^q. \tag{4.44}$$

The same argument can also be applied to the service facility of a single server queue. The time-average of the number of customers in the server is just the fraction of time that the server is busy. Denoting this fraction by ρ and the expected service time by $\overline{V}$, we get

$$\rho = \lambda \overline{V}. \tag{4.45}$$

4.5.5 Expected queueing time for an M/G/1 queue

For our last example of the use of renewal-reward processes, we consider the expected queueing time in an M/G/1 queue. We again assume that an arrival to an empty system occurs at time 0 and renewals occur on subsequent arrivals to an empty system. At any given time t, let $L^q(t)$ be the number of customers in the queue (not counting the customer in service, if any) and let $R(t)$ be the residual life of the customer in service. If no customer is in service, $R(t) = 0$, and otherwise $R(t)$ is the remaining time until the current service is completed. Let $U(t)$ be the waiting time in queue that would be experienced by a customer arriving at time t. This is often called the unfinished work in the queueing literature and represents the delay until all the customers currently in the system complete service. Thus the rv $U(t)$ is equal to $R(t)$, the residual life of the customer in service, plus the service times of each of the $L^q(t)$ customers currently waiting in the queue.

$$U(t) = \sum_{i=1}^{L^q(t)} V_{N(t)-i} + R(t), \tag{4.46}$$

where $N(t) - i$ is the customer number of the ith customer in the queue at time t. Since $L^q(t)$ is a function only of the interarrival times in $(0, t)$ and the service times of the customers that have already been served, we see that for each sample value $L^q(t) = \ell$, the rv's $\widetilde{V}_1, \ldots, \widetilde{V}_\ell$ each have the service time distribution F_V. Thus, taking expected values,

$$\mathsf{E}\left[U(t)\right] = \mathsf{E}\left[L^q(t)\right]\mathsf{E}\left[V\right] + \mathsf{E}\left[R(t)\right]. \tag{4.47}$$

Figure 4.16 illustrates how to find the time-average of $R(t)$. Viewing $R(t)$ as a reward

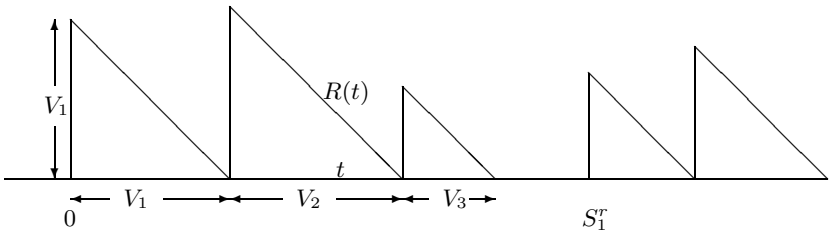


Figure 4.16: Sample value of the residual life function of customers in service.

function, we can find the accumulated reward up to time t as the sum of triangular areas.

First, consider $\int R(\tau)d\tau$ from 0 to $S^r_{N^r(t)}$, i.e., the accumulated reward up to the final renewal epoch in $[o,t]t$. Note that $S^r_{N^r(t)}$ is not only a renewal epoch for the renewal process, but also an arrival epoch for the arrival process; in particular, it is the $N(S^r_{N^r(t)})$th arrival epoch, and the $N(S^r_{N^r(t)})-1$ earlier arrivals are the customers that have received service up to time $S^r_{N^r(t)}$. Thus,

$$\int_0^{S^r_{N^r(t)}} R(\tau)\,d\tau \;=\; \sum_{i=0}^{N(S^r_{N^r(t)})-1} \frac{V_i^2}{2} \;\le\; \sum_{i=0}^{N(t)} \frac{V_i^2}{2}.$$

We can similarly upper bound the term on the right above by $\int_0^{S^r_{N^r(t)+1}} R(\tau)\,d\tau$. We also know (from going through virtually the same argument several times) that $(1/t)\int_{\tau=0}^{t} R(\tau)d\tau$ will approach a limit[15] with probability 1 as $t\to\infty$, and that the limit will be unchanged if t is replaced with $S^r_{N^r(t)}$ or $S^r_{N^r(t)+1}$. Thus, taking λ as the arrival rate,

$$\lim_{t\to\infty} \frac{\int_0^t R(\tau)\,d\tau}{t} \;=\; \lim_{t\to\infty} \frac{\sum_{i=1}^{A(t)} V_i^2}{2A(t)}\,\frac{A(t)}{t} \;=\; \frac{\lambda \mathsf{E}\left[V^2\right]}{2} \qquad \text{WP1.}$$

We will see in the next section that the time average above can be replaced with a limiting ensemble-average, so that

$$\lim_{t\to\infty} \mathsf{E}\left[R(t)\right] = \frac{\lambda \mathsf{E}\left[V^2\right]}{2}. \tag{4.48}$$

The next section also shows that there is a limiting ensemble-average form of (4.44), showing that $\lim_{t\to\infty}\mathsf{E}\left[L^q(t)\right]=\lambda\overline{W}^q$. Substituting this plus (4.48) into (4.47), we get

$$\lim_{t\to\infty} \mathsf{E}\left[U(t)\right] = \lambda \mathsf{E}\left[V\right]\,\overline{W}^q + \frac{\lambda \mathsf{E}\left[V^2\right]}{2}. \tag{4.49}$$

Thus $\mathsf{E}\left[U(t)\right]$ is asymptotically independent of t. It is now important to distinguish between $\mathsf{E}\left[U(t)\right]$ and $\overline{W}^q$. The first is the expected unfinished work at time t, which is the queueing delay that a customer would incur by arriving at t; the second is the sample-path-average expected queueing delay. For Poisson arrivals, the probability of an arrival in $(t,t+\delta]$ is independent of all earlier arrivals and service times, so it is independent of $U(t)$[16]. Thus, in the limit $t\to\infty$, each arrival faces an expected delay $\lim_{t\to\infty}\mathsf{E}\left[U(t)\right]$, so $\lim_{t\to\infty}\mathsf{E}\left[U(t)\right]$ must be equal to $\overline{W}^q$. Substituting this into (4.49), we obtain the celebrated *Pollaczek-Khinchin* formula,

$$\overline{W}^q = \frac{\lambda \mathsf{E}\left[V^2\right]}{2(1-\lambda \mathsf{E}\left[V\right])}. \tag{4.50}$$

[15] In fact, one could simply take the limit without bringing in the renewal process, since it is clear by now that the renewal process justifies the limit with probability 1.

[16] This is often called the *PASTA* property, standing for Poisson arrivals see time-averages. This holds with great generality, requiring only that time-averages exist and that the parameters of interest at a given time t are independent of future arrivals. At the same time, this property is somewhat vague, so it should be used to help the intuition rather than to prove theorems.

This queueing delay has some of the peculiar features of residual life, and in particular, if $\mathsf{E}\left[V^2\right] = \infty$, the limiting expected queueing delay is infinite even when the expected service time is less than the expected interarrival interval.

In trying to visualize why the queueing delay is so large when $\mathsf{E}\left[V^2\right]$ is large, note that while a particularly long service is taking place, numerous arrivals are coming into the system, and all are being delayed by this single long service. In other words, the number of new customers held up by a long service is proportional to the length of the service, and the amount each of them are held up is also proportional to the length of the service. This visualization is rather crude, but does serve to explain the second moment of V in (4.50). This phenomenon is sometimes called the "slow truck effect" because of the pile up of cars behind a slow truck on a single lane road.

For a G/G/1 queue, (4.49) is still valid, but arrival times are no longer independent of $U(t)$, so that typically $\mathsf{E}\left[U(t)\right] \neq \overline{W}^q$. As an example, suppose that the service time is uniformly distributed between $1-\epsilon$ and $1+\epsilon$ and that the interarrival interval is uniformly distributed between $2-\epsilon$ and $2+\epsilon$. Assuming that $\epsilon < 1/2$, the system has no queueing and $\overline{W}^q = 0$. On the other hand, for small ϵ, $\lim_{t\to\infty} \mathsf{E}\left[U(t)\right] \sim 1/4$ (i.e., the server is busy half the time with unfinished work ranging from 0 to 1).

4.6 Expected number of renewals

The purpose of this section is to evaluate $\mathsf{E}\left[N(t)\right]$, denoted $m(t)$, as a function of $t > 0$ for arbitrary renewal processes. We first find an exact expression, in the form of an integral equation, for $m(t)$. This can be easily solved by Laplace transform methods in special cases. For the general case, however, $m(t)$ becomes increasingly messy for large t, so we then find the asymptotic behavior of $m(t)$. Since $N(t)/t$ approaches $1/\overline{X}$ with probability 1, we might expect $m(t)$ to grow with a derivative $m'(t)$ that asymptotically approaches $1/\overline{X}$. This is not true in general. Two somewhat weaker results, however, are true. The first, called the elementary renewal theorem (Theorem 4.6.1), states that $\lim_{t\to\infty} m(t)/t = 1/\overline{X}$. The second result, called Blackwell's theorem (Theorem 4.6.2), states that, subject to some limitations on $\delta > 0$, $\lim_{t\to\infty}[m(t+\delta)-m(t)] = \delta/\overline{X}$. This says essentially that the expected renewal rate approaches steady state as $t \to \infty$. We will find a number of applications of Blackwell's theorem throughout the remainder of the text.

The exact calculation of $m(t)$ makes use of the fact that the expectation of a nonnegative random variable is defined as the integral of its complementary distribution function,

$$m(t) = \mathsf{E}\left[N(t)\right] = \sum_{n=1}^{\infty} \Pr\{N(t) \geq n\}\,.$$

Since the event $\{N(t) \geq n\}$ is the same as $\{S_n \leq t\}$, $m(t)$ is expressed in terms of the distribution functions of S_n, $n \geq 1$, as follows.

$$m(t) = \sum_{n=1}^{\infty} \Pr\{S_n \leq t\}\,. \tag{4.51}$$

Although this expression looks fairly simple, it becomes increasingly complex with increasing t. As t increases, there is an increasing set of values of n for which $\Pr\{S_n \leq t\}$ is significant, and $\Pr\{S_n \leq t\}$ itself is not that easy to calculate if the interarrival distribution $\mathsf{F}_X(x)$ is complicated. The main utility of (4.51) comes from the fact that it leads to an integral equation for $m(t)$. Since $S_n = S_{n-1} + X_n$ for each $n \geq 1$ (interpreting S_0 as 0), and since X_n and S_{n-1} are independent, we can use the convolution equation (1.11) to get

$$\Pr\{S_n \leq t\} = \int_{x=0}^{t} \Pr\{S_{n-1} \leq t-x\}\; d\mathsf{F}_X(x) \quad \text{for } n \geq 2.$$

Substituting this in (4.51) for $n \geq 2$ and using the fact that $\Pr\{S_1 \leq t\} = \mathsf{F}_X(t)$, we can interchange the order of integration and summation to get

$$\begin{aligned} m(t) &= \mathsf{F}_X(t) + \int_{x=0}^{t} \sum_{n=2}^{\infty} \Pr\{S_{n-1} \leq t-x\}\; d\mathsf{F}_X(x) \\ &= \mathsf{F}_X(t) + \int_{x=0}^{t} \sum_{n=1}^{\infty} \Pr\{S_n \leq t-x\}\; d\mathsf{F}_X(x) \\ &= \mathsf{F}_X(t) + \int_{x=0}^{t} m(t-x) d\mathsf{F}_X(x)\,; \qquad t \geq 0. \end{aligned} \tag{4.52}$$

An alternative derivation is given in Exercise 4.22. This integral equation is called the *renewal equation*. The following alternative form is achieved by integration by parts.[17]

$$m(t) = \mathsf{F}_X(t) + \int_{\tau=0}^{t} F_X(t-\tau)\, dm(\tau)\,; \qquad t \geq 0. \tag{4.53}$$

4.6.1 Laplace transform approach

If we assume that $X \geq 0$ has a density $\mathsf{f}_X(x)$, and that this density has a Laplace transform[18] $L_X(s) = \int_0^\infty \mathsf{f}_X(x) e^{-sx} dx$, then we can take the Laplace transform of both sides of (4.52). Note that the final term in (4.52) is the convolution of m with f_X, so that the Laplace transform of $m(t)$ satisfies

$$L_m(s) = \frac{L_X(s)}{s} + L_m(s) L_X(s).$$

Solving for $L_m(s)$,

$$L_m(s) = \frac{L_X(s)}{s[1 - L_X(s)]}. \tag{4.54}$$

[17] A mathematical subtlety with the Stieltjes integrals (4.52) and (4.53) will be discussed in Section 4.7.3.

[18] Note that $L_X(s) = \mathsf{E}\left[e^{-sX}\right] = g_X(-s)$ where g is the MGF of X. Thus the argument here could be carried out using the MGF. We use the Laplace transform since the mechanics here are so familiar to most engineering students

Example 4.6.1. As a simple example of how this can be used to calculate $m(t)$, suppose $\mathsf{f}_X(x) = (1/2)e^{-x} + e^{-2x}$ for $x \geq 0$. The Laplace transform is given by

$$L_X(s) = \frac{1}{2(s+1)} + \frac{1}{s+2} = \frac{(3/2)s+2}{(s+1)(s+2)}.$$

Substituting this into (4.54) yields

$$L_m(s) = \frac{(3/2)s+2}{s^2(s+3/2)} = \frac{4}{3s^2} + \frac{1}{9s} - \frac{1}{9(s+3/2)}.$$

We can solve for $m(t), t \geq 0$, by taking the inverse Laplace transform,

$$m(t) = \frac{4t}{3} + \frac{1-\exp[-(3/2)t]}{9}.$$

The procedure in this example can be used for any inter-renewal density $\mathsf{f}_X(x)$ for which the Laplace transform is a rational function, i.e., a ratio of polynomials. In such cases, $L_m(s)$ will also be a rational function. The Heaviside inversion formula (i.e., factoring the denominator and expressing $L_m(s)$ as a sum of individual poles as done above) can then be used to calculate $m(t)$. In the example above, there was a second order pole at $s = 0$ leading to the linear term $4t/3$ in $m(t)$, there was a first order pole at $s = 0$ leading to the constant $1/9$, and there was a pole at $s = -3/2$ leading to the exponentially decaying term.

We now show that a second order pole at $s = 0$ always occurs when $L_X(s)$ is a rational function. To see this, note that $L_X(0)$ is just the integral of $\mathsf{f}_X(x)$, which is 1; thus $1-L_X(s)$ has a zero at $s = 0$ and $L_m(s)$ has a second order pole at $s = 0$. To evaluate the residue for this second order pole, we recall that the first and second derivatives of $L_X(s)$ at $s = 0$ are $-\mathsf{E}[X]$ and $\mathsf{E}\left[X^2\right]$ respectively. Expanding $L_X(s)$ in a power series around $s = 0$ then yields $L_X(s) = 1 - s\mathsf{E}[X] + (s^2/2)\mathsf{E}\left[X^2\right]$ plus terms of order s^3 or higher. This gives us

$$L_m(s) = \frac{1 - s\overline{X} + (s^2/2)\mathsf{E}\left[X^2\right] + \cdots}{s^2\left[\overline{X} - (s/2)\mathsf{E}[X^2] + \cdots\right]} = \frac{1}{s^2\overline{X}} + \frac{1}{s}\left(\frac{\mathsf{E}\left[X^2\right]}{2\overline{X}^2} - 1\right) + \cdots. \tag{4.55}$$

The remaining terms are the other poles of $L_m(s)$ with their residues. For values of s with $\Re(s) \geq 0$, we have $|L_X(s)| = |\int \mathsf{f}_X(x)e^{-sx}dx| \leq \int \mathsf{f}_X(x)|e^{-sx}|dx \leq \int \mathsf{f}_X(x)dx = 1$ with strict inequality except for $s = 0$. Thus $L_X(s)$ cannot have any poles on the imaginary axis or the right half plane, and $1 - L_X(s)$ cannot have any zeros there other than the one at $s = 0$. It follows that all the remaining poles of $L_m(s)$ are strictly in the left half plane. This means that the inverse transforms for all these remaining poles die out as $t \to \infty$. Thus the inverse Laplace transform of $L_m(s)$ is

$$\begin{aligned} m(t) &= \frac{t}{\overline{X}} + \frac{\mathsf{E}\left[X^2\right]}{2\overline{X}^2} - 1 + \epsilon(t) \\ &= \frac{t}{\overline{X}} + \frac{\sigma^2}{2\overline{X}^2} - \frac{1}{2} + \epsilon(t) \qquad \text{for } t \geq 0, \end{aligned} \tag{4.56}$$

where $\lim_{t\to\infty} \epsilon(t) = 0$.

We have derived (4.56) only for the special case in which $\mathsf{f}_X(x)$ has a rational Laplace transform. For this case, (4.56) implies both the elementary renewal theorem ($\lim_{t\to\infty} m(t)/t = 1/\overline{X}$) and also Blackwell's theorem ($\lim_{t\to\infty}[m(t+\delta)-m(t)] = \delta/\overline{X}$). We will interpret the meaning of the constant term $\sigma^2/(2\overline{X}^2) - 1/2$ in Section 4.8.

4.6.2 The elementary renewal theorem

Theorem 4.6.1 (The elementary renewal theorem). *Let $\{N(t);\, t>0\}$ be a renewal counting process with mean inter-renewal interval $\overline{X}$. Then $\lim_{t\to\infty} \mathsf{E}[N(t)]/t = 1/\overline{X}$.*

Discussion: We have already seen that $m(t) = \mathsf{E}[N(t)]$ is finite for all $t>0$ (see Exercise 4.2). The theorem is proven by establishing a lower and upper bound to $m(t)/t$ and showing that each approaches $1/\mathsf{E}[X]$ as $t\to\infty$. The key element for each bound is (4.35), repeated below, which comes from the Wald equality.

$$m(t) = \frac{\mathsf{E}\left[S_{N(t)+1}\right]}{\overline{X}} - 1. \tag{4.57}$$

Proof: The lower bound to $m(t)/t$ comes by recognizing that $S_{N(t)+1}$ is the epoch of the first arrival after t. Thus $\mathsf{E}\left[S_{N(t)+1}\right] > t$. Substituting this into (4.57),

$$\frac{m(t)}{t} > \frac{1}{\mathsf{E}[X]} - \frac{1}{t}.$$

Clearly this lower bound approaches $1/\mathsf{E}[X]$ as $t\to\infty$. The upper bound, which is more difficult[19] and might be omitted on a first reading, is established by first truncating $X(t)$ and then applying (4.57) to the truncated process.

For an arbitrary constant $b>0$, let $\breve{X}_i = \min(b, X_i)$. Since these truncated random variables are IID, they form a related renewal counting process $\{\breve{N}(t); t>0\}$ with $\breve{m}(t) = \mathsf{E}\left[\breve{N}(t)\right]$ and $\breve{S}_n = \breve{X}_1 + \cdots + \breve{X}_n$. Since $\breve{X}_i \le X_i$ for all i, we see that $\breve{S}_n \le S_n$ for all n. Since $\{S_n \le t\} = \{N(t) \ge n\}$, it follows that $\breve{N}(t) \ge N(t)$ and thus $\breve{m}(t) \ge m(t)$. Finally, in the truncated process, $\breve{S}_{\breve{N}(t)+1} \le t+b$ and thus $\mathsf{E}\left[\breve{S}_{\breve{N}(t)+1}\right] \le t+b$. Thus, applying (4.57) to the truncated process,

$$\frac{m(t)}{t} \le \frac{\breve{m}(t)}{t} = \frac{\mathsf{E}\left[S_{\breve{N}(t)+1}\right]}{t\mathsf{E}\left[\breve{X}\right]} - \frac{1}{t} \;\le\; \frac{t+b}{t\mathsf{E}\left[\breve{X}\right]}.$$

Next, choose $b = \sqrt{t}$. Then

$$\frac{m(t)}{t} \le \frac{1}{\mathsf{E}\left[\breve{X}\right]} + \frac{1}{\sqrt{t}\,\mathsf{E}\left[\breve{X}\right]}.$$

[19] The difficulty here, and the reason for using a truncation argument, comes from the fact that the residual life, $S_{N(t)+1} - t$ at t might be arbitrarily large. We saw in Section 4.4 that the time-average residual life is infinite if $\mathsf{E}\left[X^2\right]$ is infinite. Figure 4.6 also illustrates why residual life can be so large.

Note finally that $\mathsf{E}\left[\breve{X}\right] = \int_0^b [1-\mathsf{F}_X(x)]\,dx$. Since $b = \sqrt{t}$, we have $\lim_{t\to\infty} \mathsf{E}\left[\breve{X}\right] = \mathsf{E}[X]$, completing the proof. □

Note that this theorem (and its proof) have not assumed finite variance. It can also be seen that the theorem holds when $\mathsf{E}[X]$ is infinite, since $\lim_{t\to\infty} \mathsf{E}\left[\breve{X}\right] = \infty$ in this case.

Recall that $N[t,\omega]/t$ is the average number of renewals from 0 to t for a sample function ω, and $m(t)/t$ is the average of this over ω. Combining with Theorem 4.3.1, we see that the limiting time and ensemble-average equals the time-average renewal rate for each sample function except for a set of probability 0.

Another interesting question is to determine the expected renewal rate in the limit of large t without averaging from 0 to t. That is, are there some values of t at which renewals are more likely than others for large t? If the inter-renewal intervals have an integer distribution function (i.e., each inter-renewal interval must last for an integer number of time units), then each renewal epoch S_n must also be an integer. This means that $N(t)$ can increase only at integer times and the expected rate of renewals is zero at all non-integer times.

An obvious generalization of integer valued inter-renewal intervals is that of inter-renewals that occur only at integer multiples of some real number $d > 0$. Such a distribution is called an *arithmetic distribution*. The *span* of an arithmetic distribution is the largest number λ such that this property holds. Thus, for example if X takes on only the values 0, 2, and 6, its distribution is arithmetic with span $\lambda = 2$. Similarly, if X takes on only the values 1/3 and 1/5, then the span is $\lambda = 1/15$. The remarkable thing, for our purposes, is that any inter-renewal distribution that is not an arithmetic distribution leads to a uniform expected rate of renewals in the limit of large t. This result is contained in Blackwell's renewal theorem, which we state without proof.[20]. Recall, however, that for the special case of an inter-renewal density with a rational Laplace transform, Blackwell's renewal theorem is a simple consequence of (4.56).

Theorem 4.6.2 (Blackwell). *If a renewal process has an inter-renewal distribution that is non-arithmetic, then for each $\delta > 0$,*

$$\lim_{t\to\infty} [m(t+\delta) - m(t)] = \frac{\delta}{\mathsf{E}[X]}. \tag{4.58}$$

If the inter-renewal distribution is arithmetic with span λ, then

$$\lim_{t\to\infty} [m(t+\lambda) - m(t)] = \frac{\lambda}{\mathsf{E}[X]}. \tag{4.59}$$

Eq. (4.58) says that for non-arithmetic distributions, the expected number of arrivals in the interval $(t, t+\delta]$ is equal to $\delta/\mathsf{E}[X]$ in the limit $t \to \infty$. Since the theorem is true for arbitrarily small δ, the theorem almost seems to be saying that $m(t)$ has a derivative for large t, but this is not true. One can see the reason by looking at an example where X can take on only the values 1 and π. Then no matter how large t is, $N(t)$ can only increase at discrete points of time of the form $k + j\pi$ where k and j are nonnegative integers. Thus

[20] See Theorem 1 of Section 11.1, of [8]) for a proof

$dm(t)/dt$ is either 0 or ∞ for all t. As t gets larger, the jumps in $m(t)$ become both smaller in magnitude and more closely spaced from one to the next. Thus $\big[m(t+\delta)-m(t)\big]/\delta$ can approach $1/\mathsf{E}[X]$ as $t\to\infty$ for any fixed δ (as the theorem says), but as δ gets smaller, the convergence in t gets slower. For the above example (and for all discrete non-arithmetic distributions), $\big[m(t+\delta)-m(t)\big]/\delta$ does not approach[21] $1/\mathsf{E}[X]$ for any t as $\delta\to 0$.

For an arithmetic renewal process with span λ, the asymptotic behavior of $m(t)$ as $t\to\infty$ is much simpler. Renewals can only occur at multiples of λ, and since simultaneous renewals are not allowed, either 0 or 1 renewal occurs at each time $k\lambda$. Thus for any k. we have

$$\Pr\{\text{Renewal at } \lambda k\} = m(\lambda k) - m(\lambda(k-1)), \tag{4.60}$$

where, by convention, we take $m(0)=0$. Thus (4.59) can be restated as

$$\lim_{k\to\infty} \Pr\{\text{Renewal at } k\lambda\} = \frac{\lambda}{\overline{X}}. \tag{4.61}$$

The limiting behavior of $m(t)$ is discussed further in the next section.

4.7 Renewal-reward processes; ensemble-averages

Theorem 4.4.1 showed that if a renewal-reward process has an expected inter-renewal interval $\overline{X}$ and an expected inter-renewal reward $\mathsf{E}[R_n]$, then the time-average reward is $\mathsf{E}[R_n]/\overline{X}$ with probability 1. In this section, we explore the ensemble average, $\mathsf{E}[R(t)]$, as a function of time t. It is easy to see that $\mathsf{E}[R(t)]$ typically changes with t, especially for small t, but a question of major interest here is whether $\mathsf{E}[R(t)]$ approaches a constant as $t\to\infty$.

In more concrete terms, if the arrival times of busses at a bus station forms a renewal process, then the waiting time for the next bus, *i.e.*, the residual life, starting at time t, can be represented as a reward function $R(t)$. We would like to know if the expected waiting time depends critically on t, where t is the time since the renewal process started, *i.e.*, the time since a hypothetical bus number 0 arrived. If $\mathsf{E}[R(t)]$ varies significantly with t, even as $t\to\infty$, it means that the choice of $t=0$ as the beginning of the initial interarrival interval never dies out as $t\to\infty$.

Blackwell's renewal theorem (and common sense) tell us that there is a large difference between arithmetic inter-renewal times and non-arithmetic inter-renewal times. For the arithmetic case, all renewals occur at multiples of the span λ. Thus, for example, the expected waiting time (*i.e.*, the expected residual life) decreases at rate 1 from each multiple of λ to the next, and it increases with a jump equal to the probability of an arrival at each multiple of λ. For this reason, we usually consider various reward functions only at multiples of λ. We would guess, then, that $\mathsf{E}[R(n\lambda)]$ approaches a constant as $n\to\infty$.

For the non-arithmetic case, on the other hand, the expected number of renewals in any small interval of length δ becomes independent of t as $t\to\infty$, so we might guess that

[21] This must seem like mathematical nitpicking to many readers. However, $m(t)$ is the expected number of renewals in $(0,t]$, and how $m(t)$ varies with t, is central to this chapter and keeps reappearing.

$\mathsf{E}[R(t)]$ approaches a limit as $t \to \infty$. We would also guess that these asymptotic ensemble averages are equal to the appropriate time averages from Section 4.4.

The bottom line for this section is that under very broad conditions, the above guesses are essentially correct. Thus the limit as $t \to \infty$ of a given ensemble-average reward can usually be computed simply by finding the time-average and vice-versa. Sometimes time-averages are simpler, and sometimes ensemble-averages are. The advantage of the ensemble-average approach is both the ability to find $\mathsf{E}[R(t)]$ for finite values of t and to understand the rate of convergence to the asymptotic result.

The following subsection is restricted to the arithmetic case. We will derive the joint distribution function of age and duration for any given time t, and then look at the limit as $t \to \infty$. This leads us to arbitrary reward functions (such as residual life) for the arithmetic case. We will not look specifically at generalized reward functions that depend on other processes, but this generalization is quite similar to that for time-averages.

The non-arithmetic case is analyzed in the remainder of the subsections of this section. The basic ideas are the same as the arithmetic case, but a number of subtle mathematical limiting issues arise. The reader is advised to understand the arithmetic case first, since the limiting issues in the non-arithmetic case can then be viewed within the intuitive context of the arithmetic case.

4.7.1 Age and duration for arithmetic processes

Let $\{N(t);\, t > 0\}$ be an arithmetic renewal counting process with inter-renewal intervals $X_1, X_2, \ldots$ and arrival epochs $S_1, S_2, \ldots$, where $S_n = X_1 + \cdots + X_n$. To keep the notation as uncluttered as possible, we take the span to be one and then scale to an arbitrary λ later. Thus each X_i is a positive integer-valued rv.

Recall that the age $Z(t)$ at any given $t > 0$ is $Z(t) = t - S_{N(t)}$ (where by convention $S_0 = 0$) and the duration $\widetilde{X}(t)$ is $\widetilde{X}(t) = S_{N(t)+1}(t) - S_{N(t)}$. Since arrivals occur only at integer times, we initially consider age and duration only at integer times also. If an arrival occurs at integer time t, then $S_{N(t)} = t$ and $Z(t) = 0$. Also, if $S_1 > t$, then $N(t) = 0$ and $Z(t) = t$ (*i.e.*, the age is taken to be t if no arrivals occur up to and including time t). Thus, for integer t, $Z(t)$ is an integer-valued rv taking values from $[0,\, t]$. Since $S_{N(t)+1} > t$, it follows that $\widetilde{X}(t)$ is an integer-valued rv satisfying $\widetilde{X}(t) > Z(t)$. Since both are integer valued, $\widetilde{X}(t)$ must exceed $Z(t)$ by at least 1 (or by λ in the more general case of span λ).

In order to satisfy $Z(t) = i$ and $\widetilde{X}(t) = k$ for given $i < t$, it is necessary and sufficient to have an arrival epoch at $t - i$ followed by an interarrival interval of length k, where $k \geq i + 1$. For $Z(t) = t$ and $\widetilde{X}(t) = k$, it is necessary and sufficient that $k > t$, *i.e.*, that the first inter-renewal epoch occurs at $k > t$.

Theorem 4.7.1. *Let $\{X_n;\, n \geq 1\}$ be the interarrival intervals of an arithmetic renewal process with unit span. Then the the joint PMF of the age and duration at integer time*

$t \geq 1$ is given by

$$\mathsf{p}_{Z(t),\widetilde{X}(t)}(i,k) = \begin{cases} \mathsf{p}_X(k) & \text{for } i = t,\ k > t \\ q_{t-i}\,\mathsf{p}_X(k) & \text{for } 0 \leq i < t,\ k > i \\ 0 & \text{otherwise} \end{cases} . \tag{4.62}$$

where $q_j = \Pr\{\text{arrival at time } j\}$. The limit as $t \to \infty$ for any given $0 \leq i < k$ is given by

$$\lim_{\text{integer } t \to \infty} \mathsf{p}_{Z(t),\widetilde{X}(t)}(i,k) = \frac{\mathsf{p}_X(k)}{\overline{X}}. \tag{4.63}$$

Proof: The idea is quite simple. For the upper part of (4.62), note that the age is t if and only there are no arrivals in $(0,t]$, which corresponds to $X_1 = k$ for some $k > t$. For the middle part, the age is i for a given $i < t$ if and only if there is an arrival at $t-i$ and the next arrival epoch is after t, which means that the corresponding interarrival interval k exceeds i. The probability of an arrival at $t-i$, *i.e.*, q_{t-i}, depends only on the arrival epochs up to and including time $t-i$, which should be independent of the subsequent interarrival time, leading to the product in the middle term of (4.62). To be more precise about this independence, note that for $i < t$,

$$q_{t-i} \;=\; \Pr\{\text{arrival at } t-i\} = \sum_{n \geq 1} \mathsf{p}_{S_n}(t-i). \tag{4.64}$$

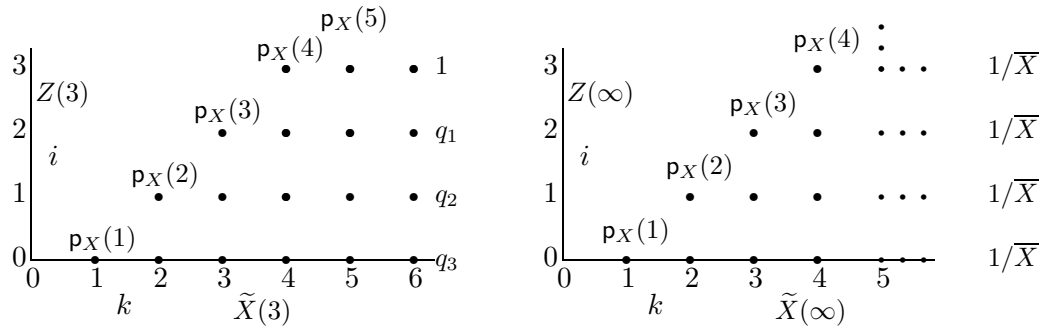
Given that $S_n = t-i$, the probability that $X_{n+1} = k$ is $\mathsf{p}_X(k)$. This is the same for all n, establishing (4.62).

For any fixed i,k with $i < k$, note that only the middle term in (4.62) is relevant as $t \to \infty$. Using Blackwell's theorem (4.61) to take the limit as $t \to \infty$, we get (4.63) □

The probabilities in the theorem, both for finite t and asymptotically as $t \to \infty$, are illustrated in Figure 4.17. The product form of the probabilities in (4.62) (as illustrated in the figure) might lead one to think that $Z(t)$ and $\widetilde{X}(t)$ are independent, but this is incorrect because of the constraint that $\widetilde{X}(t) > Z(t)$. It is curious that in the asymptotic case, (4.63) shows that, for a given duration $\widetilde{X}(t) = k$, the age is equally likely to have any integer value from 0 to $k-1$, *i.e.*, for a given duration, the interarrival interval containing t is uniformly distributed around t.

The marginal PMF for $Z(t)$ is calculated below using (4.62).

$$\mathsf{p}_{Z(t)}(i) = \begin{cases} \mathsf{F}^{\mathsf{c}}_X(i) & \text{for } i = t \\ q_{t-i}\,\mathsf{F}^{\mathsf{c}}_X(i) & \text{for } 0 \leq i < t \\ 0 & \text{otherwise} \end{cases} . \tag{4.65}$$



a) Case where $t = 3$. b) Asymptotic case where $t \to \infty$

Figure 4.17: Joint PMF, $\mathsf{p}_{\widetilde{X}(t)Z(t)}(k,i)$ of $\widetilde{X}(t)$ and $Z(t)$ in an arithmetic renewal process with span 1. In part a), $t = 3$ and the PMF at each sample point is the product of two terms, $q_{t-i} = \Pr\{\text{Arrival at } t-i\}$ and $\mathsf{p}_X(k)$. Part b) is the asymptotic case where $t \to \infty$. Here the arrival probabilities become uniform.

where $\mathsf{F}^{\mathsf{c}}_X(i) = \mathsf{p}_X(i{+}1) + \mathsf{p}_X(i{+}2) + \cdots$. The marginal PMF for $\widetilde{X}(t)$ can be calculated directly from (4.62), but it is simplified somewhat by recognizing that

$$q_j \;=\; m(j) - m(j-1). \tag{4.66}$$

Substituting this into (4.62) and summing over age,

$$\mathsf{p}_{\widetilde{X}(t)}(k) = \begin{cases} \mathsf{p}_X(k)\big[m(t) - m(t{-}k)\big] & \text{for } k < t \\ \mathsf{p}_X(k)\, m(t) & \text{for } k = t \\ \mathsf{p}_X(k)\big[m(t) + 1\big] & \text{for } k > t \end{cases} . \tag{4.67}$$

The term $+1$ in the expression for $k > t$ corresponds to the uppermost point for the given k in Figure 4.17a. This accounts for the possibility of no arrivals up to time t. It is not immediately evident that $\sum_k \mathsf{p}_{\widetilde{X}(t)}(k) = 1$, but this can be verified from the renewal equation, (4.52).

Blackwell's theorem shows that the arrival probabilities tend to $1/\overline{X}$ as $t \to \infty$, so the limiting marginal probabilities for age and duration become

$$\lim_{\text{integer } t\to\infty} \mathsf{p}_{Z(t)}(i) \;=\; \frac{\mathsf{F}^{\mathsf{c}}_X(i)}{\overline{X}}. \tag{4.68}$$

$$\lim_{\text{integer } t\to\infty} \mathsf{p}_{\widetilde{X}(t)}(k) \;=\; \frac{k\,\mathsf{p}_X(k)}{\overline{X}}. \tag{4.69}$$

The expected value of $Z(t)$ and $\widetilde{X}(t)$ can also be found for all t from (4.62) and (4.63) respectively, but they don't have a particularly interesting form. The asymptotic values as $t \to \infty$ are more simple and interesting. The asymptotic expected value for age is derived

below from (4.68).[22]

$$\begin{aligned}
\lim_{\text{integer } t\to\infty} \mathsf{E}[Z(t)] &= \sum_i i \lim_{\text{integer } t\to\infty} \mathsf{p}_{Z(t)}(i) \\
&= \frac{1}{\overline{X}} \sum_{i=1}^{\infty} \sum_{j=i+1}^{\infty} i\, \mathsf{p}_X(j) = \frac{1}{\overline{X}} \sum_{j=2}^{\infty} \sum_{i=1}^{j-1} i\, \mathsf{p}_X(j) \\
&= \frac{1}{\overline{X}} \sum_{j=2}^{\infty} \frac{j(j-1)}{2}\, \mathsf{p}_X(j) \\
&= \frac{\mathsf{E}\left[X^2\right]}{2\overline{X}} - \frac{1}{2}. \qquad (4.70)
\end{aligned}$$

This limiting ensemble average age has the same dependence on $\mathsf{E}\left[X^2\right]$ as the time average in (4.16), but, perhaps surprisingly, it is reduced from that amount by 1/2. To understand this, note that we have only calculated the expected age at integer values of t. Since arrivals occur only at integer values, the age for each sample function must increase with unit slope as t is increased from one integer to the next. The expected age thus also increases, and then at the next integer value, it drops discontinuously due to the probability of an arrival at that next integer. Thus the limiting value of $\mathsf{E}[Z(t)]$ has a saw tooth shape and the value at each discontinuity is the lower side of that discontinuity. Averaging this asymptotic expected age over a unit of time, the average is $\mathsf{E}\left[X^2\right]/2\overline{X}$, in agreement with (4.16).

As with the time average, the limiting expected age is infinite if $\mathsf{E}\left[X^2\right] = \infty$. However, for each t, $Z(t) \leq t$, so $\mathsf{E}[Z(t)] < \infty$ for all t, increasing without bound as $t \to \infty$

The asymptotic expected duration is derived in a similar way, starting from (4.69)

$$\begin{aligned}
\lim_{\text{integer } t\to\infty} \mathsf{E}\left[\widetilde{X}(t)\right] &= \sum_k \lim_{\text{integer } t\to\infty} k\mathsf{p}_{\widetilde{X}(t)}(k) \\
&= \sum_k \frac{k^2 \mathsf{p}_X(k)}{\overline{X}} = \frac{\mathsf{E}\left[X^2\right]}{\overline{X}}. \qquad (4.71)
\end{aligned}$$

This agrees with the time average in (4.17). The reduction by 1/2 seen in (4.70) is not present here, since as t is increased in the interval $[t, t+1)$, $X(t)$ remains constant.

Since the asymptotic ensemble-average age differs from the time-average age in only a trivial way, and the asymptotic ensemble-average duration is the same as the time-average duration, it might appear that we have gained little by this exploration of ensemble averages. What we have gained, however, is a set of results that apply to all t. Thus they show how (in principle) these results converge as $t \to \infty$.

[22] Finding the limiting expectations from the limiting PMF's requires interchanging a limit with an expectation. This can be justified (in both (4.70) and (4.71)) by assuming that X has a finite second moment and noting that all the terms involved are positive, that $\Pr\{\text{arrival at } j\} \leq 1$ for all j, and that $\mathsf{p}_X(k) \leq 1$ for all k.

4.7.2 Joint age and duration: non-arithmetic case

Non-arithmetic renewal processes are mathematically more complicated than arithmetic renewal processes, but the concepts are the same. We start by looking at the joint probability of the age and duration, each over an incremental interval. (see Figure 4.18 (a)).

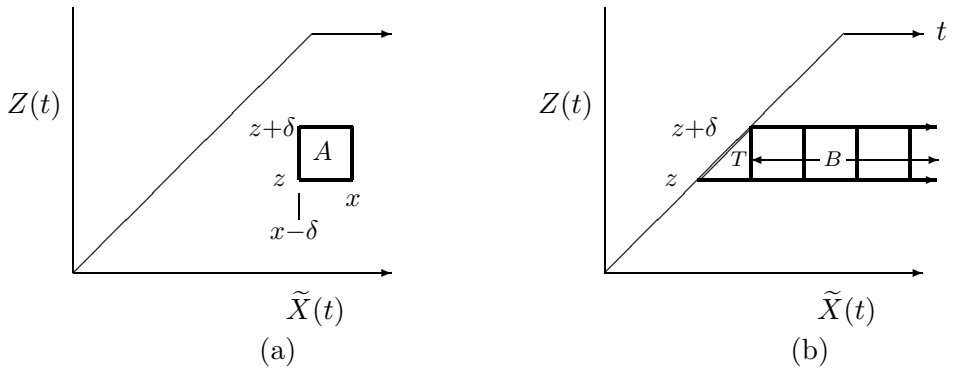


Figure 4.18: Part (a): The incremental square region A of sample values for $Z(t)$ and $\widetilde{X}(t)$ whose joint probability is specified in (4.73). The square is assumed to be inside the indicated semi-infinite trapezoidal region, *i.e.*, to satisfy $0 \leq z < z+\delta \leq t$ and $z+2\delta \leq x$.

Part (b): Summing over discrete sample regions of $\widetilde{X}(t)$ to find the marginal probability $\Pr\{z \leq Z(t) < z+\delta\} = \Pr\{T\} + \Pr\{B\}$, where T is the triangular area and B the set of indicated squares.

Theorem 4.7.2. *Consider an arbitrary renewal process with age $Z(t)$ and duration $\widetilde{X}(t)$ at any given time $t > 0$. Let A be the event*

$$A = \{z \leq Z(t) < z+\delta\} \bigcap \{x-\delta < \widetilde{X}(t) \leq x\}, \tag{4.72}$$

where $0 \leq z < z+\delta \leq t$ and $z+2\delta \leq x$. Then

$$\Pr\{A\} = \big[m(t-z) - m(t-z-\delta)\big]\,\big[\mathsf{F}_X(x) - \mathsf{F}_X(x-\delta)\big]. \tag{4.73}$$

If in addition the renewal process is non-arithmetic,

$$\lim_{t\to\infty} \Pr\{A\} = \frac{\big[\mathsf{F}_X(x) - \mathsf{F}_X(x-\delta)\big]}{\overline{X}}. \tag{4.74}$$

Proof: Note that A is the box illustrated in Figure 4.18 (a) and that under the given conditions, $\widetilde{X}(t) > Z(t)$ for all sample points in A. Recall that $Z(t) = t - S_{N(t)}$ and $\widetilde{X}(t) = S_{N(t)+1} - S_{N(t)} = X_{N(t)+1}$, so A can also be expressed as

$$A = \{t-z-\delta < S_{N(t)} \leq t - z\} \bigcap \{x - \delta < X_{N(t)+1} \leq x\}. \tag{4.75}$$

We now argue that A can also be rewritten as

$$A = \bigcup_{n=1}^{\infty} \Big\{ \{t{-}z{-}\delta < S_n \le t-z\} \bigcap \{x-\delta < X_{n+1} \le x\} \Big\}. \tag{4.76}$$

To see this, first assume that the event in (4.75) occurs. Then $N(t)$ must have some positive sample value n, so (4.76) occurs. Next assume that the event in (4.76) occurs, which means that one of the events, say the nth, in the union occurs. Since $S_{n+1} = S_n + X_{n+1} > t$, we see that n is the sample value of $S_{N(t)}$ and (4.75) must occur.

To complete the proof, we must find $\Pr\{A\}$. First note that A is a union of disjoint events. That is, although more than one arrival epoch might occur in $(t{-}z{-}\delta,\ t{-}z]$, the following arrival epoch can exceed t for only one of them. Thus

$$\Pr\{A\} = \sum_{n=1}^{\infty} \Pr\Big\{ \{t{-}z{-}\delta < S_n \le t-z\} \bigcap \{x-\delta < X_{n+1} \le x\} \Big\}. \tag{4.77}$$

For each n, X_{n+1} is independent of S_n, so

$$\begin{aligned} &\Pr\Big\{ \{t{-}z{-}\delta < S_n \le t-z\} \bigcap \{x-\delta < X_{n+1} \le x\} \Big\} \\ &\qquad = \Pr\big\{t{-}z{-}\delta < S_n \le t-z\big\} \big[\mathsf{F}^{\mathsf{c}}_X(x) - \mathsf{F}^{\mathsf{c}}_X(x{-}\delta\big]. \end{aligned}$$

Substituting this into (4.77) and using (4.51) to sum the series, we get

$$\Pr\{A\} = \big[m(t{-}z) - m(t{-}z{-}\delta)\big]\ \big[\mathsf{F}_X(x) - \mathsf{F}_X(x{-}\delta)\big].$$

This establishes (4.73). Blackwell's theorem then establishes (4.74). □

It is curious that $\Pr\{A\}$ has such a simple product expression, where one term depends only on the function $m(t)$ and the other only on the distribution function F_X. Although the theorem is most useful as $\delta \to 0$, the expression is exact for all δ such that the square region A satisfies the given constraints (*i.e.*, A lies in the indicated semi-infinite trapezoidal region).

4.7.3 Age $Z(t)$ for finite t: non-arithmetic case

In this section, we first use Theorem 4.7.2 to find bounds on the marginal incremental probability of $Z(t)$. We then find the distribution function, $\mathsf{F}_{Z(t)}(z)$, and the expected value, $\mathsf{E}[Z(t)]$ of $Z(t)$.

Corollary 4.7.1. *For $0 \le z < z{+}\delta \le t$, the following bounds hold on $\Pr\{z \le Z(t) < z{+}\delta\}$.*

$$\Pr\{z \le Z(t) < z{+}\delta\} \ \ge\ \big[m(t{-}z) - m(t{-}z{-}\delta)\big]\mathsf{F}^{\mathsf{c}}_X(z{+}\delta) \tag{4.78}$$

$$\Pr\{z \le Z(t) < z{+}\delta\} \ \le\ \big[m(t{-}z) - m(t{-}z{-}\delta)\big]\mathsf{F}^{\mathsf{c}}_X(z). \tag{4.79}$$

Proof*: As indicated in Figure 4.18 (b), $\Pr\{z \le Z(t) < z{+}\delta\} = \Pr\{T\} + \Pr\{B\}$, where T is the triangular region,

$$T = \{z \le Z(t) < z{+}\delta\} \bigcap \{Z(t) < \widetilde{X}(t) \le z{+}\delta\},$$

and B is the rectangular region

$$B = \{z \le Z(t) < z+\delta\} \bigcap \{\widetilde{X}(t) > z+\delta\}.$$

It is easy to find $\Pr\{B\}$ by summing the probabilities in (4.73) for the squares indicated in Figure 4.18 (b). The result is

$$\Pr\{B\} = \big[m(t-z) - m(t-z-\delta)\big]\mathsf{F}^{\mathsf{c}}_X(z+\delta). \tag{4.80}$$

Since $\Pr\{T\} \ge 0$, this establishes the lower bound in (4.78). We next need an upper bound to $\Pr\{T\}$ and start by finding an event that includes T.

$$\begin{aligned} T &= \{z \le Z(t) < z+\delta\} \bigcap \{Z(t) < \widetilde{X}(t) \le z+\delta\} \\ &= \bigcup_{n\ge 1} \Big[\{t-z-\delta < S_n \le t-z\} \bigcap \{t-S_n < X_{n+1} \le z+\delta\}\Big] \\ &\subseteq \bigcup_{n\ge 1} \Big[\{t-z-\delta < S_n \le t-z\} \bigcap \{z < X_{n+1} \le z+\delta\}\Big], \end{aligned}$$

where the inclusion follows from the fact that $S_n \le t-z$ for the event on the left to occur, and this implies that $z \le t - S_n$.

Using the union bound, we then have

$$\begin{aligned} \Pr\{T\} &\le \Big[\sum_{n\ge 1} \Pr\{\{t-z-\delta < S_n \le t-z\}\Big]\big[\mathsf{F}^{\mathsf{c}}_X(z) - \mathsf{F}^{\mathsf{c}}_X(z+\delta)\big] \\ &= \big[m(t-z) - m(t-z-\delta)\big]\big[\mathsf{F}^{\mathsf{c}}_X(z) - \mathsf{F}^{\mathsf{c}}_X(z+\delta)\big]. \end{aligned}$$

Combining this with (4.80), we have (4.79). □

The following corollary uses Corollary 4.7.1 to determine the distribution function of $Z(t)$. The rather strange dependence on the existence of a Stieltjes integral will be explained after the proof.

Corollary 4.7.2. *If the Stieltjes integral $\int_{t-z}^{t} \mathsf{F}^{\mathsf{c}}_X(t-\tau)\, dm(\tau)$ exists for given $t > 0$ and $0 < z < t$, then*

$$\Pr\{Z(t) \le z\} = \int_{t-z}^{t} \mathsf{F}^{\mathsf{c}}_X(t-\tau)\, dm(\tau). \tag{4.81}$$

Proof: First, partition the interval $[0, z)$ into a given number ℓ of increments, each of size $\delta = z/\ell$. Then

$$\Pr\{Z(t) < z\} = \sum_{k=0}^{\ell-1} \Pr\{k\delta \le Z(t) < k\delta + \delta\}.$$

Applying the bounds in 4.78 and 4.79 to the terms in this sum,

$$\Pr\{Z(t) < z\} \ge \sum_{k=0}^{\ell-1} \Big[m\big(t-k\delta\big) - m\big(t-k\delta-\delta\big)\Big]\mathsf{F}^{\mathsf{c}}_X(k\delta+\delta) \tag{4.82}$$

$$\Pr\{Z(t) < z\} \le \sum_{k=0}^{\ell-1} \Big[m\big(t-k\delta\big) - m\big(t-k\delta-\delta\big)\Big]\mathsf{F}^{\mathsf{c}}_X\big(k\delta\big). \tag{4.83}$$

These are, respectively, lower and upper Riemann sums for the Stieltjes integral $\int_0^z \mathsf{F}_X^{\mathsf{c}}(t-\tau)\,dm(\tau)$. Thus, if this Stieltjes integral exists, then, letting $\delta = z/\ell \to 0$,

$$\Pr\{Z(t) < z\} = \int_{t-z}^{t} \mathsf{F}_X^{\mathsf{c}}(t-\tau)\,dm(\tau).$$

This is a convolution and thus the Stieltjes integral exists unless $m(\tau)$ and $\mathsf{F}_X^{\mathsf{c}}(t-\tau)$ both have a discontinuity at some $\tau \in [0,\, z]$ (see Exercise 1.12). If no such discontinuity exists, then $\Pr\{Z(t) < z\}$ cannot have a discontinuity at z. Thus, if the Stieltjes integral exists, $\Pr\{Z(t) < z\} = \Pr\{Z(t) \le z\}$, and, for $z < t$,

$$\mathsf{F}_{Z(t)}(z) = \int_{t-z}^{t} \mathsf{F}_X^{\mathsf{c}}(t-\tau)\,dm(\tau).$$

□

The above argument showed us that the values of z at which the Stieltjes integral in (4.81) fails to exist are those at which $\mathsf{F}_{Z(t)}(z)$ has a step discontinuity. At these values we know that $\mathsf{F}_{Z(t)}(z)$ (as a distribution function) should have the value at the top of the step (thus including the discrete probability that $\Pr\{Z(t) = z\}$). In other words, at any point z of discontinuity where the Stieltjes integral does not exist, $\mathsf{F}_{Z(t)}(z)$ is the limit[23] of $\mathsf{F}_{Z(t)}(z+\epsilon)$ as $\epsilon > 0$ approaches 0. Another way of expressing this is that for $0 \le z < t$, $\mathsf{F}_{Z(t)}(z)$ is the limit of the upper Riemann sum on the right side of (4.83).

The next corollary uses an almost identical argument to find $\mathsf{E}[Z(t)]$. As we will see, the Stieltjes integral fails to exist at those values of t at which there is a discrete positive probability of arrival. The expected value at these points is the lower Riemann sum for the Stieltjes integral.

Corollary 4.7.3. *If the Stieltjes integral $\int_0^t \mathsf{F}_X^{\mathsf{c}}(t-\tau)\,dm(\tau)$ exists for given $t > 0$, then*

$$\mathsf{E}[Z(t)] = \mathsf{F}_X^{\mathsf{c}}(t) + \int_0^t (t-\tau)\mathsf{F}_X^{\mathsf{c}}(t-\tau)\,dm(\tau). \tag{4.84}$$

Proof: Note that $Z(t) = t$ if and only if $X_1 > t$, which has probability $\mathsf{F}_X^{\mathsf{c}}(t)$. For the other possible values of $Z(t)$, we divide $[0, t)$ into ℓ equal intervals of length $\delta = t/\ell$ each. Then $\mathsf{E}[Z(t)]$ can be lower bounded by

$$\begin{aligned}
\mathsf{E}[Z(t)] &\ge \mathsf{F}_X^{\mathsf{c}}(t) + \sum_{k=0}^{\ell-1} k\delta \Pr\{k\delta \le Z(t) < k\delta + \delta\}\} \\
&\ge \mathsf{F}_X^{\mathsf{c}}(t) + \sum_{k=0}^{\ell-1} k\delta \left[m(t-k\delta) - m(t-k\delta-\delta)\right] \mathsf{F}_X^{\mathsf{c}}(k\delta+\delta).
\end{aligned}$$

[23] This seems to be rather abstract mathematics, but as engineers, we often evaluate functions with step discontinuities by ignoring the values at the discontinuities or evaluating these points by adhoc means.

where we used 4.78 for the second step. Similarly, $\mathsf{E}[Z(t)]$ can be upper bounded by

$$\begin{aligned}\mathsf{E}[Z(t)] &\leq \mathsf{F}^{\mathsf{c}}_X(t) + \sum_{k=0}^{\ell-1}(k\delta+\delta)\Pr\{k\delta \leq Z(t) < k\delta+\delta\}\\ &\leq \mathsf{F}^{\mathsf{c}}_X(t) + \sum_{k=0}^{\ell-1}(k\delta+\delta)\big[m(t-k\delta)-m(t-k\delta-\delta)\big]\mathsf{F}^{\mathsf{c}}_X(k\delta).\end{aligned}$$

where we used 4.79 for the second step. These provide lower and upper Riemann sums to the Stieltjes integral in (4.81), completing the proof in the same way as the previous corollary. □

4.7.4 Age $Z(t)$ as $t \to \infty$: non-arithmetic case

Next, for non-arithmetic renewal processes, we want to find the limiting values, as $t \to \infty$, for $\mathsf{F}_{Z(t)}(z)$ and $\mathsf{E}[Z(T)]$. Temporarily ignoring any subtleties about the limit, we first view $dm(t)$ as going to $\frac{dt}{\overline{X}}$ as $t \to \infty$. Thus from (4.81),

$$\lim_{t\to\infty}\Pr\{Z(t)\leq z\} = \frac{1}{\overline{X}}\int_0^z \mathsf{F}^{\mathsf{c}}_X(\tau)d\tau. \tag{4.85}$$

If X has a PDF, this simplifies further to

$$\lim_{t\to\infty}\mathsf{f}_{Z(t)}(z) = \frac{1}{\overline{X}}\mathsf{f}_X(z). \tag{4.86}$$

Note that this agrees with the time-average result in (4.29). Taking these limits carefully requires more mathematics than seems justified here, especially since the result uses Blackwell's theorem, which was not proven here. Thus we state (without proof) another theorem, equivalent to Blackwell's theorem, called the key renewal theorem, that simplifies taking this type of limit. Essentially Blackwell's theorem is easier to interpret, but the key renewal theorem is often easier to use.

Theorem 4.7.3 (Key renewal theorem). *Let $r(x) \geq 0$ be a directly Riemann integrable function, and let $m(t) = \mathsf{E}[N(t)]$ for a non-arithmetic renewal process. Then*

$$\lim_{t\to\infty}\int_{\tau=0}^{t} r(t-\tau)dm(\tau) = \frac{1}{\overline{X}}\int_0^\infty r(x)dx. \tag{4.87}$$

We first explain what directly Riemann integrable means. If $r(x)$ is nonzero only over finite limits, say $[0,b]$, then direct Riemann integration means the same thing as ordinary Riemann integration (as learned in elementary calculus). However, if $r(x)$ is nonzero over $[0,\infty)$, then the ordinary Riemann integral (if it exists) is the result of integrating from 0 to b and then taking the limit as $b \to \infty$. The direct Riemann integral (if it exists) is the result of taking a Riemann sum over the entire half line, $[0,\infty)$ and then taking the limit as the grid becomes finer. Exercise 4.25 gives an example of a simple but bizarre function that is Riemann integrable but not directly Riemann integrable. If $r(x) \geq 0$ can be upper

bounded by a decreasing Riemann integrable function, however, then, as shown in Exercise 4.25,$r(x)$ must be directly Riemann integrable. The bottom line is that restricting $r(x)$ to be directly Riemann integrable is not a major restriction.

Next we interpret the theorem. If $m(t)$ has a derivative, then Blackwell's theorem would suggest that $dm(t)/dt \to (1/\overline{X})\,dt$, which leads to (4.87) (leaving out the mathematical details). On the other hand, if X is discrete but non-arithmetic, then $dm(t)/dt$ can be intuitively viewed as a sequence of impulses that become smaller and more closely spaced as $t \to \infty$. Then $r(t)$ acts like a smoothing filter which, as $t \to \infty$, smoothes these small impulses. The theorem says that the required smoothing occurs whenever $r(t)$ is directly Riemann integrable. The theorem does not assert that the Stieltjes integral exists for all t, but only that the limit exists. For most applications to discrete inter-renewal intervals, the Stieltjes integral does not exist everywhere. Using the key renewal theorem, we can finally determine the distribution function and expected value of $Z(t)$ as $t \to \infty$. These limiting ensemble averages are, of course, equal to the time averages found earlier.

Theorem 4.7.4. *For any non-arithmetic renewal process, the limiting distribution function and expected value of the age $Z(t)$ are given by*

$$\lim_{t\to\infty} \mathsf{F}_{Z(t)}(z) = \frac{1}{\overline{X}} \int_0^z \mathsf{F}_X^{\mathsf{c}}(x)\,dx. \tag{4.88}$$

Furthermore, if $\mathsf{E}\left[X^2\right] < \infty$, then

$$\lim_{t\to z} \mathsf{E}\left[Z(t)\right] = \frac{\mathsf{E}\left[X^2\right]}{2\overline{X}}. \tag{4.89}$$

Proof: For any given $z > 0$, let $r(x) = \mathsf{F}_X^{\mathsf{c}}(x)$ for $0 \le x \le z$ and $r(x) = 0$ elsewhere. Then (4.81) becomes

$$\Pr\{Z(t) \le z\} = \int_0^t r(t-\tau)\,dm(\tau).$$

Taking the limit as $t \to \infty$,

$$\begin{aligned} \lim_{t\to\infty} \Pr\{Z(t) \le z\} &= \lim_{t\to\infty} \int_0^t r(t-\tau)\,dm(\tau) \\ &= \frac{1}{\overline{X}} \int_0^\infty r(x)\,dx \;=\; \frac{1}{\overline{X}} \int_0^z \mathsf{F}_X^{\mathsf{c}}(x)\,dx, \end{aligned} \tag{4.90}$$

where in (4.90) we used the fact that $\mathsf{F}_X^{\mathsf{c}}(x)$ is decreasing to justify using (4.87). This establishes (4.88).

To establish (4.89), we again use the key renewal theorem, but here we let $r(x) = x\,\mathsf{F}_X^{\mathsf{c}}(x)$. Exercise 4.25 shows that $x\,\mathsf{F}_X^{\mathsf{c}}(x)$ is directly Riemann integrable if $\mathsf{E}\left[X^2\right] < \infty$. Then, taking the limit of (4.84 and then using (4.87), we have

$$\begin{aligned} \lim_{t\to\infty} \mathsf{E}\left[Z(t)\right] &= \lim_{t\to\infty} \mathsf{F}_X^{\mathsf{c}}(t) + \int_0^t r(t-\tau)\,dm(\tau) \\ &= \frac{1}{\overline{X}} \int_0^\infty r(x)\,dx \;=\; \frac{1}{\overline{X}} \int_0^\infty x\mathsf{F}_X^{\mathsf{c}}(x)\,dx. \end{aligned}$$

Integrating this by parts, we get (4.89). □

4.7.5 Arbitrary renewal-reward functions: non-arithmetic case

If we omit all the mathematical precision from the previous three subsections, we get a very simple picture. We started with (4.72), which gave the probability of an incremental square region A in the $(Z(t), \widetilde{X}(t))$ plane for given t. We then converted various sums over an increasingly fine grid of such regions into Stieltjes integrals. These integrals evaluated the distribution and expected value of age at arbitrary values of t. Finally, the key renewal theorem let us take the limit of these values as $t \to \infty$.

In this subsection, we will go through the same procedure for an arbitrary reward function, say $R(t) = \mathcal{R}(Z(t), \widetilde{X}(t))$, and show how to find $\mathsf{E}[R(T)]$. Note that $\Pr\{Z(t) \le z\} = \mathsf{E}\left[\mathbb{I}_{Z(t)\le z}\right]$ is a special case of $\mathsf{E}[R(T)]$ where $R(t)$ is chosen to be $\mathbb{I}_{Z(t)\le z}$. Similarly, finding the distribution function at a given argument for any rv can be converted to the expectation of an indicator function. Thus, having a methodology for finding the expectation of an arbitrary reward function also covers distribution functions and many other quantities of interest.

We will leave out all the limiting arguments here about converting finite incremental sums of areas into integrals, since we have seen how to do that in treating $Z(t)$. In order to make this general case more transparent, we use the following shorthand for A when it is incrementally small:

$$\Pr\{A\} = m'(t-z) f_X(x)\, dx\, dz, \tag{4.91}$$

where, if the derivatives exist, $m'(\tau) = dm(\tau)/d\tau$ and $f_X(x) = d\mathsf{F}_X(x)/dx$. If the derivatives do not exist, we can view $m'(\tau)$ and $f_X(x)$ as generalized functions including impulses, or, more appropriately, view them simply as shorthand. After using the shorthand as a guide, we can put the results in the form of Stieltjes integrals and verify the mathematical details at whatever level seems appropriate.

We do this first for the example of the distribution function of duration, $\Pr\left\{\widetilde{X}(t) \le x_0\right\}$, where we first assume that $x_0 \le t$. As illustrated in Figure 4.19, the corresponding reward function $R(t)$ is 1 in the triangular region where $\widetilde{X}(t) \le x_0$ and $Z(t) < \widetilde{X}(t)$. It is 0 elsewhere.

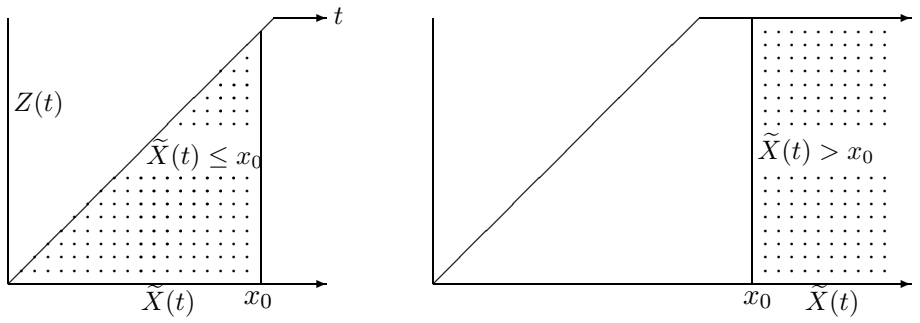


Figure 4.19: Finding $\mathsf{F}_{\widetilde{X}(t)}(x_0)$ for $x_0 \le t$ and for $x_0 > t$.

$$\begin{aligned}\Pr\Big\{\widetilde{X}(t)\le x_0\Big\} &= \int_{z=0}^{x_0}\int_{x=z}^{x_0} m'(t-z)\mathsf{f}_X(x)\,dx\,dz\\ &= \int_{z=0}^{x_0} m'(t-z)\big[\mathsf{F}_X(x_0)-\mathsf{F}_X(z)\big]\,dz\\ &= \mathsf{F}_X(x_0)\big[m(t)-m(t-x_0)\big]-\int_{t-x_0}^{t}\mathsf{F}_X(t-\tau)\,dm(\tau). \qquad (4.92)\end{aligned}$$

For the opposite case, where $x_0 > t$, it is easier to find $\Pr\Big\{\widetilde{X}(t) > x_0\Big\}$. As shown in the figure, this is the region where $0 \le Z(t) \le t$ and $\widetilde{X}(t) > x_0$. There is a subtlety here in that the incremental areas we are using are only valid for $Z(t) < t$. If the age is equal to t, then no renewals have occured in $(0, t]$, so that $\Pr\Big\{\widetilde{X}(t) > x_0; Z(t) = t\Big\} = \mathsf{F}^{\mathsf{c}}_X(x_0)$. Thus

$$\begin{aligned}\Pr\Big\{\widetilde{X}(t) > x_0\Big\} &= \mathsf{F}^{\mathsf{c}}_X(x_0)+\int_{z=0}^{t^-}\int_{x=x_0}^{\infty} m'(t-z)\mathsf{f}_X(x)\,dx\,dz\\ &= \mathsf{F}^{\mathsf{c}}_X(x_0)+m(t)\mathsf{F}^{\mathsf{c}}_X(x_0). \qquad (4.93)\end{aligned}$$

As a sanity test, the renewal equation, (4.53), can be used to show that the sum of (4.92) and (4.93) at $x_0 = t$ is equal to 1 (as they must be if the equations are correct).

We can now take the limit, $\lim_{t\to\infty}\Pr\Big\{\widetilde{X}(t)\le x_0\Big\}$. For any given x_0, (4.92) holds for sufficiently large t, and the key renewal theorem can be used since the integral has a finite range. Thus,

$$\begin{aligned}\lim_{t\to\infty}\Pr\Big\{\widetilde{X}(t)\le x_0\Big\} &= \frac{1}{\overline{X}}\Big[x_0\mathsf{F}_X(x_0)-\int_0^{x_0}\mathsf{F}_X(x)\,dx\Big]\\ &= \frac{1}{\overline{X}}\int_0^{x_0}\Big[\mathsf{F}_X(x_0)-\mathsf{F}_X(x)\Big]\,dx\\ &= \frac{1}{\overline{X}}\int_0^{x_0}\Big[\mathsf{F}^{\mathsf{c}}_X(x)-dF^{\mathsf{c}}_X(x_0)\Big]\,dx. \qquad (4.94)\end{aligned}$$

It is easy to see that the right side of (4.94) is increasing from 0 to 1 with increasing x_0, *i.e.*, it is a distribution function.

After this example, it is now straightforward to write down the expected value of an arbitrary renewal-reward function $R(t)$ whose sample value at $Z(t) = z$ and $X(t) = x$ is denoted by $\mathcal{R}(z, x)$. We have

$$\mathsf{E}\left[R(t)\right] = \int_{x=t}^{\infty}\mathcal{R}(t,x)\,d\mathsf{F}_X(x)+\int_{z=0}^{t}\int_{x=z}^{\infty}\mathcal{R}(z,x)\,d\mathsf{F}_X(x)\,dm(t-z). \qquad (4.95)$$

The first term above arises from the subtlety discussed above for the case where $Z(t) = t$. The second term is simply the integral over the semi-infinite trapezoidal area in Figure 4.19.

The analysis up to this point applies to both the arithmetic and nonarithmetic cases, but we now must assume again that the renewal process is nonarithmetic. If the inner integral,

i.e., $\int_{x=z}^{\infty} \mathcal{R}(z,x)\, d\mathsf{F}_X(x)$, as a function of z, is directly Riemann integrable, then not only can the key renewal theorem be applied to this second term, but also the first term must approach 0 as $t \to \infty$. Thus the limit of (4.95) as $t \to \infty$ is

$$\lim_{t\to\infty} \mathsf{E}\left[R(t)\right] = \frac{1}{\overline{X}} \int_{z=0}^{\infty} \int_{x=z}^{\infty} \mathcal{R}(z,x)\, d\mathsf{F}_X(x)\, dz. \tag{4.96}$$

This is the same expression as found for the time-average renewal reward in Theorem 4.4.1. Thus, as indicated earlier, we can now equate any time-average result for the nonarithmetic case with the corresponding limiting ensemble average, and the same equations have been derived in both cases.

As a simple example of (4.96), let $\mathcal{R}(z,t) = x$. Then $\mathsf{E}\left[R(t)\right] = \mathsf{E}\left[\widetilde{X}(t)\right]$ and

$$\begin{aligned}\lim_{t\to\infty} \mathsf{E}\left[\widetilde{X}(t)\right] &= \frac{1}{\overline{X}} \int_{z=0}^{\infty} \int_{x=z}^{\infty} x\, d\mathsf{F}_X(x)\, dz \;=\; \int_{x=0}^{\infty} \int_{z=0}^{x} x\, dz\, d\mathsf{F}_X(x) \\ &= \frac{1}{\overline{X}} \int_{x=0}^{\infty} x^2\, d\mathsf{F}_X(x) \;=\; \frac{\mathsf{E}\left[X^2\right]}{\overline{X}}. \end{aligned} \tag{4.97}$$

After calculating the integral above by interchanging the order of integration, we can go back and assert that the key renewal theorem applies if $\mathsf{E}\left[X^2\right]$ is finite. If it is infinite, then it is not hard to see that $\lim_{t\to\infty} \mathsf{E}\left[\widetilde{X}(t)\right]$ is infinite also.

It has been important, and theoretically reassuring, to be able to find ensemble-averages for nonarithmetic renewal-reward functions in the limit of large t and to show (not surprisingly) that they are the same as the time-average results. The ensemble-average results are quite tricky, though, and it is wise to check results achieved that way with the corresponding time-average results.

4.8 Delayed renewal processes

We have seen a certain awkwardness in our discussion of Little's theorem and the M/G/1 delay result because an arrival was assumed, but not counted, at time 0; this was necessary for the first inter-renewal interval to be statistically identical to the others. In this section, we correct that defect by allowing the epoch at which the first renewal occurs to be arbitrarily distributed. The resulting type of process is a generalization of the class of renewal processes known as *delayed renewal processes*. The word *delayed* does not necessarily imply that the first renewal epoch is in any sense larger than the other inter-renewal intervals. Rather, it means that the usual renewal process, with IID inter-renewal times, is delayed until after the epoch of the first renewal. What we shall discover is intuitively satisfying — both the time-average behavior and, in essence, the limiting ensemble behavior are not affected by the distribution of the first renewal epoch. It might be somewhat surprising, however, to find that this irrelevance of the distribution of the first renewal epoch holds even when the mean of the first renewal epoch is infinite.

To be more precise, we let $\{X_i;\, i{\geq}1\}$ be a set of independent nonnegative random variables. X_1 has a given distribution function $\mathsf{G}(x)$, whereas $\{X_i;\, i \geq 2\}$ are identically distributed

with a given distribution function $\mathsf{F}(x)$. Thus a renewal process is a special case of a delayed renewal process for which $\mathsf{G}(x) = \mathsf{F}(x)$. Let $S_n = \sum_{i=1}^{n} X_i$ be the nth renewal epoch. We first show that the SLLN still holds despite the deviant behavior of X_1.

Lemma 4.8.1. *Let $\{X_i;\ i \geq 2$ be IID with a mean $\overline{X}$ satisfying $\mathsf{E}[|X|] < \infty$ and let X_1 be a rv, independent of $\{X_i;\ i \geq 2\}$. Let $S_n = \sum_{i=1}^{n} X_i$. Then $\lim S_n/n = \overline{X}$ WP1.*

Prrof: Note that

$$\frac{S_n}{n} = \frac{X_1}{n} + \frac{\sum_{i=2}^{n} X_i}{n}.$$

Since X_1 is finite WP1, the first term above goes to 0 WP1 as $n \to \infty$. The second term goes to $\overline{X}$, proving the lemma (which is thus a trivial variation of the SLLN). □

Now, for the given delayed renewal process, let $N(t)$ be the number of renewal epochs up to and including time t. This is still determined by the fact that $\{N(t) \geq n\}$ if and only if $\{S_n \leq t\}$). $\{N(t);\ t > 0\}$ is then called a delayed renewal counting process. The following simple lemma follows from lemma 4.3.1.

Lemma 4.8.2. *Let $\{N(t);\ t > 0\}$ be a delayed renewal counting process. Then $\lim_{t\to\infty} N(t) = \infty$ with probability 1 and $\lim_{t\to\infty} \mathsf{E}[N(t)] = \infty$.*

Proof: Conditioning on $X_1 = x$, we can write $N(t) = 1 + N'(t - x)$ where $N'\{t; t \geq 0\}$ is the ordinary renewal counting process with inter-renewal intervals $X_2, X_3, \ldots$. From Lemma 4.3.1, $\lim_{t\to\infty} N'(t - x) = \infty$ with probability 1, and $\lim_{t\to\infty} \mathsf{E}[N'(t - x)] = \infty$. Since this is true for every finite $x > 0$, and X_1 is finite with probability 1, the lemma is proven. □

Theorem 4.8.1 (Strong Law for Delayed Renewal Processes). *Let $N(t);\ t > 0$ be the renewal counting process for a delayed renewal process where the inter-renewal intervals $X_2, X_3, \ldots$, have distribution function F and finite mean $\overline{X} = \int_{x=0}^{\infty} [1 - \mathsf{F}(x)]\, dx$. Then*

$$\lim_{t\to\infty} \frac{N(t)}{t} = \frac{1}{\overline{X}} \qquad \text{WP1.} \tag{4.98}$$

Proof: Using Lemma 4.8.1, the conditions for Theorem 4.3.2 are fulfilled, so the proof follows exactly as the proof of Theorem 4.3.1. □.

Next we look at the elementary renewal theorem and Blackwell's theorem for delayed renewal processes. To do this, we view a delayed renewal counting process $\{N(t);\ t > 0\}$ as an ordinary renewal counting process that starts at a random nonnegative epoch X_1 with some distribution function $\mathsf{G}(t)$. Define $N_o(t - X_1)$ as the number of renewals that occur in the interval $(X_1, t]$. Conditional on any given sample value x for X_1, $\{N_o(t-x);\ t-x > 0\}$ is an ordinary renewal counting process and thus, given $X_1 = x$, $\lim_{t\to\infty} \mathsf{E}[N_o(t - x)]/(t - x) = 1/\overline{X}$. Since $N(t) = 1 + N_o(t - X_1)$ for $t > X_1$, we see that, conditional on $X_1 = x$,

$$\lim_{t\to\infty} \frac{\mathsf{E}[N(t) \mid X_1{=}x]}{t} = \lim_{t\to\infty} \frac{\mathsf{E}[N_o(t-x)]}{t-x}\ \frac{t-x}{t} = \frac{1}{\overline{X}}. \tag{4.99}$$

Since this is true for every finite sample value x for X_1, we we have established the following theorem:

Theorem 4.8.2 (Elementary Delayed Renewal Theorem). *For a delayed renewal process with* $\mathsf{E}[X_i] = \overline{X}$ *for* $i \geq 2$*,*

$$\lim_{t\to\infty} \frac{\mathsf{E}[N(t)]}{t} = \frac{1}{\overline{X}}. \tag{4.100}$$

The same approach gives us Blackwell's theorem. Specifically, if $\{X_i;\, i{\geq}2\}$ is a sequence of IID non-arithmetic rv's, then, for any $\delta > 0$, Blackwell's theorem for ordinary renewal processes implies that

$$\lim_{t\to\infty} \frac{\mathsf{E}[N_o(t-x+\delta) - N_o(t-x)]}{\delta} = \frac{1}{\overline{X}}. \tag{4.101}$$

Thus, conditional on any sample value $X_1 = x$, $\lim_{t\to\infty} \mathsf{E}[N(t{+}\delta) - N(t) \,|\, X_1{=}x] = \delta/\overline{X}$. Taking the expected value over X_1 gives us $\lim_{t\to\infty} \mathsf{E}[N(t+\delta) - N(t)] = \delta/\overline{X}$. The case in which $\{X_i;\, i{\geq}2\}$ are arithmetic with span λ is somewhat more complicated. If X_1 is arithmetic with span λ (or a multiple of λ), then the first renewal epoch must be at some multiple of λ and $\lambda/\overline{X}$ gives the expected number of arrivals at time $i\lambda$ in the limit as $i \to \infty$. If X_1 is non-arithmetic or arithmetic with a span other than a multiple of λ, then the effect of the first renewal epoch never dies out, since all subsequent renewals occur at multiples of λ from this first epoch. We ignore this rather ugly case and state the following theorem for the nice situations.

Theorem 4.8.3 (Blackwell for Delayed Renewal). *If* $\{X_i;\, i{\geq}2\}$ *are non-arithmetic, then, for all* $\delta > 0$*,*

$$\lim_{t\to\infty} \frac{\mathsf{E}[N(t+\delta) - N(t)]}{\delta} = \frac{1}{\overline{X}}. \tag{4.102}$$

If $\{X_i; i \geq 2\}$ *are arithmetic with span* λ *and mean* $\overline{X}$ *and* X_1 *is arithmetic with span* $m\lambda$ *for some positive integer* m*, then*

$$\lim_{i\to\infty} \Pr\{\textit{renewal at } t = i\lambda\} = \frac{\lambda}{\overline{X}}. \tag{4.103}$$

4.8.1 Delayed renewal-reward processes

We have seen that the distribution of the first renewal epoch has no effect on the time or ensemble-average behavior of a renewal process (other than the ensemble dependence on time for an arithmetic process). This carries over to reward functions with almost no change. In particular, the generalized version of Theorem 4.4.1 is as follows:

Theorem 4.8.4. *Let* $\{N(t);\, t > 0\}$ *be a delayed renewal counting process where the inter-renewal intervals* $X_2, X_3, \ldots$ *have the distribution function* F*. Let* $Z(t) = t - S_{N(t)}$*, let* $\widetilde{X}(t) = S_{N(t)+1} - S_{N(t)}$*, and let* $R(t) = \mathcal{R}(Z(t), \widetilde{X}(t))$ *be a reward function. Assume that*

$$\mathsf{E}[\mathrm{R}_n] = \int_{x=0}^{\infty} \int_{z=0}^{x} \mathcal{R}(z,x)\, dz\, d\mathsf{F}(x) < \infty\,.$$

Then, with probability one,

$$\lim_{t\to\infty} \frac{1}{t} \int_{\tau=0}^{t} R(\tau)d\tau = \frac{\mathsf{E}\left[\mathrm{R}_n\right]}{\overline{X}_2} \text{ for } n \geq 2. \tag{4.104}$$

We omit the proof of this since it is a minor variation of that of theorem 4.4.1. Finally, the equivalence of time and limiting ensemble averages holds as before, yielding

$$\lim_{t\to\infty} \mathsf{E}\left[R(t)\right] = \frac{\mathsf{E}\left[\mathrm{R}_n\right]}{\overline{X}_2}. \tag{4.105}$$

4.8.2 Transient behavior of delayed renewal processes

Let $m(t) = \mathsf{E}\left[N(t)\right]$ for a delayed renewal process. As in (4.51), we have

$$m(t) = \sum_{n=1}^{\infty} \Pr\{N(t) \geq n\} = \sum_{n=1}^{\infty} \Pr\{S_n \leq t\}. \tag{4.106}$$

For $n \geq 2$, $S_n = S_{n-1} + X_n$ where X_n and S_{n-1} are independent. From the convolution equation (1.12),

$$\Pr\{S_n \leq t\} = \int_{x=0}^{t} \Pr\{S_{n-1} \leq t - x\}\ d\mathsf{F}(x) \quad \text{for } n \geq 2. \tag{4.107}$$

For $n = 1$, $\Pr\{S_n \leq t\} = \mathsf{G}(t)$. Substituting this in (4.106) and interchanging the order of integration and summation,

$$\begin{aligned} m(t) &= \mathsf{G}(t) + \int_{x=0}^{t} \sum_{n=2}^{\infty} \Pr\{S_{n-1} \leq t - x\}\, d\mathsf{F}(x) \\ &= \mathsf{G}(t) + \int_{x=0}^{t} \sum_{n=1}^{\infty} \Pr\{S_n \leq t - x\}\, d\mathsf{F}(x) \\ &= \mathsf{G}(t) + \int_{x=0}^{t} m(t-x) d\mathsf{F}(x)\,; \qquad t \geq 0. \end{aligned} \tag{4.108}$$

This is the *renewal equation* for delayed renewal processes and is a generalization of (4.52). It is shown to have a unique solution in [8], Section 11.1.

There is another useful integral equation very similar to (4.108) that arises from breaking up S_n as the sum of X_1 and $\widehat{S}_{n-1}$ where $\widehat{S}_{n-1} = X_2 + \cdots + X_n$. Letting $\widehat{m}(t)$ be the expected number of renewals in time t for an ordinary renewal process with interarrival distribution F, a similar argument to that above, starting with $\Pr\{S_n \leq t\} = \int_0^t \Pr\left\{\widehat{S}_{n-1} \leq t - x\right\}\ d\mathsf{G}(x)$ yields

$$m(t) = \mathsf{G}(t) + \int_{x=0}^{t} \widehat{m}(t-x) d\mathsf{G}(x). \tag{4.109}$$

This equation brings out the effect of the initial renewal interval clearly, and is useful in computation if one already knows $\widehat{m}(t)$.

Frequently, the most convenient way of dealing with $m(t)$ is through transforms. Following the same argument as that in (4.54), we get $L_m(r) = (1/r)L_{\mathsf{G}}(r) + L_m(r)L_{\mathsf{F}}(r)$. Solving, we get

$$L_m(r) = \frac{L_{\mathsf{G}}(r)}{r[1 - L_{\mathsf{F}}(r)]}. \tag{4.110}$$

We can find $m(t)$ from (4.110) by finding the inverse Laplace transform, using the same procedure as in Example 4.6.1. There is a second order pole at $r = 0$ again, and, evaluating the residue, it is $1/L'_{\mathsf{F}}(0) = 1/\overline{X}_2$, which is not surprising in terms of Blackwell's theorem. We can also expand numerator and denominator of (4.110) in a power series, as in (4.55). The inverse transform, corresponding to (4.56), is

$$m(t) = \frac{t}{\overline{X}} + \frac{\mathsf{E}\left[X_2^2\right]}{2\overline{X}} - \frac{\overline{X}_1}{\overline{X}} + \epsilon(t) \quad \text{for } t \to 0, \tag{4.111}$$

where $\lim_{t\to\infty} \epsilon(t) = 0$.

4.8.3 The equilibrium process

Consider an ordinary non-arithmetic renewal process with an inter-renewal interval X of distribution $\mathsf{F}(x)$. We have seen that the distribution of the interval from t to the next renewal approaches $\mathsf{F}_Y(y) = (1/\mathsf{E}[X]) \int_0^y [1 - \mathsf{F}(x)]dx$ as $t \to \infty$. This suggests that if we look at this renewal process starting at some very large t, we should see a delayed renewal process for which the distribution $\mathsf{G}(x)$ of the first renewal is equal to the residual life distribution $\mathsf{F}_Y(x)$ above and subsequent inter-renewal intervals should have the original distribution $\mathsf{F}(x)$ above. Thus it appears that such a delayed renewal process is the same as the original ordinary renewal process, except that it starts in "steady state." To verify this, we show that $m(t) = t/\overline{X}$ is a solution to (4.108) if $\mathsf{G}(t) = \mathsf{F}_Y(t)$. Substituting $(t - x)/\overline{X}$ for $m(t - x)$, the right hand side of (4.108) is

$$\frac{\int_0^t [1 - \mathsf{F}(x)]dx}{\overline{X}_2} + \frac{\int_0^t (t - x)d\mathsf{F}(x)}{\overline{X}} = \frac{\int_0^t [1 - \mathsf{F}(x)]dx}{\overline{X}} + \frac{\int_0^t \mathsf{F}(x)dx}{\overline{X}} = \frac{t}{\overline{X}},$$

where we have used integration by parts for the first equality. This particular delayed renewal process is called the *equilibrium process*, since it starts off in steady state, and thus has no transients.

4.9 Summary

Sections 4.1 to 4.7 developed the central results about renewal processes that frequently appear in subsequent chapters. The chapter starts with the strong law for renewal processes, showing that the time average rate of renewals, $N(t)/t$, approaches $1/\overline{X}$ with probability

1 as $t \to \infty$. This, combined with the strong law of large numbers, is the basis for most subsequent results about time-averages. Section 4.4 adds a reward function $R(t)$ to the underlying renewal process. These reward functions are defined to depend only on the inter-renewal interval containing t, and are used to study many surprising aspects of renewal processes such as residual life, age, and duration. For all sample paths of a renewal process (except a subset of probabiity 0), the time-average reward for a given $R(t)$ is a constant, and that constant is the expected aggregate reward over an inter-renewal interval divided by the expect length of an inter-renewal interval.

The next topic, in Section 4.5 is that of stopping trials. These have obvious applications to situations where an experiment or game is played until some desired (or undesired) outcome (based on the results up to and including the given trial) occurs. This is a basic and important topic in its right, but is also needed to understand both how the expected renewal rate $\mathsf{E}[N(t)]/t$ varies with time t and how renewal theory can be applied to queueing situations. Finally, we found that stopping rules were helpful in understanding G/G/1 queues, especially Little's theorem, and to derive an understand the Pollaczek-Khinchin expression for the expected delay in an M/G/1 queue.

This is followed, in Section 4.6, by an analysis of how $\mathsf{E}[N(t)]/t$ varies with t. This starts by using Laplace transforms to get a complete solution of the ensemble-average, $\mathsf{E}[N(t)]/t$, as a function of t, when the distribution of the inter-renewal interval has a rational Laplace transform. For the general case (where the Laplace transform is irrational or non-existant), the elementary renewal theorem shows that $\lim_{t\to\infty} \mathsf{E}[N(t)]/t = 1/\overline{X}$. The fact that the time-average (WP1) and the limiting ensemble-average are the same is not surprising, and the fact that the ensemble-average has a limit is not surprising. These results are so fundamental to other results in probability, however, that they deserve to be understood.

Another fundamental result in Section 4.6 is Blackwell's renewal theorem, showing that the distribution of renewal epochs reach a steady state as $t \to \infty$. The form of that steady state depends on whether the inter-renewal distribution is arithmetic (see (4.59)) or non-arithmetic (see (4.58)).

Section 4.7 ties together the results on rewards in 4.4 to those on ensemble averages in 4.6. Under some very minor restrictions imposed by the key renewal theorem, we found that, for non-arithmetic inter-renewal distributions, $\lim_{t\to\infty} \mathsf{E}[R(t)]$ is the same as the time-average value of reward.

Finally, all the results above were shown to apply to delayed renewal processes.

For further reading on renewal processes, see Feller,[8], Ross, [16], or Wolff, [22]. Feller still appears to be the best source for deep understanding of renewal processes, but Ross and Wolff are somewhat more accessible.

4.10 Exercises

Exercise 4.1. The purpose of this exercise is to show that for an arbitrary renewal process, $N(t)$, the number of renewals in $(0, t]$ is a (non-defective) random variable.

a) Let $X_1, X_2, \ldots$, be a sequence of IID inter-renewal rv's . Let $S_n = X_1 + \cdots + X_n$ be the corresponding renewal epochs for each $n \geq 1$. Assume that each X_i has a finite expectation $\overline{X} > 0$ and, for any given $t > 0$, use the weak law of large numbers to show that $\lim_{n\to\infty} \Pr\{S_n \leq t\} = 0$.

b) Use part a) to show that $\lim_{n\to\infty} \Pr\{N \geq n\} = 0$ and explain why this means that $N(t)$ is a rv, *i.e.*, is not defective.

c) Now suppose that the X_i do not have a finite mean. Consider truncating each X_i to $\breve{X}_i$, where for any given $b > 0$, $\breve{X}_i = \min(X_i, b)$. Let $\breve{N}(t)$ be the renewal counting process for the inter-renewal intervals $\breve{X}_i$. Show that $\breve{N}(t)$ is non-defective for each $t > 0$. Show that $N(t) \leq \breve{N}(t)$ and thus that $N(t)$ is non-defective. Note: Large inter-renewal intervals create small values of $N(t)$, and thus $\mathsf{E}[X] = \infty$ has nothing to do with potentially large values of $N(t)$, so the argument here was purely technical.

Exercise 4.2. The purpose of this exercise is to show that, for an arbitrary renewal process, $N(t)$, the number of renewals in $(0, t]$, has finite expectation.

a) Let the inter-renewal intervals have the distribution $\mathsf{F}_X(x)$, with, as usual, $\mathsf{F}_X(0) = 0$. Using whatever combination of mathematics and common sense is comfortable for you, show that numbers $\epsilon > 0$ and $\delta > 0$ must exist such that $\mathsf{F}_X(\delta) \leq 1 - \epsilon$. In other words, you are to show that a positive rv must take on some range of of values bounded away from zero with positive probability.

b) Show that $\Pr\{S_n \leq \delta\} \leq (1 - \epsilon)^n$.

c) Show that $\mathsf{E}[N(\delta)] \leq 1/\epsilon$.

d) Show that for every integer k, $\mathsf{E}[N(k\delta)] \leq k/\epsilon$ and thus that $\mathsf{E}[N(t)] \leq \frac{t+\delta}{\epsilon\delta}$ for any $t > 0$.

e) Use your result here to show that $N(t)$ is non-defective.

Exercise 4.3. Let $\{X_i; i \geq 1\}$ be the inter-renewal intervals of a renewal process generalized to allow for inter-renewal intervals of size 0 and let $\Pr\{X_i = 0\}=\alpha$, $0 < \alpha < 1$. Let $\{Y_i;\, i \geq 1\}$ be the sequence of non-zero interarrival intervals. For example, if $X_1 = x_1{>}0$, $X_2 = 0$, $X_3 = x_3{>}0, \ldots$, then $Y_1{=}x_1$, $Y_2{=}x_3, \ldots$, .

a) Find the distribution function of each Y_i in terms of that of the X_i.

b) Find the PMF of the number of arrivals of the generalized renewal process at each epoch at which arrivals occur.

c) Explain how to view the generalized renewal process as an ordinary renewal process with inter-renewal intervals $\{Y_i;\ i \geq 1\}$ and bulk arrivals at each renewal epoch.

d) When a generalized renewal process is viewed as an ordinary renewal process with bulk arrivals, what is the distribution of the bulk arrivals? (The point of this part is to illustrate that bulk arrivals on an ordinary renewal process are considerably more general than generalized renewal processes.)

Exercise 4.4. Is it true for a renewal process that:

a) $N(t) < n$ if and only if $S_n > t$?

b) $N(t) \leq n$ if and only if $S_n \geq t$?

c) $N(t) > n$ if and only if $S_n < t$?

Exercise 4.5. (This shows that convergence WP1 implies convergence in probability.) Let $\{Y_n;\, n \geq 1\}$ be a sequence of rv's that converges to 0 WP1. For any positive integers m and k, let

$$A(m,k) = \{\omega : |Y_n(\omega)| \leq 1/k \quad \text{for all } n \geq m\}.$$

a) Show that if $\lim_{n\to\infty} Y_n(\omega) = 0$ for some given ω, then (for any given k) $\omega \in A(m,k)$ for some positive integer m.

b) Show that for all $k \geq 1$

$$\Pr\Big\{\bigcup_{m=1}^{\infty} A(m,k)\Big\} = 1.$$

c) Show that, for all $m \geq 1$, $A(m,k) \subseteq A(m{+}1,\, k)$. Use this (plus (1.9)) to show that

$$\lim_{m\to\infty} \Pr\{A(m,k)\} = 1.$$

d) Show that if $\omega \in A(m,k)$, then $|Y_m(\omega)| \leq 1/k$. Use this (plus part c) to show that

$$\lim_{m\to\infty} \Pr\{|Y_m| > 1/k\} = 0.$$

Since $k \geq 1$ is arbitrary, this shows that $\{Y_n;\, n \geq 1\}$ converges in probabiity.

Exercise 4.6. In this exercise, we find an explicit expression for $\{\omega : \lim_n Y_n = 0\}$. You may use whatever level of mathematical precision you feel comfortable with.

a) Let $\{Y_n;\, n \geq 1\}$ be a sequence of rv's. Using the definition of convergence for a sequence of numbers, justify the following set equivalences:

$$\begin{aligned}
\{\omega : \lim_n Y_n(\omega) = 0\} &= \bigcap_{k=1}^{\infty} \{\omega : \text{there exists an } m \text{ such that } |Y_n(\omega)| \leq 1/k \text{ for all } n \geq m\} \\
&= \bigcap_{k=1}^{\infty} \bigcup_{m=1}^{\infty} \{\omega : Y_n(\omega) \leq 1/k \text{ for all } n \geq m\} \\
&= \bigcap_{k=1}^{\infty} \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \{\omega : Y_n(\omega) \leq 1/k\}
\end{aligned}$$

b) Explain how this shows that $\{\omega : \lim_n Y_n(\omega) = 0\}$ must be an event.

c) Use deMorgan's laws to show that the complement of the above equivalence is

$$\{\omega : \lim_n Y_n(\omega) = 0\}^{\mathsf{c}} = \bigcup_{k=1}^{\infty} \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} \{\omega : Y_n(\omega) > 1/k\}$$

d) Show that for $\{Y_n;\, n \geq 1\}$ to converge WP1, it is necessary and sufficient to satisfy

$$\Pr\Big\{\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty}\{Y_n > 1/k\}\Big\} = 0 \qquad \text{for all } k \geq 1$$

e) Show that for $\{Y_n;\, n \geq 1\}$ to converge WP1, it is necessary and sufficient to satisfy

$$\lim_{m\to\infty} \Pr\Big\{\bigcup_{n=m}^{\infty}\{Y_n > 1/k\}\Big\} = 0 \qquad \text{for all } k \geq 1$$

Hint: Use part a) of Exercise 4.7. Note: Part e) provides an equivalent condition that is often useful in establishing convergence WP1. It also brings out quite clearly the difference between convergence WP1 and convergence in probability.

Exercise 4.7. Consider the event $\bigcap_{m\geq 1}\bigcup_{n\geq m} A_n$ where $A_1, A_2, \ldots$, are arbitrary events.

a) Show that

$$\lim_{m\to\infty} \Pr\Big\{\bigcup_{n\geq m} A_n\Big\} = 0 \qquad \Longleftrightarrow \qquad \Pr\Big\{\bigcap_{m\geq 1}\bigcup_{n\geq m} A_n\Big\} = 0.$$

Hint: Apply the complement of (1.9).

b) Show that if $\sum_{m=1}^{\infty}\Pr\{A_m\} < \infty$, then $\mathsf{Pr}\big\{\bigcap_m\bigcup_{n\geq m} A_n\big\} = 0$. Hint: Recall that if $\sum_{m=1}^{\infty}\Pr\{A_m\} < \infty$, then $\lim_{m\to\infty}\mathsf{Pr}\big\{\bigcup_{n\geq m} A_n\big\} = 0$. Combine this with a). This well-known result is called the Borel-Cantelli lemma.

c) The set $\mathsf{Pr}\big\{\bigcap_m\bigcup_{n\geq m} A_n\big\}$ is often referred to as the set of ω that are contained in infinitely many of the A_n. Without trying to be precise about what this latter statement means, explain why it is a good way to think about $\mathsf{Pr}\big\{\bigcap_m\bigcup_{n\geq m} A_n\big\}$. Hint: Consider an ω that is contained in some finite number k of the sets A_n and argue that there must be an integer m such that $\omega \notin A_n$ for all $n > m$.

Exercise 4.8. Let $\{X_i;\, i\geq 1\}$ be the inter-renewal intervals of a renewal process and assume that $\mathsf{E}[X_i] = \infty$. Let $b > 0$ be an arbitrary number and $\breve{X}_i$ be a truncated random variable defined by $\breve{X}_i = X_i$ if $X_i \leq b$ and $\breve{X}_i = b$ otherwise.

a) Show that for any constant $M > 0$, there is a b sufficiently large so that $\mathsf{E}\left[\breve{X}_i\right] \geq M$.

b) Let $\{\breve{N}(t);\, t\geq 0\}$ be the renewal counting process with inter-renewal intervals $\{\breve{X}_i; i \geq 1\}$ and show that for all $t > 0$, $\breve{N}(t) \geq N(t)$.

c) Show that for all sample functions $N(t,\omega)$, except a set of probability 0, $N(t,\omega)/t < 2/M$ for all sufficiently large t. Note: Since M is arbitrary, this means that $\lim N(t)/t = 0$ with probability 1.

Exercise 4.9. Let $Y(t) = S_{N(t)+1} - t$ be the residual life at time t of a renewal process. First consider a renewal process in which the interarrival time has density $\mathsf{f}_X(x) = e^{-x}$; $x \geq 0$, and next consider a renewal process with density

$$\mathsf{f}_X(x) = \frac{3}{(x+1)^4}; \quad x \geq 0.$$

For each of the above densities, use renewal-reward theory to find:

i) the time-average of $Y(t)$

ii) the second moment in time of $Y(t)$ (i.e., $\lim_{T\to\infty} \frac{1}{T} \int_0^T Y^2(t)dt$)

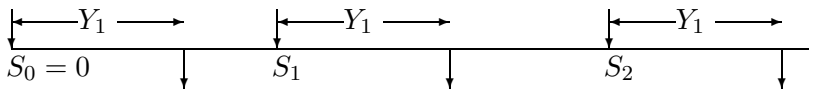
For the exponential density, verify your answers by finding $\mathsf{E}[Y(t)]$ and $\mathsf{E}\left[Y^2(t)\right]$ directly.

Exercise 4.10. Consider a variation of an M/G/1 queueing system in which there is no facility to save waiting customers. Assume customers arrive according to a Poisson process of rate λ. If the server is busy, the customer departs and is lost forever; if the server is not busy, the customer enters service with a service time distribution function denoted by $\mathsf{F}_Y(y)$.

Successive service times (for those customers that are served) are IID and independent of arrival times. Assume that customer number 0 arrives and enters service at time $t = 0$.

a) Show that the sequence of times $S_1, S_2, \ldots$ at which successive customers enter service are the renewal times of a renewal process. Show that each inter-renewal interval $X_i = S_i - S_{i-1}$ (where $S_0 = 0$) is the sum of two independent random variables, $Y_i + U_i$ where Y_i is the ith service time; find the probability density of U_i.

b) Assume that a reward (actually a cost in this case) of one unit is incurred for each customer turned away. Sketch the expected reward function as a function of time for the sample function of inter-renewal intervals and service intervals shown below; the expectation is to be taken over those (unshown) arrivals of customers that must be turned away.



c) Let $\int_0^t R(\tau)d\tau$ denote the accumulated reward (i.e., cost) from 0 to t and find the limit as $t \to \infty$ of $(1/t) \int_0^t R(\tau)d\tau$. Explain (without any attempt to be rigorous or formal) why this limit exists with probability 1.

d) In the limit of large t, find the expected reward from time t until the next renewal. Hint: Sketch this expected reward as a function of t for a given sample of inter-renewal intervals and service intervals; then find the time-average.

e) Now assume that the arrivals are deterministic, with the first arrival at time 0 and the nth arrival at time $n - 1$. Does the sequence of times $S_1, S_2, \ldots$ at which subsequent customers start service still constitute the renewal times of a renewal process? Draw a sketch of arrivals, departures, and service time intervals. Again find $\lim_{t\to\infty} \left(\int_0^t R(\tau)\, d\tau \right) /t$.

Exercise 4.11. Let $Z(t) = t - S_{N(t)}$ be the age of a renewal process and $Y(t) = S_{N(t)+1} - t$ be the residual life. Let $\mathsf{F}_X(x)$ be the distribution function of the inter-renewal interval and find the following as a function of $\mathsf{F}_X(x)$:

a) $\Pr\{Y(t){>}x \mid Z(t){=}s\}$

b) $\Pr\{Y(t){>}x \mid Z(t{+}x/2){=}s\}$

c) $\Pr\{Y(t){>}x \mid Z(t{+}x){>}s\}$ for a Poisson process.

Exercise 4.12. Let $\mathsf{F}_Z(z)$ be the fraction of time (over the limiting interval $(0,\infty)$ that the age of a renewal process is at most z. Show that $\mathsf{F}_Z(z)$ satisfies

$$\mathsf{F}_Z(z) = \frac{1}{\overline{X}} \int_{x=0}^{y} \Pr\{X > x\}\, dx \qquad \text{WP1.}$$

Hint: Follow the argument in Example 4.4.5.

Exercise 4.13. a) Let J be a stopping rule and $\mathbb{I}_{\{J\geq n\}}$ be the indicator random variable of the event $\{J \geq n\}$. Show that $J = \sum_{n\geq 1} \mathbb{I}_{\{J\geq n}$.

b) Show that $\mathbb{I}_{J\geq 1} \geq \mathbb{I}_{J\geq 2} \geq \ldots$, *i.e.*, show that for each $n > 1$, $\mathbb{I}_{J\geq n}(\omega) \geq \mathbb{I}_{J\geq n+1}(\omega)$ for each $\omega \in \Omega$ (except perhaps for a set of probability 0).

Exercise 4.14. a) Use Wald's equality to compute the expected number of trials of a Bernoulli process up to and including the kth success.

b) Use elementary means to find the expected number of trials up to and including the first success. Use this to find the expected number of trials up to and including the kth success. Compare with part a).

Exercise 4.15. A gambler with an initial finite capital of $d > 0$ dollars starts to play a dollar slot machine. At each play, either his dollar is lost or is returned with some additional number of dollars. Let X_i be his change of capital on the ith play. Assume that $\{X_i;\ i{=}1, 2, \ldots\}$ is a set of IID random variables taking on integer values $\{-1, 0, 1, \ldots\}$. Assume that $\mathsf{E}[X_i] < 0$. The gambler plays until losing all his money (i.e., the initial d dollars plus subsequent winnings).

a) Let J be the number of plays until the gambler loses all his money. Is the weak law of large numbers sufficient to argue that $\lim_{n\to\infty} \Pr\{J > n\} = 0$ (i.e., that J is a random variable) or is the strong law necessary?

b) Find $\mathsf{E}[J]$. Hint: The fact that there is only one possible negative outcome is important here.

Exercise 4.16. Let $\{X_i; i \geq 1\}$ be IID binary random variables with $P_X(0) = P_X(1) = 1/2$. Let J be a positive integer-valued random variable defined on the above sample space of binary sequences and let $S_J = \sum_{i=1}^{J} X_i$. Find the simplest example you can in which J is not a stopping trial for $\{X_i; i \geq 1\}$ and where $\mathsf{E}[X]\,\mathsf{E}[J] \neq \mathsf{E}[S_J]$. Hint: Try letting J take on only the values 1 and 2.

Exercise 4.17. Let $J = \min\{n \mid S_n{\leq}b \text{ or } S_n{\geq}a\}$, where a is a positive integer, b is a negative integer, and $S_n = X_1 + X_2 + \cdots + X_n$. Assume that $\{X_i;\ i{\geq}1\}$ is a set of zero mean IID rv's that can take on only the set of values $\{-1, 0, +1\}$, each with positive probability.

a) Is J a stopping rule? Why or why not? Hint: The more difficult part of this is to argue that J is a random variable (*i.e.*, non-defective); you do not need to construct a proof of this, but try to argue why it must be true.

b) What are the possible values of S_J?

c) Find an expression for $\mathsf{E}[S_J]$ in terms of p, a, and b, where $p = \Pr\{S_J \geq a\}$.

d) Find an expression for $\mathsf{E}[S_J]$ from Wald's equality. Use this to solve for p.

Exercise 4.18. Show that the interchange of expectation and sum in (4.32) is valid if $\mathsf{E}[J] < \infty$. Hint: First express the sum as $\sum_{n=1}^{k-1} X_n \mathbb{I}_{J\geq n} + \sum_{n=k}^{\infty}(X_n^+ + X_n^-)\mathbb{I}_{J\geq n}$ and then consider the limit as $k \to \infty$.

Exercise 4.19. Consider an amnesic miner trapped in a room that contains three doors. Door 1 leads him to freedom after two-day's travel; door 2 returns him to his room after four-day's travel; and door 3 returns him to his room after eight-day's travel. Suppose each door is equally likely to be chosen whenever he is in the room, and let T denote the time it takes the miner to become free.

a) Define a sequence of independent and identically distributed random variables $X_1, X_2, \ldots$ and a stopping rule J such that

$$T = \sum_{i=1}^{J} X_i.$$

b) Use Wald's equality to find $\mathsf{E}[T]$.

c) Compute $\mathsf{E}\left[\sum_{i=1}^{J} X_i \mid J{=}n\right]$ and show that it is not equal to $\mathsf{E}[\sum_{i=1}^{n} X_i]$.

d) Use part c) for a second derivation of $\mathsf{E}[T]$.

Exercise 4.20. a) Consider a renewal process for which the inter-renewal intervals have the PMF $p_X(1) = p_X(2) = 1/2$. Use elementary combinatorics to show that $m(1) = 1/2$, $m(2) = 5/4$, and $m(3) = 15/8$.

b) Use elementary means to show that $\mathsf{E}\left[S_{N(1)}\right] = 1/2$ and $\mathsf{E}\left[S_{N(1)+1}\right] = 9/4$. Verify (4.35) in this case (*i.e.*, for $t = 1$) and show that $N(1)$ is not a stopping trial. Note also that the expected duration, $\mathsf{E}\left[S_{N(1)+1} - S_{N(1)}\right]$ is not equal to $\overline{X}$.

c) Consider a more general form of part a) where $\Pr\{X = 1\} = 1 - p$ and $\Pr\{X = 2\} = p$. Let $\Pr\{W_n = 1\} = x_n$ and show that x_n satisfies the difference equation $x_n = 1 - px_{n-1}$ for $n \geq 1$ where by convention $x_0 = 1$. Use this to show that

$$x_n = \frac{1 - (-p)^{n+1}}{1 + p}. \tag{4.112}$$

From this, solve for $m(n)$ for $n \geq 1$.

Exercise 4.21. Let $\{N(t);\ t > 0\}$ be a renewal counting process generalized to allow for inter-renewal intervals $\{X_i\}$ of duration 0. Let each X_i have the PMF $\Pr\{X_i = 0\} = 1-\epsilon$; $\Pr\{X_i = 1/\epsilon\} = \epsilon$.

a) Sketch a typical sample function of $\{N(t);\ t > 0\}$. Note that $N(0)$ can be non-zero (i.e., $N(0)$ is the number of zero interarrival times that occur before the first non-zero interarrival time).

b) Evaluate $\mathsf{E}[N(t)]$ as a function of t.

c) Sketch $\mathsf{E}[N(t)]/t$ as a function of t.

d) Evaluate $\mathsf{E}\left[S_{N(t)+1}\right]$ as a function of t (do this directly, and then use Wald's equality as a check on your work).

e) Sketch the lower bound $\mathsf{E}[N(t)]/t \geq 1/\mathsf{E}[X] - 1/t$ on the same graph with part c).

f) Sketch $\mathsf{E}\left[S_{N(t)+1} - t\right]$ as a function of t and find the time average of this quantity.

g) Evaluate $\mathsf{E}\left[S_{N(t)}\right]$ as a function of t; verify that $\mathsf{E}\left[S_{N(t)}\right] \neq \mathsf{E}[X]\,\mathsf{E}[N(t)]$.

Exercise 4.22. Let $\{N(t);\ t > 0\}$ be a renewal counting process and let $m(t) = \mathsf{E}[N(t)]$ be the expected number of arrivals up to and including time t. Let $\{X_i; i \geq 1\}$ be the inter-renewal times and assume that $\mathsf{F}_X(0) = 0$.

a) For all $x > 0$ and $t > x$ show that $\mathsf{E}[N(t)|X_1{=}x] = \mathsf{E}[N(t-x)] + 1$.

b) Use part a) to show that $m(t) = \mathsf{F}_X(t) + \int_0^t m(t-x)d\mathsf{F}_X(x)$ for $t > 0$. This equation is the renewal equation derived differently in (4.52).

c) Suppose that X is an exponential random variable of parameter λ. Evaluate $L_m(s)$ from (4.54); verify that the inverse Laplace transform is $\lambda t;\ t{\geq}0$.

Exercise 4.23. a) Let the inter-renewal interval of a renewal process have a second order Erlang density, $\mathsf{f}_X(x) = \lambda^2 x \exp(-\lambda x)$. Evaluate the Laplace transform of $m(t) = \mathsf{E}[N(t)]$.

b) Use this to evaluate $m(t)$ for $t \geq 0$. Verify that your answer agrees with (4.56).

c) Evaluate the slope of $m(t)$ at $t = 0$ and explain why that slope is not surprising.

d) View the renewals here as being the even numbered arrivals in a Poisson process of rate λ. Sketch $m(t)$ for the process here and show one half the expected number of arrivals for the Poisson process on the same sketch. Explain the difference between the two.

Exercise 4.24. a) Let $N(t)$ be the number of arrivals in the interval $(0, t]$ for a Poisson process of rate λ. Show that the probability that $N(t)$ is even is $[1 + \exp(-2\lambda t)]/2$. Hint: Look at the power series expansion of $\exp(-\lambda t)$ and that of $\exp(\lambda t)$, and look at the sum of the two. Compare this with $\sum_{n\,\text{even}} \Pr\{N(t) = n\}$.

b) Let $\widetilde{N}(t)$ be the number of even numbered arrivals in $(0, t]$. Show that $\widetilde{N}(t) = N(t)/2 - \mathbb{I}_{\text{odd}}(t)/2$ where $\mathbb{I}_{\text{odd}}(t)$ is a random variable that is 1 if $N(t)$ is odd and 0 otherwise.

c) Use parts a) and b) to find $\mathsf{E}\left[\widetilde{N}(t)\right]$. Note that this is $m(t)$ for a renewal process with 2nd order Erlang inter-renewal intervals.

Exercise 4.25. a) Consider a function $r(z) \geq 0$ defined as follows for $0 \leq z < \infty$: For each integer $n \geq 1$ and each integer $k, 1 \leq k < n$, $r(z) = 1$ for $n + k/n \leq z \leq n + k/n + 2^{-n}$. For all other z, $r(z) = 0$. Sketch this function and show that $r(z)$ is not directly Riemann integrable.

b) Evaluate the Riemann integral $\int_0^\infty r(z)dz$.

c) Suppose $r(z)$ is decreasing, *i.e.*, that $r(z) \geq r(y)$ for all $y > z > 0$. Show tht if $r(z)$ is Riemann integrable, it is also directly Riemann integrable.

d) Suppose $f(z) \geq 0$, defined for $z \geq 0$, is decreasing and Riemann integrable and that $f(z) \geq r(z)$ for all $z \geq 0$. Show that $r(z)$ is Directly Riemann integrable.

e) Let X be a non-negative rv for which $\mathsf{E}\left[X^2\right] < \infty$. Show that $x\,\mathsf{F}^{\mathsf{c}}_X(x)$ is directly Riemann integrable. Hint: Consider $y\mathsf{F}^{\mathsf{c}}_X(y) + \int_y^\infty \mathsf{F}_X(x)\,dx$ and use Figure 1.7 (or use integration by parts) to show that this expression is decreasing in y.

Exercise 4.26. Let $Z(t), Y(t), \widetilde{X}(t)$ denote the age, residual life, and duration at time t for a renewal counting process $\{N(t);\, t > 0\}$ in which the interarrival time has a density given by $f(x)$. Find the following probability densities; assume steady state.

a) $\mathsf{f}_{Y(t)}(y \,|\, Z(t{+}s/2){=}s)$ for given $s > 0$.

b) $\mathsf{f}_{Y(t),Z(t)}(y, z)$.

c) $\mathsf{f}_{Y(t)}(y \,|\, \widetilde{X}(t){=}x)$.

d) $\mathsf{f}_{Z(t)}(z \,|\, Y(t{-}s/2){=}s)$ for given $s > 0$.

e) $\mathsf{f}_{Y(t)}(y \,|\, Z(t{+}s/2){\geq}s)$ for given $s > 0$.

Exercise 4.27. a) Find $\lim_{t\to\infty}\{\mathsf{E}\left[N(t)\right] - t/\overline{X}\}$ for a renewal counting process $\{N(t);\, t > 0\}$ with inter-renewal times $\{X_i; i \geq 1\}$. Hint: use Wald's equation.

b) Evaluate your result for the case in which X is an exponential random variable (you already know what the result should be in this case).

c) Evaluate your result for a case in which $\mathsf{E}\left[X\right] < \infty$ and $\mathsf{E}\left[X^2\right] = \infty$. Explain (very briefly) why this does not contradict the elementary renewal theorem.

Exercise 4.28. Customers arrive at a bus stop according to a Poisson process of rate λ. Independently, buses arrive according to a renewal process with the inter-renewal interval distribution $\mathsf{F}_X(x)$. At the epoch of a bus arrival, all waiting passengers enter the bus and the bus leaves immediately. Let $R(t)$ be the number of customers waiting at time t.

a) Draw a sketch of a sample function of $R(t)$.

b) Given that the first bus arrives at time $X_1 = x$, find the expected number of customers picked up; then find $\mathsf{E}\left[\int_0^x R(t)dt\right]$, again given the first bus arrival at $X_1 = x$.

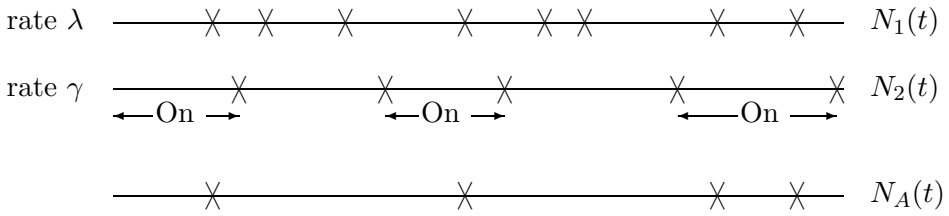
c) Find $\lim_{t\to\infty} \frac{1}{t}\int_0^t R(\tau)d\tau$ (with probability 1). Assuming that F_X is a non-arithmetic distribution, find $\lim_{t\to\infty} \mathsf{E}\left[R(t)\right]$. Interpret what these quantities mean.

d) Use part c) to find the time-average expected wait per customer.

e) Find the fraction of time that there are no customers at the bus stop. (Hint: this part is independent of a), b), and c); check your answer for $\mathsf{E}[X] \ll 1/\lambda$).

Exercise 4.29. Consider the same setup as in Exercise 4.28 except that now customers arrive according to a non-arithmetic renewal process independent of the bus arrival process. Let $1/\lambda$ be the expected inter-renewal interval for the customer renewal process. Assume that both renewal processes are in steady state (i.e., either we look only at $t \gg 0$, or we assume that they are equilibrium processes). Given that the nth customer arrives at time t, find the expected wait for customer n. Find the expected wait for customer n without conditioning on the arrival time.

Exercise 4.30. Let $\{N_1(t);\, t > 0\}$ be a Poisson counting process of rate λ. Assume that the arrivals from this process are switched on and off by arrivals from a non-arithmetic renewal counting process $\{N_2(t);\, t > 0\}$ (see figure below). The two processes are independent.



Let $\{N_A(t); t \geq 0\}$ be the switched process; that is $N_A(t)$ includes arrivals from $\{N_1(t);\, t > 0\}$ while $N_2(t)$ is even and excludes arrivals from $\{N_1(t);\, t > 0\}$ while $N_2(t)$ is odd.

a) Is $N_A(t)$ a renewal counting process? Explain your answer and if you are not sure, look at several examples for $N_2(t)$.

b) Find $\lim_{t\to\infty} \frac{1}{t} N_A(t)$ and explain why the limit exists with probability 1. Hint: Use symmetry—that is, look at $N_1(t) - N_A(t)$. To show why the limit exists, use the renewal-reward theorem. What is the appropriate renewal process to use here?

c) Now suppose that $\{N_1(t);\, t{>}0\}$ is a non-arithmetic renewal counting process but not a Poisson process and let the expected inter-renewal interval be $1/\lambda$. For any given δ, find $\lim_{t\to\infty} \mathsf{E}[N_A(t+\delta) - N_A(t)]$ and explain your reasoning. Why does your argument in (b) fail to demonstrate a time-average for this case?

Exercise 4.31. An M/G/1 queue has arrivals at rate λ and a service time distribution given by $\mathsf{F}_Y(y)$. Assume that $\lambda < 1/\mathsf{E}[Y]$. Epochs at which the system becomes empty define a renewal process. Let $\mathsf{F}_Z(z)$ be the distribution of the inter-renewal intervals and let $\mathsf{E}[Z]$ be the mean inter-renewal interval.

a) Find the fraction of time that the system is empty as a function of λ and $\mathsf{E}[Z]$. State carefully what you mean by such a fraction.

b) Apply Little's theorem, not to the system as a whole, but to the number of customers in the server (i.e., 0 or 1). Use this to find the fraction of time that the server is busy.

c) Combine your results in a) and b) to find $\mathsf{E}[Z]$ in terms of λ and $\mathsf{E}[Y]$; give the fraction of time that the system is idle in terms of λ and $\mathsf{E}[Y]$.

d) Find the expected duration of a busy period.

Exercise 4.32. Consider a sequence $X_1, X_2, \ldots$ of IID binary random variables. Let p and $1-p$ denote $\Pr\{X_m = 1\}$ and $\Pr\{X_m = 0\}$ respectively. A *renewal* is said to occur at time m if $X_{m-1} = 0$ and $X_m = 1$.

a) Show that $\{N(m); m \geq 0\}$ is a renewal counting process where $N(m)$ is the number of renewals up to and including time m and $N(0)$ and $N(1)$ are taken to be 0.

b) What is the probability that a renewal occurs at time m, $m \geq 2$?

c) Find the expected inter-renewal interval; use Blackwell's theorem here.

d) Now change the definition of renewal; a renewal now occurs at time m if $X_{m-1} = 1$ and $X_m = 1$. Show that $\{N^*_m; m \geq 0\}$ is a delayed renewal counting process where N^*_m is the number of renewals up to and including m for this new definition of renewal ($N^*_0 = N^*_1 = 0$).

e) Find the expected inter-renewal interval for the renewals of part d).

f) Given that a renewal (according to the definition in (d)) occurs at time m, find the expected time until the next renewal, conditional, first, on $X_{m+1} = 1$ and, next, on $X_{m+1} = 0$. Hint: use the result in e) plus the result for $X_{m+1} = 1$ for the conditioning on $X_{m+1} = 0$.

g) Use your result in f) to find the expected interval from time 0 to the first renewal according to the renewal definition in d).

h) Which pattern requires a larger expected time to occur: 0011 or 0101

i) What is the expected time until the first occurrence of 0111111?

Exercise 4.33. A large system is controlled by n identical computers. Each computer independently alternates between an operational state and a repair state. The duration of the operational state, from completion of one repair until the next need for repair, is a random variable X with finite expected duration $\mathsf{E}[X]$. The time required to repair a computer is an exponentially distributed random variable with density $\lambda e^{-\lambda t}$. All operating durations and repair durations are independent. Assume that all computers are in the repair state at time 0.

a) For a single computer, say the ith, do the epochs at which the computer enters the repair state form a renewal process? If so, find the expected inter-renewal interval.

b) Do the epochs at which it enters the operational state form a renewal process?

c) Find the fraction of time over which the ith computer is operational and explain what you mean by fraction of time.

d) Let $Q_i(t)$ be the probability that the ith computer is operational at time t and find $\lim_{t\to\infty} Q_i(t)$.

e) The system is in failure mode at a given time if all computers are in the repair state at that time. Do the epochs at which system failure modes begin form a renewal process?

f) Let $\Pr\{t\}$ be the probability that the the system is in failure mode at time t. Find $\lim_{t\to\infty} \Pr\{t\}$. Hint: look at part d).

g) For δ small, find the probability that the system enters failure mode in the interval $(t, t+\delta]$ in the limit as $t \to \infty$.

h) Find the expected time between successive entries into failure mode.

i) Next assume that the repair time of each computer has an arbitrary density rather than exponential, but has a mean repair time of $1/\lambda$. Do the epochs at which system failure modes begin form a renewal process?

j) Repeat part f) for the assumption in (i).

Exercise 4.34. Let $\{N_1(t);\, t{>}0\}$ and $\{N_2(t);\, t{>}0\}$ be independent renewal counting processes. Assume that each has the same distribution function $\mathsf{F}(x)$ for interarrival intervals and assume that a density $f(x)$ exists for the interarrival intervals.

a) Is the counting process $\{N_1(t) + N_2(t);\, t > 0\}$ a renewal counting process? Explain.

b) Let $Y(t)$ be the interval from t until the first arrival (from either process) after t. Find an expression for the distribution function of $Y(t)$ in the limit $t \to \infty$ (you may assume that time averages and ensemble-averages are the same).

c) Assume that a reward R of rate 1 unit per second starts to be earned whenever an arrival from process 1 occurs and ceases to be earned whenever an arrival from process 2 occurs.

Assume that $\lim_{t\to\infty}(1/t)\int_0^t R(\tau)\,d\tau$ exists with probability 1 and find its numerical value.

d) Let $Z(t)$ be the interval from t until the first time after t that $R(t)$ (as in part c) changes value. Find an expression for $\mathsf{E}[Z(t)]$ in the limit $t \to \infty$. Hint: Make sure you understand why $Z(t)$ is not the same as $Y(t)$ in part b). You might find it easiest to first find the expectation of $Z(t)$ conditional on both the duration of the $\{N_1(t);\, t > 0\}$ interarrival interval containing t and the duration of the $\{N_2(t); t \geq 0\}$ interarrival interval containing t; draw pictures!

Exercise 4.35. This problem provides another way of treating ensemble-averages for renewal-reward problems. Assume for notational simplicity that X is a continuous random variable.

a) Show that $\Pr\{\text{one or more arrivals in } (\tau, \tau+\delta)\} = m(\tau+\delta) - m(\tau) - o(\delta)$ where $o(\delta) \geq 0$ and $\lim_{\delta\to 0} o(\delta)/\delta = 0$.

b) Show that $\Pr\Big\{Z(t) \in [z, z+\delta),\ \widetilde{X}(t) \in (x, x+\delta)\Big\}$ is equal to $[m(t-z) - m(t-z-\delta) - o(\delta)][\mathsf{F}_X(x+\delta) - \mathsf{F}_X(x)]$ for $x \geq z + \delta$.

c) Assuming that $m'(\tau) = dm(\tau)/d\tau$ exists for all τ, show that the joint density of $Z(t)$, $\widetilde{X}(t)$ is $\mathsf{f}_{Z(t),\widetilde{X}(t)}(z, x) = m'(t-z)\mathsf{f}_X(x)$ for $x > z$.

d) Show that $\mathsf{E}[R(t)] = \int_{z=0}^{t}\int_{x=z}^{\infty} \mathcal{R}(z, x)\mathsf{f}_X(x)dx\, m'(t-z)dz$

Exercise 4.36. This problem is designed to give you an alternate way of looking at ensemble-averages for renewal-reward problems. First we find an exact expression for $\Pr\{S_{N(t)} > s\}$. We find this for arbitrary s and t, $0 < s < t$.

a) By breaking the event $\{S_{N(t)} > s\}$into subevents $\{S_{N(t)} > s, N(t) = n\}$, explain each of the following steps:

$$\begin{aligned}
\Pr\{S_{N(t)} > s\} &= \sum_{n=1}^{\infty} \Pr\{t \geq S_n > s, S_{n+1} > t\} \\
&= \sum_{n=1}^{\infty} \int_{y=s}^{t} \Pr\{S_{n+1}{>}t \mid S_n{=}y\} \; d\mathsf{F}_{S_n}(y) \\
&= \int_{y=s}^{t} \mathsf{F}^{\mathsf{c}}_X(t{-}y)\, d \sum_{n=1}^{\infty} \mathsf{F}_{S_n}(y) \\
&= \int_{y=s}^{t} \mathsf{F}^{\mathsf{c}}_X(t{-}y)\, dm(y) \quad \text{where } m(y) = \mathsf{E}\left[N(y)\right].
\end{aligned}$$

b) Show that for $0 < s < t < u$,

$$\Pr\{S_{N(t)} > s, S_{N(t)+1} > u\} = \int_{y=s}^{t} \mathsf{F}^{\mathsf{c}}_X(u-y)\, dm(y).$$

c) Draw a two dimensional sketch, with age and duration as the axes, and show the region of (age, duration) values corresponding to the event $\{S_N(t) > s,\ S_{N(t)+1} > u\}$.

d) Assume that for large t, $dm(y)$ can be approximated (according to Blackwell) as $(1/\overline{X})dy$, where $\overline{X} = \mathsf{E}\left[X\right]$. Assuming that X also has a density, use the result in parts b) and c) to find the joint density of age and duration.

Exercise 4.37. In this problem, we show how to calculate the residual life distribution $Y(t)$ as a transient in t. Let $\mu(t) = dm(t)/dt$ where $m(t) = \mathsf{E}\left[N(t)\right]$, and let the interarrival distribution have the density $\mathsf{f}_X(x)$. Let $Y(t)$ have the density $\mathsf{f}_{Y(t)}(y)$.

a) Show that these densities are related by the integral equation

$$\mu(t+y) = \mathsf{f}_{Y(t)}(y) + \int_{u=0}^{y} \mu(t+u)\mathsf{f}_X(y-u)du.$$

b) Let $L_{\mu,t}(r) = \int_{y\geq 0} \mu(t+y)e^{-ry}dy$ and let $L_{Y(t)}(r)$ and $L_X(r)$ be the Laplace transforms of $\mathsf{f}_{Y(t)}(y)$ and $\mathsf{f}_X(x)$ respectively. Find $L_{Y(t)}(r)$ as a function of $L_{\mu,t}$ and L_X.

c) Consider the inter-renewal density $\mathsf{f}_X(x) = (1/2)e^{-x} + e^{-2x}$ for $x \geq 0$ (as in Example 4.6.1). Find $L_{\mu,t}(r)$ and $L_{Y(t)}(r)$ for this example.

d) Find $\mathsf{f}_{Y(t)}(y)$. Show that your answer reduces to that of (4.28) in the limit as $t \to \infty$.

e) Explain how to go about finding $\mathsf{f}_{Y(t)}(y)$ in general, assuming that f_X has a rational Laplace transform.

Exercise 4.38. Show that for a G/G/1 queue, the time-average wait in the system is the same as $\lim_{n\to\infty} \mathsf{E}[W_n]$. Hint: Consider an integer renewal counting process $\{M(n); n \geq 0\}$ where $M(n)$ is the number of renewals in the G/G/1 process of Section 4.5.3 that have occurred by the nth arrival. Show that this renewal process has a span of 1. Then consider $\{W_n; n \geq 1\}$ as a reward within this renewal process.

Exercise 4.39. If one extends the definition of renewal processes to include inter-renewal intervals of duration 0, with $\Pr\{X{=}0\} = \alpha$, show that the expected number of simultaneous renewals at a renewal epoch is $1/(1-\alpha)$, and that, for a non-arithmetic process, the probability of 1 or more renewals in the interval $(t, t+\delta]$ tends to $(1-\alpha)\delta/\mathsf{E}[X] + o(\delta)$ as $t \to \infty$.

Exercise 4.40. The purpose of this exercise is to show why the interchange of expectation and sum in the proof of Wald's equality is justified when $\mathsf{E}[J] < \infty$ but not otherwise. Let $X_1, X_2, \ldots,$ be a sequence of IID rv's, each with the distribution F_X. Assume that $\mathsf{E}[|X|] < \infty$.

a) Show that $S_n = X_1 + \cdots + X_n$ is a rv for each integer $n > 0$. Note: S_n is obviously a mapping from the sample space to the real numbers, but you must show that it is finite with probability 1. Hint: Recall the additivity axiom for the real numbers.

b) Let J be a stopping trial for $X_1, X_2, \ldots$. Show that $S_J = X_1 + \cdots X_J$ is a rv. Hint: Represent $\Pr\{S_J\}$ as $\sum_{n=1}^{\infty} \Pr\{J = n\}\, S_n$.

c) For the stopping trial J above, let $J^{(k)} = \min(J, k)$ be the stopping trial J truncated to integer k. Explain why the interchange of sum and expectation in the proof of Wald's equality is justified in this case, so $\mathsf{E}[S_{J^{(k)}}] = \overline{X}\mathsf{E}\left[J^{(k)}\right]$.

d) In parts d), e), and f), assume, in addition to the assumptions above, that $\mathsf{F}_X(0) = 0$, *i.e.*, that the X_i are positive rv's. Show that $\lim_{k\to\infty} \mathsf{E}[S_{J^{(k)}}] < \infty$ if $\mathsf{E}[J] < \infty$ and $\lim_{k\to\infty} \mathsf{E}[S_{J^{(k)}}] = \infty$ if $\mathsf{E}[J] = \infty$.

e) Show that

$$\Pr\{S_{J^{(k)}} > x\} \leq \Pr\{S_J > x\}$$

for all k, x.

f) Show that $\mathsf{E}[S_J] = \overline{X}\mathsf{E}[J]$ if $\mathsf{E}[J] < \infty$ and $\mathsf{E}[S_J] = \infty$ if $\mathsf{E}[J] = \infty$.

g) Now assume that X has both negative and positive values with nonzero probability and let $X^+ = \max(0, X)$ and $X^- = \min(X, 0)$. Express S_J as $S_J^+ + S_J^-$ where $S_J^+ = \sum_{i=1}^{J} X_i^+$ and $S_J^- = \sum_{i=1}^{J} X_i^-$. Show that $\mathsf{E}[S_J] = \overline{X}\mathsf{E}[J]$ if $\mathsf{E}[J] < \infty$ and that $\mathsf{E}[S_j]$ is undefined otherwise.

Exercise 4.41. This is a very simple exercise designed to clarify confusion about the roles of past, present, and future in stopping rules. Let $\{X_n;\, n \geq 1\}$ be a sequence of IID binary

rv's , each with the pmf $\mathsf{p}_X(1) = 1/2$, $\mathsf{p}_X(0) = 1/2$. Let J be a positive integer-valued rv that takes on the sample value n of the first trial for which $X_n = 1$. That is, for each $n \geq 1$,

$$\{J = n\} = \{X_1{=}0,\ X_2{=}0, \ldots, X_{n-1}{=}0,\ X_n{=}1\}.$$

a) Use the definition of stopping trial, Definition 4.5.1 in the text, to show that J is a stopping trial for $\{X_n;\, n \geq 1\}$.

b) Show that for any given n, the rv's X_n and $\mathbb{I}_{J=n}$ are *statistically dependent*.

c) Show that for every $m > n$, X_n and $\mathbb{I}_{J=m}$ are *statistically dependent*.

d) Show that for every $m < n$, X_n and $\mathbb{I}_{J=m}$ are *statistically independent*.

e) Show that X_n and $\mathbb{I}_{J\geq n}$ are *statistically independent*. Give the simplest characterization you can of the event $\{J \geq n\}$.

f) Show that X_n and $\mathbb{I}_{J>n}$ are *statistically dependent*.

Note: The results here are characteristic of most sequences of IID rv's. For most people, this requires some realignment of intuition, since $\{J{\geq}n\}$ is the union of $\{J = m\}$ for all $m \geq n$, and all of these events are highly dependent on X_n. The right way to think of this is that $\{J{\geq}n\}$ is the complement of $\{J{<}n\}$, which is determined by $X_1, \ldots, X_{n-1}$. Thus $\{J{\geq}n\}$ is also determined by $X_1, \ldots, X_{n-1}$ and is thus independent of X_n. The moral of the story is that thinking of stopping rules as rv's independent of the future is very tricky, even in totally obvious cases such as this.

Exercise 4.42. Assume a friend has developed an excellent program for finding the steady-state probabilities for finite-state Markov chains. More precisely, given the transition matrix [P], the program returns $\lim_{n\to\infty} P_{ii}^n$ for each i. Assume all chains are aperiodic.

a) You want to find the expected time to first reach a given state k starting from a different state m for a Markov chain with transition matrix $[P]$. You modify the matrix to $[P']$ where $P'_{km} = 1$, $P'_{kj} = 0$ for $j \neq m$, and $P'_{ij} = P_{ij}$ otherwise. How do you find the desired first-passage time from the program output given $[P']$ as an input? (Hint: The times at which a Markov chain enters any given state can be considered as renewals in a (perhaps delayed) renewal process).

b) Using the same $[P']$ as the program input, how can you find the expected number of returns to state m before the first passage to state k?

c) Suppose, for the same Markov chain $[P]$ and the same starting state m, you want to find the probability of reaching some given state n before the first-passage to k. Modify $[P]$ to some $[P'']$ so that the above program with P'' as an input allows you to easily find the desired probability.

d) Let $\Pr\{X(0) = i\} = Q_i$, $1 \leq i \leq \mathsf{M}$ be an arbitrary set of initial probabilities for the same Markov chain $[P]$ as above. Show how to modify $[P]$ to some $[P''']$ for which the steady-state probabilities allow you to easily find the expected time of the first-passage to state k.

Exercise 4.43. Consider a ferry that carries cars across a river. The ferry holds an integer number k of cars and departs the dock when full. At that time, a new ferry immediately appears and begins loading newly arriving cars ad infinitum. The ferry business has been good, but customers complain about the long wait for the ferry to fill up.

a) Assume that cars arrive according to a renewal process. The IID interarrival times have mean $\overline{X}$, variance σ^2 and moment generating function $g_X(r)$. Does the sequence of departure times of the ferries form a renewal process? Explain carefully.

b) Find the expected time that a customer waits, starting from its arrival at the ferry terminal and ending at the departure of its ferry. Note 1: Part of the problem here is to give a reasonable definition of the expected customer waiting time. Note 2: It might be useful to consider $k = 1$ and $k = 2$ first.

c Is there a 'slow truck' phenomenon (a dependence on $\mathsf{E}\left[X^2\right]$) here? Give an intuitive explanation. Hint: Look at $k = 1$ and $k = 2$ again.

d) In an effort to decrease waiting, the ferry managers institute a policy where no customer ever has to wait more than one hour. Thus, the first customer to arrive after a ferry departure waits for either one hour or the time at which the ferry is full, whichever comes first, and then the ferry leaves and a new ferry starts to accumulate new customers. Does the sequence of ferry departures form a renewal process under this new system? Does the sequence of times at which each successive empty ferry is entered by its first customer form a renewal process? You can assume here that $t = 0$ is the time of the first arrival to the first ferry. Explain carefully.

e) Give an expression for the expected waiting time of the first new customer to enter an empty ferry under this new strategy.

Chapter 5

COUNTABLE-STATE MARKOV CHAINS

5.1 Introduction and classification of states

Markov chains with a countably-infinite state space (more briefly, *countable-state Markov chains*) exhibit some types of behavior not possible for chains with a finite state space. With the exception of the first example to follow and the section on branching processes, we label the states by the nonnegative integers. This is appropriate when modeling things such as the number of customers in a queue, and causes no loss of generality in other cases.

The following two examples give some insight into the new issues posed by countable state spaces.

Example 5.1.1. Consider the familiar Bernoulli process $\{S_n = X_1 + \cdots X_n;\ n \geq 1\}$ where $\{X_n;\ n \geq 1\}$ is an IID binary sequence with $\mathsf{p}_X(1) = p$ and $\mathsf{p}_X(-1) = (1-p) = q$. The sequence $\{S_n;\ n \geq 1\}$ is a sequence of integer random variables (rv's) where $S_n = S_{n-1} + 1$ with probability p and $S_n = S_{n-1} - 1$ with probability q. This sequence can be modeled by the Markov chain in Figure 5.1.

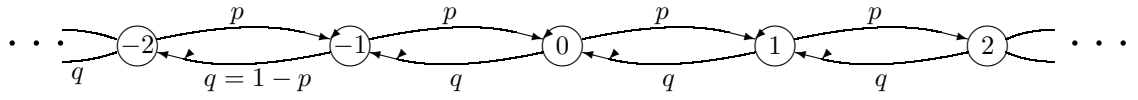


Figure 5.1: A Markov chain with a countable state space modeling a Bernoulli process. If $p > 1/2$, then as time n increases, the state X_n becomes large with high probability, *i.e.*, $\lim_{n\to\infty} \Pr\{X_n \geq j\} = 1$ for each integer j. Similarly, for $p < 1/2$, the state becomes highly negative.

Using the notation of Markov chains, P^n_{0j} is the probability of being in state j at the end of the nth transition, conditional on starting in state 0. The final state j is the number of positive transitions k less the number of negative transitions $n - k$, *i.e.*, $j = 2k - n$. Thus,

using the binomial formula,

$$P_{0j}^n = \binom{n}{k} p^k q^{n-k} \qquad \text{where } k = \frac{j+n}{2}; \quad j+n \text{ even.} \tag{5.1}$$

All states in this Markov chain communicate with all other states, and are thus in the same class. The formula makes it clear that this class, *i.e.*, the entire set of states in the Markov chain, is periodic with period 2. For n even, the state is even and for n odd, the state is odd.

What is more important than the periodicity, however, is what happens to the state probabilities for large n. As we saw in (1.88) while proving the central limit theorem for the binomial case,

$$P_{0j}^n \sim \frac{1}{\sqrt{2\pi npq}} \exp\left[\frac{-(k-np)^2}{2pqn}\right]\} \qquad \text{where } k = \frac{j+n}{2}; \quad j+n \text{ even.} \tag{5.2}$$

In other words, P_{0j}^n, as a function of j, looks like a quantized form of the Gaussian density for large n. The significant terms of that distribution are close to $k = np$, *i.e.*, to $j = n(2p-1)$. For $p > 1/2$, the state increases with increasing n. Its distribution is centered at $n(2p-1)$, but the distribution is also spreading out as $\sqrt{n}$. For $p < 1/2$, the state similarly decreases and spreads out. The most interesting case is $p = 1/2$, where the distribution remains centered at 0, but due to the spreading, the PMF approaches 0 as $1/\sqrt{n}$ for all j.

For this example, then, the probability of each state approaches zero as $n \to \infty$, and this holds for all choices of p, $0 < p < 1$. If we attempt to define a steady-state probability as 0 for each state, then these probabilities do not sum to 1, so they cannot be viewed as a steady-state distribution. Thus, for countable-state Markov chains, the notions of recurrence and steady-state probabilities will have to be modified from that with finite-state Markov chains. The same type of situation occurs whenever $\{S_n;\, n \geq 1\}$ is a sequence of sums of arbitrary IID integer-valued rv's.

Most countable-state Markov chains that are useful in applications are quite different from Example 5.1.1, and instead are quite similar to finite-state Markov chains. The following example bears a close resemblance to Example 5.1.1, but at the same time is a countable-state Markov chain that will keep reappearing in a large number of contexts. It is a special case of a birth-death process, which we study in Section 5.2.

Example 5.1.2. Figure 5.2 is similar to Figure 5.1 except that the negative states have been eliminated. A sequence of IID binary rv's $\{X_n;\, n \geq 1\}$, with $\mathsf{p}_X(1) = p$ and $\mathsf{p}_X(-1) = q = 1-p$, controls the state transitions. Now, however, $S_n = \max(0, S_{n-1} + X_n$, so that S_n is a nonnegative rv. All states again communicate, and because of the self transition at state 0, the chain is aperiodic.

For $p > 1/2$, transitions to the right occur with higher frequency than transitions to the left. Thus, reasoning heuristically, we expect the state S_n at time n to drift to the right with increasing n. Given $S_0 = 0$, the probability P_{0j}^n of being in state j at time n, should then tend to zero for any fixed j with increasing n. As in Example 5.1.1, we see that a

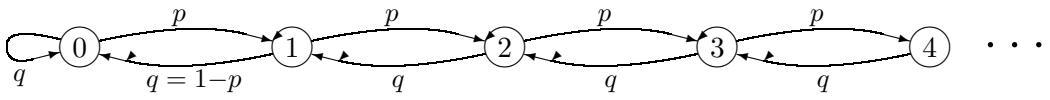


Figure 5.2: A Markov chain with a countable state space. If $p > 1/2$, then as time n increases, the state X_n becomes large with high probability, *i.e.*, $\lim_{n\to\infty} \Pr\{X_n \geq j\} = 1$ for each integer j.

steady state does not exist. In more poetic terms, the state wanders off into the wild blue yonder.

One way to understand this chain better is to look at what happens if the chain is truncated The truncation of Figure 5.2 to k states is analyzed in Exercise 3.9. The solution there defines $\rho = p/q$ and shows that if $\rho \neq 1$, then $\pi_i = (1-\rho)\rho^i/(1-\rho^k)$ for each i, $0 \leq i < k$. For $\rho = 1$, $\pi_i = 1/k$ for each i. For $\rho < 1$, the limiting behavior as $k \to \infty$ is $\pi_i = (1-\rho)\rho^i$. Thus for $\rho < 1$ ($p < 1/2$), the steady state probabilities for the truncated Markov chain approaches a limit which we later interpret as the steady state probabilities for the untruncated chain. For $\rho > 1$ ($p > 1/2$), on the other hand, the steady-state probabilities for the truncated case are geometrically decreasing from the right, and the states with significant probability keep moving to the right as k increases. Although the probability of each fixed state j approaches 0 as k increases, the truncated chain never resembles the untruncated chain.

Perhaps the most interesting case is that where $p = 1/2$. The nth order transition probabilities, P^n_{0j} can be calculated exactly for this case (see Exercise 5.3) and are very similar to those of Example 5.1.1. In particular,

$$P^n_{0j} = \begin{cases} \binom{n}{(j+n)/2} 2^{-n} & \text{for } j \geq 0, (j+n) \text{ even} \\ \} & \\ \binom{n}{(j+n+1)/2} 2^{-n} & \text{for } j \geq 0, (j+n) \text{ odd} \end{cases} \tag{5.3}$$

$$\sim \sqrt{\frac{2}{\pi n}} \exp\left[\frac{-j^2}{2n}\right] \qquad \text{for } j \geq 0. \tag{5.4}$$

We see that P^n_{0j} for large n is approximated by the positive side of a quantized Gaussian distribution. It looks like the positive side of the PMF of (5.1) except that it is no longer periodic. For large n, P^n_{0j} is concentrated in a region of width $\sqrt{n}$ around $j = 0$, and the PMF goes to 0 as $1/\sqrt{n}$ for each j as $n \to \infty$.

Fortunately, the strange behavior of Figure 5.2 when $p \geq q$ is not typical of the Markov chains of interest for most applications. For typical countable-state Markov chains, a steady state does exist, and the steady-state probabilities of all but a finite number of states (the number depending on the chain and the application) can almost be ignored for numerical calculations.

5.1.1 Using renewal theory to classify and analyze Markov chains

The matrix approach used to analyze finite-state Markov chains does not generalize easily to the countable-state case. Fortunately, renewal theory is ideally suited for this purpose, especially for analyzing the long term behavior of countable-state Markov chains. We must first revise the definition of recurrent states. The definition for finite-state Markov chains does not apply here, and we will see that, under the new definition, the Markov chain in Figure 5.2 is recurrent for $p \leq 1/2$ and transient for $p > 1/2$. For $p = 1/2$, the chain is called null-recurrent, as explained later.

In general, we will find that for a recurrent state j, the sequence of subsequent entries to state j, conditional on starting in j, forms a renewal process. The renewal theorems then specify the time-average relative-frequency of state j, the limiting probability of j with increasing time and a number of other relations.

We also want to understand the sequence of epochs at which one state, say j, is entered, conditional on starting the chain at some other state, say i. We will see that, subject to the classification of states i and j, this gives rise to a delayed renewal process. In preparing to study these renewal processes and delayed renewal process, we need to understand the inter-renewal intervals. The probability mass functions (PMF's) of these intervals are called *first-passage-time probabilities* in the notation of Markov chains.

Definition 5.1.1. *The first-passage-time probability, $f_{ij}(n)$, of a Markov chain is the probability, conditional on $X_0 = i$, that the first subsequent entry to state j occurs at discrete epoch n. That is, $f_{ij}(1) = P_{ij}$ and for $n \geq 2$,*

$$f_{ij}(n) = \Pr\{X_n{=}j, X_{n-1}{\neq}j, X_{n-2}{\neq}j, \ldots, X_1{\neq}j | X_0{=}i\}. \tag{5.5}$$

The distinction between $f_{ij}(n)$ and $P_{ij}^n = \Pr\{X_n = j | X_0 = i\}$ is that $f_{ij}(n)$ is the probability that the *first* entry to j (after time 0) occurs at time n, whereas P_{ij}^n is the probability that *any* entry to j occurs at time n, both conditional on starting in state i at time 0. The definition in (5.5) also applies for $j = i$; $f_{ii}(n)$ is thus the probability, given $X_0 = i$, that the first occurrence of state i after time 0 occurs at time n. Since the transition probabilities are independent of time, $f_{kj}(n-1)$ is also the probability, given $X_1 = k$, that the first subsequent occurrence of state j occurs at time n. Thus we can calculate $f_{ij}(n)$ from the iterative relations

$$f_{ij}(n) = \sum_{k \neq j} P_{ik} f_{kj}(n-1); \quad n > 1; \qquad f_{ij}(1) = P_{ij}. \tag{5.6}$$

Note that the sum excludes $k = j$, since $P_{ij} f_{jj}(n{-}1)$ is the probability that state j occurs first at epoch 1 and next at epoch n. Note also from the Chapman-Kolmogorov equation that $P_{ij}^n = \sum_k P_{ik} P_{kj}^{n-1}$. In other words, the only difference between the iterative expressions to calculate $f_{ij}(n)$ and P_{ij}^n is the exclusion of $k = j$ in the expression for $f_{ij}(n)$.

With this iterative approach, the first-passage-time probabilities $f_{ij}(n)$ for a given n must be calculated for all i before proceeding to calculate them for the next larger value of n. This also gives us $f_{jj}(n)$, although $f_{jj}(n)$ is not used in the iteration.

Let $\mathsf{F}_{ij}(n)$, for $n \geq 1$, be the probability, given $X_0 = i$, that state j occurs at some time between 1 and n inclusive. Thus,

$$\mathsf{F}_{ij}(n) = \sum_{m=1}^{n} \} f_{ij}(m). \tag{5.7}$$

For each i, j, $\mathsf{F}_{ij}(n)$ is non-decreasing in n and (since it is a probability) is upper bounded by 1. Thus $\mathsf{F}_{ij}(\infty)$ (*i.e.*, $\lim_{n\to\infty} \mathsf{F}_{ij}(n)$) must exist, and is the probability, given $X_0 = i$, that state j will ever occur. If $\mathsf{F}_{ij}(\infty) = 1$, then, given $X_0 = i$, it is certain (with probability 1) that the chain will eventually enter state j. In this case, we can define a random variable (rv) T_{ij}, conditional on $X_0 = i$, as the *first-passage time* from i to j. Then $f_{ij}(n)$ is the PMF of T_{ij} and $\mathsf{F}_{ij}(n)$ is the distribution function of T_{ij}. If $\mathsf{F}_{ij}(\infty) < 1$, then T_{ij} is a defective rv, since, with some non-zero probability, there is no first-passage to j. Defective rv's are not considered to be rv's (in the theorems here or elsewhere), but they do have many of the properties of rv's.

The first-passage time T_{jj} from a state j back to itself is of particular importance. It has the PMF $f_{jj}(n)$ and the distribution function $\mathsf{F}_{jj}(n)$. It is a rv (as opposed to a defective rv) if $\mathsf{F}_{jj}(\infty) = 1$, *i.e.*, if the state eventually returns to state j with probability 1 given that it starts in state j. This leads to the definition of recurrence.

Definition 5.1.2. *A state j in a countable-state Markov chain is* recurrent *if $\mathsf{F}_{jj}(\infty) = 1$. It is* transient *if $\mathsf{F}_{jj}(\infty) < 1$.*

Thus each state j in a countable-state Markov chain is either recurrent or transient, and is recurrent if and only if an eventual return to j (conditional on $X_0 = j$) occurs with probability 1. Equivalently, j is recurrent if and only if T_{jj}, the time of first return to j, is a rv. Note that for the special case of finite-state Markov chains, this definition is consistent with the one in Chapter 3. For a countably-infinite state space, however, the earlier definition is not adequate. An example is provided by the case $p > 1/2$ in Figure 5.2. Here i and j communicate for all states i and j, but it is intuitively obvious (and shown in Exercise 5.2, and further explained in Section 5.2) that each state is transient.

If the initial state X_0 of a Markov chain is a recurrent state j, then T_{jj} is the integer time of the first recurrence of state j. At that recurrence, the Markov chain is in the same state j as it started in, and the discrete interval from T_{jj} to the next occurrence of state j, say $T_{jj,2}$ has the same distribution as T_{jj} and is clearly independent of T_{jj}. Similarly, the sequence of successive recurrence intervals, $T_{jj}, T_{jj,2}, T_{jj,3}, \ldots$ is a sequence of IID rv's. This sequence of recurrence intervals[1] is then the sequence of inter-renewal intervals of a renewal process, where each renewal interval has the distribution of T_{jj}. These inter-renewal intervals have the PMF $f_{jj}(n)$ and the distribution function $\mathsf{F}_{jj}(n)$.

Since results about Markov chains depend very heavily on whether states are recurrent or transient, we will look carefully at the probabilities $\mathsf{F}_{ij}(n)$. Substituting (5.6) into (5.7), we

[1]Note that in Chapter 4 the inter-renewal intervals were denoted $X_1, X_2, \ldots$, whereas here $X_0, X_1, \ldots$, is the sequence of states in the Markov chain and $T_{jj}, T_{jj,2}, \ldots$, is the sequence of inter-renewal intervals.

obtain

$$\mathsf{F}_{ij}(n) = P_{ij} + \sum_{k\neq j} P_{ik}\mathsf{F}_{kj}(n-1); \quad n>1; \qquad \mathsf{F}_{ij}(1) = P_{ij}. \tag{5.8}$$

To understand the expression $P_{ij} + \sum_{k\neq j} P_{ik}\mathsf{F}_{kj}(n-1)$, note that the first term, P_{ij}, is $f_{ij}(1)$ and the second term, $\sum_{k\neq j} P_{ik}\mathsf{F}_{kj}(n-1)$, is equal to $\sum_{\ell=2}^{n} f_{ij}(\ell)$.

We have seen that $\mathsf{F}_{ij}(n)$ is non-decreasing in n and upper bounded by 1, so the limit $\mathsf{F}_{ij}(\infty)$ must exist. Similarly, $\sum_{k\neq j} P_{ik}\mathsf{F}_{kj}(n-1)$ is non-decreasing in n and upper bounded by 1, so it also has a limit, equal to $\sum_{k\neq j} P_{ik}\mathsf{F}_{kj}(\infty)$. Thus

$$\mathsf{F}_{ij}(\infty) = P_{ij} + \sum_{k\neq j} P_{ik}\mathsf{F}_{kj}(\infty). \tag{5.9}$$

For any given j, (5.9) can be viewed as a set of linear equations in the variables $\mathsf{F}_{ij}(\infty)$ for each state i. There is not always a unique solution to this set of equations. In fact, the set of equations

$$x_{ij} = P_{ij} + \sum_{k\neq j} P_{ik}x_{kj}; \qquad \text{all states } i \tag{5.10}$$

always has a solution in which $x_{ij} = 1$ for all i. If state j is transient, however, there is another solution in which x_{ij} is the true value of $\mathsf{F}_{ij}(\infty)$ and $\mathsf{F}_{jj}(\infty) < 1$. Exercise 5.1 shows that if (5.10) is satisfied by a set of nonnegative numbers $\{x_{ij}; 1 \leq i \leq J\}$, then $\mathsf{F}_{ij}(\infty) \leq x_{ij}$ for each i.

We have defined a state j to be recurrent if $\mathsf{F}_{jj}(\infty) = 1$ and have seen that if j is recurrent, then the returns to state j, given $X_0 = j$ form a renewal process. All of the results of renewal theory can then be applied to the random sequence of integer times at which j is entered. The main results from renewal theory that we need are stated in the following lemma.

Lemma 5.1.1. *Let $\{N_{jj}(t); t \geq 0\}$ be the counting process for occurrences of state j up to time t in a Markov chain with $X_0 = j$. The following conditions are then equivalent.*

1. *state j is recurrent.*
2. $\lim_{t\to\infty} N_{jj}(t) = \infty$ *with probability* 1.
3. $\lim_{t\to\infty} \mathsf{E}\left[N_{jj}(t)\right] = \infty$.
4. $\lim_{t\to\infty} \sum_{1\leq n\leq t} P_{jj}^n = \infty$.

Proof: First assume that j is recurrent, *i.e.*, that $\mathsf{F}_{jj}(\infty) = 1$. This implies that the inter-renewal times between occurrences of j are IID rv's, and consequently $\{N_{jj}(t);\, t \geq 1\}$ is a renewal counting process. Recall from Lemma 4.3.1 of Chapter 4 that, whether or not the expected inter-renewal time $\mathsf{E}\left[T_{jj}\right]$ is finite, $\lim_{t\to\infty} N_{jj}(t) = \infty$ with probability 1 and $\lim_{t\to\infty} \mathsf{E}\left[N_{jj}(t)\right] = \infty$.

Next assume that state j is transient. In this case, the inter-renewal time T_{jj} is not a rv, so $\{N_{jj}(t); t \geq 0\}$ is not a renewal process. An eventual return to state j occurs only with probability $\mathsf{F}_{jj}(\infty) < 1$, and, since subsequent returns are independent, the total number of returns to state j is a geometric rv with mean $\mathsf{F}_{jj}(\infty)/[1-\mathsf{F}_{jj}(\infty)]$. Thus the total number of returns is finite with probability 1 and the expected total number of returns is finite. This establishes the first three equivalences.

Finally, note that P^n_{jj} , the probability of a transition to state j at integer time n, is equal to the expectation of a transition to j at integer time n (*i.e.*, a single transition occurs with probability P^n_{jj} and 0 occurs otherwise). Since $N_{jj}(t)$ is the sum of the number of transitions to j over times 1 to t, we have

$$\mathsf{E}\left[N_{jj}(t)\right] = \sum_{1\leq n\leq t} P^n_{jj},$$

which establishes the final equivalence. □

Our next objective is to show that all states in the same class as a recurrent state are also recurrent. Recall that two states are in the same class if they communicate, *i.e.*, each has a path to the other. For finite-state Markov chains, the fact that either all states in the same class are recurrent or all transient was relatively obvious, but for countable-state Markov chains, the definition of recurrence has been changed and the above fact is no longer obvious.

Lemma 5.1.2. *If state j is recurrent and states i and j are in the same class, i.e., i and j communicate, then state i is also recurrent.*

Proof: From Lemma 5.1.1, state j satisfies $\lim_{t\to\infty} \sum_{1\leq n\leq t} P^n_{jj} = \infty$. Since j and i communicate, there are integers m and k such that $P^m_{ij} > 0$ and $P^k_{ji} > 0$. For every walk from state j to j in n steps, there is a corresponding walk from i to i in $m+n+k$ steps, going from i to j in m steps, j to j in n steps, and j back to i in k steps. Thus

$$\begin{aligned} P^{m+n+k}_{ii} &\geq P^m_{ij}\, P^n_{jj}\, P^k_{ji} \\ \sum_{n=1}^{\infty} P^n_{ii} &\geq \sum_{n=1}^{\infty} P^{m+n+k}_{ii} \geq P^m_{ij}\, P^k_{ji} \sum_{n=1}^{\infty} P^n_{jj} = \infty. \end{aligned}$$

Thus, from Lemma 5.1.1, i is recurrent, completing the proof. □

Since each state in a Markov chain is either recurrent or transient, and since, if one state in a class is recurrent, all states in that class are recurrent, we see that if one state in a class is transient, they all are. Thus we can refer to each class as being recurrent or transient. This result shows that Theorem 3.2.1 also applies to countable-state Markov chains. We state this theorem separately here to be specific.

Theorem 5.1.1. *For a countable-state Markov chain, either all states in a class are transient or all are recurrent.*

We next look at the delayed counting process $\{N_{ij}(n);\, n \geq 1\}$. If this is a delayed *renewal* counting process, then we can use delayed renewal processes to study whether the effect of

the initial state eventually dies out. If state j is recurrent, we know that $\{N_{jj}(n);\, n \geq 1\}$ is a renewal counting process. In addition, in order for $\{N_{ij}(n);\, n \geq 1\}$ to be a delayed renewal counting process, it is necessary for the first-passage time to be a rv, *i.e.*, for $\mathsf{F}_{ij}(\infty)$ to be 1.

Lemma 5.1.3. *Let states i and j be in the same recurrent class. Then $\mathsf{F}_{ij}(\infty) = 1$.*

Proof: Since i is recurrent, the number of visits to i by time t, given $X_0 = i$, is a renewal counting process $N_{ii}(t)$. There is a path from i to j, say of probability $\alpha > 0$. Thus the probability that the first return to i occurs before visiting j is at most $1-\alpha$. The probability that the second return occurs before visiting j is thus at most $(1-\alpha)^2$ and the probability that the nth occurs without visiting j is at most $(1-\alpha)^n$. Since i is visited infinitely often with probability 1 as $n \to \infty$, the probability that j is never visited is 0. Thus $\mathsf{F}_{ij}(\infty) = 1$. □

Lemma 5.1.4. *Let $\{N_{ij}(t); t \geq 0\}$ be the counting process for transitions into state j up to time t for a Markov chain given $X_0 = i \neq j$. Then if i and j are in the same recurrent class, $\{N_{ij}(t); t \geq 0\}$ is a delayed renewal process.*

Proof: From Lemma 5.1.3, T_{ij}, the time until the first transition into j, is a rv. Also T_{jj} is a rv by definition of recurrence, and subsequent intervals between occurrences of state j are IID, completing the proof. □

If $\mathsf{F}_{ij}(\infty) = 1$, we have seen that the first-passage time from i to j is a rv, *i.e.*, is finite with probability 1. In this case, the mean time $\overline{T}_{ij}$ to first enter state j starting from state i is of interest. Since T_{ij} is a nonnegative random variable, its expectation is the integral of its complementary distribution function,

$$\overline{T}_{ij} = 1 + \sum_{n=1}^{\infty} (1 - \mathsf{F}_{ij}(n)). \tag{5.11}$$

It is possible to have $\mathsf{F}_{ij}(\infty) = 1$ but $\overline{T}_{ij} = \infty$. As will be shown in Section 5.2, the chain in Figure 5.2 satisfies $\mathsf{F}_{ij}(\infty) = 1$ and $\overline{T}_{ij} < \infty$ for $p < 1/2$ and $\mathsf{F}_{ij}(\infty) = 1$ and $\overline{T}_{ij} = \infty$ for $p = 1/2$. As discussed before, $\mathsf{F}_{ij}(\infty) < 1$ for $p > 1/2$. This leads us to the following definition.

Definition 5.1.3. *A state j in a countable-state Markov chain is* positive-recurrent *if $\mathsf{F}_{jj}(\infty) = 1$ and $\overline{T}_{jj} < \infty$. It is* null-recurrent *if $\mathsf{F}_{jj}(\infty) = 1$ and $\overline{T}_{jj} = \infty$.*

Each state of a Markov chain is thus classified as one of the following three types — positive-recurrent, null-recurrent, or transient. For the example of Figure 5.2, null-recurrence lies on a boundary between positive-recurrence and transience, and this is often a good way to look at null-recurrence. Part f) of Exercise 6.3 illustrates another type of situation in which null-recurrence can occur.

Assume that state j is recurrent and consider the renewal process $\{N_{jj}(t); t \geq 0\}$. The limiting theorems for renewal processes can be applied directly. From the strong law for

renewal processes, Theorem 4.3.1,

$$\lim_{t\to\infty} N_{jj}(t)/t = 1/\overline{T}_{jj} \qquad \text{with probability 1.} \tag{5.12}$$

From the elementary renewal theorem, Theorem 4.6.1,

$$\lim_{t\to\infty} \mathsf{E}\left[N_{jj}(t)/t\right] = 1/\overline{T}_{jj}. \tag{5.13}$$

Equations (5.12) and (5.13) are valid whether j is positive-recurrent or null-recurrent.

Next we apply Blackwell's theorem to $\{N_{jj}(t); t \geq 0\}$. Recall that the period of a given state j in a Markov chain (whether the chain has a countable or finite number of states) is the greatest common divisor of the set of integers $n > 0$ such that $P^n_{jj} > 0$. If this period is d, then $\{N_{jj}(t); t \geq 0\}$ is arithmetic with span λ (*i.e.*, renewals occur only at times that are multiples of λ). From Blackwell's theorem in the arithmetic form of (4.61),

$$\lim_{n\to\infty} \Pr\{X_{n\lambda} = j \mid X_0 = j\} = \lambda/\overline{T}_{jj}. \tag{5.14}$$

If state j is aperiodic (*i.e.*, $\lambda = 1$), this says that $\lim_{n\to\infty} \Pr\{X_n = j \mid X_0 = j\} = 1/\overline{T}_{jj}$. Equations (5.12) and (5.13) suggest that $1/\overline{T}_{jj}$ has some of the properties associated with a steady-state probability of state j, and (5.14) strengthens this if j is aperiodic. For a Markov chain consisting of a single class of states, all positive-recurrent, we will strengthen this association further in Theorem 5.1.4 by showing that there is a unique *steady-state distribution*, $\{\pi_j, j \geq 0\}$ such that $\pi_j = 1/\overline{T}_{jj}$ for all j and such that $\pi_j = \sum_i \pi_i P_{ij}$ for all $j \geq 0$ and $\sum_j \pi_j = 1$. The following theorem starts this development by showing that (5.12-5.14) are independent of the starting state.

Theorem 5.1.2. *Let j be a recurrent state in a Markov chain and let i be any state in the same class as j. Given $X_0 = i$, let $N_{ij}(t)$ be the number of transitions into state j by time t and let $\overline{T}_{jj}$ be the expected recurrence time of state j (either finite or infinite). Then*

$$\lim_{t\to\infty} N_{ij}(t)/t = 1/\overline{T}_{jj} \qquad \textit{with probability 1} \tag{5.15}$$

$$\lim_{t\to\infty} \mathsf{E}\left[N_{ij}(t)/t\right] = 1/\overline{T}_{jj}. \tag{5.16}$$

If j is also aperiodic, then

$$\lim_{n\to\infty} \Pr\{X_n = j \mid X_0 = i\} = 1/\overline{T}_{jj}. \tag{5.17}$$

Proof: Since i and j are recurrent and in the same class, Lemma 5.1.4 asserts that $\{N_{ij}(t); t \geq 0\}$ is a delayed renewal process for $j \neq i$. Thus (5.15) and (5.16) follow from Theorems 4.8.1 and 4.8.2 of Chapter 4. If j is aperiodic, then $\{N_{ij}(t); t \geq 0\}$ is a delayed renewal process for which the inter-renewal intervals T_{jj} have span 1 and T_{ij} has an integer span. Thus, (5.17) follows from Blackwell's theorem for delayed renewal processes, Theorem 4.8.3. For $i = j$, (5.15-5.17) follow from (5.12-5.14), completing the proof. □

Theorem 5.1.3. *All states in the same class of a Markov chain are of the same type — either all positive-recurrent, all null-recurrent, or all transient.*

Proof: Let j be a recurrent state. From Theorem 5.1.1, all states in a class are recurrent or all are transient. Next suppose that j is positive-recurrent, so that $1/\overline{T}_{jj} > 0$. Let i be in the same class as j, and consider the renewal-reward process on $\{N_{jj}(t); t \geq 0\}$ for which $R(t) = 1$ whenever the process is in state i (*i.e.*, if $X_n = i$, then $R(t) = 1$ for $n \leq t < n+1$). The reward is 0 whenever the process is in some state other than i. Let $\mathsf{E}[\mathrm{R}_n]$ be the expected reward in an inter-renewal interval; this must be positive since i is accessible from j. From the strong law for renewal-reward processes, Theorem 4.4.1,

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t R(\tau)d\tau = \frac{\mathsf{E}[\mathrm{R}_n]}{\overline{T}_{jj}} \qquad \text{with probability 1.}$$

The term on the left is the time-average number of transitions into state i, given $X_0 = j$, and this is $1/\overline{T}_{ii}$ from (5.15). Since $\mathsf{E}[\mathrm{R}_n] > 0$ and $\overline{T}_{jj} < \infty$, we have $1/\overline{T}_{ii} > 0$, so i is positive-recurrent. Thus if one state is positive-recurrent, the entire class is, completing the proof. □

If all of the states in a Markov chain are in a null-recurrent class, then $1/\overline{T}_{jj} = 0$ for each state, and one might think of $1/\overline{T}_{jj} = 0$ as a "steady-state" probability for j in the sense that 0 is both the time-average rate of occurrence of j and the limiting probability of j. However, these "probabilities" do not add up to 1, so a steady-state probability *distribution* does not exist. This appears rather paradoxical at first, but the example of Figure 5.2, with $p = 1/2$ will help to clarify the situation. As time n increases (starting in state i, say), the random variable X_n spreads out over more and more states around i, and thus is less likely to be in each individual state. For each j, $\lim_{n\to\infty} P_{ij}(n) = 0$. Thus, $\sum_j \{\lim_{n\to\infty} P_{ij}^n\} = 0$. On the other hand, for every n, $\sum_j P_{ij}^n = 1$. This is one of those unusual examples where a limit and a sum cannot be interchanged.

In Chapter 3, we defined the steady-state distribution of a finite-state Markov chain as a probability vector $\boldsymbol{\pi}$ that satisfies $\boldsymbol{\pi} = \boldsymbol{\pi}[P]$. Here we define $\{\pi_i; i \geq 0\}$ in the same way, as a set of numbers that satisfy

$$\pi_j = \sum_i \}\pi_i P_{ij} \quad \text{for all } j; \qquad \pi_j \geq 0 \quad \text{for all } j; \qquad \sum_j \pi_j = 1. \tag{5.18}$$

Suppose that a set of numbers $\{\pi_i; i \geq 0\}$ satisfying (5.18) is chosen as the initial probability distribution for a Markov chain, *i.e.*, if $\Pr\{X_0 = i\} = \pi_i$ for all i. Then $\Pr\{X_1 = j\} = \sum_i \pi_i P_{ij} = \pi_j$ for all j, and, by induction, $\Pr\{X_n = j\} = \pi_j$ for all j and all $n \geq 0$. The fact that $\Pr\{X_n = j\} = \pi_j$ for all j motivates the definition of steady-state distribution above. Theorem 5.1.2 showed that $1/\overline{T}_{jj}$ is a 'steady-state' probability for state j, both in a time-average and a limiting ensemble-average sense. The following theorem brings these ideas together. An irreducible Markov chain is a Markov chain in which all pairs of states communicate. For finite-state chains, irreducibility implied a single class of recurrent states, whereas for countably infinite chains, an irreducible chain is a single class that can be transient, null-recurrent, or positive-recurrent.

Theorem 5.1.4. *Assume an irreducible Markov chain with transition probabilities $\{P_{ij}\}$. If (5.18) has a solution, then the solution is unique, $\pi_i = 1/\overline{T}_{ii} > 0$ for all $i \geq 0$, and the states are positive-recurrent. Also, if the states are positive-recurrent then (5.18) has a solution.*

Proof*: Let $\{\pi_j; j \geq 0\}$ satisfy (5.18) and be the initial distribution of the Markov chain, *i.e.*, $\Pr\{X_0{=}j\} = \pi_j$, $j \geq 0$. Then, as shown above, $\Pr\{X_n{=}j\} = \pi_j$ for all $n \geq 0$, $j \geq 0$. Let $\tilde{N}_j(t)$ be the number of occurrences of any given state j from time 1 to t. Equating $\Pr\{X_n{=}j\}$ to the expectation of an occurrence of j at time n, we have,

$$(1/t)\mathsf{E}\left[\tilde{N}_j(t)\right] = (1/t)\sum_{1\leq n\leq t}\Pr\{X_n{=}j\} = \pi_j \qquad \text{for all integers } t \geq 1.$$

Conditioning this on the possible starting states i, and using the counting processes $\{N_{ij}(t); t \geq 0\}$ defined earlier,

$$\pi_j = (1/t)\mathsf{E}\left[\widetilde{N}_j(t)\right] = \sum_i \pi_i \mathsf{E}\left[N_{ij}(t)/t\right] \qquad \text{for all integer } t \geq 1. \tag{5.19}$$

For any given state i, let T_{ij} be the time of the first occurrence of state j given $X_0 = i$. Then if $T_{ij} < \infty$, we have $N_{ij}(t) \leq N_{ij}(T_{ij} + t)$. Thus, for all $t \geq 1$,

$$\mathsf{E}\left[N_{ij}(t)\right] \leq \mathsf{E}\left[N_{ij}(T_{ij} + t)\right] = 1 + \mathsf{E}\left[N_{jj}(t)\right]. \tag{5.20}$$

The last step follows since the process is in state j at time T_{ij}, and the expected number of occurrences of state j in the next t steps is $\mathsf{E}\left[N_{jj}(t)\right]$.

Substituting (5.20) in (5.19) for each i, $\pi_j \leq 1/t + \mathsf{E}\left[N_{jj}(t)/t\right]$. Taking the limit as $t \to \infty$ and using (5.16), $\pi_j \leq \lim_{t\to\infty} \mathsf{E}\left[N_{jj}(t)/t\right]$. Since $\sum_i \pi_i = 1$, there is at least one value of j for which $\pi_j > 0$, and for this j, $\lim_{t\to\infty} \mathsf{E}\left[N_{jj}(t)/t\right] > 0$, and consequently $\lim_{t\to\infty} \mathsf{E}\left[N_{jj}(t)\right] = \infty$. Thus, from Lemma 5.1.1, state j is recurrent, and from Theorem 5.1.2, j is positive-recurrent. From Theorem 5.1.3, all states are then positive-recurrent. For any j and any integer M, (5.19) implies that

$$\pi_j \geq \sum_{i\leq M} \pi_i \mathsf{E}\left[N_{ij}(t)/t\right] \quad \text{for all } t. \tag{5.21}$$

From Theorem 5.1.2, $\lim_{t\to\infty} \mathsf{E}\left[N_{ij}(t)/t\right] = 1/\overline{T}_{jj}$ for all i. Substituting this into (5.21), we get $\pi_j \geq 1/\overline{T}_{jj} \sum_{i\leq M} \pi_i$. Since M is arbitrary, $\pi_j \geq 1/\overline{T}_{jj}$. Since we already showed that $\pi_j \leq \lim_{t\to\infty} \mathsf{E}\left[N_{jj}(t)/t\right] = 1/\overline{T}_{jj}$, we have $\pi_j = 1/\overline{T}_{jj}$ for all j. This shows both that $\pi_j > 0$ for all j and that the solution to (5.18) is unique. Exercise 5.5 completes the proof by showing that if the states are positive-recurrent, then choosing $\pi_j = 1/\overline{T}_{jj}$ for all j satisfies (5.18). □

In practice, it is usually easy to see whether a chain is irreducible. We shall also see by a number of examples that the steady-state distribution can often be calculated from (5.18). Theorem 5.1.4 then says that the calculated distribution is unique and that its existence guarantees that the chain is positive recurrent.

Example 5.1.3. Age of a renewal process: Consider a renewal process $\{N(t); t > 0\}$ in which the inter-renewal random variables $\{W_n; n \geq 1\}$ are arithmetic with span 1. We will use a Markov chain to model the age of this process (see Figure 5.3). The probability that a renewal occurs at a particular integer time depends on the past only through the integer time back to the last renewal. The state of the Markov chain during a unit interval

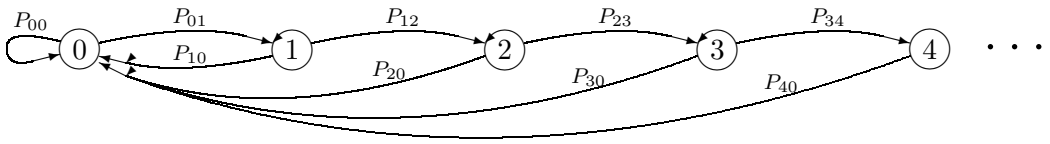


Figure 5.3: A Markov chain model of the age of a renewal process.

will be taken as the age of the renewal process at the beginning of the interval. Thus, each unit of time, the age either increases by one or a renewal occurs and the age decreases to 0 (*i.e.*, if a renewal occurs at time t, the age at time t is 0).

$\Pr\{W > n\}$ is the probability that an inter-renewal interval lasts for more than n time units. We assume that $\Pr\{W > 0\} = 1$, so that each renewal interval lasts at least one time unit. The probability $P_{n,0}$ in the Markov chain is the probability that a renewal interval has duration $n + 1$, given that the interval exceeds n. Thus, for example, P_{00} is the probability that the renewal interval is equal to 1. $P_{n,n+1}$ is $1 - P_{n,0}$, which is $\Pr\{W > n + 1\} / \Pr\{W > n\}$. We can then solve for the steady state probabilities in the chain: for $n > 0$,

$$\pi_n = \pi_{n-1} P_{n-1,n} = \pi_{n-2} P_{n-2,n-1} P_{n-1,n} = \pi_0 P_{0,1} P_{1,2} \dots P_{n-1,n}.$$

The first equality above results from the fact that state n, for $n > 0$ can be entered only from state $n - 1$. The subsequent equalities come from substituting in the same expression for π_{n-1}, then p_{n-2}, and so forth.

$$\pi_n = \pi_0 \frac{\Pr\{W > 1\} \Pr\{W > 2\}}{\Pr\{W > 0\} \Pr\{W > 1\}} \cdots \frac{\Pr\{W > n\}}{\Pr\{W > n - 1\}} = \pi_0 \Pr\{W > n\}. \tag{5.22}$$

We have cancelled out all the cross terms above and used the fact that $\Pr\{W > 0\} = 1$. Another way to see that $\pi_n = \pi_0 \Pr\{W > n\}$ is to observe that state 0 occurs exactly once in each inter-renewal interval; state n occurs exactly once in those inter-renewal intervals of duration n or more.

Since the steady-state probabilities must sum to 1, (5.22) can be solved for π_0 as

$$\pi_0 = \frac{1}{\sum_{n=0}^{\infty} \Pr\{W > n\}} = \frac{1}{\mathsf{E}[W]}. \tag{5.23}$$

The second equality follows by expressing $\mathsf{E}[W]$ as the integral of the complementary distribution function of W. Combining this with (5.22), the steady-state probabilities for $n \geq 0$ are

$$\pi_n = \frac{\Pr\{W > n\}}{\mathsf{E}[W]}. \tag{5.24}$$

In terms of the renewal process, π_n is the probability that, at some large integer time, the age of the process will be n. Note that if the age of the process at an integer time is n, then the age increases toward $n + 1$ at the next integer time, at which point it either drops

to 0 or continues to rise. Thus π_n can be interpreted as the fraction of time that the age of the process is between n and $n+1$. Recall from (4.28) (and the fact that residual life and age are equally distributed) that the distribution function of the time-average age is given by $\mathsf{F}_Z(n) = \int_0^n \Pr\{W > w\}\, dw/\mathsf{E}[W]$. Thus, the probability that the age is between n and $n+1$ is $\mathsf{F}_Z(n+1) - \mathsf{F}_Z(n)$. Since W is an integer random variable, this is $\Pr\{W > n\}/\mathsf{E}[W]$ in agreement with our result here.

The analysis here gives a new, and intuitively satisfying, explanation of why the age of a renewal process is so different from the inter-renewal time. The Markov chain shows the ever increasing loops that give rise to large expected age when the inter-renewal time is heavy tailed (*i.e.*, has a distribution function that goes to 0 slowly with increasing time). These loops can be associated with the isosceles triangles of Figure 4.7. The advantage here is that we can associate the states with steady-state probabilities if the chain is recurrent. Even when the Markov chain is null-recurrent (*i.e.*, the associated renewal process has infinite expected age), it seems easier to visualize the phenomenon of infinite expected age.

5.2 Birth-death Markov chains

A *birth-death Markov chain* is a Markov chain in which the state space is the set of nonnegative integers; for all $i \geq 0$, the transition probabilities satisfy $P_{i,i+1} > 0$ and $P_{i+1,i} > 0$, and for all $|i-j| > 1$, $P_{ij} = 0$ (see Figure 5.4). A transition from state i to $i+1$ is regarded as a birth and one from $i+1$ to i as a death. Thus the restriction on the transition probabilities means that only one birth or death can occur in one unit of time. Many applications of birth-death processes arise in queueing theory, where the state is the number of customers, births are customer arrivals, and deaths are customer departures. The restriction to only one arrival or departure at a time seems rather peculiar, but usually such a chain is a finely sampled approximation to a continuous-time process, and the time increments are then small enough that multiple arrivals or departures in a time increment are unlikely and can be ignored in the limit.

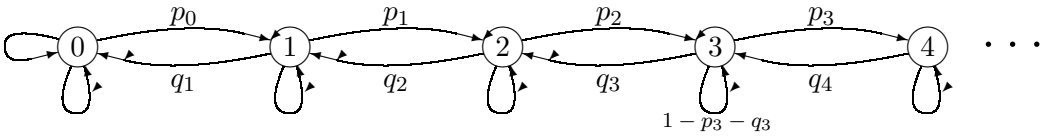


Figure 5.4: Birth-death Markov chain.

We denote $P_{i,i+1}$ by p_i and $P_{i,i-1}$ by q_i. Thus $P_{ii} = 1 - p_i - q_i$. There is an easy way to find the steady-state probabilities of these birth-death chains. In any sample function of the process, note that the number of transitions from state i to $i+1$ differs by at most 1 from the number of transitions from $i+1$ to i. If the process starts to the left of i and ends to the right, then one more $i \to i+1$ transition occurs than $i+1 \to i$, etc. Thus if we visualize a renewal-reward process with renewals on occurrences of state i and unit reward on transitions from state i to $i+1$, the limiting time-average number of transitions per unit time is $\pi_i\, p_i$. Similarly, the limiting time-average number of transitions per unit time from

$i+1$ to i is $\pi_{i+1}q_{i+1}$. Since these two must be equal in the limit,

$$\pi_i p_i = \pi_{i+1} q_{i+1} \qquad \text{for } i \geq 0. \tag{5.25}$$

The intuition in (5.25) is simply that the rate at which downward transitions occur from $i+1$ to i must equal the rate of upward transitions. Since this result is very important, both here and in our later study of continuous-time birth-death processes, we show that (5.25) also results from using the steady-state equations in (5.18):

$$\pi_i = p_{i-1}\pi_{i-1} + (1-p_i-q_i)\pi_i + q_{i+1}\pi_{i+1}; \quad i>0 \tag{5.26}$$

$$\pi_0 = (1-p_0)\pi_0 + q_1\pi_1. \tag{5.27}$$

From (5.27), $p_0\pi_0 = q_1\pi_1$. To see that (5.25) is satisfied for $i>0$, we use induction on i, with $i=0$ as the base. Thus assume, for a given i, that $p_{i-1}\pi_{i-1} = q_i\pi_i$. Substituting this in (5.26), we get $p_i\pi_i = q_{i+1}\pi_{i+1}$, thus completing the inductive proof.

It is convenient to define ρ_i as p_i/q_{i+1}. Then we have $\pi_{i+1} = \rho_i\pi_i$, and iterating this,

$$\pi_i = \pi_0 \prod_{j=0}^{i-1} \rho_j; \quad \pi_0 = \frac{1}{1+\sum_{i=1}^{\infty} \prod_{j=0}^{i-1} \rho_j}. \tag{5.28}$$

If $\sum_{i\geq 1}\prod_{0\leq j<i}\rho_j < \infty$, then π_0 is positive and all the states are positive-recurrent. If this sum of products is infinite, then no state is positive-recurrent. If ρ_j is bounded below 1, say $\rho_j \leq 1-\epsilon$ for some fixed $e>0$ and all sufficiently large j, then this sum of products will converge and the states will be positive-recurrent.

For the simple birth-death process of Figure 5.2, if we define $\rho = q/p$, then $\rho_j = \rho$ for all j. For $\rho<1$, (5.28) simplifies to $\pi_i = \pi_o\rho^i$ for all $i\geq 0$, $\pi_0 = 1-\rho$, and thus $\pi_i = (1-\rho)\rho^i$ for $i\geq 0$. Exercise 5.2 shows how to find $\mathsf{F}_{ij}(\infty)$ for all i,j in the case where $\rho\geq 1$. We have seen that the simple birth-death chain of Figure 5.2 is transient if $\rho>1$. This is not necessarily so in the case where self-transitions exist, but the chain is still either transient or null-recurrent. An example of this will arise in Exercise 6.3.

5.3 Reversible Markov chains

Many important Markov chains have the property that, in steady state, the sequence of states looked at backwards in time, *i.e.*,$\ldots X_{n+1}, X_n, X_{n-1}, \ldots$, has the same probabilistic structure as the sequence of states running forward in time. This equivalence between the forward chain and backward chain leads to a number of results that are intuitively quite surprising and that are quite difficult to derive without using this equivalence. We shall study these results here and then extend them in Chapter 6 to Markov processes with a discrete state space. This set of ideas, and its use in queueing and queueing networks, has been an active area of queueing research over many years . It leads to many simple results for systems that initially look very complex. We only scratch the surface here and refer the interested reader to [13] for a more comprehensive treatment. Before going into reversibility, we describe the backward chain for an arbitrary Markov chain.

The defining characteristic of a Markov chain $\{X_n; n \geq 0\}$ is that for all $n \geq 0$,

$$\Pr\{X_{n+1} \mid X_n, X_{n-1}, \ldots, X_0\} = \Pr\{X_{n+1} \mid X_n\}. \tag{5.29}$$

For homogeneous chains, which we have been assuming throughout, $\Pr\{X_{n+1} = j \mid X_n = i\} = P_{ij}$, independent of n. For any $k > 1$, we can extend (5.29) to get

$$\begin{aligned} &\Pr\{X_{n+k}, X_{n+k-1}, \ldots, X_{n+1} \mid X_n, X_{n-1}, \ldots, X_0\} \\ &\quad = \Pr\{X_{n+k} \mid X_{n+k-1}\} \Pr\{X_{n+k-1} \mid X_{n+k-2}\} \cdots \Pr\{X_{n+1} \mid X_n\} \\ &\quad = \Pr\{X_{n+k}, X_{n+k-1}, \ldots, X_{n+1} \mid X_n\}. \end{aligned} \tag{5.30}$$

By letting A^+ be any event defined on the states X_{n+1} to X_{n+k} and letting A^- be any event defined on X_0 to X_{n-1}, this can be written more succinctly as

$$\Pr\{A^+ \mid X_n, A^-\} = \Pr\{A^+ \mid X_n\}. \tag{5.31}$$

This says that, given state X_n, any future event A^+ is statistically independent of any past event A^-. This result, namely that past and future are independent given the present state, is equivalent to (5.29) for defining a Markov chain, but it has the advantage of showing the symmetry between past and future. This symmetry is best brought out by multiplying both sides of (5.31) by $\Pr\{A^- \mid X_n\}$, obtaining[2]

$$\Pr\{A^+, A^- \mid X_n\} = \Pr\{A^+ \mid X_n\} \Pr\{A^- \mid X_n\}. \tag{5.32}$$

This symmetric form says that, conditional on the current state, the past and future states are statistically independent. Dividing both sides by $\Pr\{A^+ \mid X_n\}$ then yields

$$\Pr\{A^- \mid X_n, A^+\} = \Pr\{A^- \mid X_n\}. \tag{5.33}$$

By letting A^- be X_{n-1} and A^+ be $X_{n+1}, X_{n+2}, \ldots, X_{n+k}$, this becomes

$$\Pr\{X_{n-1} \mid X_n, X_{n+1}, \ldots, X_{n+k}\} = \Pr\{X_{n-1} \mid X_n\}.$$

This is the equivalent form to (5.29) for the backward chain, and says that the backward chain is also a Markov chain. By Bayes' law, $\Pr\{X_{n-1} \mid X_n\}$ can be evaluated as

$$\Pr\{X_{n-1} \mid X_n\} = \frac{\Pr\{X_n \mid X_{n-1}\} \Pr\{X_{n-1}\}}{\Pr\{X_n\}}. \tag{5.34}$$

Since the distribution of X_n can vary with n, $\Pr\{X_{n-1} \mid X_n\}$ can also depend on n. *Thus the backward Markov chain is not necessarily homogeneous.* This should not be surprising, since the forward chain was defined with some arbitrary distribution for the initial state at time 0. This initial distribution was not relevant for equations (5.29) to (5.31), but as soon as $\Pr\{A^- \mid X_n\}$ was introduced, the initial state implicitly became a part of each equation and destroyed the symmetry between past and future. For a chain in steady state, however, $\Pr\{X_n = j\} = \Pr\{X_{n-1} = j\} = \pi_j$ for all j, and we have

$$\Pr\{X_{n-1} = j \mid X_n = i\} = P_{ji}\pi_j/\pi_i. \tag{5.35}$$

[2]Much more broadly, any 3 events, say A^-, X_0, A^+ are said to be Markov if $\Pr\{A^+ \mid X_0 A^-\} = \Pr\{A^+ \mid X_0\}$, and this implies the more symmetric form $\Pr\{A^- A^+ \mid X_0\} = \Pr\{A^- \mid X_0\} \Pr\{A^+ \mid X_0\}$.

Thus the backward chain is homogeneous if the forward chain is in steady state. For a chain with steady-state probabilities $\{\pi_i;\, i \geq 0\}$, we define the backward transition probabilities P^*_{ij} as

$$\pi_i P^*_{ij} = \pi_j P_{ji}. \tag{5.36}$$

From (5.34), the backward transition probability P^*_{ij}, for a Markov chain in steady state, is then equal to $\Pr\{X_{n-1} = j \mid X_n = i\}$, the probability that the previous state is j given that the current state is i.

Now consider a new Markov chain with transition probabilities $\{P^*_{ij}\}$. Over some segment of time for which both this new chain and the old chain are in steady state, the set of states generated by the new chain is statistically indistinguishable from the backward running sequence of states from the original chain. It is somewhat simpler, in talking about forward and backward running chains, however, to visualize Markov chains running in steady state from $t = -\infty$ to $t = +\infty$. If one is uncomfortable with this, one can also visualize starting the Markov chain at some very negative time with the initial distribution equal to the steady-state distribution.

Definition 5.3.1. *A Markov chain that has steady-state probabilities $\{\pi_i;\, i \geq 0\}$ is reversible if $P_{ij} = \pi_j P_{ji}/\pi_i$ for all i, j,* i.e.*, if $P^*_{ij} = P_{ij}$ for all i, j.*

Thus the chain is reversible if, in steady state, the backward running sequence of states is statistically indistinguishable from the forward running sequence. Comparing (5.36) with the steady-state equations (5.25) that we derived for birth-death chains, we have the important theorem:

Theorem 5.3.1. *Every birth-death chain with a steady-state probability distribution is reversible.*

We saw that for birth-death chains, the equation $\pi_i P_{ij} = \pi_j P_{ji}$ (which only had to be considered for $|i-j| \leq 1$) provided a very simple way of calculating the steady-state probabilities. Unfortunately, it appears that we must first calculate the steady-state probabilities in order to show that a chain is reversible. The following simple theorem gives us a convenient escape from this dilemma.

Theorem 5.3.2. *Assume that an irreducible Markov chain has transition probabilities $\{P_{ij}\}$. Suppose $\{\pi_i\}$ is a set of positive numbers summing to* 1 *and satisfying*

$$\pi_i P_{ij} = \pi_j P_{ji}; \quad \text{all } i, j. \tag{5.37}$$

then, first, $\{\pi_i; i \geq 0\}$ is the steady-state distribution for the chain, and, second, the chain is reversible.

Proof: Given a solution to (5.37) for all i and j, we can sum this equation over i for each j.

$$\sum_i \} \pi_i P_{ij} = \pi_j \sum_i \} P_{ji} = \pi_j. \tag{5.38}$$

Thus the solution to (5.37), along with the constraints $\pi_i > 0$, $\sum_i \pi_i = 1$, satisfies the steady-state equations, (5.18), and, from Theorem 5.1.4, this is the unique steady-state distribution. Since (5.37) is satisfied, the chain is also reversible.

It is often possible, sometimes by using an educated guess, to find a solution to (5.37). If this is successful, then we are assured both that the chain is reversible and that the actual steady-state probabilities have been found.

Note that the theorem applies to periodic chains as well as to aperiodic chains. If the chain is periodic, then the steady-state probabilities have to be interpreted as average values over the period, but from Theorem 5.1.4 shows that (5.38) still has a unique solution (assuming an irreducible chain). On the other hand, for a chain with period $d > 1$, there are d subclasses of states and the sequence $\{X_n\}$ must rotate between these classes in a fixed order. For this same order to be followed in the backward chain, the only possibility is $d = 2$. Thus periodic chains with periods other than 2 cannot be reversible.

There are several simple tests that can be used to show that some given irreducible chain is not reversible. First, the steady-state probabilities must satisfy $\pi_i > 0$ for all i, and thus, if $P_{ij} > 0$ but $P_{ji} = 0$ for some i, j, then (5.37) cannot be satisfied and the chain is not reversible. Second, consider any set of three states, i, j, k. If $P_{ji}P_{ik}P_{kj}$ is unequal to $P_{jk}P_{ki}P_{ij}$ then the chain cannot be reversible. To see this, note that (5.37) requires that

$$\pi_i = \pi_j P_{ji}/P_{ij} = \pi_k P_{ki}/P_{ik}.$$

Thus, $\pi_j P_{ji} P_{ik} = \pi_k P_{ki} P_{ij}$. Equation (5.37) also requires that $\pi_j P_{jk} = \pi_k P_{kj}$. Taking the ratio of these equations, we see that $P_{ji}P_{ik}P_{kj} = P_{jk}P_{ki}P_{ij}$. Thus if this equation is not satisfied, the chain cannot be reversible. In retrospect, this result is not surprising. What it says is that for any cycle of three states, the probability of three transitions going around the cycle in one direction must be the same as the probability of going around the cycle in the opposite (and therefore backwards) direction.

It is also true (see [16] for a proof), that a necessary and sufficient condition for a chain to be reversible is that the product of transition probabilities around any cycle of arbitrary length must be the same as the product of transition probabilities going around the cycle in the opposite direction. This doesn't seem to be a widely useful way to demonstrate reversibility.

There is another result, generalizing Theorem 5.3.2, for finding the steady-state probabilities of an arbitrary Markov chain and simultaneously finding the transition probabilities of the backward chain.

Theorem 5.3.3. *Assume that an irreducible Markov chain has transition probabilities $\{P_{ij}\}$. Suppose $\{\pi_i\}$ is a set of positive numbers summing to 1 and that $\{P^*_{ij}\}$ is a set of transition probabilities satisfying*

$$\pi_i P_{ij} = \pi_j P^*_{ji}; \quad \text{all } i, j. \tag{5.39}$$

Then $\{\pi_i\}$ is the steady-state distribution and $\{P^*_{ij}\}$ is the set of transition probabilities for the backward chain.

Proof: Summing (5.39) over i, we get the steady-state equations for the Markov chain, so the fact that the given $\{\pi_i\}$ satisfy these equations asserts that they are the steady-state probabilities. Equation (5.39) then asserts that $\{P^*_{ij}\}$ is the set of transition probabilities for the backward chain.

The following two sections illustrate some important applications of reversibility.

5.4 The M/M/1 sample-time Markov chain

The M/M/1 Markov chain is a sampled-time model of the M/M/1 queueing system. Recall that the M/M/1 queue has Poisson arrivals at some rate λ and IID exponentially distributed service times at some rate μ. We assume throughout this section that $\lambda < \mu$ (this is required to make the states positive-recurrent). For some given small increment of time δ, we visualize observing the state of the system at the sample times $n\delta$. As indicated in Figure 5.5, the probability of an arrival in the interval from $(n-1)\delta$ to $n\delta$ is modeled as $\lambda\delta$, independent of the state of the chain at time $(n-1)\delta$ and thus independent of all prior arrivals and departures. Thus the arrival process, viewed as arrivals in subsequent intervals of duration δ, is Bernoulli, thus approximating the Poisson arrivals. This is a sampled-time approximation to the Poisson arrival process of rate λ for a continuous-time M/M/1 queue.

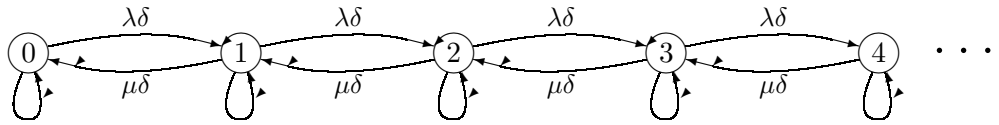


Figure 5.5: Sampled-time approximation to M/M/1 queue for time increment δ.

When the system is non-empty (*i.e.*, the state of the chain is one or more), the probability of a departure in the interval $(n-1)\delta$ to $n\delta$ is $\mu\delta$, thus modelling the exponential service times. When the system is empty, of course, departures cannot occur.

Note that in our sampled-time model, there can be at most one arrival or departure in an interval $(n-1)\delta$ to $n\delta$. As in the Poisson process, the probability of more than one arrival, more than one departure, or both an arrival and a departure in an increment δ is of order δ^2 for the actual continuous-time M/M/1 system being modeled. Thus, for δ very small, we expect the sampled-time model to be relatively good. At any rate, we can now analyze the model with no further approximations.

Since this chain is a birth-death chain, we can use (5.28) to determine the steady-state probabilities; they are

$$\pi_i = \pi_0 \rho^i \ ; \ \rho = \lambda/\mu < 1.$$

Setting the sum of the π_i to 1, we find that $\pi_0 = 1-\rho$, so

$$\pi_i = (1-\rho)\rho^i \ ; \qquad \text{all } i \geq 0. \tag{5.40}$$

Thus the steady-state probabilities exist and the chain is a birth-death chain, so from Theorem 5.3.1, it is reversible. We now exploit the consequences of reversibility to find some rather surprising results about the M/M/1 chain in steady state. Figure 5.6 illustrates a sample path of arrivals and departures for the chain. To avoid the confusion associated with the backward chain evolving backward in time, we refer to the original chain as the chain moving to the right and to the backward chain as the chain moving to the left.

There are two types of correspondence between the right-moving and the left-moving chain:

1. The left-moving chain has the same Markov chain description as the right-moving chain, and thus can be viewed as an M/M/1 chain in its own right. We still label the sampled-time intervals from left to right, however, so that the left-moving chain makes transitions from X_{n+1} to X_n to X_{n-1}. Thus, for example, if $X_n = i$ and $X_{n-1} = i+1$, the left-moving chain has an arrival in the interval from $n\delta$ to $(n-1)\delta$.

2. Each sample function $\ldots x_{n-1}, x_n, x_{n+1} \ldots$ of the right-moving chain corresponds to the same sample function $\ldots x_{n+1}, x_n, x_{n-1} \ldots$ of the left-moving chain, where $X_{n-1} = x_{n-1}$ is to the left of $X_n = x_n$ for both chains. With this correspondence, an arrival to the right-moving chain in the interval $(n-1)\delta$ to $n\delta$ is a departure from the left-moving chain in the interval $n\delta$ to $(n-1)\delta$, and a departure from the right-moving chain is an arrival to the left-moving chain. Using this correspondence, each event in the left-moving chain corresponds to some event in the right-moving chain.

In each of the properties of the M/M/1 chain to be derived below, a property of the left-moving chain is developed through correspondence 1 above, and then that property is translated into a property of the right-moving chain by correspondence 2.

Property 1: Since the arrival process of the right-moving chain is Bernoulli, the arrival process of the left-moving chain is also Bernoulli (by correspondence 1). Looking at a sample function x_{n+1}, x_n, x_{n-1} of the left-moving chain (*i.e.*, using correspondence 2), an arrival in the interval $n\delta$ to $(n-1)\delta$ of the left-moving chain is a departure in the interval $(n-1)\delta$ to $n\delta$ of the right-moving chain. Since the arrivals in successive increments of the left-moving chain are independent and have probability $\lambda\delta$ in each increment δ, we conclude that departures in the right-moving chain are similarly Bernoulli.

The fact that the departure process is Bernoulli with departure probability $\lambda\delta$ in each increment is surprising. Note that the probability of a departure in the interval $(n\delta - \delta, n\delta]$ is $\mu\delta$ conditional on $X_{n-1} \geq 1$ and is 0 conditional on $X_{n-1} = 0$. Since $\Pr\{X_{n-1} \geq 1\} = 1 - \Pr\{X_{n-1} = 0\} = \rho$, we see that the unconditional probability of a departure in the interval $(n\delta - \delta, n\delta]$ is $\rho\mu\delta = \lambda\delta$ as asserted above. The fact that successive departures are independent is much harder to derive without using reversibility (see exercise 5.13).

Property 2: In the original (right-moving) chain, arrivals in the time increments after $n\delta$ are independent of X_n. Thus, for the left-moving chain, arrivals in time increments to the left of $n\delta$ are independent of the state of the chain at $n\delta$. From the correspondence between sample paths, however, a left chain arrival is a right chain departure, so that for the right-moving chain, departures in the time increments prior to $n\delta$ are independent of X_n, which is equivalent to saying that the state X_n is independent of the prior departures.

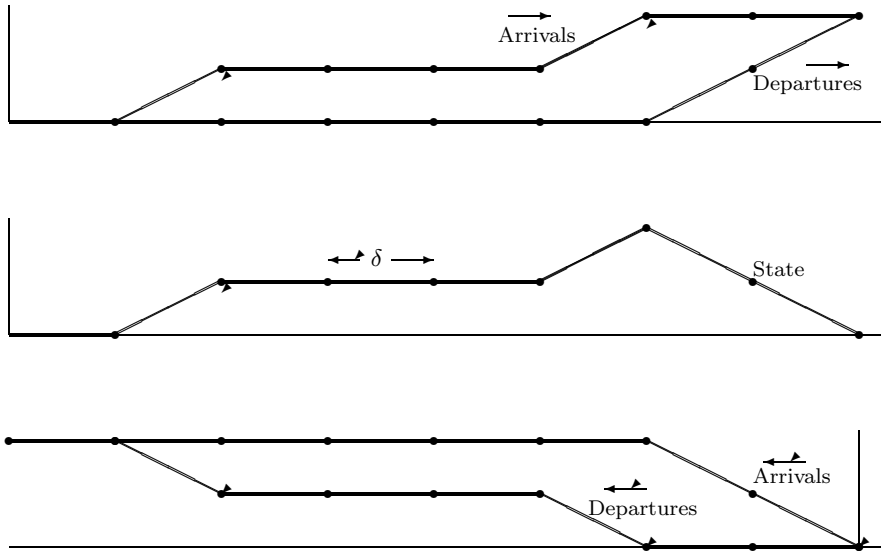


Figure 5.6: Sample function of M/M/1 chain over a busy period and corresponding arrivals and departures for right and left-moving chains. Arrivals and departures are viewed as occurring between the sample times, and an arrival in the left-moving chain between time $n\delta$ and $(n+1)\delta$ corresponds to a departure in the right-moving chain between $(n+1)\delta$ and $n\delta$.

This means that if one observes the departures prior to time $n\delta$, one obtains no information about the state of the chain at $n\delta$. This is again a surprising result. To make it seem more plausible, note that an unusually large number of departures in an interval from $(n-m)\delta$ to $n\delta$ indicates that a large number of customers were probably in the system at time $(n-m)\delta$, but it doesn't appear to say much (and in fact it says exactly nothing) about the number remaining at $n\delta$.

The following theorem summarizes these results.

Theorem 5.4.1 (Burke's theorem for sampled-time). *Given an M/M/1 Markov chain in steady state with $\lambda < \mu$,*

a) the departure process is Bernoulli,

b) the state X_n at any time $n\delta$ is independent of departures prior to $n\delta$.

The proof of Burke's theorem above did not use the fact that the departure probability is the same for all states except state 0. Thus these results remain valid for any birth-death chain with Bernoulli arrivals that are independent of the current state (*i.e.*, for which $P_{i,i+1} = \lambda\delta$ for all $i \geq 0$). One important example of such a chain is the sampled time approximation to an M/M/m queue. Here there are m servers, and the probability of departure from state i in an increment δ is $\mu i\delta$ for $i \leq m$ and $\mu m\delta$ for $i > m$. For the states to be recurrent, and thus for a steady state to exist, λ must be less than μm. Subject to this restriction, properties a) and b) above are valid for sampled-time M/M/m queues.

5.5 Branching processes

Branching processes provide a simple model for studying the population of various types of individuals from one generation to the next. The individuals could be photons in a photo-multiplier, particles in a cloud chamber, micro-organisms, insects, or branches in a data structure.

Let X_n be the number of individuals in generation n of some population. Each of these X_n individuals, independently of each other, produces a random number of offspring, and these offspring collectively make up generation $n+1$. More precisely, a *branching process* is a Markov chain in which the state X_n at time n models the number of individuals in generation n. Denote the individuals of generation n as $\{1, 2, ..., X_n\}$ and let $Y_{k,n}$ be the number of offspring of individual k. The random variables $Y_{k,n}$ are defined to be IID over k and n, with a PMF $p_j = \Pr\{Y_{k,n} = j\}$. The state at time $n+1$, namely the number of individuals in generation $n+1$, is

$$X_{n+1} = \sum_{k=1}^{X_n} Y_{k,n}. \tag{5.41}$$

Assume a given distribution (perhaps deterministic) for the initial state X_0. The transition probability, $P_{ij} = \Pr\{X_{n+1} = j \mid X_n = i\}$, is just the probability that $Y_{1,n}+Y_{2,n}+\cdots+Y_{i,n} = j$. The zero state (i e., the state in which there are *no* individuals) is a trapping state (*i.e.*, $P_{00} = 1$) since no future offspring can arise in this case.

One of the most important issues about a branching process is the probability that the population dies out eventually. Naturally, if p_0 (the probability that an individual has no offspring) is zero, then each generation must be at least as large as the generation before, and the population cannot die out unless $X_0 = 0$. We assume in what follows that $p_0 > 0$ and $X_0 > 0$. Recall that $\mathsf{F}_{ij}(n)$ was defined as the probability, given $X_0 = i$, that state j is entered between times 1 and n. From (5.8), this satisfies the iterative relation

$$\mathsf{F}_{ij}(n) = P_{ij} + \sum_{k \neq j} P_{ik}\mathsf{F}_{kj}(n-1), n > 1; \quad \mathsf{F}_{ij}(1) = P_{ij}. \tag{5.42}$$

The probability that the process dies out by time n or before, given $X_0 = i$, is thus $\mathsf{F}_{i0}(n)$. For the n^{th} generation to die out, starting with an initial population of i individuals, the descendants of each of those i individuals must die out. Since each individual generates descendants independently, we have $\mathsf{F}_{i0}(n) = [\mathsf{F}_{10}(n)]^i$ for all i and n. Because of this relationship, it is sufficient to find $\mathsf{F}_{10}(n)$, which can be determined from (5.42). Observe that P_{1k} is just p_k, the probability that an individual will have k offspring. Thus, (5.42) becomes

$$\mathsf{F}_{10}(n) = p_0 + \sum_{k=1}^{\infty} p_k[\mathsf{F}_{10}(n-1)]^k = \sum_{k=0}^{\infty} p_k[\mathsf{F}_{10}(n-1)]^k. \tag{5.43}$$

Let $h(z) = \sum_k p_k z^k$ be the z transform of the number of an individual's offspring. Then (5.43) can be written as

$$\mathsf{F}_{10}(n) = h(\mathsf{F}_{10}(n-1)). \tag{5.44}$$

This iteration starts with $\mathsf{F}_{10}(1) = p_0$. Figure 5.7 shows a graphical construction for evaluating $\mathsf{F}_{10}(n)$. Having found $\mathsf{F}_{10}(n)$ as an ordinate on the graph for a given value of n, we find the same value as an abscissa by drawing a horizontal line over to the straight line of slope 1; we then draw a vertical line back to the curve $h(z)$ to find $h(\mathsf{F}_{10}(n)) = \mathsf{F}_{10}(n+1)$.

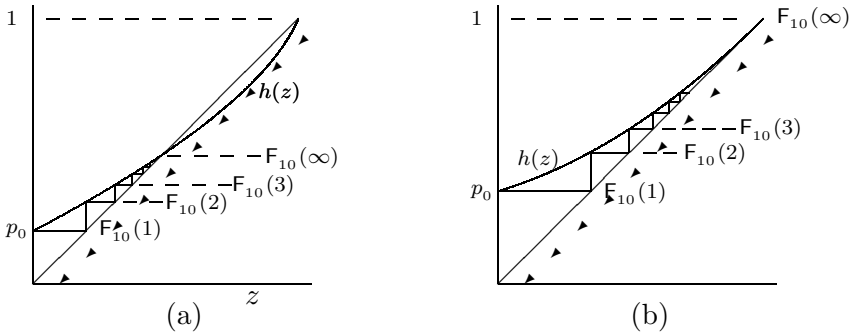


Figure 5.7: Graphical construction to find the probability that a population dies out. Here $\mathsf{F}_{10}(n)$ is the probability that a population starting with one member at generation 0 dies out by generation n or before. Thus $\mathsf{F}_{10}(\infty)$ is the probability that the population ever dies out.

For the two subfigures shown, it can be seen that $\mathsf{F}_{10}(\infty)$ is equal to the smallest root of the equation $h(z) - z = 0$. We next show that these two figures are representative of all possibilities. Since $h(z)$ is a z transform, we know that $h(1) = 1$, so that $z = 1$ is one root of $h(z) - z = 0$. Also, $h'(1) = \overline{Y}$, where $\overline{Y} = \sum_k kp_k$ is the expected number of an individual's offspring. If $\overline{Y} > 1$, as in Figure 5.7a, then $h(z) - z$ is negative for z slightly smaller than 1. Also, for $z = 0$, $h(z) - z = h(0) = p_0 > 0$. Since $h''(z) \geq 0$, there is exactly one root of $h(z) - z = 0$ for $0 < z < 1$, and that root is equal to $\mathsf{F}_{10}(\infty)$. By the same type of analysis, it can be seen that if $\overline{Y} \leq 1$, as in Figure 5.7b, then there is no root of $h(z) - z = 0$ for $z < 1$, and $\mathsf{F}_{10}(\infty) = 1$.

As we saw earlier, $\mathsf{F}_{i0}(\infty) = [\mathsf{F}_{10}(\infty)]^i$, so that for any initial population size, there is a probability strictly between 0 and 1 that successive generations eventually die out for $\overline{Y} > 1$, and probability 1 that successive generations eventually die out for $\overline{Y} \leq 1$. Since state 0 is accessible from all i, but $\mathsf{F}_{0i}(\infty) = 0$, it follows from Lemma 5.1.3 that all states other than state 0 are transient.

We next evaluate the expected number of individuals in a given generation. Conditional on $X_{n-1} = i$, (5.41) shows that the expected value of X_n is $i\overline{Y}$. Taking the expectation over X_{n-1}, we have

$$\mathsf{E}[X_n] = \overline{Y}\mathsf{E}[X_{n-1}]. \tag{5.45}$$

Iterating this equation, we get

$$\mathsf{E}[X_n] = \overline{Y}^n\mathsf{E}[X_0]. \tag{5.46}$$

Thus, if $\overline{Y} > 1$, the expected number of individuals in a generation increases exponentially with n, and $\overline{Y}$ gives the rate of growth. Physical processes do not grow exponentially

forever, so branching processes are appropriate models of such physical processes only over some finite range of population. Even more important, the model here assumes that the number of offspring of a single member is independent of the total population, which is highly questionable in many areas of population growth. The advantage of an oversimplified model such as this is that it explains what would happen under these idealized conditions, thus providing insight into how the model should be changed for more realistic scenarios.

It is important to realize that, for branching processes, the mean number of individuals is not a good measure of the actual number of individuals. For $\overline{Y} = 1$ and $X_0 = 1$, the expected number of individuals in each generation is 1, but the probability that $X_n = 0$ approaches 1 with increasing n; this means that as n gets large, the n^{th} generation contains a large number of individuals with a very small probability and contains no individuals with a very large probability. For $\overline{Y} > 1$, we have just seen that there is a positive probability that the population dies out, but the expected number is growing exponentially.

A surprising result, which is derived from the theory of martingales in Chapter 7, is that if $X_0 = 1$ and $\overline{Y} > 1$, then the sequence of random variables $X_n/\overline{Y}^n$ has a limit with probability 1. This limit is a random variable; it has the value 0 with probability $\mathsf{F}_{10}(\infty)$, and has larger values with some given distribution. Intuitively, for large n, X_n is either 0 or very large. If it is very large, it tends to grow in an orderly way, increasing by a multiple of $\overline{Y}$ in each subsequent generation.

5.6 Round-robin and processor sharing

Typical queueing systems have one or more servers who each serve customers in FCFS order, serving one customer completely while other customers wait. These typical systems have larger average delay than necessary. For example, if two customers with service requirements of 10 and 1 units respectively are waiting when a single server becomes empty, then serving the first before the second results in departures at times 10 and 11, for an average delay of 10.5. Serving the customers in the opposite order results in departures at times 1 and 11, for an average delay of 6. Supermarkets have recognized this for years and have special express checkout lines for customers with small service requirements.

Giving priority to customers with small service requirements, however, has some disadvantages; first, customers with high service requirements can feel discriminated against, and second, it is not always possible to determine the service requirements of customers before they are served. The following alternative to priorities is popular both in the computer and data network industries. When a processor in a computer system has many jobs to accomplish, it often serves these jobs on a time-shared basis, spending a small increment of time on one, then the next, and so forth. In data networks, particularly high-speed networks, messages are broken into small fixed-length packets, and then the packets from different messages can be transmitted on an alternating basis between messages.

A *round-robin* service system is a system in which, if there are m customers in the system, say $c_1, c_2, \ldots, c_m$, then c_1 is served for an incremental interval δ, followed by c_2 being served for an interval δ, and so forth up to c_m. After c_m is served for an interval δ, the server returns and starts serving c_1 for an interval δ again. Thus the customers are served in a

cyclic, or "round-robin" order, each getting a small increment of service on each visit from the server. When a customer's service is completed, the customer leaves the system, m is reduced, and the server continues rotating through the now reduced cycle as before. When a new customer arrives, m is increased and the new customer must be inserted into the cycle of existing customers in a way to be discussed later.

Processor sharing is the limit of round-robin service as the increment δ goes to zero. Thus, with processor sharing, if m customers are in the system, all are being served simultaneously, but each is being served at $1/m$ times the basic server rate. For the example of two customers with service requirement 1 and 10, each customer is initially served at rate 1/2, so one customer departs at time 2. At that time, the remaining customer is served at rate 1 and departs at time 11. For round-robin service with an increment of 1, the customer with unit service requirement departs at either time 1 or 2, depending on the initial order of service. With other increments of service, the results are slightly different.

We first analyze round-robin service and then go to the processor-sharing limit as $\delta \to 0$. As the above example suggests, the results are somewhat cleaner in the limiting case, but more realistic in the round-robin case. Round robin provides a good example of the use of backward transition probabilities to find the steady-state distribution of a Markov chain. The techniques used here are quite similar to those used in the next chapter to analyze queueing networks.

Assume a Bernoulli arrival process in which the probability of an arrival in an interval δ is $\lambda\delta$. Assume that the i^{th} arriving customer has a service requirement W_i. The random variables W_i, $i \geq 1$, are IID and independent of the arrival epochs. Thus, in terms of the arrival process and the service requirements, this is the same as an M/G/1 queue (see Section 4.5.5), but with M/G/1 queues, the server serves each customer completely before going on to the next customer. We shall find that the round-robin service here avoids the "slow truck effect" identified with the M/G/1 queue.

For simplicity, assume that W_i is arithmetic with span δ, taking on only values that are positive integer multiples of δ. Let $f(j) = \Pr\{W_i = j\delta\}, j \geq 1$ and let $\overline{F}(j) = \Pr\{W_i > j\delta\}$. Note that if a customer has already received j increments of service, then the probability that that customer will depart after 1 more increment is $f(j+1)/\overline{F}(j)$. This probability of departure on the next service increment after the j^{th} is denoted by

$$g(j) = f(j+1)/\overline{F}(j); j \geq 1. \tag{5.47}$$

The state $\boldsymbol{s}$ of a round-robin system can be expressed as the number, m, of customers in the system, along with an ordered listing of how many service increments each of those m customers have received, *i.e.*,

$$\boldsymbol{s} = (m, z_1, z_2, \ldots, z_m), \tag{5.48}$$

where $z_1\delta$ is the amount of service already received by the customer at the front of the queue, $z_2\delta$ is the service already received by the next customer in order, etc. In the special case of an idle queue, $\boldsymbol{s} = (0)$, which we denote as ϕ.

Given that the state X_n at time $n\delta$ is $\boldsymbol{s} \neq \phi$, the state X_{n+1} at time $n\delta + \delta$ evolves as follows:

- A new arrival enters with probability $\lambda\delta$ and is placed at the front of the queue;
- The customer at the front of the queue receives an increment δ of service;
- The customer departs if service is complete.
- Otherwise, the customer goes to the back of the queue

It can be seen that the state transition depends, first, on whether a new arrival occurs (an event of probability $\lambda\delta$), and, second, on whether a departure occurs. If no arrival and no departure occurs, then the queue simply rotates. The new state is $\boldsymbol{s}' = r(\boldsymbol{s})$, where the rotation operator $r(\boldsymbol{s})$ is defined by $r(\boldsymbol{s}) = (m, z_2, \ldots, z_m, z_1 + 1)$. If a departure but no arrival occurs, then the customer at the front of the queue receives its last unit of service and departs. The new state is $\boldsymbol{s}' = \delta(\boldsymbol{s})$, where the departure operator $\delta(\boldsymbol{s})$ is defined by $\delta(\boldsymbol{s}) = (m-1, z_2, \ldots, z_m)$.

If an arrival occurs, the new customer receives one unit of service and goes to the back of the queue if more than one unit of service is required. In this case, the new state is $\boldsymbol{s}' = a(\boldsymbol{s})$ where the arrival operator $a(\boldsymbol{s})$ is defined by $a(\boldsymbol{s}) = (m+1, z_1, z_2, \ldots, z_m, 1)$. If only one unit of service is required by a new arrival, the arrival departs and $\boldsymbol{s}' = \boldsymbol{s}$. In the special case of an empty queue, $\boldsymbol{s} = \phi$, the state is unchanged if either no arrival occurs or an arrival requiring one increment of service arrives. Otherwise, the new state is $\boldsymbol{s} = (1, 1)$, *i.e.*, the one customer in the system has received one increment of service.

We next find the probability of each transition for $\boldsymbol{s} \neq \phi$. The probability of no arrival is $1 - \lambda\delta$. Given no arrival, and given a non-empty system, $\boldsymbol{s} \neq \phi$, the probability of a departure is $g(z_1) = f(z_1 + 1)/\overline{F}(z_1)$, i e., the probability that one more increment of service allows the customer at the front of the queue to depart. Thus the probability of a departure is $(1 - \lambda\delta)g(z_1)$ and the probability of a rotation is $(1 - \lambda\delta)[1 - g(z_1)]$. Finally, the probability of an arrival is $\lambda\delta$, and given an arrival, the new arrival will leave the system after one unit of service with probability $g(0) = f(1)$. Thus the probability of an arrival and no departure is $\lambda\delta[1 - f(1)]$ and the probability of an unchanged system is $\lambda\delta f(1)$. To summarize, for $\boldsymbol{s} \neq \phi$,

$$\begin{aligned}
P_{\boldsymbol{s},r(\boldsymbol{s})} &= (1-\lambda\delta)[1-g(z_1)]; \quad && r(\boldsymbol{s}) = (m, z_2, \ldots, z_m, z_1+1) \\
P_{\boldsymbol{s},d(\boldsymbol{s})} &= (1-\lambda\delta)g(z_1); && d(\boldsymbol{s}) = (m-1, z_2, \ldots, z_m) \\
P_{\boldsymbol{s},a(\boldsymbol{s})} &= \lambda\delta[1-f(1)]; && a(\boldsymbol{s}) = (m+1, z_1, z_2, \ldots, z_m, 1) \\
P_{\boldsymbol{s},s} &= \lambda\delta f(1). && \qquad (5.49)
\end{aligned}$$

For the special case of the idle state, $P_{\phi,\phi} = (1 - \lambda\delta) + \lambda\delta f(1)$ and $P_{\phi,(1,1)} = \lambda\delta(1 - f(1))$.

We now find the steady-state distribution for this Markov chain by looking at the backward Markov chain. We will hypothesize backward transition probabilities, and then use Theorem 5.3.3 to verify that the hypothesis is correct. Consider the backward transitions corresponding to each of the forward transitions in (5.49). A rotation in forward time causes the elements $z_1, \ldots, z_m$ in the state $\boldsymbol{s} = (m, z_1, \ldots, z_m)$ to rotate left, and the left most element (corresponding to the front of the queue) is incremented while rotating to the right end. The backward transition from $r(\boldsymbol{s})$ to $\boldsymbol{s}$ corresponds to the elements $z_2, \ldots, z_m, z_1 + 1$

rotating to the right, with the right most element being decremented while rotating to the left end. If we view the transitions in backward time as a kind of round-robin system, we see that the rotation is in the opposite direction from the forward time system.

In the backward time system, we view the numbers $z_1, \ldots, z_m$ in the state as the remaining service required before the corresponding customers can depart. Thus, these numbers decrease in the backward moving system. Also, since the customer rotation in the backward moving system is opposite to that in the forward moving system, z_m is the remaining service of the customer at the front of the queue, and z_1 is the remaining service of the customer at the back of the queue. We also view departures in forward time as arrivals in backward time. Thus the backward transition from $d(\boldsymbol{s}) = (m-1, z_2, \ldots, z_m)$ to $\boldsymbol{s} = (m, z_1, \ldots, z_m)$ corresponds to an arrival requiring $z_1 + 1$ units of service; the arrival goes to the front of the queue, receives one increment of service, and then goes to the back of the queue with z_1 increments of remaining service.

The nicest thing we could now hope for is that the arrivals in backward time are Bernoulli. This is a reasonable hypothesis to make, partly because it is plausible, and partly because it is easy to check via Theorem 5.3.3. Fortunately, we shall find that it is valid. According to this hypothesis, the backward transition probability $P^*_{r(\boldsymbol{s}),\boldsymbol{s}}$ is given by $1 - \lambda\delta$; that is, given that X_{n+1} is $r(\boldsymbol{s}) = (m, z_2, \ldots, z_m, z_1+1)$, and given that there is no arrival in the backward system at time $(n+1)\delta$, then the only possible state at time n is $\boldsymbol{s} = (m, z_1, \ldots, z_n)$. Next consider a backward transition from $d(\boldsymbol{s}) = (m-1, z_2, \ldots, z_n)$ to $\boldsymbol{s} = (m, z_1, z_2, \ldots, z_m)$. This corresponds to an arrival in the backward moving system; the arrival requires $z_1 + 1$ increments of service, one of which is provided immediately, leaving the arrival at the back of the queue with z_1 required increments of service remaining. The probability of this transition is $P^*_{d(\boldsymbol{s}),\boldsymbol{s}} = \lambda\delta f(z_1 + 1)$. Calculating the other backward transitions in the same way, the hypothesized backward transition probabilities are given by

$$\begin{array}{rclll} P^*_{r(\boldsymbol{s}),\boldsymbol{s}} & = & 1 - \lambda\delta & P^*_{d(\boldsymbol{s}),\boldsymbol{s}} = \lambda\delta f(z_1 + 1) & \\ P^*_{a(\boldsymbol{s}),\boldsymbol{s}} & = & 1 - \lambda\delta & P^*_{\boldsymbol{s},\boldsymbol{s}} = \lambda\delta f(1). & (5.50) \end{array}$$

One should view (5.50) as an hypothesis for the backward transition probabilities. The arguments leading up to (5.50) are simply motivation for this hypothesis. If the hypothesis is correct, we can combine (5.49) and (5.50) to express the steady-state equations of Theorem 5.3.3 (for $\boldsymbol{s} \neq f$) as

$$\begin{array}{rclll} \pi_{\boldsymbol{s}} P_{\boldsymbol{s},r(\boldsymbol{s})} & = & \pi_{r(\boldsymbol{s})} P^*_{r(\boldsymbol{s}),\boldsymbol{s}}; & (1-\lambda\delta)[1 - g(z_1)]\pi_{\boldsymbol{s}} = (1-\lambda\delta)\pi_{r(\boldsymbol{s})} & (5.51) \\ \pi_{\boldsymbol{s}} P_{\boldsymbol{s},d(\boldsymbol{s})} & = & \pi_{d(\boldsymbol{s})} P^*_{d(\boldsymbol{s}),\boldsymbol{s}}; & (1-\lambda\delta)g(z_1)\pi_{\boldsymbol{s}} = \lambda\delta f(z_1+1)\pi_{d(\boldsymbol{s})} & (5.52) \\ \pi_{\boldsymbol{s}} P_{\boldsymbol{s},a(\boldsymbol{s})} & = & \pi_{a(\boldsymbol{s})} P^*_{a(\boldsymbol{s}),\boldsymbol{s}}; & \lambda\delta[1 - f(1)]\pi_{\boldsymbol{s}} = (1-\lambda\delta)\pi_{a(\boldsymbol{s})} & (5.53) \\ \pi_{\boldsymbol{s}} P_{\boldsymbol{s},\boldsymbol{s}} & = & \pi_{\boldsymbol{s}} P^*_{\boldsymbol{s},\boldsymbol{s}}; & \lambda\delta f(1)\pi_{\boldsymbol{s}} = \lambda\delta f(1)\pi_{\boldsymbol{s}}. & (5.54) \end{array}$$

We next show that (5.52), applied repeatedly, will allow us to solve for $\pi_{\boldsymbol{s}}$ (if λ is small enough for the states to be positive recurrent). Verifying that the solution also satisfies (5.51) and (5.53), will then verify the hypothesis. Since $f(z_1 + 1)/g(z_1)$ is $\overline{F}(z_1)$ from (5.47), we have

$$\pi_{\boldsymbol{s}} = \frac{\lambda\delta}{1 - \lambda\delta}\, \overline{F}(z_1)\pi_{d(\boldsymbol{s})}. \tag{5.55}$$

For $m > 1$, $d(\boldsymbol{s}) = (m-1, z_2, \ldots, z_m)$, so we can apply (5.55) to $\pi_{d(\boldsymbol{s})}$, and substitute the result back into (5.55), yielding

$$\pi_{\boldsymbol{s}} = \left(\frac{\lambda\delta}{1-\lambda\delta}\right)^2 \overline{F}(z_1)\overline{F}(z_2)\pi_{d(d(\boldsymbol{s}))}, \tag{5.56}$$

where $d(d(\boldsymbol{s})) = (m-2, z_3, \ldots, z_m)$. Applying (5.55) repeatedly to $\pi_{d(d(\boldsymbol{s}))}, \pi_{d(d(d(\boldsymbol{s})))}$, and so forth, we eventually get

$$\pi_{\boldsymbol{s}} = \left(\frac{\lambda\delta}{1-\lambda\delta}\right)^m \left(\prod_{j=1}^{m} \overline{F}(z_j)\right) \pi_\phi. \tag{5.57}$$

Before this can be accepted as a steady-state distribution, we must verify that it satisfies (5.51) and (5.53). The left hand side of (5.51) is $(1-\lambda\delta)[1-g(z_1)]\pi_{\boldsymbol{s}}$, and, from (5.47), $1-g(z_1) = [\overline{F}(z_1) - f(z_1+1)]/\overline{F}(z_1) = \overline{F}(z_1+1)/(z_1)$. Thus using (5.57), the left side of (5.51) is

$$(1-\lambda\delta)\ \frac{\overline{F}(z_1+1)}{\overline{F}(z_1)}\quad \left(\frac{\lambda\delta}{1-\lambda\delta}\right)^m \left(\prod_{j=1}^{m} \overline{F}(z_j)\right) \pi_\phi = (1-\lambda\delta)\quad \left(\frac{\lambda\delta}{1-\lambda\delta}\right)^m \left(\prod_{j=2}^{m}\}\overline{F}(z_j)\right) \overline{F}(z_1+1)\pi_\phi.$$

This is equal to $(1-\lambda\delta)\pi_{r(\boldsymbol{s})}$, verifying (5.51). Equation (5.53) is verified in the same way. We now have to find whether there is a solution for pf such that these probabilities sum to 1. First define $P_m = \sum z_1, \ldots, z_m \pi(m, z_1, \ldots, z_m)$. This is the probability of m customers in the system. Whenever a new customer enters the system, it receives one increment of service immediately, so each $z_i \geq 1$. Using the hypothesized solution in (5.57),

$$P_m = \left(\frac{\lambda\delta}{1-\lambda\delta}\right)^m \left(\prod_{j=1}^{m} \sum_{i=1}^{\infty} \overline{F}(i)\right) \pi_\phi. \tag{5.58}$$

Since $\overline{F}(i) = \Pr\{W > i\delta\}$, since W is arithmetic with span δ, and since the mean of a nonnegative random variable is the integral of its complementary distribution function, we have

$$\delta \sum_{i=1}^{\infty}\}\overline{F}(i) = \mathsf{E}[W] - \delta \tag{5.59}$$

$$P_m = \left(\frac{\lambda}{1-\lambda\delta}\right)^m \Big(\mathsf{E}[W] - \delta\Big)^m \pi_\phi. \tag{5.60}$$

Defining $\rho = [\lambda/(1-\lambda\delta)]\{\mathsf{E}[W] - \delta\}$, we see $P_m = \rho^m \pi_\phi$. If $\rho < 1$, then $\pi_\phi = 1-\rho$, and

$$P_m = (1-\rho)\rho^m; \quad m \geq 0. \tag{5.61}$$

The condition $r < 1$ is required for the states to be positive-recurrent. The expected number of customers in the system for a round-robin queue is $\sum_m mP_m = \rho/(1-\rho)$, and using Little's theorem, Theorem 4.5.3, the expected delay is $\rho/[\lambda(1-\rho)]$. In using Little's

theorem here, however, we are viewing the time a customer spends in the system as starting when the number m in the state increases; that is, if a customer arrives at time $n\delta$, it goes to the front of the queue and receives one increment of service, and then, assuming it needs more than one increment, the number m in the state increases at time $(n+1)\delta$. Thus the actual expected delay, including the original d when the customer is being served but not counted in the state, is $\delta + \rho/[\lambda(1-\rho)]$.

The relation between ρ and $\lambda \mathsf{E}[W]$ is shown in Figure 5.8, and it is seen that $\rho < 1$ for $\lambda \mathsf{E}[W] < 1$. The extreme case where $\lambda\delta = \lambda \mathsf{E}[W]$ is the case for which each customer requires exactly one unit of service. Since at most one customer can arrive per time increment, the state always remains at $\boldsymbol{s} = \phi$, and the delay is δ, *i.e.*, the original increment of service received when a customer arrives.

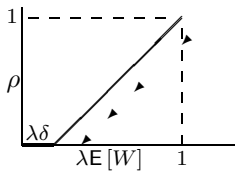


Figure 5.8: ρ as a function of $\lambda \mathsf{E}[W]$ for given $\lambda\delta$.

Note that (5.61) is the same as the distribution of customers in the system for the M/M/1 Markov chain in (5.40), except for the anomaly in the definition of ρ here. We then have the surprising result that if round-robin queueing is used rather than FCFS, then the distribution of the number of customers in the system is approximately the same as that for an M/M/1 queue. In other words, the slow truck effect associated with the M/G/1 queue has been eliminated.

Another remarkable feature of round-robin systems is that one can also calculate the expected delay for a customer conditional on the required service of that customer. This is done in Exercise 5.16, and it is found that the expected delay is linear in the required service.

Next we look at processor sharing by going to the limit as $\delta \to 0$. We first eliminate the assumption that the service requirement distribution is arithmetic with span δ. Assume that the server always spends an increment of time δ on the customer at the front of the queue, and if service is finished before the interval of length δ ends, the server is idle until the next sample time. The analysis of the steady-state distribution above is still valid if we define $\overline{F}(j) = \Pr\{W > j\delta\}$, and $f(j) = \overline{F}(j) - \overline{F}(j+1)$. In this case $\delta \sum_{i=1}^{\infty} \overline{F}(i)$ lies between $\mathsf{E}[W] - \delta$ and $\mathsf{E}[W]$. As $\delta \to 0$, $\rho = \lambda \mathsf{E}[W]$, and distribution of time in the system becomes identical to that of the M/M/1 system.

5.7 Summary

This chapter extended the finite-state Markov chain results of Chapter 3 to the case of countably-infinite state spaces. It also provided an excellent example of how renewal pro-

cesses can be used for understanding other kinds of processes. In Section 5.1, the first-passage-time random variables were used to construct renewal processes with renewals on successive transitions to a given state. These renewal processes were used to rederive the basic properties of Markov chains using renewal theory as opposed to the algebraic Perron-Frobenius approach of Chapter 3. The central result of this was Theorem 5.1.4, which showed that, for an irreducible chain, the states are positive-recurrent if and only if the steady-state equations, (5.18), have a solution. Also if (5.18) has a solution, it is positive and unique. We also showed that these steady-state probabilities are, with probability 1, time-averages for sample paths, and that, for an ergodic chain, they are limiting probabilities independent of the starting state.

We found that the major complications that result from countable state spaces are, first, different kinds of transient behavior, and second, the possibility of null-recurrent states. For finite-state Markov chains, a state is transient only if it can reach some other state from which it can't return. For countably infinite chains, there is also the case, as in Figure 5.2 for $p > 1/2$, where the state just wanders away, never to return. Null recurrence is a limiting situation where the state wanders away and returns with probability 1, but with an infinite expected time. There is not much engineering significance to null recurrence; it is highly sensitive to modeling details over the entire infinite set of states. One usually uses countably infinite chains to simplify models; for example, if a buffer is very large and we don't expect it to overflow, we assume it is infinite. Finding out, then, that the chain is transient or null-recurrent simply means that the modeling assumption is not very good.

We next studied birth-death Markov chains and reversibility. Birth-death chains are widely used in queueing theory as sample time approximations for systems with Poisson arrivals and various generalizations of exponentially distributed service times. Equation (5.28) gives their steady-state probabilities if positive-recurrent, and shows the conditions under which they are positive-recurrent. We showed that these chains are reversible if they are positive-recurrent.

Theorems 5.3.2 and 5.3.3 provides a simple way to find the steady-state distribution of reversible chains and also of chains where the backward chain behavior could be hypothesized or deduced. We used reversibility to show that M/M/1 and M/M/m Markov chains satisfy Burke's theorem for sampled-time — namely that the departure process is Bernoulli, and that the state at any time is independent of departures before that time.

Branching processes were introduced in Section 5.5 as a model to study the growth of various kinds of elements that reproduce. In general, for these models (assuming $p_0 > 0$), there is one trapping state and all other states are transient. Figure 5.7 showed how to find the probability that the trapping state is entered by the nth generation, and also the probability that it is entered eventually. If the expected number of offspring of an element is at most 1, then the population dies out with probability 1, and otherwise, the population dies out with some given probability q, and grows without bound with probability $1 - q$.

Round-robin queueing was then used as a more complex example of how to use the backward process to deduce the steady-state distribution of a rather complicated Markov chain; this also gave us added insight into the behavior of queueing systems and allowed us to show that, in the processor-sharing limit, the distribution of number of customers is the same as

that in an M/M/1 queue.

For further reading on Markov chains with countably-infinite state spaces, see [8], [16], or [22]. Feller [8] is particularly complete, but Ross [16] and Wolff [22] are somewhat more accessible. Harris, [12] is the standard reference on branching processes and Kelly, [13] is the standard reference on reversibility. The material on round-robin systems is from [24] and is generalized there.

5.8 Exercises

Exercise 5.1. Let $\{P_{ij}; i, j \geq 0\}$ be the set of transition probabilities for an countable-state Markov chain. For each i, j, let $\mathsf{F}_{ij}(n)$ be the probability that state j occurs sometime between time 1 and n inclusive, given $X_0 = i$. For some given j, assume that $\{x_k; k \geq 0\}$ is a set of nonnegative numbers satisfying $x_i = P_{ij} + \sum_{k \neq j} P_{ik} x_k$. Show that $x_i \geq \mathsf{F}_{ij}(n)$ for all n and i, and hence that $x_i \geq \mathsf{F}_{ij}(\infty)$ for all i. Hint: use induction.

Exercise 5.2. a) For the Markov chain in Figure 5.2, show that, for $p \geq 1/2$, $\mathsf{F}_{00}(\infty) = 2(1-p)$ and show that $\mathsf{F}_{i0}(\infty) = [(1-p)/p]^i$ for $i \geq 1$. Hint: first show that this solution satisfies (5.9) and then show that (5.9) has no smaller solution (see Exercise 5.1). Note that you have shown that the chain is transient for $p > 1/2$ and that it is recurrent for $p = 1/2$.

b) Under the same conditions as part a), show that $\mathsf{F}_{ij}(\infty)$ equals $2(1-p)$ for $j = i$, equals $[(1-p)/p]^{i-j}$ for $i > j$, and equals 1 for $i < j$.

Exercise 5.3. a): Show that the nth order transition probabilities, starting in state 0, for the Markov chain in Figure 5.2 satisfy

$$P^n_{0j} = pP^{n-1}_{0,i-1} + qP^{n-1}_{0,i+1} \quad j \neq 0; \qquad P^n_{00} = qP^{n-1}_{00} + qP^{n-1}_{01}.$$

Hint: Use the Chapman-Kolmogorov equality, (3.8).

b) For $p = 1/2$, use this equation to calculate P^n_{0j} iteratively for $n = 1, 2, 3, 4$. Verify (5.3) for these values and then use induction to verify (5.3) in general. Note: this becomes an absolute mess for $p \neq 1/2$, so don't attempt this in general.

c) As a more interesting approach, which brings out the relationship of Figures 5.2 and 5.1, note that (5.3), with $j + n$ even, is the probability that $S_n = j$ for the chain in 5.1 and (5.3) with $j + n$ odd is the probability that $S_n = -j - 1$ for the chain in 5.1. By viewing each transition over the self loop at state 0 as a sign reversal for the chain in 5.1, explain why this surprising result is true. (Again, this doesn't work for $p \neq 1/2$, since the sign reversals also reverse the +1, -1 transitions.)

Exercise 5.4. Let j be a transient state in a Markov chain and let j be accessible from i. Show that i is transient also. Interpret this as a form of Murphy's law (if something bad can happen, it will, where the bad thing is the lack of an eventual return). Note: give a direct demonstration rather than using Lemma 5.1.3.

Exercise 5.5. Consider an irreducible positive-recurrent Markov chain. Consider the renewal process $\{N_{jj}(t); t \geq 0\}$ where, given $X_0 = j$, $N_{jj}(t)$ is the number of times that state j is visited from time 1 to t. For each $i \geq 0$, consider a renewal-reward function $R_i(t)$ equal to 1 whenever the chain is in state i and equal to 0 otherwise. Let π_i be the time-average reward.

a) Show that $\pi_i = 1/\overline{T}_{ii}$ for each i with probability 1.

b) Show that $\sum_i \pi_i = 1$. Hint: consider $\sum_{i \leq M} \pi_i$ for any integer M.

c) Consider a renewal-reward function $R_{ij}(t)$ that is 1 whenever the chain is in state i and the next state is state j. $R_{ij}(t) = 0$ otherwise. Show that the time-average reward is equal to $\pi_i P_{ij}$ with probability 1. Show that $p_k = \sum_i \pi_i P_{ik}$ for all k.

Exercise 5.6. Let $\{X_n; n \geq 0\}$ be a branching process with $X_0 = 1$. Let $\overline{Y}$, σ^2 be the mean and variance of the number of offspring of an individual.

a) Argue that $\lim_{n\to\infty} X_n$ exists with probability 1 and either has the value 0 (with probability $\mathsf{F}_{10}(\infty)$) or the value ∞ (with probability $1 - \mathsf{F}_{10}(\infty)$).

b) Show that $\mathsf{VAR}[X_n] = \sigma^2 \overline{Y}^{n-1}(\overline{Y}^n - 1)/(\overline{Y} - 1)$ for $\overline{Y} \neq 1$ and $\mathsf{VAR}[X_n] = n\sigma^2$ for $\overline{Y} = 1$.

Exercise 5.7. There are n states and for each pair of states i and j, a positive number $d_{ij} = d_{ji}$ is given. A particle moves from state to state in the following manner: Given that the particle is in any state i, it will next move to any $j \neq i$ with probability P_{ij} given by

$$P_{ij} = \frac{d_{ij}}{\sum_{j \neq i} d_{ij}}.$$

Assume that $P_{ii} = 0$ for all i. Show that the sequence of positions is a reversible Markov chain and find the limiting probabilities.

Exercise 5.8. Consider a reversible Markov chain with transition probabilities P_{ij} and limiting probabilities π_i. Also consider the same chain truncated to the states $0, 1, \ldots, M$. That is, the transition probabilities $\{P'_{ij}\}$ of the truncated chain are

$$P'_{ij} = \begin{cases} \frac{P_{ij}}{\sum_{k=0}^{m} P_{ik}} & ; \quad 0 \leq i, j \leq M \\ 0 & ; \quad \text{elsewhere.} \end{cases}$$

Show that the truncated chain is also reversible and has limiting probabilities given by

$$\overline{\pi}_i = \frac{\pi_i \sum_{j=0}^{M} P_{ij}}{\sum_{k=0}^{M} \pi_i \sum_{m=0}^{M} P_{km}}.$$

Exercise 5.9. A Markov chain (with states $\{0, 1, 2, \ldots, J-1\}$ where J is either finite or infinite) has transition probabilities $\{P_{ij}; i, j \geq 0\}$. Assume that $P_{0j} > 0$ for all $j > 0$ and $P_{j0} > 0$ for all $j > 0$. Also assume that for all i, j, k, we have $P_{ij}P_{jk}P_{ki} = P_{ik}P_{kj}P_{ji}$.

a) Assuming also that all states are positive recurrent, show that the chain is reversible and find the steady-state probabilities $\{\pi_i\}$ in simplest form.

b) Find a condition on $\{P_{0j}; j \geq 0\}$ and $\{P_{j0}; j \geq 0\}$ that is sufficient to ensure that all states are positive recurrent.

Exercise 5.10. a) Use the birth and death model described in figure 5.4 to find the steady-state probability mass function for the number of customers in the system (queue plus service facility) for the following queues:

i) M/M/1 with arrival probability $\lambda\delta$, service completion probability $\mu\delta$.

ii) M/M/m with arrival probability $\lambda\delta$, service completion probability $i\mu\delta$ for i servers busy, $1 \leq i \leq m$.

iii) M/M/∞ with arrival probability $\lambda\delta$, service probability $i\mu\delta$ for i servers. Assume δ so small that $i\mu\delta < 1$ for all i of interest.

Assume the system is positive recurrent.

b) For each of the queues above give necessary conditions (if any) for the states in the chain to be i) transient, ii) null recurrent, iii) positive recurrent.

c) For each of the queues find:

L = (steady-state) mean number of customers in the system.

L^q = (steady-state) mean number of customers in the queue.

W = (steady-state) mean waiting time in the system.

W^q = (steady-state) mean waiting time in the queue.

Exercise 5.11. a) Given that an arrival occurs in the interval $(n\delta, (n+1)\delta)$ for the sampled-time M/M/1 model in figure 5, find the conditional PMF of the state of the system at time $n\delta$ (assume n arbitrarily large and assume positive recurrence).

b) For the same model, again in steady state but not conditioned on an arrival in $(n\delta, (n+1)\delta)$, find the probability $Q(i, j)(i \geq j > 0)$ that the system is in state i at $n\delta$ and that $i-j$ departures occur before the next arrival.

c) Find the expected number of customers seen in the system by the first arrival after time $n\delta$. (Note: the purpose of this exercise is to make you cautious about the meaning of "the state seen by a random arrival").

Exercise 5.12. Find the backward transition probabilities for the Markov chain model of age in figure 2. Draw the graph for the backward Markov chain, and interpret it as a model for residual life.

Exercise 5.13. Consider the sample time approximation to the M/M/1 queue in figure 5.5.

a) Give the steady-state probabilities for this chain (no explanations or calculations required–just the answer).

In parts b) to g) do not use reversibility and do not use Burke's theorem. Let X_n be the state of the system at time $n\delta$ and let D_n be a random variable taking on the value 1 if a departure occurs between $n\delta$ and $(n+1)\delta$, and the value 0 if no departure occurs. Assume that the system is in steady state at time $n\delta$.

b) Find $\Pr\{X_n = i, D_n = j\}$ for $i \geq 0$, $j = 0, 1$

c) Find $\Pr\{D_n = 1\}$

d) Find $\Pr\{X_n = i \mid D_n = 1\}$ for $i \geq 0$

e) Find $\Pr\{X_{n+1} = i \mid D_n = 1\}$ and show that X_{n+1} is statistically independent of D_n. Hint: Use part d); also show that $\Pr\{X_{n+1} = i\} = \Pr\{X_{n+1} = i \mid D_n = 1\}$ for all $i \geq 0$ is sufficient to show independence.

f) Find $\Pr\{X_{n+1} = i, D_{n+1} = j \mid D_n\}$ and show that the pair of variables (X_{n+1}, D_{n+1}) is statistically independent of D_n.

g) For each $k > 1$, find $\Pr\{X_{n+k} = i, D_{n+k} = j \mid D_{n+k-1}, D_{n+k-2}, \ldots, D_n\}$ and show that the pair (X_{n+k}, D_{n+k}) is statistically independent of $(D_{n+k-1}, D_{n+k-2}, \ldots, D_n)$. Hint: use induction on k; as a substep, find $\Pr\{X_{n+k} = i \mid D_{n+k-1} = 1, D_{n+k-2}, \ldots, D_n\}$ and show that X_{n+k} is independent of $D_{n+k-1}, D_{n+k-2}, \ldots, D_n$.

h) What do your results mean relative to Burke's theorem.

Exercise 5.14. Let $\{X_n, n \geq 1\}$ denote a irreducible recurrent Markov chain having a countable state state space. Now consider a new stochastic process $\{Y_n, n \geq 0\}$ that only accepts values of the Markov chain that are between 0 and some integer m. For instance, if $m = 3$ and $X_1 = 1$, $X_2 = 3$, $X_3 = 5$, $X_4 = 6$, $X_5 = 2$, then $Y_1 = 1$, $Y_2 = 3$, $Y_3 = 2$.

a) Is $\{Y_n, n \geq 0\}$ a Markov chain? Explain briefly.

b) Let p_j denote the proportion of time that $\{X_n, n \geq 1\}$ is in state j. If $p_j > 0$ for all j, what proportion of time is $\{Y_n, n \geq 0\}$ in each of the states $0, 1, \ldots, m$?

c) Suppose $\{X_n\}$ is null-recurrent and let $p_i(m), i = 0, 1, \ldots, m$ denote the long-run proportions for $\{Y_n, n \geq 0\}$. Show that $p_j(m) = p_i(m)E[\text{time the } X \text{ process spends in } j \text{ between returns to } i]$, $j \neq i$.}

Exercise 5.15. Verify that (5.53) is satisfied by the hypothesized solution to p in (5.57). Also show that the equations involving the idle state f are satisfied.

Exercise 5.16. Replace the state $\mathbf{m} = (m, z_1, \ldots, z_m)$ in Section 5.6 with an expanded state $\mathbf{m} = (m, z_1, w_1, z_2, w_2, \ldots, z_m, w_m)$ where m and $\{z_i; 1 \leq i \leq m\}$ are as before and $w_1, w_2, \ldots, w_m$ are the original service requirements of the m customers.

a) Hypothesizing the same backward round-robin system as hypothesized in Section 5.6, find the backward transition probabilities and give the corresponding equations to (5.51-5.54) for the expanded state description.

b) Solve the resulting equations to show that

$$\pi_{\mathbf{m}} = \pi + \phi\Big(\frac{\lambda\delta}{1-\lambda\delta}\Big)^m \prod_{j=1}^{m} f(w_j).$$

c) Show that the probability that there are m customers in the system, and that those customers have original service requirements given by $w_1, \ldots, w_m$, is

$$\Pr\{m, w_1, \ldots, w_m\} = \pi_\phi \left(\frac{\lambda\delta}{1-\lambda\delta}\right)^m \prod_{j=1}^{m} (w_j - 1) f(w_j).$$

d) Given that a customer has original service requirement w, find the expected time that customer spends in the system.

Chapter 6

MARKOV PROCESSES WITH COUNTABLE STATE SPACES

6.1 Introduction

Recall that a Markov chain is a discrete-time process $\{X_n;\, n \geq 0\}$ for which the state at each time $n \geq 1$ is an integer-valued random variable (rv) that is statistically dependent on $X_0, \ldots X_{n-1}$ only through X_{n-1}. A *countable-state Markov process*[1] (Markov process for short) is a generalization of a Markov chain in the sense that, along with the Markov chain $\{X_n;\, n \geq 1\}$, there is a randomly-varying holding interval in each state which is exponentially distributed with a parameter determined by the current state.

To be more specific, let $X_0{=}i,\ X_1{=}j,\ X_2 = k, \ldots$, denote a sample path of the sequence of states in the Markov chain (henceforth called the *embedded Markov chain*). Then the *holding interval* U_n between the time that state $X_{n-1} = \ell$ is entered and X_n is entered is a nonnegative exponential rv with parameter ν_ℓ, *i.e.*, for all $u \geq 0$,

$$\Pr\{U_n \leq u \mid X_{n-1} = \ell\} = 1 - \exp(-\nu_\ell u). \tag{6.1}$$

Furthermore, U_n, conditional on X_{n-1}, is jointly independent of X_m for all $m \neq n-1$ and of U_m for all $m \neq n$.

If we visualize starting this process at time 0 in state $X_0 = i$, then the first transition of the embedded Markov chain enters state $X_1 = j$ with the transition probability P_{ij} of the embedded chain. This transition occurs at time U_1, where U_1 is independent of X_1 and exponential with rate ν_i. Next, conditional on $X_1 = j$, the next transition enters state $X_2 = k$ with the transition probability P_{jk}. This transition occurs after an interval U_2, *i.e.*, at time $U_1 + U_2$, where U_2 is independent of X_2 and exponential with rate ν_j. Subsequent transitions occur similarly, with the new state, say $X_n = i$, determined from the old state, say $X_{n-1} = \ell$, via $P_{\ell i}$, and the new holding interval U_n determined via the exponential rate ν_ℓ. Figure 6.1 illustrates the statistical dependencies between the rv's $\{X_n;\, n \geq 0\}$ and $\{U_n;\, n \geq 1\}$.

[1] These processes are often called *continuous-time* Markov chains.

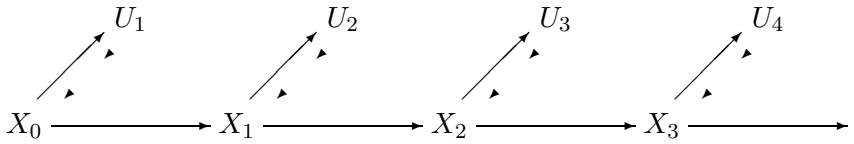


Figure 6.1: The statistical dependencies between the rv's of a Markov process. Each holding interval U_i, conditional on the current state X_{i-1}, is independent of all other states and holding intervals.

The epochs at which successive transitions occur are denoted $S_1, S_2, \ldots$, so we have $S_1 = U_1$, $S_2 = U_1 + U_2$, and in general $S_n = \sum_{m=1}^{n} U_m$ for $n \geq 1$ wug $S_0 = 0$. The state of a Markov process at any time $t > 0$ is denoted by $X(t)$ and is given by

$$X(t) = X_n \quad \text{for } S_n \leq t < S_{n+1} \qquad \text{for each } n \geq 0.$$

This defines a stochastic process $\{X(t);\, t \geq 0\}$ in the sense that each sample point $\omega \in \Omega$ maps to a sequence of sample values of $\{X_n;\, n \geq 0\}$ and $\{S_n;\, n \geq 1\}$, and thus into a sample function of $\{X(t);\, t \geq 0\}$. This stochastic process is what is usually referred to as a Markov process, but it is often simpler to view $\{X_n;\, n \geq 0\}, \{S_n;\, n \geq 1\}$ as a characterization of the process. Figure 6.2 illustrates the relationship between all these quantities.

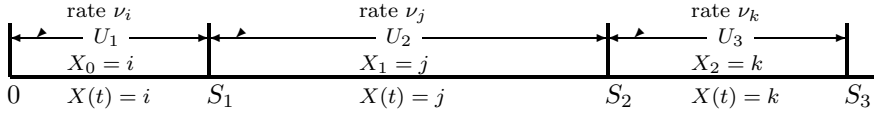


Figure 6.2: The relationship of the holding intervals $\{U_n; n \geq 1\}$ and the epochs $\{S_n; n \geq 1\}$ at which state changes occur. The state $X(t)$ of the Markov process and the corresponding state of the embedded Markov chain are also illustrated. Note that if $X_n = i$, then $X(t) = i$ for $S_n \leq t < S_{n+1}$

This can be summarized in the following definition.

Definition 6.1.1. *A countable-state Markov process $\{X(t);\, t \geq 0\}$ is a stochastic process mapping each nonnegative real number t to the nonnegative integer-valued rv $X(t)$ in such a way that for each $t \geq 0$,*

$$X(t) = X_n \quad \textit{for } S_n \leq t < S_{n+1}; \quad S_0 = 0;\ S_n = \sum_{m=1}^{n} U_m \quad \textit{for } n \geq 1, \tag{6.2}$$

where $\{X_n;\, n \geq 0\}$ is a Markov chain with a countably infinite or finite state space and each U_n, given $X_{n-1} = i$, is exponential with rate $\nu_i > 0$ and is conditionally independent of all other U_m and X_m.

The tacit assumptions that the state space is the set of nonnegative integers and that the process starts at $t = 0$ are taken only for notational simplicity but will serve our needs here.

We assume throughout this chapter (except in a few places where specified otherwise) that the embedded Markov chain has no self transitions, *i.e.*, $P_{ii} = 0$ for all states i. One

reason for this is that such transitions are invisible in $\{X(t);\, t \geq 0\}$. Another is that with this assumption, the sample functions of $\{X(t);\, t \geq 0\}$ and the joint sample functions of $\{X_n;\, n \geq 0\}$ and $\{U_n;\, n \geq 1\}$ uniquely specify each other.

We are not interested for the moment in exploring the probability distribution of $X(t)$ for given values of t, but one important feature of this distribution is that for any times $t > \tau > 0$ and any states i, j,

$$\Pr\{X(t)=j \mid X(\tau)=i, \{X(s) = x(s);\, s < \tau\}\} = \Pr\{X(t-\tau)=j \mid X(0)=i\}\,. \tag{6.3}$$

This property arises because of the memoryless property of the exponential distribution. That is, if $X(\tau) = i$, it makes no difference how long the process has been in state i before τ; the time to the next transition is still exponential with rate ν_i and subsequent states and holding intervals are determined as if the process starts in state i at time 0. This will be seen more clearly in the following exposition. This property is the reason why these processes are called Markov, and is often taken as the defining property of Markov processes.

Example 6.1.1. The M/M/1 queue: An M/M/1 queue has Poisson arrivals at a rate denoted by λ and has a single server with an exponential service distribution of rate $\mu > \lambda$ (see Figure 6.3). Successive service times are independent, both of each other and of arrivals. The state $X(t)$ of the queue is the total number of customers either in the queue or in service. When $X(t) = 0$, the time to the next transition is the time until the next arrival, *i.e.*, $\nu_0 = \lambda$. When $X(t) = i$, $i \geq 1$, the server is busy and the time to the next transition is the time until either a new arrival occurs or a departure occurs. Thus $\nu_i = \lambda + \mu$. For the embedded Markov chain, $P_{01} = 1$ since only arrivals are possible in state 0, and they increase the state to 1. In the other states, $P_{i,i-1} = \mu/(\lambda+\mu)$ and $P_{i,i+1} = \lambda/(\lambda+\mu)$.

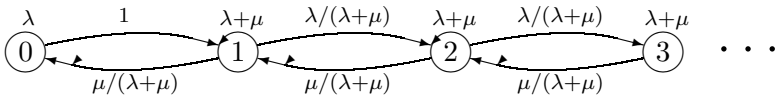


Figure 6.3: The embedded Markov chain for an M/M/1 queue. Each node i is labeled with the corresponding rate ν_i of the exponentially distributed holding interval to the next transition. Each transition, say i to j, is labeled with the corresponding transition probability P_{ij} in the embedded Markov chain.

The embedded Markov chain is a Birth-death chain, and its steady state probabilities can be calculated easily using (5.25). The result is

$$\begin{aligned} \pi_0 &= \frac{1-\rho}{2} \qquad \text{where } \rho = \frac{\lambda}{\mu} \\ \pi_n &= \frac{1-\rho^2}{2}\rho^{n-1} \qquad \text{for } n \geq 1. \end{aligned} \tag{6.4}$$

Note that if $\lambda << \mu$, then π_0 and π_1 are each close to 1/2 (*i.e.*, the embedded chain mostly alternates between states 0 and 1, and higher ordered states are rarely entered), whereas because of the large holding interval in state 0, the process spends most of its time in state 0 waiting for arrivals. The steady-state probability π_i of state i in the embedded chain

is the long-term fraction of the total transitions that go to state i. We will shortly learn how to find the long term *fraction of time* spent in state i as opposed to this fraction of transitions, but for now we return to the general study of Markov processes.

The evolution of a Markov process can be visualized in several ways. We have already looked at the first, in which for each state $X_{n-1} = i$ in the embedded chain, the next state X_n is determined by the probabilities $\{P_{ij};\, j \geq 0\}$ of the embedded Markov chain, and the holding interval U_n is independently determined by the exponential distribution with rate ν_i.

For a second viewpoint, suppose an independent Poisson process of rate $\nu_i > 0$ is associated with each state i. When the Markov process enters a given state i, the next transition occurs at the next arrival epoch in the Poisson process for state i. At that epoch, a new state is chosen according to the transition probabilities P_{ij}. Since the choice of next state, given state i, is independent of the interval in state i, this view describes the same process as the first view.

For a third visualization, suppose, for each pair of states i and j, that an independent Poisson process of rate $\nu_i P_{ij}$ is associated with a possible transition to j conditional on being in i. When the Markov process enters a given state i, both the time of the next transition and the choice of the next state are determined by the set of i to j Poisson processes over all possible next states j. The transition occurs at the epoch of the first arrival, for the given i, to any of the i to j processes, and the next state is the j for which that first arrival occurred. Since such a collection of independent Poisson processes is equivalent to a single process of rate ν_i followed by an independent selection according to the transition probabilities P_{ij}, this view again describes the same process as the other views.

It is convenient in this third visualization to define the rate from any state i to any other state j as

$$q_{ij} = \nu_i P_{ij}.$$

If we sum over j, we see that ν_i and P_{ij} are also uniquely determined by $\{q_{ij};\, i, j \geq 0\}$ as

$$\nu_i = \sum_j q_{ij}; \qquad P_{ij} = q_{ij}/\nu_i. \tag{6.5}$$

This means that the fundamental characterization of the Markov process in terms of the P_{ij} and the ν_i can be replaced by a characterization in terms of the set of transition rates q_{ij}. In many cases, this is a more natural approach. For the M/M/1 queue, for example, $q_{i,i+1}$ is simply the arrival rate λ. Similarly, $q_{i,i-1}$ is the departure rate μ when there are customers to be served, *i.e.*, when $i > 0$. Figure 6.4 shows Figure 6.3 incorporating this notational simplification.

Note that the interarrival density for the Poisson process, from any given state i to other state j, is given by $q_{ij} \exp(-q_{ij}x)$. On the other hand, given that the process is in state i, the probability density for the interval until the next transition, whether conditioned on the next state or not, is $\nu_i \exp(-\nu_i x)$ where $\nu_i = \sum_j q_{ij}$. One might argue, incorrectly, that, conditional on the next transition being to state j, the time to that transition has density $q_{ij} \exp(-q_{ij}x)$. Exercise 6.1 uses an M/M/1 queue to provide a guided explanation of why this argument is incorrect.

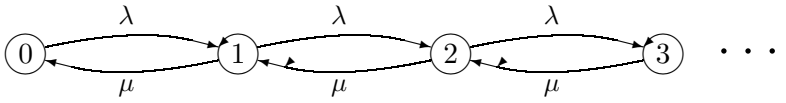


Figure 6.4: The Markov process for an M/M/1 queue. Each transition (i, j) is labelled with the corresponding transition rate q_{ij}.

6.1.1 The sampled-time approximation to a Markov process

As yet another way to visualize a Markov process, consider approximating the process by viewing it only at times separated by a given increment size δ. The Poisson processes above are then approximated by Bernoulli processes where the transition probability from i to j in the sampled-time chain is defined to be $q_{ij}\delta$ for all $j \neq i$.

The Markov process is then approximated by a Markov chain. Since each δq_{ij} decreases with decreasing δ, there is an increasing probability of no transition out of any given state in the time increment δ. These must be modeled with self-transition probabilities, say $P_{ii}(\delta)$ which must satisfy

$$P_{ii}(\delta) = 1 - \sum_j q_{ij}\delta = 1 - \nu_i\delta \qquad \text{for each } i \geq 0.$$

This is illustrated in Figure 6.5 for the M/M/1 queue. Recall that this sampled-time M/M/1 Markov chain was analyzed in Section 5.4 and the steady-state probabilities were shown to be

$$p_i(\delta) = (1-\rho)\rho^i \qquad \text{for all } i \geq 0 \quad \text{where } \rho = \lambda/\mu. \tag{6.6}$$

We have denoted the steady-state probabilities here by $p_i(\delta)$ to avoid confusion with the steady-state probabilities for the embedded chain. As discussed later, the steady-state probabilities in (6.6) do not depend on δ, so long as δ is small enough that the self-transition probabilities are nonnegative.

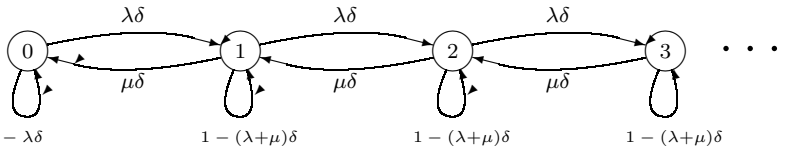


Figure 6.5: Approximating an M/M/1 queue by a sampled-time Markov chain.

This sampled-time approximation is an approximation in two ways. First, transitions occur only at integer multiples of the increment δ, and second, $q_{ij}\delta$ is an approximation to $\Pr\{X(\delta)=j \mid X(0)=i\}$. From (6.3), $\Pr\{X(\delta)=j \mid X(0)=i\} = q_{ij}\delta + o(\delta)$, so this second approximation is increasingly good as $\delta \to 0$.

As observed above, the steady-state probabilities for the sampled-time approximation to an M/M/1 queue do not depend on δ. As seen later, whenever the embedded chain is positive

recurrent and a sampled-time approximation exists for a Markov process, then the steady-state probability of each state i is independent of δ and represents the limiting fraction of time spent in state i with probability 1.

Figure 6.6 illustrates the sampled-time approximation of a generic Markov process. Note that $P_{ii}(\delta)$ is equal to $1-\delta\nu_i$ for each i in any such approximation, and thus it is necessary for δ to be small enough to satisfy $\nu_i\delta \leq 1$ for all i. For a finite state space, this is satisfied for any $\delta \leq [\max_i \nu_i]^{-1}$. For a countably infinite state space, however, the sampled-time approximation requires the existence of some finite B such that $\nu_i \leq B$ for all i. The consequences of having no such bound are explored in the next section.

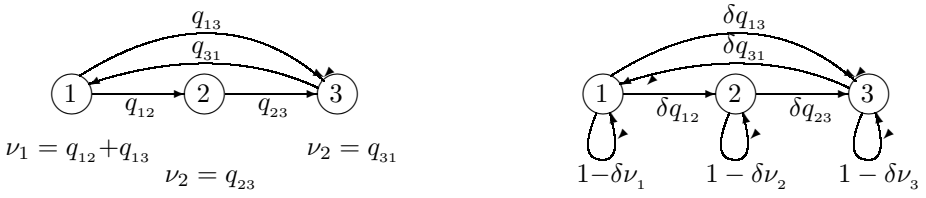


Figure 6.6: Approximating a generic Markov process by its sampled-time Markov chain.

6.2 Steady-state behavior of irreducible Markov processes

Definition 6.2.1 (Irreducible Markov processes). *An irreducible Markov process is a Markov process for which the embedded Markov chain is irreducible (*i.e.*, all states are in the same class).*

The analysis in this chapter is restricted almost entirely to irreducible Markov processes. The reason for this restriction is not that Markov processes with multiple classes of states are unimportant, but rather that they can usually be best understood by first looking at the embedded Markov chains for the various classes making up that overall chain.

We will ask the same types of steady-state questions for Markov processes as we asked about Markov chains. In particular, under what conditions is there a set of steady-state probabilities, $p_0, p_1, \ldots$ with the property that for any given starting state i, the limiting fraction of of time spent in any given state j is p_j with probability 1? Do these probabilities also have the property that $p_j = \lim_{t\to\infty} \Pr\{X(t) = j \mid X_0 = i\}$?

We will find that simply having a positive-recurrent embedded Markov chain is not quite enough to ensure that such a set of probabilities exists. It is also necessary for the embedded-chain steady-state probabilities $\{\pi_i;\, i \geq 0\}$ and the holding-interval parameters $\{\nu_i; i \geq 0\}$ to satisfy $\sum_i \pi_i/\nu_i < \infty$. We will interpret this latter condition as asserting that the limiting long-term rate at which transitions occur must be strictly positive. Finally we will show that when these conditions are satisfied, the steady-state probabilities for the process are

related to those of the embedded chain by

$$p_j = \frac{\pi_j/\nu_j}{\sum_k \pi_k/\nu_k}. \tag{6.7}$$

Definition 6.2.2. *The steady-state process probabilities, $p_0, p_1, \ldots$ for a Markov process are a set of numbers satisfying (6.7), where $\{\pi_i;\ i \geq 0\}$ and $\{\nu_i;\ i \geq 0\}$ are the steady-state probabilities for the embedded chain and the holding-interval rates respectively.*

As one might guess, the appropriate approach to answering these questions comes from applying renewal theory to various renewal processes associated with the Markov process. Many of the needed results for this have already been developed in looking at the steady-state behavior of countable-state Markov chains.

We start with a very technical lemma that will perhaps appear obvious, and the reader is welcome to ignore the proof until perhaps questioning the issue later. The lemma is not restricted to irreducible processes, although we only use it in that case.

Lemma 6.2.1. *Consider a Markov process for which the embedded chain starts in some given state i. Then the holding time intervals, $U_1, U_2, \ldots$ are all rv's. Let $M_i(t)$ be the number of transitions made by the process up to and including time t. Then with probability 1 (WP1),*

$$\lim_{t\to\infty} M_i(t) = \infty. \tag{6.8}$$

Proof: The first holding interval U_1 is exponential with rate $\nu_i > 0$, so it is clearly a rv (*i.e.*, not defective). In general the state after the $(n-1)$th transition has the PMF P_{ij}^{n-1}, so the complementary distribution function of U_n is

$$\begin{aligned}\Pr\{U_n > u\} &= \lim_{k\to\infty} \sum_{j=1}^{k} P_{ij}^{n-1} \exp(-\nu_j u). \\ &\leq \sum_{j=1}^{k} P_{ij}^{n-1} \exp(-\nu_j u) + \sum_{j=k+1}^{\infty} P_{ij}^{n+1} \qquad \text{for every } k.\end{aligned}$$

For each k, the first sum above approaches 0 with increasing u and the second sum approaches 0 with increasing k so the limit as $u \to \infty$ must be 0 and U_n is a rv.

It follows that each $S_n = U_1 + \cdots + U_n$ is also a rv. Now $\{S_n;\ n \geq 1\}$ is the sequence of arrival epochs in an arrival process, so we have the set equality $\{S_n \leq t\} = \{M_i(t) \geq n\}$ for each choice of n. Since S_n is a rv, we have $\lim_{t\to\infty} \Pr\{S_n \leq t\} = 1$ for each n. Thus $\lim_{t\to\infty} \Pr\{M_i(t) \geq n\} = 1$ for all n. This means that the set of sample points ω for which $\lim_{t\to\infty} M_i(t,\omega) < n$ has probability 0 for all n, and thus $\lim_{t\to\infty} M_i(t,\omega) = \infty$ WP1. □

6.2.1 Renewals on successive entries to a given state

For an irreducible Markov process with $X_0 = i$, let $M_{ij}(t)$ be the number of transitions into state j over the interval $(0, t]$. We want to find when this is a delayed renewal counting

process. It is clear that the sequence of epochs at which state j is entered form renewal points, since they form renewal points in the embedded Markov chain and the holding intervals between transitions depend only on the current state. The questions are whether the first entry to state j must occur within some finite time, and whether recurrences to j must occur within finite time. The following lemma answers these questions for the case where the embedded chain is recurrent (either positive recurrent or null recurrent).

Lemma 6.2.2. *Consider a Markov process with an irreducible recurrent embedded chain $\{X_n;\, n \geq 0\}$. Given $X_0 = i$, let $\{M_{ij}(t);\, t \geq 0\}$ be the number of transitions into a given state j in the interval $(0, t]$. Then $\{M_{ij}(t);\, t \geq 0\}$ is a delayed renewal counting process (or an ordinary renewal counting process if $j = i$).*

Proof: Given $X_0 = i$, let $N_{ij}(n)$ be the number of transitions into state j that occur in the embedded Markov chain by the n^{th} transition of the embedded chain. From Lemma 5.1.4, $\{N_{ij}(n);\, n \geq 0\}$ is a delayed renewal process, so from Lemma 4.8.2, $\lim_{n\to\infty} N_{ij}(n) = \infty$ with probability 1. Note that $M_{ij}(t) = N_{ij}(M_i(t))$, where $M_i(t)$ is the total number of state transitions (between all states) in the interval $(0, t]$. Thus, with probability 1,

$$\lim_{t\to\infty} M_{ij}(t) = \lim_{t\to\infty} N_{ij}(M_i(t)) = \lim_{n\to\infty} N_{ij}(n) = \infty.$$

where we have used Lemma 6.2.1, which asserts that $\lim_{t\to\infty} M_i(t) = \infty$ with probability 1.

It follows that the interval W_1 until the first transition to state j, and the subsequent interval W_2 until the next transition to state j, are both finite with probability 1. Subsequent intervals have the same distribution as W_2, and all intervals are independent, so $\{M_{ij}(t);\, t \geq 0\}$ is a delayed renewal process with inter-renewal intervals $\{W_k;\, k \geq 1\}$. If $i = j$, then all W_k are identically distributed and we have an ordinary renewal process, completing the proof. □

The inter-renewal intervals $W_2, W_3, \ldots$ for $\{M_{ij}(t);\, t \geq 0\}$ above are well-defined nonnegative IID rv's whose distribution depends on j but not i. They either have an expectation as a finite number or have an infinite expectation. In either case, this expectation is denoted as $\mathsf{E}[W(j)] = \overline{W}(j)$. This is the mean time between successive entries to state j, and we will see later that in some cases this mean time can be infinite.

6.2.2 The limiting fraction of time in each state

In order to study the fraction of time spent in state j, we define a delayed renewal-reward process, based on $\{M_{ij}(t);\, t \geq 0\}$, for which unit reward is accumulated whenever the process is in state j. That is (given $X_0 = i$), $R_{ij}(t) = 1$ when $X(t) = j$ and $R_{ij}(t) = 0$ otherwise. (see Figure 6.7). Given that transition $n-1$ of the embedded chain enters state j, the interval U_n is exponential with rate ν_j, so $\mathsf{E}[U_n \mid X_{n-1}{=}j] = 1/\nu_j$.

Let $p_j(i)$ be the limiting time-average fraction of time spent in state j. We will see later that such a limit exists WP1, that the limit does not depend on i, and that it is equal to

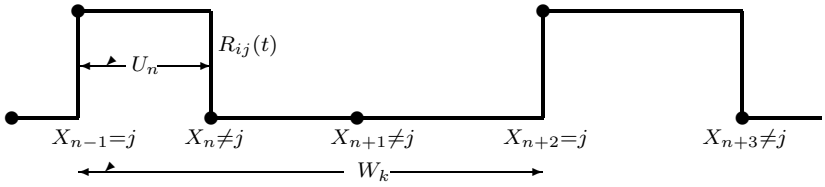


Figure 6.7: The delayed renewal-reward process $\{R_{ij}(t);\, t \geq 0\}$ for time in state j. The reward is one whenever the process is in state j, *i.e.*, $R_{ij}(t) = \mathbb{I}_{\{X(t)=j\}}$. A renewal occurs on each entry to state j, so the reward starts at each such entry and continues until a state transition, assumed to enter a state other than j. The reward then ceases until the next renewal, *i.e.*, the next entry to state j. The figure illustrates the kth inter-renewal interval, of duration W_k, which is assumed to start on the $n-1$st state transition. The expected interval over which a reward is accumulated is ν_j and the expected duration of the inter-renewal interval is $\overline{W}(j)$.

the steady-state probability p_j in (6.7). Since $\overline{U}(j) = 1/\nu_j$, Theorems 4.4.1 and 4.8.4, for ordinary and delayed renewal-reward processes respectively, state that[2]

$$p_j(i) = \lim_{t\to\infty} \frac{\int_0^t R_{ij}(\tau)d\tau}{t} \qquad \text{WP1} \tag{6.9}$$

$$= \frac{\overline{U}(j)}{\overline{W}(j)} = \frac{1}{\nu_j \overline{W}(j)}. \tag{6.10}$$

This shows that the limiting time average, $p_j(i)$, exists with probability 1 and is independent of the starting state i. We show later that it is the steady-state process probability given by (6.7). We can also investigate the limit, as $t \to \infty$, of the probability that $X(t) = j$. This is equal to $\lim_{t\to\infty} \mathsf{E}[R(t)]$ for the renewal-reward process above. Because of the exponential holding intervals, the inter-renewal times are non-arithmetic, and from Blackwell's theorem, in the form of (4.105),

$$\lim_{t\to\infty} \Pr\{X(t) = j\} = \frac{1}{\nu_j \overline{W}(j)} = p_j(i). \tag{6.11}$$

We summarize these results in the following lemma.

Lemma 6.2.3. *Consider an irreducible Markov process with a recurrent embedded Markov chain starting in $X_0 = i$. Then with probability 1, the limiting time average in state j is given by $p_j(i) = \frac{1}{\nu_j \overline{W}(j)}$. This is also the limit, as $t \to \infty$, of $\Pr\{X(t) = j\}$.*

6.2.3 Finding $\{p_j(i);\, j \geq 0\}$ in terms of $\{\pi_j;\, j \geq 0\}$

Next we must express the mean inter-renewal time, $\overline{W}(j)$, in terms of more accessible quantities that allow us to show that $p_j(i) = p_j$ where p_j is the steady-state process probability

[2]Theorems 4.4.1 and 4.8.4 do not cover the case where $\overline{W}(j) = \infty$, but, since the expected reward per renewal interval is finite, it is not hard to verify (6.9) in that special case.

of Definition 6.2.2. We now assume that the embedded chain is not only recurrent but also positive recurrent with steady-state probabilities $\{\pi_j;\ j \geq 0\}$. We continue to assume a given starting state $X_0 = i$. Applying the strong law for delayed renewal processes (Theorem 4.8.1) to the Markov process,

$$\lim_{t\to\infty} M_{ij}(t)/t = 1/\overline{W}(j) \qquad \text{WP1.} \tag{6.12}$$

As before, $M_{ij}(t) = N_{ij}(M_i(t))$. Since $\lim_{t\to\infty} M_i(t) = \infty$ with probability 1,

$$\lim_{t\to\infty} \frac{M_{ij}(t)}{M_i(t)} = \lim_{t\to\infty} \frac{N_{ij}(M_i(t))}{M_i(t)} = \lim_{n\to\infty} \frac{N_{ij}(n)}{n} = \pi_j \quad \text{WP1.} \tag{6.13}$$

In the last step, we applied the same strong law to the embedded chain. Combining (6.12) and (6.13), the following equalities hold with probability 1.

$$\begin{aligned} \frac{1}{\overline{W}(j)} &= \lim_{t\to\infty} \frac{M_{ij}(t)}{t} \\ &= \lim_{t\to\infty} \frac{M_{ij}(t)}{M_i(t)} \frac{M_i(t)}{t} \\ &= \pi_j \lim_{t\to\infty} \frac{M_i(t)}{t}. \end{aligned} \tag{6.14}$$

This tells us that $\overline{W}(j)\pi_j$ is the same for all j. Also, since $\pi_j > 0$ for a positive recurrent chain, it tell us that if $\overline{W}(j) < \infty$ for one state j, it is finite for all states. Also these expected recurrence times are finite if and only if $\lim_{t\to\infty} M_i(t)/t > 0$. Finally, it says implicitly that $\lim_{t\to\infty} M_i(t)/t$ exists WP1 and has the same value for all starting states i.

There is relatively little left to do, and the following theorem does most of it.

Theorem 6.2.1. *Consider an irreducible Markov process with a positive recurrent embedded Markov chain. Let $\{\pi_j;\ j \geq 0\}$ be the steady-state probabilities of the embedded chain and let $X_0 = i$ be the starting state. Then, with probability 1, the limiting time-average fraction of time spent in any arbitrary state j is the steady-state process probability in (6.7),* i.e.*,*

$$p_j(i) = p_j = \frac{\pi_j/\nu_j}{\sum_k \pi_k/\nu_k}. \tag{6.15}$$

The expected time between returns to state j is

$$\overline{W}(j) = \frac{\sum_k \pi_k/\nu_k}{\pi_j}, \tag{6.16}$$

and the limiting rate at which transitions take place is independent of the starting state and given by

$$\lim_{t\to\infty} \frac{M_i(t)}{t} = \frac{1}{\sum_k \pi_k/\nu_k} \qquad \text{WP1.} \tag{6.17}$$

Discussion: Recall that $p_j(i)$ was defined as a time average WP1, and we saw earlier that this time average existed with a value independent of i. The theorem states that this time average (and the limiting ensemble average) is given by the steady-state process probabilities in (6.7). Thus, after the proof, we can stop distinguishing these quantities.

At a superficial level, the theorem is almost obvious from what we have done. In particular, substituting (6.14) into (6.9), we see that

$$p_j(i) = \frac{\pi_j}{\nu_j} \lim_{t\to\infty} \frac{M_i(t)}{t} \qquad \text{WP1.} \tag{6.18}$$

Since $p_j(i) = \lim_{t\to\infty} \Pr\{X(t) = j\}$, and since $X(t)$ is in some state at all times, we would conjecture (and even insist if we didn't read on) that $\sum_j p_j(i) = 1$. Adding that condition to normalize (6.18), we get (6.15), and (6.16) and (6.17) follow immediately. The trouble is that if $\sum_j \pi_j/\nu_j = \infty$, then (6.15) says that $p_j = 0$ for all j, and (6.17) says that $\lim M_i(t)/t = 0$, *i.e.*, the process 'gets tired' with increasing t and the rate of transitions decreases toward 0. The rather technical proof to follow deals with these limits more carefully.

Proof: We have seen in (6.14) that $\lim_{t\to\infty} M_i(t)/t$ is equal to a constant, say α, with probability 1 and that this constant is the same for all starting states i. We first consider the case where $\alpha > 0$. In this case, from (6.14), $\overline{W}(j) < \infty$ for all j. Choosing any given j and any positive integer ℓ, consider a renewal-reward process with renewals on transitions to j and a reward $R^\ell_{ij}(t) = 1$ when $X(t) \le \ell$. This reward is independent of j and equal to $\sum_{k=1}^{\ell} R_{ik}(t)$. Thus, from (6.9), we have

$$\lim_{t\to\infty} \frac{\int_0^t R^\ell_{ij}(\tau)d\tau}{t} = \sum_{k=1}^{\ell} p_k(i). \tag{6.19}$$

If we let $\mathsf{E}\left[\mathrm{R}^\ell_j\right]$ be the expected reward over a renewal interval, then, from Theorem 4.8.4,

$$\lim_{t\to\infty} \frac{\int_0^t R^\ell_{ij}(\tau)d\tau}{t} = \frac{\mathsf{E}\left[\mathrm{R}^\ell_j\right]}{\overline{U}_j}. \tag{6.20}$$

Note that $\mathsf{E}\left[\mathrm{R}^\ell_j\right]$ above is non-decreasing in ℓ and goes to the limit $\overline{W}(j)$ as $\ell \to \infty$. Thus, combining (6.19) and (6.20), we see that

$$\lim_{\ell\to\infty} \sum_{k=1}^{\ell} p_k(i) = 1.$$

With this added relation, (6.15), (6.16), and (6.17) follow as in the discussion. This completes the proof for the case where $\alpha > 0$.

For the remaining case, where $\lim_{t\to\infty} M_i(t)/t = \alpha = 0$, (6.14) shows that $\overline{W}(j) = \infty$ for all j and (6.18) then shows that $p_j(i) = 0$ for all j. We give a guided proof in Exercise 6.6 that, for $\alpha = 0$, we must have $\sum_i \pi_i/\nu_i = \infty$. It follows that (6.15), (6.16), and (6.17) are all satisfied. □

This has been quite a difficult proof for something that might seem almost obvious for simple examples. However, the fact that these time-averages are valid over all sample points with probability 1 is not obvious and the fact that $\pi_j \overline{W}(j)$ is independent of j is certainly not obvious.

The most subtle thing here, however, is that if $\sum_i \pi_i/\nu_i = \infty$, then $p_j = 0$ for all states j. This is strange because the time-average state probabilities do not add to 1, and also strange because the embedded Markov chain continues to make transitions, and these transitions, in steady state for the Markov chain, occur with the probabilities π_i. Example 6.2.1 and Exercise 6.3 give some insight into this. Some added insight can be gained by looking at the embedded Markov chain starting in steady state, *i.e.*, with probabilities $\{\pi_i;\, i \geq 0\}$. Given $X_0 = i$, the expected time to a transition is $1/\nu_i$, so the unconditional expected time to a transition is $\sum_i \pi_i/\nu_i$, which is infinite for the case under consideration. This is not a phenomenon that can be easily understood intuitively, but Example 6.2.1 and Exercise 6.3 will help.

6.2.4 Solving for the steady-state process probabilities directly

Let us return to the case where $\sum_k \pi_k/\nu_k < \infty$, which is the case of virtually all applications. We have seen that a Markov process can be specified in terms of the time-transitions $q_{ij} = \nu_i P_{ij}$, and it is useful to express the steady-state equations for p_j directly in terms of q_{ij} rather than indirectly in terms of the embedded chain. As a useful prelude to this, we first express the π_j in terms of the p_j. Denote $\sum_k \pi_k/\nu_k$ as $\beta < \infty$. Then, from (6.15), $p_j = \pi_j/\nu_j\beta$, so $\pi_j = p_j\nu_j\beta$. Expressing this along with the normalization $\sum_k \pi_k = 1$, we obtain

$$\pi_i = \frac{p_i\nu_i}{\sum_k\ p_k\nu_k}. \tag{6.21}$$

Thus, $\beta = 1/\sum_k p_k\nu_k$, so

$$\sum_k \pi_k/\nu_k = \frac{1}{\sum_k p_k\nu_k}. \tag{6.22}$$

We can now substitute π_i as given by (6.21) into the steady-state equations for the embedded Markov chain, *i.e.*, $\pi_j = \sum_i \pi_i P_{ij}$ for all j, obtaining

$$p_j\nu_j = \sum_i p_i\nu_i P_{ij}$$

for each state j. Since $\nu_i P_{ij} = q_{ij}$,

$$p_j\nu_j = \sum_i p_i q_{ij}; \qquad \sum_i p_i = 1. \tag{6.23}$$

This set of equations is known as the steady-state equations for the Markov process. The normalization condition $\sum_i p_i = 1$ is a consequence of (6.22) and also of (6.15). Equation (6.23) has a nice interpretation in that the term on the left is the steady-state rate at which

transitions occur out of state j and the term on the right is the rate at which transitions occur into state j. Since the total number of entries to j must differ by at most 1 from the exits from j for each sample path, this equation is not surprising.

The embedded chain is positive recurrent, so its steady-state equations have a unique solution with all $\pi_i > 0$. Thus (6.23) also has a unique solution with all $p_i > 0$ under the the added condition that $\sum_i \pi_i/\nu_i < \infty$. However, we would like to solve (6.23) directly without worrying about the embedded chain.

If we find a solution to (6.23), however, and if $\sum_i p_i\nu_i < \infty$ in that solution, then the corresponding set of π_i from (6.21) must be the unique steady-state solution for the embedded chain. Thus the solution for p_i must be the corresponding steady-state solution for the Markov process. This is summarized in the following theorem.

Theorem 6.2.2. *Assume an irreducible Markov process and let $\{p_i;\, i \geq 0\}$ be a solution to (6.23). If $\sum_i\ p_i\nu_i < \infty$, then, first, that solution is unique, second, each p_i is positive, and third, the embedded Markov chain is positive recurrent with steady-state probabilities satisfying (6.21). Also, if the embedded chain is positive recurrent, and $\sum_i \pi_i/\nu_i < \infty$ then the set of p_i satisfying (6.15) is the unique solution to (6.23).*

6.2.5 The sampled-time approximation again

For an alternative view of the probabilities $\{p_i;\, i \geq 0\}$, consider the special case (but the typical case) where the transition rates $\{\nu_i;\, i \geq 0\}$ are bounded. Consider the sampled-time approximation to the process for a given increment size $\delta \leq [\max_i \nu_i]^{-1}$ (see Figure 6.6). Let $\{p_i(\delta);\, i \geq 0\}$ be the set of steady-state probabilities for the sampled-time chain, assuming that they exist. These steady-state probabilities satisfy

$$p_j(\delta) = \sum_{i\neq j} p_i(\delta)q_{ij}\delta + p_j(\delta)(1-\nu_j\delta); \quad p_j(\delta) \geq 0; \quad \sum_j p_j(\delta) = 1. \tag{6.24}$$

The first equation simplifies to $p_j(\delta)\nu_j = \sum_{i\neq j} p_i(\delta)q_{ij}$, which is the same as (6.23). It follows that the steady-state probabilities $\{p_i;\, i \geq 0\}$ for the process are the same as the steady-state probabilities $\{p_i(\delta);\, i \geq 0\}$ for the sampled-time approximation. Note that this is not an approximation; $p_i(\delta)$ is exactly equal to p_i for all values of $\delta \leq 1/\sup_i \nu_i$. We shall see later that the dynamics of a Markov process are not quite so well modeled by the sampled time approximation except in the limit $\delta \to 0$.

6.2.6 Pathological cases

Example 6.2.1 (Zero transition rate). Consider the Markov process with a positive-recurrent embedded chain in Figure 6.8. This models a variation of an M/M/1 queue in which the server becomes increasingly rattled and slow as the queue builds up, and the customers become almost equally discouraged about entering. The downward drift in the transitions is more than overcome by the slow-down in large numbered states. Transitions continue to occur, but the number of transitions per unit time goes to 0 with increasing

time. Although the embedded chain has a steady-state solution, the process can not be viewed as having any sort of steady state. Exercise 6.3 gives some added insight into this type of situation.

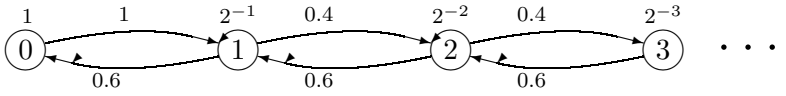


Figure 6.8: The Markov process for a variation on M/M/1 where arrivals and services get slower with increasing state. Each node i has a rate $\nu_i = 2^{-i}$. The embedded chain transition probabilities are $P_{i,i+1} = 0.4$ for $i \geq 1$ and $P_{i,i-1} = 0.6$ for $i \geq 1$, thus ensuring that the embedded Markov chain is positive recurrent. Note that $q_{i,i+1} > q_{i+1,i}$, thus ensuring that the Markov process drifts to the right.

It is also possible for (6.23) to have a solution for $\{p_i;\ i \geq 0\}$ with $\sum_i\ p_i = 1$, but $\sum_i\ p_i\nu_i = \infty$. This is not possible for a positive recurrent embedded chain, but is possible both if the embedded Markov chain is transient and if it is null recurrent. A transient chain means that there is a positive probability that the embedded chain will *never* return to a state after leaving it, and thus there can be no sensible kind of steady-state behavior for the process. These processes are characterized by arbitrarily large transition rates from the various states, and these allow the process to transit through an infinite number of states in a finite time.

Processes for which there is a non-zero probability of passing through an infinite number of states in a finite time are called *irregular*. Exercises 6.8 and 6.7 give some insight into irregular processes. Exercise 6.9 gives an example of a process that is not irregular, but for which (6.23) has a solution with $\sum_i p_i = 1$ and the embedded Markov chain is null recurrent. We restrict our attention in what follows to irreducible Markov chains for which (6.23) has a solution, $\sum\}p_i = 1$, and $\sum\}p_i\nu_i < \infty$. This is slightly more restrictive than necessary, but processes for which $\sum_i\ p_i\nu_i = \infty$ (see Exercise 6.9) are not very robust.

6.3 The Kolmogorov differential equations

Let $P_{ij}(t)$ be the probability that a Markov process $\{X(t);\ t \geq 0\}$ is in state j at time t given that $X(0) = i$,

$$P_{ij}(t) = P\{X(t){=}j \mid X(0){=}i\}. \tag{6.25}$$

$P_{ij}(t)$ is analogous to the n^{th} order transition probabilities P_{ij}^n for Markov chains. We have already seen that $\lim_{t\to\infty} P_{ij}(t) = p_j$ for the case where the embedded chain is positive recurrent and $\sum_i\ \pi_i/\nu_i < \infty$. Here we want to find the transient behavior, and we start by deriving the Chapman-Kolmogorov equations for Markov processes. Let s and t be

arbitrary times, $0 < s < t$. By including the state at time s, we can rewrite (6.25) as

$$\begin{aligned} P_{ij}(t) &= \sum_k \Pr\{X(t)=j,\ X(s)=k \mid X(0)=i\} \\ &= \sum_k \Pr\{X(s)=k \mid X(0)=i\}\ \Pr\{X(t)=j \mid X(s)=k\}\,; \quad \text{all } i,j, \end{aligned} \tag{6.26}$$

where we have used the Markov condition, (6.3). Given that $X(s) = k$, the residual time until the next transition after s is exponential with rate ν_k, and thus the process starting at time s in state k is statistically identical to that starting at time 0 in state k. Thus, for any s, $0 \le s \le t$, we have

$$P\{X(t)=j \mid X(s)=k\} = P_{kj}(t-s).$$

Substituting this into (6.26), we have the *Chapman-Kolmogorov equations* for a Markov process,

$$P_{ij}(t) = \sum_k P_{ik}(s)P_{kj}(t-s). \tag{6.27}$$

These equations correspond to (3.8) for Markov chains. We now use these equations to derive two types of sets of differential equations for $P_{ij}(t)$. The first are called the *Kolmogorov forward differential equations*, and the second the *Kolmogorov backward differential equations*. The forward equations are obtained by letting s approach t from below, and the backward equations are obtained by letting s approach 0 from above. First we derive the forward equations.

For $t - s$ small and positive, $P_{kj}(t-s)$ in (6.27) can be expressed as $(t-s)q_{kj} + \mathrm{o}(t-s)$ for $k \neq j$. Similarly, $P_{jj}(t-s)$ can be expressed as $1 - (t-s)\nu_j + \mathrm{o}(s)$. Thus (6.27) becomes

$$P_{ij}(t) = \sum_{k\neq j} [P_{ik}(s)(t-s)q_{kj}] + P_{ij}(s)[1-(t-s)\nu_j] + \mathrm{o}(t-s) \tag{6.28}$$

We want to express this, in the limit $s \to t$, as a differential equation. To do this, subtract $P_{ij}(s)$ from both sides and divide by $t - s$.

$$\frac{P_{ij}(t) - P_{ij}(s)}{t-s} = \sum_{k\neq j} (P_{ik}(s)q_{kj}) \ - P_{ij}(s)\nu_j + \frac{\mathrm{o}(s)}{s}$$

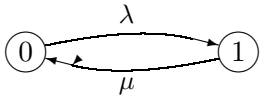
Taking the limit as $s \to t$ from below,[3] we get the Kolmogorov forward equations,

$$\frac{dP_{ij}(t)}{dt} = \sum_{k\neq j} (P_{ik}(t)q_{kj}) - P_{ij}(t)\nu_j, \tag{6.29}$$

The first term on the right side of (6.29) is the rate at which transitions occur into state j at time t and the second term is the rate at which transitions occur out of state j. Thus the difference of these terms is the net rate at which transitions occur into j, which is the rate at which $P_{ij}(t)$ is increasing at time t.

[3]We have assumed that the sum and the limit in (6.3) can be interchanged. This is certainly valid if the state space is finite, which is the only case we analyze in what follows.

Example 6.3.1 (A queueless M/M/1 queue). Consider the following 2-state Markov process where $q_{01} = \lambda$ and $q_{10} = \mu$.



This can be viewed as a model for an M/M/1 queue with no storage for waiting customers. When the system is empty (state 0), memoryless customers arrive at rate λ, and when the server is busy, an exponential server operates at rate μ, with the system returning to state 0 when service is completed.

To find $P_{01}(t)$, the probability of state 1 at time t conditional on state 0 at time 0, we use the Kolmogorov forward equations for $P_{01}(t)$, getting

$$\frac{dP_{01}(t)}{dt} \;=\; P_{00}(t)q_{01} - P_{01}(t)\nu_1 \;=\; P_{00}(t)\lambda - P_{01}(t)\mu.$$

Using the fact that $P_{00}(t) = 1 - P_{01}(t)$, this becomes

$$\frac{dP_{01}(t)}{dt} = \lambda - P_{01}(t)(\lambda + \mu).$$

Using the boundary condition $P_{01}(0) = 0$, the solution is

$$P_{01}(t) = \frac{\lambda}{\lambda + \mu}\left[1 - e^{-(\lambda+\mu)t}\right]\} \tag{6.30}$$

Thus $P_{01}(t)$ is 0 at $t = 0$ and increases as $t \to \infty$ to its steady-state value in state 1, which is $\lambda/(\lambda + \mu)$.

In general, for any given starting state i in a Markov process with M states, (6.29) provides a set of M simultaneous linear differential equations, one for each j, $1 \le j \le \mathsf{M}$. As we saw in the example, one of these is redundant because $\sum_{j=1}^{\mathsf{M}} P_{ij}(t) = 1$. This leaves $\mathsf{M} - 1$ simultaneous linear differential equations to be solved.

For more than 2 or 3 states, it is more convenient to express (6.29) in matrix form. Let $[P(t)]$ (for each $t > 0$) be an M by M matrix whose i, j element is $P_{ij}(t)$. Let $[Q]$ be an M by M matrix whose i, j element is q_{ij} for each $i \neq j$ and $-\nu_j$ for $i = j$. Then (6.29) becomes

$$\frac{d[P(t)]}{dt} = [P(t)][Q]. \tag{6.31}$$

For Example 6.3.1, $P_{ij}(t)$ can be calculated for each i, j as in (6.30), resulting in

$$[P(t)] = \begin{bmatrix} \}\frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu}e^{-(\lambda+\mu)t} & \frac{\lambda}{\lambda+\mu} - \frac{\lambda}{\lambda+\mu}e^{-(\lambda+\mu)t} \\ \}\frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu}e^{-(\lambda+\mu)t} & \frac{\lambda}{\lambda+\mu} - \frac{\lambda}{\lambda+\mu}e^{-(\lambda+\mu)t} \end{bmatrix}\} \qquad [Q] = \begin{bmatrix} \}-\lambda & \lambda \\ \} \mu & -\mu \end{bmatrix}\}$$

In order to provide some insight into the general solution of (6.31), we go back to the sampled time approximation of a Markov process. With an increment of size δ between samples, the probability of a transition from i to j, $i \neq j$, is $q_{ij}\delta + o(\delta)$, and the probability

of remaining in state i is $1-\nu_i\delta+\mathrm{o}(\delta)$. Thus, in terms of the matrix $[Q]$ of transition rates, the transition probability matrix in the sampled time model is $[I]+\delta[Q]$, where $[I]$ is the identity matrix. We denote this matrix by $[W_\delta]=[I]+\delta[Q]$. Note that λ is an eigenvalue of $[Q]$ if and only if $1+\lambda\delta$ is an eigenvalue of $[W_\delta]$. Also the eigenvectors of these corresponding eigenvalues are the same. That is, if $\boldsymbol{\nu}$ is a right eigenvector of $[Q]$ with eigenvalue λ, then $\boldsymbol{\nu}$ is a right eigenvector of $[W_\delta]$ with eigenvalue $1+\lambda\delta$, and conversely. Similarly, if $\boldsymbol{p}$ is a left eigenvector of $[Q]$ with eigenvalue λ, then $\boldsymbol{p}$ is a left eigenvectorof $[W_\delta]$ with eigenvalue $1+\lambda\delta$, and conversely.

We saw in Section 6.2.5 that the steady-state probability vector $\boldsymbol{p}$ of a Markov process is the same as that of any sampled-time approximation. We have now seen that, in addition, *all* the eigenvectors are the same and the eigenvalues are simply related. Thus study of these differential equations can be largely replaced by studying the sampled-time approximation.

The following theorem uses our knowledge of the eigenvalues and eigenvectors of transition matrices such as $[W_\delta]$ in Section 3.4, to be more specific about the properties of $[Q]$.

Theorem 6.3.1. *Consider an irreducible finite-state Markov process with M states. Then the matrix $[Q]$ for that process has an eigenvalue λ equal to 0. That eigenvalue has a right eigenvector $\boldsymbol{e}=(1,1,\ldots,1)^{\mathsf{T}}$ which is unique within a scale factor. It has a left eigenvector $\boldsymbol{p}=(p_1,\ldots,p_{\mathsf{M}})$ that is positive, sums to 1, satisfies (6.23), and is unique within a scale factor. All the other eigenvalues of $[Q]$ have strictly negative real parts.*

Proof: Since all M states communicate, the sampled time chain is recurrent. From Theorem 3.4.1, $[W_\delta]$ has a unique eigenvalue $\lambda=1$. The corresponding right eigenvector is $\boldsymbol{e}$ and the left eigenvector is the steady-state probability vector $\boldsymbol{p}$ as given in (3.9). Since $[W_\delta]$ is recurrent, the components of $\boldsymbol{p}$ are strictly positive. From the equivalence of (6.23) and (6.24), $\boldsymbol{p}$, as given by (6.23), is the steady-state probability vector of the process. Each eigenvalue λ_δ of $[W_\delta]$ corresponds to an eigenvalue λ of $[Q]$ with the correspondence $\lambda_\delta=1+\lambda\delta$, *i.e.*, $\lambda=(\lambda_\delta-1)/\delta$. Thus the eigenvalue 1 of $[W_\delta]$ corresponds to the eigenvalue 0 of $[Q]$. Since $|\lambda_\delta|\leq 1$ and $\lambda_\delta\neq 1$ for all other eigenvalues, the other eigenvalues of $[Q]$ all have strictly negative real parts, completing the proof. □

We complete this section by deriving the Komogorov backward equations. For s small and positive, the Chapman-Kolmogorov equations in (6.27) become

$$\begin{aligned} P_{ij}(t) &= \sum_k P_{ik}(s)P_{kj}(t-s) \\ &= \sum_{k\neq i} sq_{ik}P_{kj}(t-s)+(1-s\nu_i)P_{ij}(t-s)+\mathrm{o}(s) \end{aligned}$$

Subtracting $P_{ij}(t-s)$ from both sides and dividing by s,

$$\begin{aligned} \frac{P_{ij}(t)-P_{ij}(t-s)}{s} &= \sum_{k\neq i} q_{ik}P_{kj}(t-s)-\nu_iP_{ij}(t-s)+\frac{\mathrm{o}(s)}{s} \\ \frac{dP_{ij}(t)}{dt} &= \sum_{k\neq i} q_{ik}P_{kj}(t)-\nu_iP_{ij}(t) \end{aligned} \tag{6.32}$$

In matrix form, this is expressed as

$$\frac{d[P(t)]}{dt} = [Q][P(t)] \tag{6.33}$$

By comparing (6.33 and (6.31), we see that $[Q][P(t)] = [P(t)][Q]$, *i.e.*, that the matrices $[Q]$ and $[P(t)]$ commute. Simultaneous linear differential equations appear in so many applications that we leave the further exploration of these forward and backward equations as simple differential equation topics rather than topics that have special properties for Markov processes.

6.4 Uniformization

Up until now, we have discussed Markov processes under the assumption that $q_{ii} = 0$ (*i.e.*, no transitions from a state into itself are allowed). We now consider what happens if this restriction is removed. Suppose we start with some Markov process defined by a set of transition rates q_{ij} with $q_{ii} = 0$, and we modify this process by some arbitrary choice of $q_{ii} \geq 0$ for each state i. This modification changes the embedded Markov chain, since ν_i is increased from $\sum_{k\neq i} q_{ik}$ to $\sum_{k\neq i} q_{ik} + q_{ii}$. From (6.5), P_{ij} is changed to q_{ij}/ν_i for the new value of ν_i for each i, j. Thus the steady-state probabilities π_i for the embedded chain are changed. The Markov process $\{X(t);\, t \geq 0\}$ is not changed, since a transition from i into itself does not change $X(t)$ and does not change the distribution of the time until the next transition to a different state. The steady-state probabilities for the process still satisfy

$$p_j\nu_j = \sum_k \}p_k q_{kj} \quad ; \quad \sum_i \}p_i = 1. \tag{6.34}$$

The addition of the new term q_{jj} increases ν_j by q_{jj}, thus increasing the left hand side by $p_j\, q_{jj}$. The right hand side is similarly increased by $p_j\, q_{jj}$, so that the solution is unchanged (as we already determined it must be).

A particularly convenient way to add self-transitions is to add them in such a way as to make the transition rate ν_j the same for all states. Assuming that the transition rates $\{\nu_i;\, i \geq 0\}$ are bounded, we define ν^* as $\sup_j\ \nu_j$ for the original transition rates. Then we set $q_{jj} = \nu^* - \sum_{k\neq j} q_{jk}$ for each j. With this addition of self-transitions, all transition rates become ν^*. From (6.21), we see that the new steady state probabilities, π_i^*, in the embedded Markov chain become equal to the steady-state process probabilities, p_i. Naturally, we could also choose any ν greater than ν^* and increase each q_{jj} to make all transition rates equal to that value of ν. When the transition rates are changed in this way, the resulting embedded chain is called a *uniformized chain* and the Markov process is called the *uniformized process*. The uniformized process is the same as the original process, except that quantities like the number of transitions over some interval are different because of the self transitions.

Assuming that all transition rates are made equal to ν^*, the new transition probabilities in the embedded chain become $P_{ij}^* = q_{ij}/\nu^*$. Let $N(t)$ be the total number of transitions that occur from 0 to t in the uniformized process. Since the rate of transitions is the same from all states and the inter-transition intervals are independent and identically exponentially

distributed, $N(t)$ is a Poisson counting process of rate ν^*. Also, $N(t)$ is independent of the sequence of transitions in the embedded uniformized Markov chain. Thus, given that $N(t) = n$, the probability that $X(t) = j$ given that $X(0) = i$ is just the probability that the embedded chain goes from i to j in ν steps, *i.e.*, P_{ij}^{*n}. This gives us another formula for calculating $P_{ij}(t)$, (*i.e.*, the probability that $X(t) = j$ given that $X(0) = i$).

$$P_{ij}(t) = \sum_{n=0}^{\infty} \} P_{ij}^{*n} \ \frac{e^{-\nu^* t}(\nu^* t)^n}{n!}. \tag{6.35}$$

Another situation where the uniformized process is useful is in extending Markov decision theory to Markov processes, but we do not pursue this.

6.5 Birth-death processes

Birth-death processes are very similar to the birth-death Markov chains that we studied earlier. Here transitions occur only between neighboring states, so it is convenient to define λ_i as $q_{i,i+1}$ and μ_i as $q_{i,i-1}$ (see Figure 6.9). Since the number of transitions from i to $i+1$ is within 1 of the number of transitions from $i+1$ to i for every sample path, we conclude that

$$p_i \lambda_i = p_{i+1} \mu_{i+1}. \tag{6.36}$$

This can also be obtained inductively from (6.23) using the same argument that we used earlier for birth-death Markov chains.

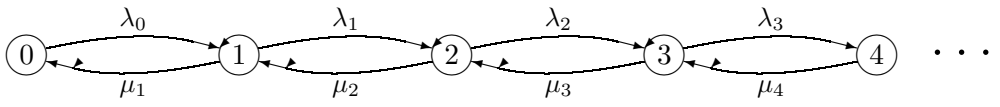


Figure 6.9: Birth-death process.

Define ρ_i as λ_i / μ_{i+1}. Then applying (6.36) iteratively, we obtain the steady-state equations

$$p_i = p_0 \ \prod_{j=0}^{i-1} \} \rho_j \ ; \quad i \geq 1. \tag{6.37}$$

We can solve for p_0 by substituting (6.37) into $\sum_i \ p_i$, yielding

$$p_0 = \frac{1}{1 + \sum_{i=1}^{\infty} \ \prod_{j=0}^{i-1} \rho_j}. \tag{6.38}$$

For the M/M/1 queue, the state of the Markov process is the number of customers in the system (*i.e.*, customers either in queue or in service). The transitions from i to $i+1$ correspond to arrivals, and since the arrival process is Poisson of rate λ, we have $\lambda_i = \lambda$ for all $i \geq 0$. The transitions from i to $i-1$ correspond to departures, and since the service

time distribution is exponential with parameter μ, say, we have $\mu_i = \mu$ for all $i \geq 1$. Thus, (6.38) simplifies to $p_0 = 1 - \rho$, where $\rho = \lambda/\mu$ and thus

$$p_i = (1 - \rho)\rho^i; \qquad i \geq 0. \tag{6.39}$$

We assume that $\rho < 1$, which is required for positive recurrence. The probability that there are i or more customers in the system in steady state is then given by $P\{X(t) \geq i\} = \rho^i$ and the expected number of customers in the system is given by

$$\mathsf{E}[X(t)] = \sum_{i=1}^{\infty} P\{X(t) \geq i\} = \frac{\rho}{1 - \rho}. \tag{6.40}$$

The expected time that a customer spends in the system in steady state can now be determined by Little's formula (Theorem 4.5.3).

$$\mathsf{E}[\text{System time}] = \frac{\mathsf{E}[X(t)]}{\lambda} = \frac{\rho}{\lambda(1 - \rho)} = \frac{1}{\mu - \lambda}. \tag{6.41}$$

The expected time that a customer spends in the queue (*i.e.*, before entering service) is just the expected system time less the expected service time, so

$$\mathsf{E}[\text{Queueing time}] = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda}. \tag{6.42}$$

Finally, the expected number of customers in the queue can be found by applying Little's formula to (6.42),

$$\mathsf{E}[\text{Number in queue}] = \frac{\lambda\rho}{\mu - \lambda}. \tag{6.43}$$

Note that the expected number of customers in the system and in the queue depend only on ρ, so that if the arrival rate and service rate were both speeded up by the same factor, these expected values would remain the same. The expected system time and queueing time, however would decrease by the factor of the rate increases. Note also that as ρ approaches 1, all these quantities approach infinity as $1/(1 - \rho)$. At the value $\rho = 1$, the embedded Markov chain becomes null-recurrent and the steady-state probabilities (both $\{\pi_i;\ i \geq 0\}$ and $\{p_i;\ i \geq 0\}$) can be viewed as being all 0 or as failing to exist.

There are many types of queueing systems that can be modeled as birth-death processes. For example the arrival rate could vary with the number in the system and the service rate could vary with the number in the system. All of these systems can be analyzed in steady state in the same way, but (6.37) and (6.38) can become quite messy in these more complex systems. As an example, we analyze the M/M/m system. Here there are m servers, each with exponentially distributed service times with parameter μ. When i customers are in the system, there are i servers working for $i < m$ and all m servers are working for $i \geq m$. With i servers working, the probability of a departure in an incremental time δ is $i\mu\delta$, so that μ_i is $i\mu$ for $i < m$ and $m\mu$ for $i \geq m$ (see Figure 6.10).

Define $\rho = \lambda/(m\mu)$. Then in terms of our general birth-death process notation, $\rho_i = m\rho/(i+1)$ for $i < m$ and $\rho_i = \rho$ for $i \geq m$. From (6.37), we have

$$p_i = p_0 \frac{m\rho}{1}\frac{m\rho}{2}\cdots\frac{m\rho}{i} = \frac{p_0(m\rho)^i}{i!}; \qquad i \leq m \tag{6.44}$$

$$p_i = \frac{p_0\rho^i m^m}{m!} \quad ; \qquad i \geq m. \tag{6.45}$$

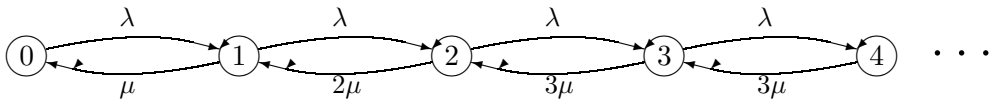


Figure 6.10: M/M/m queue for $m = 3$..

We can find p_0 by summing p_i and setting the result equal to 1; a solution exists if $\rho < 1$. Nothing simplifies much in this sum, except that $\sum_{i\geq m} p_i = p_0(\rho m)^m/[m!(1-\rho)]$, and the solution is

$$p_0 = \left[\frac{(m\rho)^m}{m!(1-\rho)} + \sum_{i=0}^{m-1}\frac{(m\rho)^i}{i!}\right]^{-1}. \tag{6.46}$$

6.6 Reversibility for Markov processes

In Section 5.3 on reversibility for Markov chains, (5.37) showed that the backward transition probabilities P^*_{ij} in steady state satisfy

$$\pi_i P^*_{ij} = \pi_j P_{ji}. \tag{6.47}$$

These equations are then valid for the embedded chain of a Markov process. Next, consider backward transitions in the process itself. Given that the process is in state i, the probability of a transition in an increment δ of time is $\nu_i\delta + o(\delta)$, and transitions in successive increments are independent. Thus, if we view the process running backward in time, the probability of a transition in each increment δ of time is also $\nu_i\delta + o(\delta)$ with independence between increments. Thus, going to the limit $\delta \to 0$, the distribution of the time backward to a transition is exponential with parameter ν_i. This means that the process running backwards is again a Markov process with transition probabilities P^*_{ij} and transition rates ν_i. Figure 6.11 helps to illustrate this.

Since the steady-state probabilities $\{p_i;\, i \geq 0\}$ for the Markov process are determined by

$$p_i = \frac{\pi_i/\nu_i}{\sum_k \pi_k/\nu_k}, \tag{6.48}$$

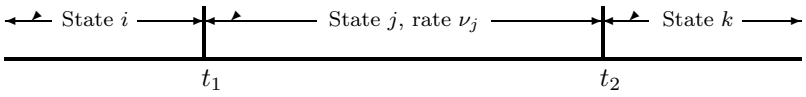


Figure 6.11: The forward process enters state j at time t_1 and departs at t_2. The backward process enters state j at time t_2 and departs at t_1. In any sample function, as illustrated, the interval in a given state is the same in the forward and backward process. Given $X(t) = j$, the time forward to the next transition and the time backward to the previous transition are each exponential with rate ν_j.

and since $\{\pi_i;\ i \geq 0\}$ and $\{\nu_i;\ i \geq 0\}$ are the same for the forward and backward processes, we see that the steady-state probabilities in the backward Markov process are the same as the steady-state probabilities in the forward process. This result can also be seen by the correspondence between sample functions in the forward and backward processes.

The *transition rates* in the backward process are defined by $q^*_{ij} = \nu_i P^*_{ij}$. Using (6.47), we have

$$q^*_{ij} = \nu_j P^*_{ij} = \frac{\nu_i \pi_j P_{ji}}{\pi_i} = \frac{\nu_i \pi_j q_{ji}}{\pi_i \nu_j}. \tag{6.49}$$

From (6.48), we note that $p_j = \alpha \pi_j / \nu_j$ and $p_i = \alpha \pi_i / \nu_i$ for the same value of α. Thus the ratio of π_j/ν_j to π_i/ν_i is p_j/p_i. This simplifies (6.49) to $q^*_{ij} = p_j q_{ji}/p_i$, and

$$p_i q^*_{ij} = p_j q_{ji}. \tag{6.50}$$

This equation can be used as an alternate definition of the backward transition rates. To interpret this, let δ be a vanishingly small increment of time and assume the process is in steady state at time t. Then $\delta p_j q_{ji} \approx \Pr\{X(t) = j\} \Pr\{X(t+\delta) = i \mid X(t) = j\}$ whereas $\delta p_i q^*_{ij} \approx \Pr\{X(t+\delta) = i\} \Pr\{X(t) = j \mid X(t+\delta) = i\}$.

A Markov process is defined to be *reversible* if $q^*_{ij} = q_{ij}$ for all i, j. If the embedded Markov chain is reversible, (*i.e.*, $P^*_{ij} = P_{ij}$ for all i, j), then one can repeat the above steps using P_{ij} and q_{ij} in place of P^*_{ij} and q^*_{ij} to see that $p_i q_{ij} = p_j q_{ji}$ for all i, j. Thus, if the embedded chain is reversible, the process is also. Similarly, if the Markov process is reversible, the above argument can be reversed to see that the embedded chain is reversible. Thus, we have the following useful lemma.

Lemma 6.6.1. *Assume that steady-state probabilities $\{p_i;\ i \geq 0\}$ exist in an irreducible Markov process (*i.e.*, (6.23) has a solution and $\sum p_i \nu_i < \infty$). Then the Markov process is reversible if and only if the embedded chain is reversible.*

One can find the steady-state probabilities of a reversible Markov process and simultaneously show that it is reversible by the following useful theorem (which is directly analogous to Theorem 5.3.2 of chapter 5).

Theorem 6.6.1. *For an irreducible Markov process, assume that $\{p_i;\, i \geq 0\}$ is a set of nonnegative numbers summing to 1, satisfying $\sum_i p_i \nu_i \leq \infty$, and satisfying*

$$p_i q_{ij} = p_j q_{ji} \qquad \text{for all } i, j. \tag{6.51}$$

Then $\{p_i;\, i \geq 0\}$ is the set of steady-state probabilities for the process, $p_i > 0$ for all i, the process is reversible, and the embedded chain is positive recurrent.

Proof: Summing (6.51) over i, we obtain

$$\sum_i p_i q_{ij} = p_j \nu_j \qquad \text{for all } j.$$

These, along with $\sum_i p_i = 1$ are the steady-state equations for the process. These equations have a solution, and by Theorem 6.2.2, $p_i > 0$ for all i, the embedded chain is positive recurrent, and $p_i = \lim_{t\to\infty} \Pr\{X(t) = i\}$. Comparing (6.51) with (6.50), we see that $q_{ij} = q^*_{ij}$, so the process is reversible. □

There are many irreducible Markov processes that are not reversible but for which the backward process has interesting properties that can be deduced, at least intuitively, from the forward process. Jackson networks (to be studied shortly) and many more complex networks of queues fall into this category. The following simple theorem allows us to use whatever combination of intuitive reasoning and wishful thinking we desire to guess both the transition rates q^*_{ij} in the backward process and the steady-state probabilities, and to then verify rigorously that the guess is correct. One might think that guessing is somehow unscientific, but in fact, the art of educated guessing and intuitive reasoning leads to much of the best research.

Theorem 6.6.2. *For an irreducible Markov process, assume that a set of positive numbers $\{p_i;\, i \geq 0\}$ satisfy $\sum_i p_i = 1$ and $\sum_i p_i \nu_i < \infty$. Also assume that a set of nonnegative numbers $\{q^*_{ij}\}$ satisfy the two sets of equations*

$$\sum_j q_{ij} = \sum_j q^*_{ij} \quad \text{for all } i \tag{6.52}$$

$$p_i q_{ij} = p_j q^*_{ji} \quad \text{for all } i, j. \tag{6.53}$$

Then $\{p_i\}$ is the set of steady-state probabilities for the process, $p_i > 0$ for all i, the embedded chain is positive recurrent, and $\{q^*_{ij}\}$ is the set of transition rates in the backward process.

Proof: Sum (6.53) over i. Using the fact that $\sum_j q_{ij} = \nu_i$ and using (6.52), we obtain

$$\sum_i p_i q_{ij} = p_j \nu_j \qquad \text{for all} j. \tag{6.54}$$

These, along with $\sum_i p_i = 1$, are the steady-state equations for the process. These equations thus have a solution, and by Theorem 6.2.2, $p_i > 0$ for all i, the embedded chain is positive recurrent, and $p_i = \lim_{t\to\infty} \Pr\{X(t) = i\}$. Finally, q^*_{ij} as given by (6.53) is the backward transition rate as given by (6.50) for all i, j. □

We see that Theorem 6.6.1 is just a special case of Theorem 6.6.2 in which the guess about q^*_{ij} is that $q^*_{ij} = q_{ij}$.

Birth-death processes are all reversible if the steady-state probabilities exist. To see this, note that Equation (6.36) (the equation to find the steady-state probabilities) is just (6.51) applied to the special case of birth-death processes. Due to the importance of this, we state it as a theorem.

Theorem 6.6.3. *For a birth-death process, if there is a solution $\{p_i;\, i \geq 0\}$ to (6.36) with $\sum_i p_i = 1$ and $\sum_i p_i \nu_i < \infty$, then the process is reversible, and the embedded chain is positive recurrent and reversible.*

Since the M/M/1 queueing process is a birth-death process, it is also reversible. Burke's theorem, which was given as Theorem 5.4.1 for sampled-time M/M/1 queues, can now be established for continuous-time M/M/1 queues. Note that the theorem here contains an extra part, part c).

Theorem 6.6.4 (Burke's theorem). *Given an M/M/1 queueing system in steady state with $\lambda < \mu$,*

a) the departure process is Poisson with rate λ,

b) the state $X(t)$ at any time t is independent of departures prior to t, and

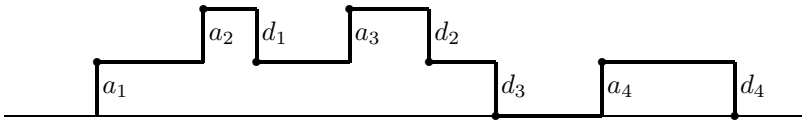
c) for FCFS service, given that a customer departs at time t, the arrival time of that customer is independent of the departures prior to t.

Proof: The proofs of parts a) and b) are the same as the proof of Burke's theorem for sampled-time, Theorem 5.4.1, and thus will not be repeated. For part c), note that with FCFS service, the m^{th} customer to arrive at the system is also the m^{th} customer to depart. Figure 6.12 illustrates that the association between arrivals and departures is the same in the backward system as in the forward system (even though the customer ordering is reversed in the backward system). In the forward, right moving system, let τ be the epoch of some given arrival. The customers arriving after τ wait behind the given arrival in the queue, and have no effect on the given customer's service. Thus the interval from τ to the given customer's service completion is independent of arrivals after τ.

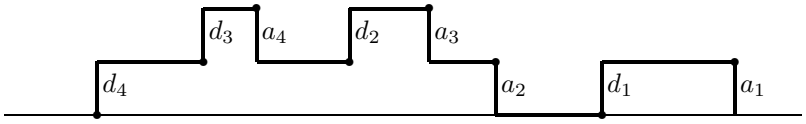
Since the backward, left moving, system is also an M/M/1 queue, the interval from a given backward arrival, say at epoch t, moving left until the corresponding departure, is independent of arrivals to the left of t. From the correspondence between sample functions in the right moving and left moving systems, given a departure at epoch t in the right moving system, the departures before time t are independent of the arrival epoch of the given customer departing at t; this completes the proof. □

Part c) of Burke's theorem does not apply to sampled-time M/M/1 queues because the sampled time model does not allow for both an arrival and departure in the same increment of time.

Note that the proof of Burke's theorem (including parts a and b from Section 5.4) does not make use of the fact that the transition rate $q_{i,i-1} = \mu$ for $i \geq 1$ in the M/M/1 queue. Thus



Right moving (forward) M/M/1 process



Left moving (backward) M/M/1 process

Figure 6.12: FCFS arrivals and departures in right and left moving M/M/1 processes.

Burke's theorem remains true for any birth-death Markov process in steady state for which $q_{i,i+1} = \lambda$ for all $i \geq 0$. For example, parts a and b are valid for M/M/m queues; part c is also valid (see [22]), but the argument here is not adequate since the first customer to enter the system might not be the first to depart.

We next show how Burke's theorem can be used to analyze a tandem set of queues. As shown in Figure 6.13, we have an M/M/1 queueing system with Poisson arrivals at rate λ and service at rate μ_1. The departures from this queueing system are the arrivals to a second queueing system, and we assume that a departure from queue 1 at time t instantaneously enters queueing system 2 at the same time t. The second queueing system has a single server and the service times are IID and exponentially distributed with rate μ_2. The successive service times at system 2 are also independent of the arrivals to systems 1 and 2, and independent of the service times in system 1. Since we have already seen that the departures from the first system are Poisson with rate λ, the arrivals to the second queue are Poisson with rate λ. Thus the second system is also M/M/1.

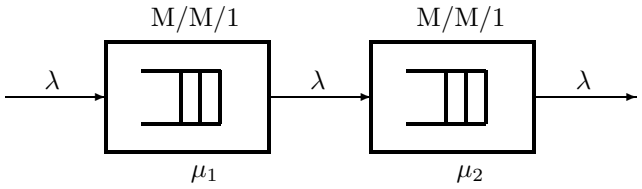


Figure 6.13: A tandem queueing system. Assuming that $\lambda > \mu_1$ and $\lambda > \mu_2$, the departures from each queue are Poisson of rate λ.

Let X(t) be the state of queueing system 1 and $Y(t)$ be the state of queueing system 2. Since $X(t)$ at time t is independent of the departures from system 1 prior to t, $X(t)$ is independent of the arrivals to system 2 prior to time t. Since $Y(t)$ depends only on the arrivals to system 2 prior to t and on the service times that have been completed prior to t, we see that $X(t)$ is independent of $Y(t)$. This leaves a slight nit-picking question about what happens at the instant of a departure from system 1. We have considered the state $X(t)$ at the instant of a departure to be the number of customers remaining in system 1

not counting the departing customer. Also the state $Y(t)$ is the state in system 2 including the new arrival at instant t. The state $X(t)$ then is independent of the departures up to and including t, so that $X(t)$ and $Y(t)$ are still independent.

Next assume that both systems use FCFS service. Consider a customer that leaves system 1 at time t. The time at which that customer arrived at system 1, and thus the waiting time in system 1 for that customer, is independent of the departures prior to t. This means that the state of system 2 immediately before the given customer arrives at time t is independent of the time the customer spent in system 1. It therefore follows that the time that the customer spends in system 2 is independent of the time spent in system 1. Thus the total system time that a customer spends in both system 1 and system 2 is the sum of two independent random variables.

This same argument can be applied to more than 2 queueing systems in tandem. It can also be applied to more general networks of queues, each with single servers with exponentially distributed service times. The restriction here is that there can not be any cycle of queueing systems where departures from each queue in the cycle can enter the next queue in the cycle. The problem posed by such cycles can be seen easily in the following example of a single queueing system with feedback (see Figure 6.14).

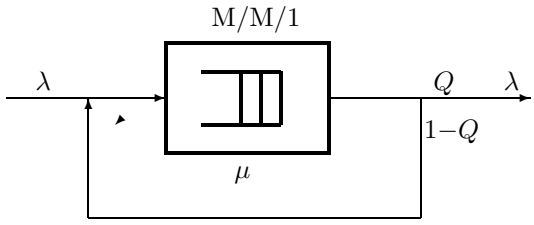


Figure 6.14: A queue with feedback. Assuming that $\mu > \lambda/Q$, the exogenous output is Poisson of rate λ.

We assume that the queueing system in Figure 6.14 has a single server with IID exponentially distributed service times that are independent of arrival times. The exogenous arrivals from outside the system are Poisson with rate λ. With probability Q, the departures from the queue leave the entire system, and, alternatively, with probability $1 - Q$, they return instantaneously to the input of the queue. Successive choices between leaving the system and returning to the input are IID and independent of exogenous arrivals and of service times. Figure 6.15 shows a sample function of the arrivals and departures in the case in which the service rate μ is very much greater than the exogenous arrival rate λ. Each exogenous arrival spawns a geometrically distributed set of departures and simultaneous re-entries. Thus the overall arrival process to the queue, counting both exogenous arrivals and feedback from the output, is not Poisson. Note, however, that if we look at the Markov process description, the departures that are fed back to the input correspond to self loops from one state to itself. Thus the Markov process is the same as one without the self loops with a service rate equal to μQ. Thus, from Burke's theorem, the exogenous departures are Poisson with rate λ. Also the steady-state distribution of $X(t)$ is $P\{X(t) = i\} = (1 - \rho)\rho^i$ where $\rho = \lambda/(\mu Q)$ (assuming, of course, that $\rho < 1$).



Figure 6.15: Sample path of arrivals and departures for queue with feedback.

The tandem queueing system of Figure 6.13 can also be regarded as a combined Markov process in which the state at time t is the pair $(X(t),\ Y(t))$. The transitions in this process correspond to, first, exogenous arrivals in which $X(t)$ increases, second, exogenous departures in which $Y(t)$ decreases, and third, transfers from system 1 to system 2 in which $X(t)$ decreases and $Y(t)$ simultaneously increases. The combined process is not reversible since there is no transition in which $X(t)$ increases and $Y(t)$ simultaneously decreases. In the next section, we show how to analyze these combined Markov processes for more general networks of queues.

6.7 Jackson networks

In many queueing situations, a customer has to wait in a number of different queues before completing the desired transaction and leaving the system. For example, when we go to the registry of motor vehicles to get a driver's license, we must wait in one queue to have the application processed, in another queue to pay for the license, and in yet a third queue to obtain a photograph for the license. In a multiprocessor computer facility, a job can be queued waiting for service at one processor, then go to wait for another processor, and so forth; frequently the same processor is visited several times before the job is completed. In a data network, packets traverse multiple intermediate nodes; at each node they enter a queue waiting for transmission to other nodes.

Such systems are modeled by a network of queues, and Jackson networks are perhaps the simplest models of such networks. In such a model, we have a network of k interconnected queueing systems which we call nodes. Each of the k nodes receives customers (*i.e.*, tasks or jobs) both from outside the network (exogenous inputs) and from other nodes within the network (endogenous inputs). It is assumed that the exogenous inputs to each node i form a Poisson process of rate r_i and that these Poisson processes are independent of each other. For analytical convenience, we regard this as a single Poisson input process of rate λ_0, with each input independently going to each node i with probability $Q_{0i} = r_i/\lambda_0$.

Each node i contains a single server, and the successive service times at node i are IID random variables with an exponentially distributed service time of rate μ_i. The service times at each node are also independent of the service times at all other nodes and independent of the exogenous arrival times at all nodes. When a customer completes service at a given

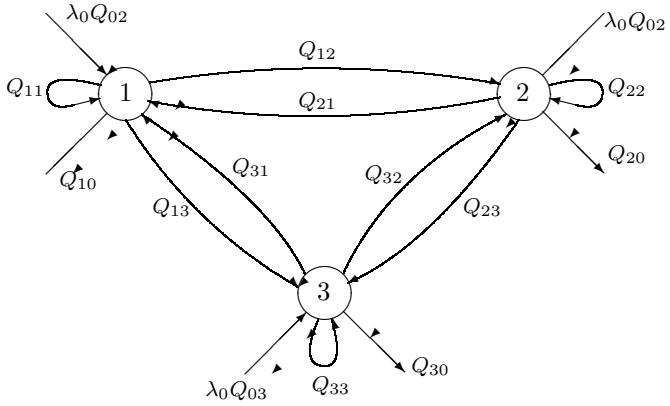


Figure 6.16: A Jackson network with 3 nodes. Given a departure from node i, the probability that departure goes to node j (or, for $j = 0$, departs the system) is Q_{ij}. Note that a departure from node i can re-enter node i with probability Q_{ii}. The overall exogenous arrival rate is λ_0, and, conditional on an arrival, the probability the arrival enters node i is Q_{0i}.

node i, that customer is routed to node j with probability Q_{ij} (see Figure 6.16). It is also possible for the customer to depart from the network entirely (called an exogenous departure), and this occurs with probability $Q_{i0} = 1 - \sum_{j \geq 1} Q_{ij}$. For a customer departing from node i, the next node j is a random variable with PMF $\{Q_{ij}, 0 \leq j \leq k\}$.

Successive choices of the next node for customers at node i are IID, independent of the customer routing at other nodes, independent of all service times, and independent of the exogenous inputs. Notationally, we are regarding the outside world as a fictitious node 0 from which customers appear and to which they disappear.

When a customer is routed from node i to node j, it is assumed that the routing is instantaneous; thus at the epoch of a departure from node i, there is a simultaneous endogenous arrival at node j. Thus a node j receives Poisson exogenous arrivals from outside the system at rate $\lambda_0 Q_{0j}$ and receives endogenous arrivals from other nodes according to the probabilistic rules just described. We can visualize these combined exogenous and endogenous arrivals as being served in FCFS fashion, but it really makes no difference in which order they are served, since the customers are statistically identical and simply give rise to service at node j at rate μ_j whenever there are customers to be served.

The Jackson queueing network, as just defined, is fully described by the exogenous input rate λ_0, the service rates $\{\mu_i\}$, and the routing probabilities $\{Q_{ij}; 0 \leq i, j \leq k\}$. The network as a whole is a Markov process in which the state is a vector $\boldsymbol{m} = (m_1, m_2, \ldots, m_k)$, where $m_i, 1 \leq i \leq k$, is the number of customers at node i. State changes occur upon exogenous arrivals to the various nodes, exogenous departures from the various nodes, and departures from one node that enter another node. In a vanishingly small interval δ of time, given that the state at the beginning of that interval is $\boldsymbol{m}$, an exogenous arrival at node j occurs in the interval with probability $\lambda_0 Q_{0j} \delta$ and changes the state to $\boldsymbol{m}' = \boldsymbol{m} + \boldsymbol{e}_j$ where $\boldsymbol{e}_j$ is a unit vector with a one in position j. If $m_i > 0$, an exogenous departure from node i occurs

in the interval with probability $\mu_i Q_{i0}\delta$ and changes the state to $\boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i$. Finally, if $m_i > 0$, a departure from node i entering node j occurs in the interval with probability $\mu_i Q_{ij}\delta$ and changes the state to $\boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j$. Thus, the transition rates are given by

$$\begin{aligned} q_{\boldsymbol{m},\boldsymbol{m}'} &= \lambda_0 Q_{0j} \qquad \text{for } \boldsymbol{m}' = \boldsymbol{m} + \boldsymbol{e}_j, \quad 1 \le i \le k && (6.55)\\ &= \mu_i Q_{i0} \qquad \text{for } \boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{m}_i, \quad m_i > 0, \quad 1 \le i \le k && (6.56)\\ &= \mu_i Q_{ij} \qquad \text{for } \boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j, \quad m_i > 0, \quad 1 \le i, j \le k && (6.57)\\ &= 0 \qquad \text{for all other choices of } \boldsymbol{m}'. \end{aligned}$$

Note that a departure from node i that re-enters node i causes a transition from state $\boldsymbol{m}$ back into state $\boldsymbol{m}$; we disallowed such transitions in sections 6.1 and 6.2, but showed that they caused no problems in our discussion of uniformization. It is convenient to allow these self-transitions here, partly for the added generality and partly to illustrate that the single node network with feedback of Figure 6.14 is an example of a Jackson network.

Our objective is to find the steady-state probabilities $p(\boldsymbol{m})$ for this type of process, and our plan of attack is in accordance with Theorem 6.6.2; that is, we shall guess a set of transition rates for the backward Markov process, use these to guess $p(\boldsymbol{m})$, and then verify that the guesses are correct. Before making these guesses, however, we must find out a little more about how the system works, so as to guide the guesswork. Let us define λ_i for each i, $1 \le i \le k$, as the time-average overall rate of arrivals to node i, including both exogenous and endogenous arrivals. Since λ_0 is the rate of exogenous inputs, we can interpret λ_i/λ_0 as the expected number of visits to node i per exogenous input. The endogenous arrivals to node i are not necessarily Poisson, as the example of a single queue with feedback shows, and we are not even sure at this point that such a time-average rate exists in any reasonable sense. However, let us assume for the time being that such rates exist and that the time-average rate of departures from each node equals the time-average rate of arrivals (*i.e.*, the queue sizes do not grow linearly with time). Then these rates must satisfy the equation

$$\lambda_j = \sum_{j=0}^{k} \}\lambda_i Q_{ij}; \qquad 1 \le j \le k. \tag{6.58}$$

To see this, note that $\lambda_0 Q_{0j}$ is the rate of exogenous arrivals to j. Also λ_i is the time-average rate at which customers depart from queue i, and $\lambda_i Q_{ij}$ is the rate at which customers go from node i to node j. Thus, the right hand side of (6.58) is the sum of the exogenous and endogenous arrival rates to node j. Note the distinction between the time-average rate of customers going from i to j in (6.58) and the rate $q_{\boldsymbol{m},\boldsymbol{m}'} = \mu_i Q_{ij}$ for $\boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j, m_i > 0$ in (6.57). The rate in (6.57) is conditioned on a state $\boldsymbol{m}$ with $m_i > 0$, whereas that in (6.58) is the overall time-average rate, averaged over all states.

Note that $\{Q_{ij};\, 0 \le i, j \le k\}$ forms a stochastic matrix and (6.58) is formally equivalent to the equations for steady-state probabilities (except that steady-state probabilities sum to 1). The usual equations for steady-state probabilities include an equation for $j = 0$, but that equation is redundant. Thus we know that, if there is a path between each pair of nodes (including the fictitious node 0), then (6.58) has a solution for $\{\lambda_i;\, 0 \le i \le k$, and that solution is unique within a scale factor. The known value of λ_0 determines this scale factor and makes the solution unique. Note that we don't have to be careful at this point

about whether these rates are time-averages in any nice sense, since this will be verified later; we do have to make sure that (6.58) has a solution, however, since it will appear in our solution for $p(\boldsymbol{m})$. Thus we assume in what follows that a path exists between each pair of nodes, and thus that (6.58) has a unique solution as a function of λ_0.

We now make the final necessary assumption about the network, which is that $\mu_i > \lambda_i$ for each node i. This will turn out to be required in order to make the process positive recurrent. We also define ρ_i as λ_i/μ_i. We shall find that, even though the inputs to an individual node i are not Poisson in general, there is a steady-state distribution for the number of customers at i, and that distribution is the same as that of an M/M/1 queue with the parameter ρ_i.

Now consider the backward time process. We have seen that only three kinds of transitions are possible in the forward process. First, there are transitions from $\boldsymbol{m}$ to $\boldsymbol{m}' = \boldsymbol{m} + \boldsymbol{e}_j$ for any j, $1 \leq j \leq k$. Second, there are transitions from $\boldsymbol{m}$ to $\boldsymbol{m} - \boldsymbol{e}_i$ for any i, $1 \geq i \leq k$, such that $m_i > 0$. Third, there are transitions from $\boldsymbol{m}$ to $\boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j$ for $1 \leq i, j \leq k$ with $m_i > 0$. Thus in the backward process, transitions from $\boldsymbol{m}'$ to $\boldsymbol{m}$ are possible only for the $\boldsymbol{m}$, $\boldsymbol{m}'$ pairs above. Corresponding to each arrival in the forward process, there is a departure in the backward process; for each forward departure, there is a backward arrival; and for each forward passage from i to j, there is a backward passage from j to i.

We now make the conjecture that the backward process is itself a Jackson network with Poisson exogenous arrivals at rates $\{\lambda_0 Q^*_{0j}\}$, service times that are exponential with rates $\{\mu_i\}$, and routing probabilities $\{Q^*_{ij}\}$. The backward routing probabilities $\{Q^*_{ij}\}$ must be chosen to be consistent with the transition rates in the forward process. Since each transition from i to j in the forward process must correspond to a transition from j to i in the backward process, we should have

$$\lambda_i Q_{ij} = \lambda_j Q^*_{ji} \quad ; \quad 0 \leq i, j \leq k. \tag{6.59}$$

Note that $\lambda_i Q_{ij}$ represents the rate at which forward transitions go from i to j, and λ_i represents the rate at which forward transitions *leave* node i. Equation (6.59) takes advantage of the fact that λ_i is also the rate at which forward transitions *enter* node i, and thus the rate at which backward transitions *leave* node i. Using the conjecture that the backward time system is a Jackson network with routing probabilities $\{Q^*_{ij};\, 0 \leq i, j \leq k\}$, we can write down the backward transition rates in the same way as (6.55-6.57),

$$\begin{aligned}
q^*_{\boldsymbol{m},\boldsymbol{m}'} &= \lambda_0 Q^*_{0j} \qquad \text{for } \boldsymbol{m}' = \boldsymbol{m} + \boldsymbol{e}_j && (6.60)\\
&= \mu_i Q^*_{i0} \qquad \text{for } \boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i, m_i > 0,\ 1 \leq i \leq k && (6.61)\\
&= \mu_i Q^*_{ij} \qquad \text{for } \boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j,\quad m - i > 0,\quad 1 \leq i,\quad j \leq k. && (6.62)
\end{aligned}$$

If we substitute (6.59) into (6.60)-(6.62), we obtain

$$\begin{aligned}
q^*_{\boldsymbol{m},\boldsymbol{m}'} &= \lambda_j Q_{j0} \qquad\quad \text{for } \boldsymbol{m}' = \boldsymbol{m} + \boldsymbol{e}_j,\quad 1 \leq j \leq k && (6.63)\\
&= (\mu_i/\lambda_i)\lambda_0 Q_{0i} \quad \text{for } \boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i, m_i > 0,\quad 1 \leq i \leq k && (6.64)\\
&= (\mu_i/\lambda_i)\lambda_j Q_{ji} \quad \text{for } \boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j,\qquad m_i > 0,\ 1 \leq i, j \leq k. && (6.65)
\end{aligned}$$

This gives us our hypothesized backward transition rates in terms of the parameters of the original Jackson network. To use theorem 6.6.2, we must verify that there is a set of positive

numbers, $p(\boldsymbol{m})$, satisfying $\sum_{\boldsymbol{m}} p(\boldsymbol{m}) = 1$ and $\sum_{\boldsymbol{m}} \nu_{\boldsymbol{m}} p_{\boldsymbol{m}} < \infty$, and a set of nonnegative numbers $q^*_{\boldsymbol{m}',\boldsymbol{m}}$ satisfying the following two sets of equations:

$$p(\boldsymbol{m})q_{\boldsymbol{m},\boldsymbol{m}'} = p(\boldsymbol{m}')q^*_{\boldsymbol{m}',\boldsymbol{m}} \qquad \text{for all } \boldsymbol{m}, \boldsymbol{m}' \tag{6.66}$$

$$\sum_{\boldsymbol{m}} q_{\boldsymbol{m},\boldsymbol{m}'} = \sum_{\boldsymbol{m}'} q^*_{\boldsymbol{m},\boldsymbol{m}'} \qquad \text{for all } \boldsymbol{m}. \tag{6.67}$$

We verify (6.66) by substituting (6.55)-(6.57) on the left side of (6.66) and (6.63)-(6.65) on the right side. Recalling that ρ_i is defined as λ_i/μ_i, and cancelling out common terms on each side, we have

$$p(\boldsymbol{m}) = p(\boldsymbol{m}')/\rho_j \qquad \text{for } \boldsymbol{m}' = \boldsymbol{m} + \boldsymbol{e}_j \tag{6.68}$$

$$p(\boldsymbol{m}) = p(\boldsymbol{m}')\rho_i \qquad \text{for } \boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i, m_i > 0 \tag{6.69}$$

$$p(\boldsymbol{m}) = p(\boldsymbol{m}')\rho_i/\rho_j \qquad \text{for } \boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j, \qquad m_i > 0. \tag{6.70}$$

Looking at the case $\boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i$, and using this equation repeatedly to get from state $(0, 0, \ldots, 0)$ up to an arbitrary m, we obtain

$$p(\boldsymbol{m}) = p(0, 0, \ldots, 0) \prod_{i=1}^{k} \}\rho_i^{m_i}. \tag{6.71}$$

It is easy to verify that (6.71) satisfies (6.68)-(6.70) for all possible transitions. Summing over all $\boldsymbol{m}$ to solve for $p(0, 0, \ldots, 0)$, we get

$$\begin{aligned} 1 = \sum_{m_1, m_2, \ldots, m_k} p(\boldsymbol{m}) &= p(0, 0, \ldots, 0) \sum_{m_1} \rho_1^{m_1} \sum_{m_2} \rho_2^{m_2} \cdots \sum_{m_k} \rho_k^{m_k} \\ &= p(0, 0 \ldots, 0)(1-\rho_1)^{-1}(1-\rho_2)^{-1} \ldots (1-\rho_k)^{-1}. \end{aligned}$$

Thus, $p(0, 0, \ldots, 0) = (1-\rho_1)(1-\rho_2)\ldots(1-\rho_k)$, and substituting this in (6.71), we get

$$p(\boldsymbol{m}) = \prod_{i=1}^{k} \}p_i(m_i) = \prod_{i=1}^{k} \Big[(1-\rho_i)\rho_i^{m_i}\Big]. \tag{6.72}$$

where $p_i(m) = (1-\rho_i)\rho_i^m$ is the steady-state distribution of a single M/M/1 queue. Now that we have found the steady-state distribution implied by our assumption about the backward process being a Jackson network, our remaining task is to verify (6.67)

To verify (6.67), *i.e.*, $\sum_{\boldsymbol{m}'} q_{\boldsymbol{m},\boldsymbol{m}'} = \sum_{\boldsymbol{m}'} q^*_{\boldsymbol{m},\boldsymbol{m}'}$, first consider the right side. Using (6.60) to sum over all $\boldsymbol{m}' = \boldsymbol{m} + \boldsymbol{e}_j$, then (6.61) to sum over $\boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i$ (for i such that $m_i > 0$), and finally (6.62) to sum over $\boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j$, (again for i such that $m_i > 0$), we get

$$\sum_{\boldsymbol{m}'} q^*_{\boldsymbol{m},\boldsymbol{m}'} = \sum_{j=1}^{k} \}\lambda_0 Q^*_{0j} + \sum_{i:m_j>0} \mu_i Q^*_{i0} + \sum_{i:m_i>0} \mu_i \sum_{j=1}^{k} Q^*_{ij}. \tag{6.73}$$

Using the fact Q^* is a stochastic matrix, then,

$$\sum_{\boldsymbol{m}'} q^*_{\boldsymbol{m},\boldsymbol{m}'} = \lambda_0 + \sum_{i:m_i>0} \mu_i. \tag{6.74}$$

The left hand side of (6.67) can be summed in the same way to get the result on the right side of (6.74), but we can see that this must be the result by simply observing that λ_0 is the rate of exogenous arrivals and $\sum_{i:m_i>0} \mu_i$ is the overall rate of service completions in state $\boldsymbol{m}$. Note that this also verifies that $\nu_{\boldsymbol{m}} = \sum_{\boldsymbol{m}'} q_{\boldsymbol{m},\boldsymbol{m}'} \geq \lambda_0 + \sum_i \mu_i$, and since $\nu_{\boldsymbol{m}}$ is bounded, $\sum_{\boldsymbol{m}} \nu_{\boldsymbol{m}} p(\boldsymbol{m}) < \infty$. Since all the conditions of Theorem 6.6.2 are satisfied, $p(\boldsymbol{m})$, as given in (6.72), gives the steady-state probabilities for the Jackson network. This also verifies that the backward process is a Jackson network, and hence the exogenous departures are Poisson and independent.

Although the exogenous arrivals and departures in a Jackson network are Poisson, the endogenous processes of customers travelling from one node to another are typically not Poisson if there are feedback paths in the network. Also, although (6.72) shows that the numbers of customers at the different nodes are independent random variables at any given time in steady-state, it is not generally true that the number of customers at one node at one time is independent of the number of customers at another node at another time.

There are many generalizations of the reversibility arguments used above, and many network situations in which the nodes have independent states at a common time. We discuss just two of them here and refer to Kelly, [13], for a complete treatment.

For the first generalization, assume that the service time at each node depends on the number of customers at that node, *i.e.*, μ_i is replaced by μ_{i,m_i}. Note that this includes the M/M/m type of situation in which each node has several independent exponential servers. With this modification, the transition rates in (6.56) and (6.57) are modified by replacing μ_i with μ_{i,m_i}. The hypothesized backward transition rates are modified in the same way, and the only effect of these changes is to replace ρ_i and ρ_j for each i and j in (6.68)-(6.70) with $\rho_{i,m_i} = \lambda_i/\mu_{i,m_i}$ and $\rho_{j,m_j} = \lambda_j/\mu_{j,m_j}$. With this change, (6.71) becomes

$$p(\boldsymbol{m}) = \prod_{i=1}^{k} p_i(m_i) = \prod_{i=1}^{k} p_i(0) \prod_{j=0}^{m_i} \rho_{i,j} \tag{6.75}$$

$$p_i(0) = \left[1 + \sum_{m=1}^{\infty} \prod_{j=0}^{m-1} \rho_{i,j} \right]^{-1} . \tag{6.76}$$

Thus, $p(\boldsymbol{m})$ is given by the product distribution of k individual birth-death systems.

6.7.1 Closed Jackson networks

The second generalization is to a network of queues with a fixed number M of customers in the system and with no exogenous inputs or outputs. Such networks are called *closed Jackson* networks, whereas the networks analyzed above are often called *open Jackson* networks. Suppose a k node closed network has routing probabilities Q_{ij}, $1 \leq i, j \leq k$, where $\sum_j Q_{ij} = 1$, and has exponential service times of rate μ_i (this can be generalized to μ_{i,m_i} as above). We make the same assumptions as before about independence of service variables and routing variables, and assume that there is a path between each pair of nodes. Since $\{Q_{ij};\, 1 \leq i, j \leq k\}$ forms an irreducible stochastic matrix, there is a one dimensional

set of solutions to the steady-state equations

$$\lambda_j = \sum_i \lambda_i Q_{ij}; \qquad 1 \le j \le k. \tag{6.77}$$

We interpret λ_i as the time-average rate of transitions that go into node i. Since this set of equations can only be solved within an unknown multiplicative constant, and since this constant can only be determined at the end of the argument, we define $\{\pi_i;\, 1 \le i \le k\}$ as the particular solution of (6.77) satisfying

$$\pi_j = \sum_i \pi_i Q_{ij};\ 1 \le j \le k; \quad \sum_i \pi_i = 1. \tag{6.78}$$

Thus, for all i, $\lambda_i = \alpha\pi_i$, where α is some unknown constant. The state of the Markov process is again taken as $\boldsymbol{m} = (m_1, m_2, \ldots, m_k)$ with the condition $\sum_i\ m_i = M$. The transition rates of the Markov process are the same as for open networks, except that there are no exogenous arrivals or departures; thus (6.55)-(6.57) are replaced by

$$q_{\boldsymbol{m},\boldsymbol{m}'} = \mu_i Q_{ij} \quad \text{for } \boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j, \qquad m_i > 0,\ \ 1 \le i, j \le k. \tag{6.79}$$

We hypothesize that the backward time process is also a closed Jackson network, and as before, we conclude that if the hypothesis is true, the backward transition rates should be

$$q^*_{\boldsymbol{m},\boldsymbol{m}'} = \mu_i Q^*_{ij} \quad \text{for } \boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j,\ m_i > 0,\ 1 \le i, j \le k \tag{6.80}$$

$$\text{where } \lambda_i Q_{ij} = \lambda_j Q^*_{ji} \qquad \text{for } 1 \le i, j \le k. \tag{6.81}$$

In order to use Theorem 6.6.2 again, we must verify that a PMF $p(\boldsymbol{m})$ exists satisfying $p(\boldsymbol{m})q_{\boldsymbol{m},\boldsymbol{m}'} = p(\boldsymbol{m}')q^*_{\boldsymbol{m}',\boldsymbol{m}}$ for all possible states and transitions, and we must also verify that $\sum_{\boldsymbol{m}'} q_{\boldsymbol{m},\boldsymbol{m}'} = \sum_{\boldsymbol{m}'} q^*_{\boldsymbol{m},\boldsymbol{m}'}$ for all possible $\boldsymbol{m}$. This latter verification is virtually the same as before and is left as an exercise. The former verification, with the use of (72), (73), and (74), becomes

$$p(\boldsymbol{m})(\mu_i/\lambda_i) = p(\boldsymbol{m}')(\mu_j/\lambda_j) \quad \text{for } \boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j, \qquad m_i > 0. \tag{6.82}$$

Using the open network solution to guide our intuition, we see that the following choice of $p(\boldsymbol{m})$ satisfies (6.82) for all possible $\boldsymbol{m}$ (*i.e.*, all $\boldsymbol{m}$ such that $\sum_i\ m_i = M$)

$$p(\boldsymbol{m}) = A \prod_{i=1}^{k} (\lambda_i/\mu_i)^{m_i}; \quad \text{for } \boldsymbol{m} \text{ such that } \sum_i m_i = M. \tag{6.83}$$

The constant A is a normalizing constant, chosen to make $p(\boldsymbol{m})$ sum to unity. The problem with (6.83) is that we do not know λ_i (except within a multiplicative constant independent of i). Fortunately, however, if we substitute π_i/α for λ_i, we see that α is raised to the power $-M$, independent of the state $\boldsymbol{m}$. Thus, letting $A' = A\alpha - M$, our solution becomes

$$p(\boldsymbol{m}) = A' \prod_{i=1}^{K} (\pi_i/\mu_i)^{m_i}; \ \text{ for } \boldsymbol{m} \text{ such that } \sum_i m_i = M. \tag{6.84}$$

$$\frac{1}{A'} = \sum_{\boldsymbol{m}:\, \sum_i m_i = M} \prod_{i=1}^{k} \frac{\pi_i}{\mu_i}\Bigg)^{m_i}. \tag{6.85}$$

Note that the steady-state distribution of the closed Jackson network has been found without solving for the time-average transition rates. Note also that the steady-state distribution looks very similar to that for an open network; that is, it is a product distribution over the nodes with a geometric type distribution within each node. This is somewhat misleading, however, since the constant A' can be quite difficult to calculate. It is surprising at first that the parameter of the geometric distribution can be changed by a constant multiplier in (6.84) and (6.85) (*i.e.*, π_i could be replaced with λ_i) and the solution does not change; the important quantity is the relative values of π_i/μ_i from one value of i to another rather than the absolute value.

In order to find λ_i (and this is important, since it says how quickly the system is doing its work), note that $\lambda_i = \mu_i \Pr\{m_i > 0\})$. Solving for $\Pr\{m_i > 0\}$ requires finding the constant A' in (6.79). In fact, the major difference between open and closed networks is that the relevant constants for closed networks are tedious to calculate (even by computer) for large networks and large M.

6.8 Semi-Markov processes

Semi-Markov processes are generalizations of Markov processes in which the time intervals between transitions have an arbitrary distribution rather than an exponential distribution. To be specific, there is an embedded Markov chain, $\{X_n;\ n \geq 0\}$ with a finite or countably infinite state space, and a sequence $\{U_n;\, n \geq 1\}$ of holding intervals between state transitions. The epochs at which state transitions occur are then given, for $n \geq 1$, as $S_n = \sum_{m=1}^{n} U_n$. The process starts at time 0 with S_0 defined to be 0. The semi-Markov process is then the continuous-time process $\{X(t);\, t \geq 0\}$ where, for each $n \geq 0$, $X(t) = X_n$ for t in the interval $S_n \leq X_n < S_{n+1}$. Initially, $X_0 = i$ where i is any given element of the state space.

The holding intervals $\{U_n;\, n \geq 1\}$ are nonnegative rv's that depend only on the current state X_{n-1} and the next state X_n. More precisely, given $X_{n-1} = j$ and $X_n = k$,say, the interval U_n is independent of $\{U_m;\ m < n\}$ and independent of $\{X_m;\, m < n-1\}$ The conditional distribution function for such an interval U_n is denoted by $G_{ij}(u)$, *i.e.*,

$$\Pr\{U_n \leq u \mid X_{n-1} = j, X_n = k\} = G_{jk}(u). \tag{6.86}$$

The dependencies between the rv's $\{X_n;\, n \geq 0\}$ and $\{U_n;\, n \geq 1$ is illustrated in Figure 6.17.

The conditional mean of U_n, conditional on $X_{n-1} = j$, $X_n = k$, is denoted $\overline{U}(j,k)$, i.e.,

$$\overline{U}(i,j) = \mathsf{E}\left[U_n \mid X_{n-1} = i, X_n = j\right] = \int_{u \geq 0} [1 - G_{ij}(u)]du. \tag{6.87}$$

A semi-Markov process evolves in essentially the same way as a Markov process. Given an initial state, $X_0 = i$ at time 0, a new state $X_1 = j$ is selected according to the embedded chain with probability P_{ij}. Then $U_1 = S_1$ is selected using the distribution $G_{ij}(u)$. Next a new state $X_2 = k$ is chosen according to the probability P_{jk}; then, given $X_1 = j$ and

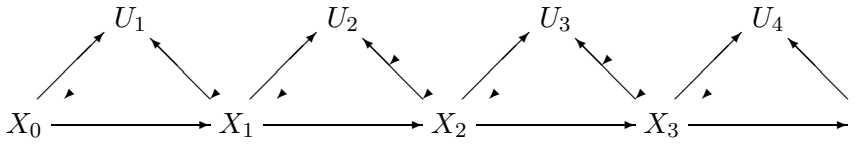


Figure 6.17: The statistical dependencies between the rv's of a semi-Markov process. Each holding interval U_n, conditional on the current state X_{n-1} and next state X_n, is independent of all other states and holding intervals. Note that, conditional on X_n, the holding intervals $U_n, U_{n-1}, \ldots$, are statistically independent of $U_{n+1}, X_{n+2}, \ldots$.

$X_2 = k$, the interval U_2 is selected with distribution function $G_{jk}(u)$. Successive state transitions and transition times are chosen in the same way.

The steady-state behavior of semi-Markov processes can be analyzed in virtually the same way as Markov processes. We outline this in what follows, and often omit proofs where they are the same as the corresponding proof for Markov processes. First, since the holding intervals, U_n, are rv's, the transition epochs, $S_n = \sum_{m=1}^{n} U_n$, are also rv's. The following lemma then follows in the same way as Lemma 6.2.1 for Markov processes.

Lemma 6.8.1. *Let $M_i(t)$ be the number of transitions in a semi-Markov process in the interval $(0, t]$ for some given initial state $X_0 = i$. Then $\lim_{t\to\infty} M_i(t) = \infty$ WP1.*

In what follows, we assume that the embedded Markov chain is irreducible and positive-recurrent. We want to find the limiting fraction of time that the process spends in any given state, say j. We will find that this limit exists WP1, and will find that it depends only on the steady-state probabilities of the embedded Markov chain and on the expected holding interval in each state. This is given by

$$\overline{U}(j) = \mathsf{E}\left[U_n \mid X_{n-1} = j\right] = \sum_k P_{jk} \mathsf{E}\left[U_n \mid X_{n-1} = j, X_n = k\right] = \sum_k P_{jk} \overline{U}(j,k), \qquad (6.88)$$

where $\overline{U}(j,k)$ is given in 6.87. The steady-state probabilities $\{\pi_i;\, i \geq 0\}$ for the embedded chain tell us the fraction of transitions that enter any given state i. Since $\overline{U}(i)$ is the expected holding interval in i per transition into i, we would guess that the fraction of time spent in state i should be proportional to $\pi_i \overline{U}(i)$. Normalizing, we would guess that the time-average probability of being in state i should be

$$p_j = \frac{\pi_j \overline{U}(j)}{\sum_k \pi_k \overline{U}(k)}. \qquad (6.89)$$

Identifying the mean holding interval, $\overline{U}_j$ with $1/\nu_j$, this is the same result that we established for the Markov process case. Using the same arguments, we find this is valid for the semi-Markov case. It is valid both in the conventional case where each p_j is positive and $\sum_j p_j = 1$, and also in the case where $\sum_k \pi_k \overline{U}(k) = \infty$, where each $p_j = 0$. The analysis is based on the fact that successive transitions to some given state, say j, given $X_0 = i$, form a delayed renewal process.

Lemma 6.8.2. *Consider a semi-Markov process with an irreducible recurrent embedded chain $\{X_n; n \geq 0\}$. Given $X_0 = i$, let $\{M_{ij}(t); t \geq 0\}$ be the number of transitions into a given state j in the interval $(0, t]$. Then $\{M_{ij}(t); t \geq 0\}$ is a delayed renewal process (or, if $j = i$, is an ordinary renewal process).*

This is the same as Lemma 6.2.2, but it is not quite so obvious that successive intervals between visits to state j are statistically independent. This can be seen, however, from Figure 6.17, which makes it clear that, given $X_n = j$, the future holding intervals, $U_n, U_{n+1}, \ldots$, are independent of the past intervals $U_{n-1}, U_{n-2}, \ldots$.

Next, using the same renewal reward process as in Lemma 6.2.3, assigning reward 1 whenever $X(t) = j$, we define W_n as the interval between the $n-1$st and the nth entry to state j and get the following lemma:

Lemma 6.8.3. *Consider a semi-Markov process with an irreducible, recurrent, embedded Markov chain starting in $X_0 = i$. Then with probability 1, the limiting time-average in state j is given by $p_j(i) = \frac{\overline{U}_j}{\overline{W}(j)}$.*

This lemma has omitted any assertion about the limiting ensemble probability of state j, *i.e.*, $\lim_{t\to\infty} \Pr\{X(t) = j\}$. This follows easily from Blackwell's theorem, but depends on whether the successive entries to state j, *i.e.*, $\{W_n;\ n \geq 1\}$, are arithmetic or non-arithmetic. This is explored in Exercise 6.33. The lemma shows (as expected) that the limiting time-average in each state is independent of the starting state, so we henceforth replace $p_j(i)$ with p_j.

Next, let $M_i(t)$ be the total number of transitions in the semi-Markov process up to and including time n, given $X_0 = i$. This is not a renewal counting process, but, as with Markov processes, it provides a way to combine the time-average results for all states j. The following theorem is the same as that for Markov processes, except for the omission of ensemble average results.

Theorem 6.8.1. *Consider a semi Markov process with an irreducible, positive-recurrent, embedded Markov chain. Let $\{\pi_j;\ j \geq 0\}$ be the steady-state probabilities of the embedded chain and let $X_0 = i$ be the starting state. Then, with probability 1, the limiting time-average fraction of time spent in any arbitrary state j is given by*

$$p_j = \frac{\pi_j \overline{U}(j)}{\sum_k \pi_k \overline{U}(j)}. \tag{6.90}$$

The expected time between returns to state j is

$$\overline{W}(j) = \frac{\sum_k \pi_k \overline{U}(k)}{\pi_j}, \tag{6.91}$$

and the rate at which transitions take place is independent of X_0 and given by

$$\lim_{t\to\infty} \frac{M_i(t)}{t} = \frac{1}{\sum_k \pi_k \overline{U}(k)} \qquad \text{WP1.} \tag{6.92}$$

For a semi-Markov process, knowing the steady state probability of $X(t) = j$ for large t does not completely specify the steady-state behavior. Another important steady state question is to determine the fraction of time involved in i to j transitions. To make this notion precise, define $Y(t)$ as the residual time until the next transition after time t (i e., $t + Y(t)$ is the epoch of the next transition after time t). We want to determine the fraction of time t over which $X(t) = i$ and $X(t + Y(t)) = j$. Equivalently, for a non-arithmetic process, we want to determine $\Pr\{X(t) = i, X(t + Y(t) = j)\}$ in the limit as $t \to \infty$. Call this limit $Q(i, j)$.

Consider a renewal process, starting in state i and with renewals on transitions to state i. Define a reward $R(t) = 1$ for $X(t) = i$, $X(t + Y(t)) = j$ and $R(t) = 0$ otherwise (see Figure 6.18). That is, for each n such that $X(S_n) = i$ and $X(S_{n+1}) = j$, $R(t) = 1$ for $S_n \leq t < S_{n+1}$. The expected reward in an inter-renewal interval is then $P_{ij}\overline{U}(i, j)$. It follows that $Q(i, j)$ is given by

$$Q(i,j) = \lim_{t\to\infty} \frac{\int_0^t R(\tau)(d\tau}{t} = \frac{P_{ij}\overline{U}(i,j)}{\overline{W}(i)} = \frac{p_i P_{ij}\overline{U}(i,j)}{\overline{U}(i)}. \tag{6.93}$$

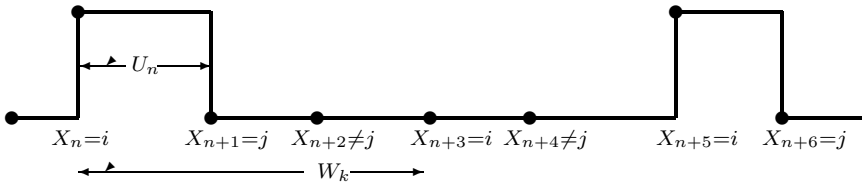


Figure 6.18: The renewal-reward process for i to j transitions. The expected value of U_n if $X_n = i$ and $X_{n+1} = j$ is $\overline{U}(i, j)$ and the expected interval between entries to i is $\overline{W}(i)$.

6.8.1 Example — the M/G/1 queue

As one example of a semi-Markov chain, consider an M/G/1 queue. Rather than the usual interpretation in which the state of the system is the number of customers in the system, we view the state of the system as changing only at departure times; the new state at a departure time is the number of customers left behind by the departure. This state then remains fixed until the next departure. New customers still enter the system according to the Poisson arrival process, but these new customers are not considered as part of the state until the next departure time. The number of customers in the system at arrival epochs does not in general constitute a "state" for the system, since the age of the current service is also necessary as part of the statistical characterization of the process.

One purpose of this example is to illustrate that it is often more convenient to visualize the transition interval $U_n = S_n - S_{n-1}$ as being chosen first and the new state X_n as being chosen second rather than choosing the state first and the transition time second. For the M/G/1 queue, first suppose that the state is some $i > 0$. In this case, service begins on

the next customer immediately after the old customer departs. Thus, U_n, conditional on $X_n = i$ for $i > 0$, has the distribution of the service time, say $G(u)$. The mean interval until a state transition occurs is

$$\overline{U}(i) = \int_0^\infty [1 - G(u)]du; \quad i > 0. \tag{6.94}$$

Given the interval u for a transition from state $i > 0$, the number of arrivals in that period is a Poisson random variable with mean λu, where λ is the Poisson arrival rate. Since the next state j is the old state i, plus the number of new arrivals, minus the single departure,

$$\Pr\{X_{n+1} = j \mid X_n = i, U_n = u\} = \frac{(\lambda u)^{j+i+1} \exp(-\lambda u)}{(j-i+1)!}. \tag{6.95}$$

for $j \geq i - 1$. For $j < i - 1$, the probability above is 0. The unconditional probability P_{ij} of a transition from i to j can then be found by multiplying the right side of (6.95) by the probability density $g(u)$ of the service time and integrating over u.

$$P_{ij} = \int_0^\infty \frac{G(u)(\lambda u)^{j-i+1} \exp(-\lambda u)}{(j-i+1)} \, du; \quad j \geq i-1, i > 0. \tag{6.96}$$

For the case $i = 0$, the server must wait until the next arrival before starting service. Thus the expected time from entering the empty state until a service completion is

$$\overline{U}(0) = (1/\lambda) + \int_0^\infty [1 - G(u)]du. \tag{6.97}$$

We can evaluate P_{0j} by observing that the departure of that first arrival leaves j customers in this system if and only if j customers arrive during the service time of that first customer; *i.e.*, the new state doesn't depend on how long the server waits for a new customer to serve, but only on the arrivals while that customer is being served. Letting $g(u)$ be the density of the service time,

$$P_{0j} = \int_0^\infty \frac{g(u)(\lambda u)^j \exp(-\lambda u)}{j!} \, du; \quad j \geq 0. \tag{6.98}$$

6.9 Summary

We have seen that Markov processes with countable state spaces are remarkably similar to Markov chains with countable state spaces, and throughout the chapter, we frequently made use of both the embedded chain corresponding to the process and to the sampled time approximation to the process.

For irreducible processes, the steady-state equations, (6.23) and $\sum_i p_i \nu_i = 1$, were found to specify the steady-state probabilities, p_i, which have significance both as time-averages and as limiting probabilities. If the transition rates ν_i are bounded, then the sampled-time approximation exists and has the same steady-state probabilities as the Markov process itself. If the transition rates ν_i are unbounded but $\sum_i \; p_i \nu_i < \infty$, then the embedded chain

is positive recurrent and has steady-state probabilities, but the sampled-time approximation does not exist. We assumed throughout the remainder of the chapter that $\sum_i p_i\nu_i < \infty$. This ruled out irregular processes in which there is no meaningful steady state, and also some peculiar processes such as that in Exercise 6.9 where the embedded chain is null recurrent.

Section 6.3 developed the Kolmogoroff backward and forward differential equations for the transient probabilities $P_{ij}(t)$ of being in state j at time t given state i at time 0. We showed that for finite-state processes, these equations can be solved by finding the eigenvalues and eigenvectors of the transition rate matrix Q. There are close analogies between this analysis and the algebraic treatment of finite-state Markov chains in chapter 3, and exercise 6.7 showed how the transients of the process are related to the transients of the sampled time approximation.

For irreducible processes with bounded transition rates, uniformization was introduced as a way to simplify the structure of the process. The addition of self transitions does not change the process itself, but can be used to adjust the transition rates ν_i to be the same for all states. This changes the embedded Markov chain, and the steady-state probabilities for the embedded chain become the same as those for the process. The epochs at which transitions occur then form a Poisson process which is independent of the set of states entered. This yields a separation between the transition epochs and the sequence of states.

The next two sections analyzed birth-death processes and reversibility. The results about birth-death Markov chains and reversibility for Markov chains carried over almost without change to Markov processes. These results are central in queueing theory, and Burke's theorem allowed us to look at simple queueing networks with no feedback and to understand how feedback complicates the problem.

Jackson networks were next discussed. These are important in their own right and also provide a good example of how one can solve complex queueing problems by studying the reverse time process and making educated guesses about the steady-state behavior. The somewhat startling result here is that in steady state, and at a fixed time, the number of customers at each node is independent of the number at each other node and satisfies the same distribution as for an M/M/1 queue. Also the exogenous departures from the network are Poisson and independent from node to node. We emphasized that the number of customers at one node at one time is often dependent on the number at other nodes at other times. The independence holds only when all nodes are viewed at the same time.

Finally, semi-Markov processes were introduced. Renewal theory again provided the key to analyzing these systems. Theorem 6.8.1 showed how to find the steady-state probabilities of these processes, and it was shown that these probabilities could be interpreted both as time-averages and, in the case of non-arithmetic transition times, as limiting probabilities in time.

For further reading on Markov processes, see [13], [16], [22], and [8].

6.10 Exercises

Exercise 6.1. Consider an M/M/1 queue as represented in Figure 6.4. Assume throughout that $X_0 = i$ where $i > 0$. The purpose of this exercise is to understand the relationship between the holding interval until the next state transition and the interval until the next arrival to the M/M/1 queue. Your explanations in the following parts can and should be very brief.

a) Explain why the expected holding interval $\mathsf{E}[U_1|X_0 = i]$ until the next state transition is $1/(\lambda + \mu)$.

b) Explain why the expected holding interval U_1, conditional on $X_0 = i$ and $X_1 = i + 1$, is

$$\mathsf{E}[U_1|X_0 = i,\, X_1 = i + 1] = 1/(\lambda + \mu).$$

Show that $\mathsf{E}[U_1|X_0 = i, X_1 = i - 1]$ is the same.

c) Let V be the time of the first arrival after time 0 (this may occur either before or after the time W of the first departure.) Show that

$$\begin{aligned}\mathsf{E}[V|X_0 = i,\, X_1 = i + 1] &= \frac{1}{\lambda + \mu}\\ \mathsf{E}[V|X_0 = i,\, X_1 = i - 1] &= \frac{1}{\lambda + \mu} + \frac{1}{\lambda}.\end{aligned}$$

Hint: In the second equation, use the memorylessness of the exponential rv and the fact that V under this condition is the time to the first departure plus the remaining time to an arrival.

d) Use your solution to part c) plus the probability of moving up or down in the Markov chain to show that $\mathsf{E}[V] = 1/\lambda$. (Note: you already know that $\mathsf{E}[V] = 1/\lambda$. The purpose here is to show that your solution to part c) is consistent with that fact.)

Exercise 6.2. Consider a Markov process for which the embedded Markov chain is a birth-death chain with transition probabilities $P_{i,i+1} = 2/5$ for all $i \geq 0$, $P_{i,i-1} = 3/5$ for all $i \geq 1$, $P_{01} = 1$, and $P_{ij} = 0$ otherwise.

a) Find the steady-state probabilities $\{\pi_i;\, i \geq 0\}$ for the embedded chain.

b) Assume that the transition rate ν_i out of state i, for $i \geq 0$, is given by $\nu_i = 2^i$. Find the transition rates $\{q_{ij}\}$ between states and find the steady-state probabilities $\{p_i\}$ for the Markov process. Explain heuristically why $\pi_i \neq p_i$.

c) Explain why there is no sampled-time approximation for this process. Then truncate the embedded chain to states 0 to m and find the steady-state probabilities for the sampled-time approximation to the truncated process.

d) Show that as $m \to \infty$, the steady-state probabilities for the sequence of sampled-time approximations approaches the probabilities p_i in part b).

Exercise 6.3. Consider a Markov process for which the embedded Markov chain is a birth-death chain with transition probabilities $P_{i,i+1} = 2/5$ for all $i \geq 1$, $P_{i,i-1} = 3/5$ for all $i \geq 1$, $P_{01} = 1$, and $P_{ij} = 0$ otherwise.

a) Find the steady-state probabilities $\{\pi_i;\ i \geq 0\}$ for the embedded chain.

b) Assume that the transition rate out of state i, for $i \geq 0$, is given by $\nu_i = 2^{-i}$. Find the transition rates $\{q_{ij}\}$ between states and show that there is no probability vector solution $\{p_i;\ i \geq 0\}$ to (6.23).

e) Argue that the expected time between visits to any given state i is infinite. Find the expected number of transitions between visits to any given state i. Argue that, starting from any state i, an eventual return to state i occurs with probability 1.

f) Consider the sampled-time approximation of this process with $\delta = 1$. Draw the graph of the resulting Markov chain and argue why it must be null-recurrent.

Exercise 6.4. Consider the Markov process for the M/M/1 queue, as given in Figure 6.4.

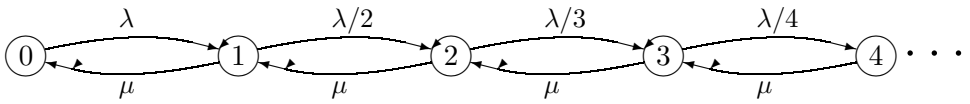
a) Find the steady state process probabilities (as a function of $\rho = \lambda/\mu$) from (6.15) and also as the solution to (6.23). Verify that the two solutions are the same.

b) For the remaining parts of the exercise, assume that $\rho = 0.01$, thus ensuring (for aiding intuition) that states 0 and 1 are much more probable than the other states. Assume that the process has been running for a very long time and is in steady state. Explain in your own words the difference between π_1 (the steady-state probability of state 1 in the embedded chain) and p_1 (the steady-state probability that the process is in state 1. More explicitly, what experiments could you perform (repeatedly) on the process to measure π_1 and p_1.

c) Now suppose you want to start the process in steady state. Show that it is impossible to choose initial probabilities so that both the process and the embedded chain start in steady state. Which version of steady state is closest to your intuitive view? (There is no correct answer here, but it is important to realiize that the notion of steady state is not quite as simple as you might imagine).

d) Let $M(t)$ be the number of transitions (counting both arrivals and departures) that take place by time t in this Markov process and assume that the embedded Markov chain starts in steady state at time 0. Let $U_1, U_2, \ldots$, be the sequence of holding intervals between transitions (with U_1 being the time to the first transition). Show that these rv's are identically distributed. Show by example that they are not independent (*i.e.*, $M(t)$ is not a renewal process).

Exercise 6.5. Consider the Markov process illustrated below. The transitions are labelled by the rate q_{ij} at which those transitions occur. The process can be viewed as a single server queue where arrivals become increasingly discouraged as the queue lengthens. The word *time-average* below refers to the limiting time-average over each sample-path of the process, except for a set of sample paths of probability 0.



a) Find the time-average fraction of time p_i spent in each state $i > 0$ in terms of p_0 and then solve for p_0. Hint: First find an equation relating p_i to p_{i+1} for each i. It also may help to recall the power series expansion of e^x.

b) Find a closed form solution to $\sum_i p_i \nu_i$ where ν_i is the departure rate from state i. Show that the process is positive recurrent for all choices of $\lambda > 0$ and $\mu > 0$ and explain intuitively why this must be so.

c) For the embedded Markov chain corresponding to this process, find the steady-state probabilities π_i for each $i \geq 0$ and the transition probabilities P_{ij} for each i, j.

d) For each i, find both the time-average interval and the time-average number of overall state transitions between successive visits to i.

Exercise 6.6. (Detail from proof of Theorem 6.2.1) **a)** Let U_n be the nth holding interval for a Markov process starting in steady state, with $\Pr\{X_0 = i\} = \pi_i$. Show that $\mathsf{E}[U_n] = \sum_k \pi_k/\nu_k$ for each integer n.

b) Let S_n be the epoch of the nth transition. Show that $\Pr\{S_n \geq n\beta\} \leq \frac{\sum_k \pi_k/\nu_k}{\beta}$ for all $\beta > 0$.

c) Let $M(t)$ be the number of transitions of the Markov process up to time t, given that X_0 is in steady state. Show that $\Pr\{M(n\beta) < n\} > 1 - \frac{\sum_k \pi_k/\nu_k}{\beta}$.

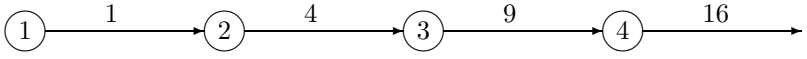
d) Show that if $\sum_k \pi_k/\nu_k$ is finite, then $\lim_{t\to\infty} M(t)/t = 0$ WP1 is impossible. (Note that this is equivalent to showing that $\lim_{t\to\infty} M(t)/t = 0$ WP1 implies $\sum_k \pi_k/\nu_k = \infty$).

e) Let $M_i(t)$ be the number of transitions by time t starting in state i. Show that if $\sum_k \pi_k/\nu_k$ is finite, then $\lim_{t\to\infty} M_i(t)/t = 0$ WP1 is impossible.

Exercise 6.7. a) Consider the process in the figure below. The process starts at $X(0) = 1$, and for all $i \geq 1$, $P_{i,i+1} = 1$ and $\nu_i = i^2$ for all i. Let S_n be the epoch when the n^{th} transition occurs. Show that

$$\mathsf{E}[S_n] = \sum_{i=1}^{n} \}i^{-2} < 2 \text{ for all } n.$$

Hint: Upper bound the sum from $i = 2$ by integrating x^{-2} from $x = 1$.



b) Use the Markov inequality to show that $\Pr\{S_n > 4\} \leq 1/2$ for all n. Show that the probability of an infinite number of transitions by time 4 is at least 1/2.

Exercise 6.8. a) Consider a Markov process with the set of states $\{0, 1, \dots\}$ in which the transition rates $\{q_{ij}\}$ between states are given by $q_{i,i+1} = (3/5)2^i$ for $i \geq 0$, $q_{i,i-1} = (2/5)2^i$

for $i \geq 1$, and $q_{ij} = 0$ otherwise. Find the transition rate ν_i out of state i for each $i \geq 0$ and find the transition probabilities $\{P_{ij}\}$ for the embedded Markov chain.

b) Find a solution $\{p_i;\ i \geq 0\}$ with $\sum_i p_i = 1$ to (6.23).

c) Show that all states of the embedded Markov chain are transient.

d) Explain in your own words why your solution to part b) is not in any sense a set of steady-state probabilities

Exercise 6.9. Let $q_{i,i+1} = 2^{i-1}$ for all $i \geq 0$ and let $q_{i,i-1} = 2^{i-1}$ for all $i \geq 1$. All other transition rates are 0.

a) Solve the steady-state equations and show that $p_i = 2^{-i-1}$ for all $i \geq 0$.

b) Find the transition probabilities for the embedded Markov chain and show that the chain is null recurrent.

c) For any state i, consider the renewal process for which the Markov process starts in state i and renewals occur on each transition to state i. Show that, for each $i \geq 1$, the expected inter-renewal interval is equal to 2. Hint: Use renewal-reward theory.

d) Show that the expected number of transitions between each entry into state i is infinite. Explain why this does *not* mean that an infinite number of transitions can occur in a finite time.

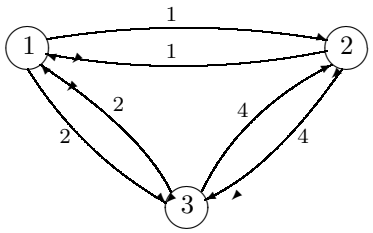
Exercise 6.10. a) Consider the two state Markov process of Example 6.3.1 with $q_{01} = \lambda$ and $q_{10} = \mu$. Find the eigenvalues and eigenvectors of the transition rate matrix $[Q]$.

b) If $[Q]$ has M distinct eigenvalues, the differential equation $d[P(t)]/dt = [Q][P(t)]$ can be solved by the equation

$$[P(t)] = \sum_{i=1}^{\mathsf{M}} \boldsymbol{\nu}_i e^{t\lambda_i} \boldsymbol{p}^{\mathsf{T}}{}_i,$$

where $\boldsymbol{p}_i$ and $\boldsymbol{\nu}_i$ are the left and right eigenvectors of eigenvalue λ_i. Show that this equation gives the same solution as that given for Example 6.3.1.

Exercise 6.11. Consider the three state Markov process below; the number given on edge (i, j) is q_{ij}, the transition rate from i to j. Assume that the process is in steady state.



a) Is this process reversible?

b) Find p_i, the time-average fraction of time spent in state i for each i.

c) Given that the process is in state i at time t, find the mean delay from t until the process leaves state i.

d) Find π_i, the time-average fraction of all transitions that go into state i for each i.

e) Suppose the process is in steady state at time t. Find the steady state probability that the next state to be entered is state 1.

f) Given that the process is in state 1 at time t, find the mean delay until the process first returns to state 1.

g) Consider an arbitrary irreducible finite-state Markov process in which $q_{ij} = q_{ji}$ for all i, j. Either show that such a process is reversible or find a counter example.

Exercise 6.12. a) Consider an M/M/1 queueing system with arrival rate λ, service rate μ, $\mu > \lambda$. Assume that the queue is in steady state. Given that an arrival occurs at time t, find the probability that the system is in state i immediately *after* time t.

b) Assuming FCFS service, and conditional on i customers in the system immediately after the above arrival, characterize the time until the above customer departs as a sum of random variables.

c) Find the unconditional probability density of the time until the above customer departs. **Hint:** You know (from splitting a Poisson process) that the sum of a geometrically distributed number of IID exponentially distributed random variables is exponentially distributed. Use the same idea here.

Exercise 6.13. a) Consider an M/M/1 queue in steady state. Assume $\rho = \lambda/\mu < 1$. Find the probability $Q(i, j)$ for $i \geq j > 0$ that the system is in state i at time t and that $i - j$ departures occur before the next arrival.

b) Find the PMF of the state immediately before the first arrival after time t.

c) There is a well-known queueing principle called PASTA, standing for "Poisson arrivals see time-averages". Given your results above, give a more precise statement of what that principle means in the case of the M/M/1 queue.

Exercise 6.14. A small bookie shop has room for at most two customers. Potential customers arrive at a Poisson rate of 10 customers per hour; they enter if there is room and are turned away, never to return, otherwise. The bookie serves the admitted customers in order, requiring an exponentially distributed time of mean 4 minutes per customer.

a) Find the steady-state distribution of number of customers in the shop.

b) Find the rate at which potential customers are turned away.

c) Suppose the bookie hires an assistant; the bookie and assistant, working together, now serve each customer in an exponentially distributed time of mean 2 minutes, but there is

only room for one customer (*i.e.*, the customer being served) in the shop. Find the new rate at which customers are turned away.

Exercise 6.15. This exercise explores a continuous time version of a simple branching process.

Consider a population of primitive organisms which do nothing but procreate and die. In particular, the population starts at time 0 with one organism. This organism has an exponentially distributed lifespan T_0 with rate μ (*i.e.*, $\Pr\{T_0 \geq \tau\} = e^{-\mu\tau}$). While this organism is alive, it gives birth to new organisms according to a Poisson process of rate λ. Each of these new organisms, while alive, gives birth to yet other organisms. The lifespan and birthrate for each of these new organisms are independently and identically distributed to those of the first organism. All these and subsequent organisms give birth and die in the same way, again indepedently of all other organisms.

a) Let $X(t)$ be the number of (live) organisms in the population at time t. Show that $\{X(t);\ t \geq 0\}$ is a Markov process and specify the transition rates between the states.

b) Find the embedded Markov chain $\{X_n;\ n \geq 0\}$ corresponding to the Markov process in part a). Find the transition probabilities for this Markov chain.

c) Explain why the Markov process and Markov chain above are not irreducible. Note: The majority of results you have seen for Markov processes assume the process is irreducible, so be careful not to use those results in this exercise.

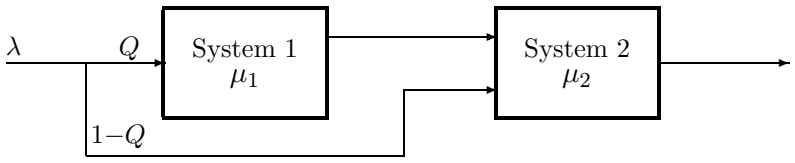
d) For purposes of analysis, add an additional transition of rate λ from state 0 to state 1. Show that the Markov process and the embedded chain are irreducible. Find the values of λ and μ for which the modified chain is positive recurrent, null-recurrent, and transient.

e) Assume that $\lambda < \mu$. Find the steady state process probabilities for the modified Markov process.

f) Find the mean recurrence time between visits to state 0 for the modified Markov process.

g) Find the mean time $\overline{T}$ for the population in the original Markov process to die out. Note: We have seen before that removing transitions from a Markov chain or process to create a trapping state can make it easier to find mean recurrence times. This is an example of the opposite, where adding an exit from a trapping state makes it easy to find the recurrence time.

Exercise 6.16. Consider the job sharing computer system illustrated below. Incoming jobs arrive from the left in a Poisson stream. Each job, independently of other jobs, requires pre-processing in system 1 with probability Q. Jobs in system 1 are served FCFS and the service times for successive jobs entering system 1 are IID with an exponential distribution of mean $1/\mu_1$. The jobs entering system 2 are also served FCFS and successive service times are IID with an exponential distribution of mean $1/\mu_2$. The service times in the two systems are independent of each other and of the arrival times. Assume that $\mu_1 > \lambda Q$ and that $\mu_2 > \lambda$. Assume that the combined system is in steady state.



a Is the input to system 1 Poisson? Explain.

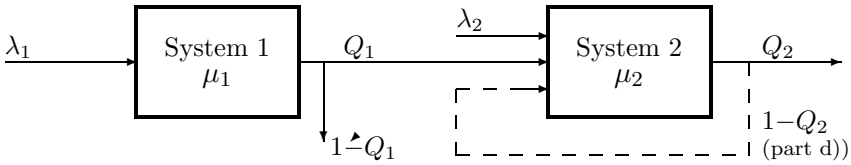
b) Are each of the two input processes coming into system 2 Poisson? Explain.

d) Give the joint steady-state PMF of the number of jobs in the two systems. Explain briefly.

e) What is the probability that the first job to leave system 1 after time t is the same as the first job that entered the entire system after time t?

f) What is the probability that the first job to leave system 2 after time t both passed through system 1 and arrived at system 1 after time t?

Exercise 6.17. Consider the following combined queueing system. The first queue system is M/M/1 with service rate μ_1. The second queue system has IID exponentially distributed service times with rate μ_2. Each departure from system 1 independently $1 - Q_1$. System 2 has an additional Poisson input of rate λ_2, independent of inputs and outputs from the first system. Each departure from the second system independently leaves the combined system with probability Q_2 and re-enters system 2 with probability $1 - Q_2$. For parts a), b), c) assume that $Q_2 = 1$ (*i.e.*, there is no feedback).



a) Characterize the process of departures from system 1 that enter system 2 and characterize the overall process of arrivals to system 2.

b) Assuming FCFS service in each system, find the steady state distribution of time that a customer spends in each system.

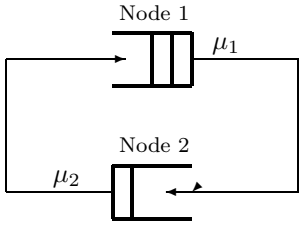
c) For a customer that goes through both systems, show why the time in each system is independent of that in the other; find the distribution of the combined system time for such a customer.

d) Now assume that $Q_2 < 1$. Is the departure process from the combined system Poisson? Which of the three input processes to system 2 are Poisson? Which of the input processes are independent? Explain your reasoning, but do not attempt formal proofs.

Exercise 6.18. Suppose a Markov chain with transition probabilities $\{P_{ij}\}$ is reversible. Suppose we change the transition probabilities out of state 0 from $\{P_{0j};\, j \geq 0\}$ to $\{P'_{0j};\, j \geq$

0}. Assuming that all P_{ij} for $i \neq 0$ are unchanged, what is the most general way in which we can choose $\{P'_{0j};\, j \geq 0\}$ so as to maintain reversibility? Your answer should be explicit about how the steady-state probabilities $\{\pi_i;\, i \geq 0\}$ are changed. Your answer should also indicate what this problem has to do with uniformization of reversible Markov processes, if anything. Hint: Given P_{ij} for all i, j, a single additional parameter will suffice to specify P'_{0j} for all j.

Exercise 6.19. Consider the closed queueing network in the figure below. There are three customers who are doomed forever to cycle between node 1 and node 2. Both nodes use FCFS service and have exponentially distributed IID service times. The service times at one node are also independent of those at the other node and are independent of the customer being served. The server at node i has mean service time $1/\mu_i$, $i = 1, 2$. Assume to be specific that $\mu_2 < \mu_1$.



a) The system can be represented by a four state Markov process. Draw its graphical representation and label it with the individual states and the transition rates between them.

b) Find the steady-state probability of each state.

c) Find the time-average rate at which customers leave node 1.

d) Find the time-average rate at which a given customer cycles through the system.

e) Is the Markov process reversible? Suppose that the backward Markov process is interpreted as a closed queueing network. What does a departure from node 1 in the forward process correspond to in the backward process? Can the transitions of a single customer in the forward process be associated with transitions of a single customer in the backward process?

Exercise 6.20. Consider an M/G/1 queueing system with last come first serve (LCFS) pre-emptive resume service. That is, customers arrive according to a Poisson process of rate λ. A newly arriving customer interrupts the customer in service and enters service itself. When a customer is finished, it leaves the system and the customer that had been interrupted by the departing customer resumes service from where it had left off. For example, if customer 1 arrives at time 0 and requires 2 units of service, and customer 2 arrives at time 1 and requires 1 unit of service, then customer 1 is served from time 0 to 1; customer 2 is served from time 1 to 2 and leaves the system, and then customer 1 completes service from time 2 to 3. Let X_i be the service time required by the i^{th} customer; the X_i are IID random variables with expected value $\mathsf{E}[X]$; they are independent of customer arrival times. Assume $\lambda\, \mathsf{E}[X] < 1$.

a) Find the mean time between busy periods (*i.e.*, the time until a new arrival occurs after the system becomes empty).

b) Find the time-average fraction of time that the system is busy.

c) Find the mean duration, $\mathsf{E}[B]$, of a busy period. Hint: use a) and b).

d) Explain briefly why the customer that starts a busy period remains in the system for the entire busy period; use this to find the expected system time of a customer given that that customer arrives when the system is empty.

e) Is there any statistical dependence between the system time of a given customer (*i.e.*, the time from the customer's arrival until departure) and the number of customers in the system when the given customer arrives?

f) Show that a customer's expected system time is equal to $\mathsf{E}[B]$. Hint: Look carefully at your answers to **d)** and **e)**.

g) Let C be the expected system time of a customer conditional on the service time X of that customer being 1. Find (in terms of C) the expected system time of a customer conditional on $X = 2$; (Hint: compare a customer with $X = 2$ to two customers with $X = 1$ each); repeat for arbitrary $X = x$.

h) Find the constant C. Hint: use **f)** and **g)**; don't do any tedious calculations.

Exercise 6.21. Consider a queueing system with two classes of customers, type A and type B. Type A customers arrive according to a Poisson process of rate λ_A and customers of type B arrive according to an independent Poisson process of rate λ_B. The queue has a FCFS server with exponentially distributed IID service times of rate $\mu > \lambda_A + \lambda_B$. Characterize the departure process of class A customers; explain carefully. Hint: Consider the combined arrival process and be judicious about how to select between A and B types of customers.

Exercise 6.22. Consider a pre-emptive resume last come first serve (LCFS) queueing system with two classes of customers. Type A customer arrivals are Poisson with rate λ_A and Type B customer arrivals are Poisson with rate λ_B. The service time for type A customers is exponential with rate μ_A and that for type B is exponential with rate μ_B. Each service time is independent of all other service times and of all arrival epochs.

Define the "state" of the system at time t by the string of customer types in the system at t, in order of arrival. Thus state AB means that the system contains two customers, one of type A and the other of type B; the type B customer arrived later, so is in service. The set of possible states arising from transitions out of AB is as follows:

ABA if another type A arrives.

ABB if another type B arrives.

A if the customer in service (B) departs.

Note that whenever a customer completes service, the next most recently arrived resumes service, so the state changes by eliminating the final element in the string.

a) Draw a graph for the states of the process, showing all states with 2 or fewer customers and a couple of states with 3 customers (label the empty state as E). Draw an arrow, labelled by the rate, for each state transition. Explain why these are states in a Markov process.

b) Is this process reversible. Explain. Assume positive recurrence. Hint: If there is a transition from one state S to another state S', how is the number of transitions from S to S' related to the number from S' to S?

c) Characterize the process of type A departures from the system (*i.e.*, are they Poisson?; do they form a renewal process?; at what rate do they depart?; etc.)

d) Express the steady-state probability $\Pr\{A\}$ of state A in terms of the probability of the empty state $\Pr\{E\}$. Find the probability $\Pr\{AB\}$ and the probability $\Pr\{ABBA\}$ in terms of $\Pr\{E\}$. Use the notation $\rho_A = \lambda_A/\mu_A$ and $\rho_B = \lambda_B/\mu_B$.

e) Let Q_n be the probability of n customers in the system, as a function of $Q_0 = \Pr\{E\}$. Show that $Q_n = (1-\rho)\rho^n$ where $\rho = \rho_A + \rho_B$.

Exercise 6.23. a) Generalize Exercise 6.22 to the case in which there are m types of customers, each with independent Poisson arrivals and each with independent exponential service times. Let λ_i and μ_i be the arrival rate and service rate respectively of the i^{th} user. Let $\rho_i = \lambda_i/\mu_i$ and assume that $\rho = \rho_1 + \rho_2 + \cdots + \rho_m < 1$. In particular, show, as before that the probability of n customers in the system is $Q_n = (1-\rho)\rho^n$ for $0 \leq n < \infty$.

b) View the customers in part a) as a single type of customer with Poisson arrivals of rate $\lambda = \sum_i \lambda_i$ and with a service density $\sum_i (\lambda_i/\lambda)\mu_i \exp(-\mu_i x)$. Show that the expected service time is ρ/λ. Note that you have shown that, if a service distribution can be represented as a weighted sum of exponentials, then the distribution of customers in the system for LCFS service is the same as for the M/M/1 queue with equal mean service time.

Exercise 6.24. Consider a k node Jackson type network with the modification that each node i has s servers rather than one server. Each server at i has an exponentially distributed service time with rate μ_i. The exogenous input rate to node i is $\rho_i = \lambda_0 Q_{0i}$ and each output from i is switched to j with probability Q_{ij} and switched out of the system with probability Q_{i0} (as in the text). Let $\lambda_i, 1 \leq i \leq k$, be the solution, for given λ_0, to

$$\lambda_j = \sum_{i=0}^{k} \}\lambda_i Q_{ij};$$

$1 \leq j \leq k$ and assume that $\lambda_i < s\mu_i$; $1 \leq i \leq k$. Show that the steady-state probability of state $\boldsymbol{m}$ is

$$\Pr\{\boldsymbol{m}\} = \prod_{i=1}^{k} \}p_i(m_i),$$

where $p_i(m_i)$ is the probability of state m_i in an (M, M, s) queue. Hint: simply extend the argument in the text to the multiple server case.

Exercise 6.25. Suppose a Markov process with the set of states A is reversible and has steady-state probabilities p_i; $i \in A$. Let B be a subset of A and assume that the process is changed by setting $q_{ij} = 0$ for all $i \in B, j \notin B$. Assuming that the new process (starting in B and remining in B) is irreducible, show that the new process is reversible and find its steady-state probabilities.

Exercise 6.26. Consider a queueing system with two classes of customers. Type A customer arrivals are Poisson with rate λ_A and Type B customer arrivals are Poisson with rate λ_B. The service time for type A customers is exponential with rate μ_A and that for type B is exponential with rate μ_B. Each service time is independent of all other service times and of all arrival epochs.

a) First assume there are infinitely many identical servers, and each new arrival immediately enters an idle server and begins service. Let the state of the system be (i, j) where i and j are the numbers of type A and B customers respectively in service. Draw a graph of the state transitions for $i \leq 2, j \leq 2$. Find the steady-state PMF, $\{p(i, j);\, i, j \geq 0\}$, for the Markov process. Hint: Note that the type A and type B customers do not interact.

b) Assume for the rest of the exercise that there is some finite number m of servers. Customers who arrive when all servers are occupied are turned away. Find the steady-state PMF, $\{p(i, j);\, i, j \geq 0, i + j \leq m\}$, in terms of $p(0, 0)$ for this Markov process. Hint: Combine part a) with the result of Exercise 6.25.

c) Let Q_n be the probability that there are n customers in service at some given time in steady state. Show that $Q_n = p(0, 0)\rho_n/n!$ for $0 \leq n \leq m$ where $\rho = \rho_A + \rho_B$, $\rho_A = \lambda_A/\mu_A$, and $\rho_B = \lambda_B/\mu_B$. Solve for $p(0, 0)$.

Exercise 6.27. a) Generalize Exercise 6.26 to the case in which there are K types of customers, each with independent Poisson arrivals and each with independent exponential service times. Let λ_k and μ_k be the arrival rate and service rate respectively for the k^h user type, $1 \leq k \leq K$. Let $\rho_k = \lambda_k/\mu_k$ and $\rho = \rho_1 + \rho_2 + \cdots + \rho_K$. In particular, show, as before, that the probability of n customers in the system is $Q_n = p(0, \ldots, 0)\rho_n/n!$ for $0 \leq n \leq m$.

b) View the customers in part a) as a single type of customer with Poisson arrivals of rate $\lambda = \sum_k \lambda_k$ and with a service density $\sum_k (\lambda_k/\lambda)\mu_k \exp(-\mu_k x)$. Show that the expected service time is ρ/λ. Note that what you have shown is that, if a service distribution can be represented as a weighted sum of exponentials, then the distribution of customers in the system is the same as for the M/M/m,m queue with equal mean service time.

Exercise 6.28. Consider a sampled-time M/D/m/m queueing system. The arrival process is Bernoulli with probability $\lambda << 1$ of arrival in each time unit. There are m servers; each arrival enters a server if a server is not busy and otherwise the arrival is discarded. If an arrival enters a server, it keeps the server busy for d units of time and then departs; d is some integer constant and is the same for each server.

Let n, $0 \leq n \leq m$ be the number of customers in service at a given time and let x_i be the number of time units that the i^{th} of those n customers (in order of arrival) has been

in service. Thus the state of the system can be taken as $(n, \boldsymbol{x}) = (n, x_1, x_2, \ldots, x_n)$ where $0 \leq n \leq m$ and $1 \leq x_1 < x_2 < \ldots < x_n \leq d$.

Let $A(n, \boldsymbol{x})$ denote the next state if the present state is $(n, \boldsymbol{x})$ and a new arrival enters service. That is,

$$A(n, \boldsymbol{x}) = (n+1, 1, x_1+1, x_2+1, \ldots, x_n+1) \quad \text{for } n < m \text{ and } x_n < d \quad (6.99)$$
$$A(n, \boldsymbol{x}) = (n, 1, x_1+1, x_2+1, \ldots, x_{n-1}+1) \quad \text{for } n \leq m \text{ and } x_n = d. \quad (6.100)$$

That is, the new customer receives one unit of service by the next state time, and all the old customers receive one additional unit of service. If the oldest customer has received d units of service, then it leaves the system by the next state time. Note that it is possible for a customer with d units of service at the present time to leave the system and be replaced by an arrival at the present time (*i.e.*, (6.100) with $n = m$, $x_n = d$). Let $B(n, \boldsymbol{x})$ denote the next state if either no arrival occurs or if a new arrival is discarded.

$$B(n, \boldsymbol{x}) = (n, x_1+1, x_2+1, \ldots, x_n+1) \quad \text{for } x_n < d \quad (6.101)$$
$$B(n, \boldsymbol{x}) = (n-1, x_1+1, x_2+1, \ldots, x_{n-1}+1) \quad \text{for } x_n = d. \quad (6.102)$$

a) Hypothesize that the *backward* chain for this system is also a sampled-time M/D/m/m queueing system, but that the state $(n, x_1, \ldots, x_n)(0 \leq n \leq m, 1 \leq x_1 < x_2 < \ldots < x_n \leq d)$ has a different interpretation: n is the number of customers as before, but x_i is now the remaining service required by customer i. Explain how this hypothesis leads to the following steady-state equations:

$$\lambda \pi_{n,\boldsymbol{x}} = (1-\lambda)\pi_{A(n,\boldsymbol{x})} \quad ; n < m, x_n < d \quad (6.103)$$
$$\lambda \pi_{n,\boldsymbol{x}} = \lambda \pi_{A(n,\boldsymbol{x})} \quad ; n \leq m, x_n = d \quad (6.104)$$
$$(1-\lambda)\pi_{n,\boldsymbol{x}} = \lambda \pi_{B(n,\boldsymbol{x})} \quad ; n \leq m, x_n = d \quad (6.105)$$
$$(1-\lambda)\pi_{n,\boldsymbol{x}} = (1-\lambda)\pi_{B(n,\boldsymbol{x})} \quad ; n \leq m, x_n < d. \quad (6.106)$$

b) Using this hypothesis, find $\pi_{n,\boldsymbol{x}}$ in terms of π_0, where π_0 is the probability of an empty system. Hint: Use (6.105) and (6.106); your answer should depend on n, but not $\boldsymbol{x}$.

c) Verify that the above hypothesis is correct.

d) Find an expression for π_0.

e) Find an expression for the steady-state probability that an arriving customer is discarded.

Exercise 6.29. A taxi alternates between three locations. When it reaches location 1 it is equally likely to go next to either 2 or 3. When it reaches 2 it will next go to 1 with probability 1/3 and to 3 with probability 2/3. From 3 it always goes to 1. The mean time between locations i and j are $t_{12} = 20$, $t_{13} = 30$, $t_{23} = 30$. Assume $t_{ij} = t_{ji}$).

Find the (limiting) probability that the taxi's most recent stop was at location i, $i = 1, 2, 3$?

What is the (limiting) probability that the taxi is heading for location 2?

What fraction of time is the taxi traveling from 2 to 3. Note: Upon arrival at a location the taxi immediately departs.

Exercise 6.30. (Semi-Markov continuation of Exercise 6.5) **a)** Assume that the Markov process in Exercise 6.5 is changed in the following way: whenever the process enters state 0, the time spent before leaving state 0 is now a *uniformly distributed* rv, taking values from 0 to $2/\lambda$. All other transitions remain the same. For this new process, determine whether the successive epochs of entry to state 0 form renewal epochs, whether the successive epochs of exit from state 0 form renewal epochs, and whether the successive entries to any other given state i form renewal epochs.

a) For each i, find both the time-average interval and the time-average number of overall state transitions between successive visits to i.

b) Is this modified process a Markov process in the sense that $\Pr\{X(t) = i \mid X(\tau) = j, X(s) = k\} = \Pr\{X(t) = i \mid X(\tau) = j\}$ for all $0 < s < \tau < t$ and all i, j, k? Explain.

Exercise 6.31. Consider an M/G/1 queueing system with Poisson arrivals of rate λ and expected service time $\mathsf{E}[X]$. Let $\rho = \lambda \mathsf{E}[X]$ and assume $\rho < 1$. Consider a semi-Markov process model of the M/G/1 queueing system in which transitions occur on departures from the queueing system and the state is the number of customers immediately following a departure.

a) Suppose a colleague has calculated the steady-state probabilities $\{p_i\}$ of being in state i for each $i \geq 0$. For each $i \geq 0$, find the steady-state probability pi_i of state i in the embedded Markov chain. Give your solution as a function of ρ, π_i, and p_0.

b) Calculate p_0 as a function of ρ.

c) Find π_i as a function of ρ and p_i.

d) Is p_i the same as the steady-state probability that the queueing system contains i customers at a given time? Explain carefully.

Exercise 6.32. Consider an M/G/1 queue in which the arrival rate is λ and the service time distributin is uniform $(0, 2W)$ with $\lambda W < 1$. Define a semi-Markov chain following the framework for the M/G/1 queue in Section 6.8.1.

a) Find $P_{0j}; j \geq 0$.

b) Find P_{ij} for $i > 0$; $j \geq i - 1$.

Exercise 6.33. Consider a semi-Markov process for which the embedded Markov chain is irreducible and positive-recurrent. Assume that the distribution of inter-renewal intervals for one state j is arithmetic with span d. Show that the distribution of inter-renewal intervals for all states is arithmetic with the same span.

Chapter 7

RANDOM WALKS, LARGE DEVIATIONS, AND MARTINGALES

7.1 Introduction

Definition 7.1.1. *Let $\{X_i;\, i \geq 1\}$ be a sequence of IID random variables, and let $S_n = X_1 + X_2 + \cdots + X_n$. The integer-time stochastic process $\{S_n;\, n \geq 1\}$ is called a random walk, or, more precisely, the one-dimensional random walk based on $\{X_i;\, i \geq 1\}$.*

For any given n, S_n is simply a sum of IID random variables, but here the behavior of the entire random walk *process*, $\{S_n;\, n \geq 1\}$, is of interest. Thus, for a given real number $\alpha > 0$, we might want to find the probability that the sequence $\{S_n;\ n \geq 1\}$ contains any term for which $S_n \geq \alpha$ (*i.e.*, that a threshold at α is crossed) or to find the distribution of the smallest n for which $S_n \geq \alpha$.

We know that S_n/n essentially tends to $\mathsf{E}[X] = \overline{X}$ as $n \to \infty$. Thus if $\overline{X} < 0$, S_n will tend to drift downward and if $\overline{X} > 0$, S_n will tend to drift upward. This means that the results to be obtained depend critically on whether $\overline{X} < 0$, $\overline{X} > 0$, or $\overline{X} = 0$. Since results for $\overline{X} > 0$ can be easily found from results for $\overline{X} < 0$ by considering $\{-S_n;\, n \geq 1\}$, we usually focus on the case $\overline{X} < 0$.

As one might expect, both the results and the techniques have a very different flavor when $\overline{X} = 0$, since here S_n/n essentially tends to 0 and we will see that the random walk typically wanders around in a rather aimless fashion.[1] With increasing n, σ_{S_n} increases as $\sqrt{n}$ (for X both zero-mean and non-zero-mean), and this is often called diffusion.[2]

[1] When $\overline{X}$ is very close to 0, its behavior for small n resembles that for $\overline{X} = 0$, but for large enough n the drift becomes significant, and this is reflected in the major results.

[2] If we view S_n as our winnings in a zero-mean game, the fact that $S_n/n \to 0$ makes it easy to imagine that a run of bad luck will probably be followed by a run of good luck. However, this is a fallacy here, since the X_n are assumed to be independent. Adjusting one's intuition to understand this at a gut level should be one of the reader's goals in this chapter.

The following three subsections discuss three special cases of random walks. The first two, simple random walks and integer random walks, will be useful throughout as examples, since they can be easily visualized and analyzed. The third special case is that of renewal processes, which we have already studied and which will provide additional insight into the general study of random walks.

After this, Sections 7.2 and 7.3 show how two major application areas, G/G/1 queues and hypothesis testing, can be viewed in terms of random walks. These sections also show why questions related to threshold crossings are so important in random walks.

Section 7.4 then develops the theory of threshold crossings for general random walks and Section 7.5 extends and in many ways simplifies these results through the use of stopping rules and a powerful generalization of Wald's equality known as Wald's identity.

The remainder of the chapter is devoted to a rather general type of stochastic process called martingales. The topic of martingales is both a subject of interest in its own right and also a tool that provides additional insight Rdensage into random walks, laws of large numbers, and other basic topics in probability and stochastic processes.

7.1.1 Simple random walks

Suppose $X_1, X_2, \ldots$ are IID binary random variables, each taking on the value 1 with probability p and -1 with probability $q = 1-p$. Letting $S_n = X_1 + \cdots + X_n$, the sequence of sums $\{S_n;\, n \geq 1\}$, is called a *simple random walk*. Note that S_n is the difference between positive and negative occurrences in the first n trials, and thus a simple random walk is little more than a notational variation on a Bernoulli process. For the Bernoulli process, X takes on the values 1 and 0, whereas for a simple random walk X takes on the values 1 and -1. For the random walk, if $X_m = 1$ for m out of n trials, then $S_n = 2m - n$, and

$$\Pr\{S_n = 2m - n\} = \frac{n!}{m!(n-m)!}\, p^m (1-p)^{n-m}. \tag{7.1}$$

This distribution allows us to answer questions about S_n for any given n, but it is not very helpful in answering such questions as the following: for any given integer $k > 0$, what is the probability that the sequence $S_1, S_2, \ldots$ ever reaches or exceeds k? This probability can be expressed as[3] $\Pr\{\bigcup_{n=1}^{\infty}\{S_n \geq k\}\}$ and is referred to as the probability that the random walk *crosses a threshold* at k. Exercise 7.1 demonstrates the surprisingly simple result that for a simple random walk with $p \leq 1/2$, this threshold crossing probability is

$$\Pr\left\{\bigcup_{n=1}^{\infty}\{S_n \geq k\}\right\} = \left(\frac{p}{1-p}\right)^k. \tag{7.2}$$

[3]This same probability is often expressed as as $\Pr\{\sup_{n=1} S_n \geq k\}$. For a general random walk, the event $\bigcup_{n\geq 1}\{S_n \geq k\}$ is slightly different from $\sup_{n\geq 1} S_n \geq k$, since $\sup_{n\geq 1} S_n \geq k$ can include sample sequences $s_1, s_2, \ldots$ in which a subsequence of values s_n approach k as a limit but never quite reach k. This is impossible for a simple random walk since all s_k must be integers. It is possible, but can be shown to have probability zero, for general random walks. It is simpler to avoid this unimportant issue by not using the sup notation to refer to threshold crossings.

Sections 7.4 and 7.5 treat this same question for general random walks, but the results are far less simple. They also treat questions such as the overshoot given a threshold crossing, the time at which the threshold is crossed given that it is crossed, and the probability of crossing such a positive threshold before crossing any given negative threshold.

7.1.2 Integer-valued random walks

Suppose next that $X_1, X_2, \ldots$ are arbitrary IID integer-valued random variables. We can again ask for the probability that such an integer-valued random walk crosses a threshold at k, *i.e.*, that the event $\bigcup_{n=1}^{\infty}\{S_n \geq k\}$ occurs, but the question is considerably harder than for simple random walks. Since this random walk takes on only integer values, it can be represented as a Markov chain with the set of integers forming the state space. In the Markov chain representation, threshold crossing problems are first passage-time problems. These problems can be attacked by the Markov chain tools we already know, but the special structure of the random walk provides new approaches and simplifications that will be explained in Sections 7.4 and 7.5.

7.1.3 Renewal processes as special cases of random walks

If $X_1, X_2, \ldots$ are IID positive random variables, then $\{S_n;\, n \geq 1\}$ is both a special case of a random walk and also the sequence of arrival epochs of a renewal counting process, $\{N(t);\, t > 0\}$. In this special case, the sequence $\{S_n;\, n \geq 1\}$ must eventually cross a threshold at any given positive value α, and the question of whether the threshold is ever crossed becomes uninteresting. However, the trial on which a threshold is crossed and the overshoot when it is crossed are familiar questions from the study of renewal theory. For the renewal counting process, $N(\alpha)$ is the largest n for which $S_n \leq \alpha$ and $N(\alpha)+1$ is the smallest n for which $S_n > \alpha$, *i.e.*, the smallest n for which the threshold at α is strictly exceeded. Thus the trial at which α is crossed is a central issue in renewal theory. Also the overshoot, which is $S_{N(\alpha)+1} - \alpha$ is familiar as the residual life at α.

Figure 7.1 illustrates the difference between general random walks and positive random walks, *i.e.*, renewal processes. Note that the renewal process in part b) is illustrated with the axes reversed from the conventional renewal process representation. We usually view each renewal epoch as a time (epoch) and view $N(\alpha)$ as the number of renewals up to time α, whereas with random walks, we usually view the number of trials as a discrete-time variable and view the sum of rv's as some kind of amplitude or cost. There is no mathematical difference between these viewpoints, and each viewpoint is often helpful.

7.2 The queueing delay in a G/G/1 queue:

Before analyzing random walks in general, we introduce two important problem areas that are often best viewed in terms of random walks. In this section, the queueing delay in a G/G/1 queue is represented as a threshold crossing problem in a random walk. In the

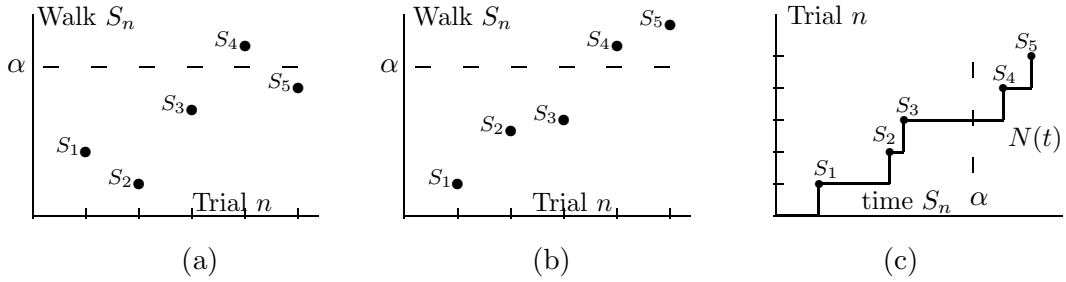


Figure 7.1: The sample function in (a) above illustrates a random walk $S_1, S_2, \dots$, with arbitrary (positive and negative) step sizes $\{X_i; i \geq 1\}$. The sample function in (b) illustrates a random walk, $S_1, S_2, \dots$, with only positive step sizes $\{X_i > 0; i \geq 1\}$. Thus, $S_1, S_2, \dots$, in (b) are sample renewal points in a renewal process. Note that the axes in (b) are reversed from the usual depiction of a renewal process. The usual depiction, illustrated in (c) for the same sample points, also shows the corresponding counting process. The random walks in parts a) and b) each illustrate a threshold at α, which in each case is crossed on trial 4 with an overshoot $S_4 - \alpha$.

next section, the error probability in a standard type of detection problem is represented as a random walk problem. This detection problem will then be generalized to a sequential detection problem based on threshold crossings in a random walk.

Consider a G/G/1 queue with first-come-first-serve (FCFS) service. We shall associate the probability that a customer must wait more than some given time α in the queue with the probability that a certain random walk crosses a threshold at α. Let $X_1, X_2, \dots$ be the interarrival times of a G/G/1 queueing system; thus these variables are IID with an arbitrary distribution function $\mathsf{F}_X(x) = \Pr\{X_i \leq x\}$. Assume that arrival 0 enters an empty system at time 0, and thus $S_n = X_1 + X_2 + \cdots + X_n$ is the epoch of the n^{th} arrival after time 0. Let $Y_0, Y_1, \dots$, be the service times of the successive customers. These are independent of $\{X_i; i \geq 1\}$ and are IID with some given distribution function $\mathsf{F}_Y(y)$. Figure 7.2 shows the arrivals and departures for an illustrative sample path of the process and illustrates the queueing delay for each arrival.

Let W_n be the queueing delay for the nth customer, $n \geq 1$. The system time for customer n is then defined as the queueing delay W_n plus the service time Y_n. As illustrated in Figure 7.2, customer $n \geq 1$ arrives X_n time units after the beginning of customer $n-1$'s system time. If $X_n < W_{n-1} + Y_{n-1}$, *i.e.*, if customer n arrives before the end of customer $n-1$'s system time, then customer n must wait in the queue until n finishes service (in the figure, for example, customer 2 arrives while customer 1 is still in the queue). Thus

$$W_n = W_{n-1} + Y_{n-1} - X_n \qquad \text{if } X_n \leq W_{n-1} + Y_{n-1}. \tag{7.3}$$

On the other hand, if $X_n > W_{n-1} + Y_{n-1}$, then customer $n-1$ (and all earlier customers) have departed when n arrives. Thus n starts service immediately and $W_n = 0$. This is the case for customer 3 in the figure. These two cases can be combined in the single equation

$$W_n = \max[W_{n-1} + Y_{n-1} - X_n,\ 0]; \quad \text{for } n \geq 1; \quad W_0 = 0. \tag{7.4}$$

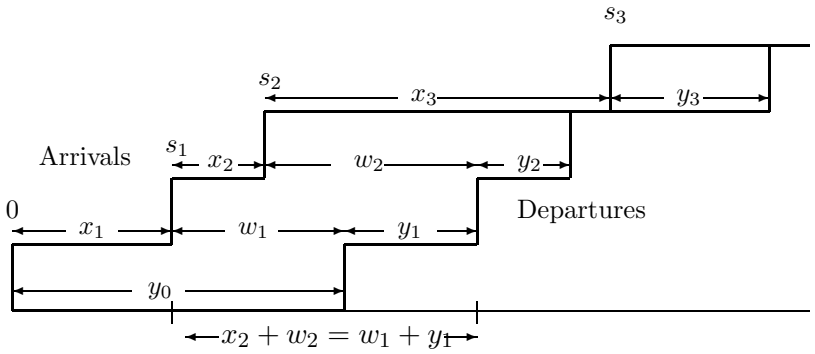


Figure 7.2: Sample path of arrivals and departures from a G/G/1 queue. Customer 0 arrives at time 0 and enters service immediately. Customer 1 arrives at time $s_1 = x_1$. For the case shown above, customer 0 has not yet departed, *i.e.*, $x_1 < y_0$, so customer 1's time in queue is $w_1 = y_0 - x_1$. As illustrated, customer 1's system time (queueing time plus service time) is $w_1 + y_1$.

Customer 2 arrives at $s_2 = x_1 + x_2$. For the case shown above, this is before customer 1 departs at $y_0 + y_1$. Thus, customer 2's wait in queue is $w_2 = y_0 + y_1 - x_1 - x_2$. As illustrated above, $x_2 + w_2$ is also equal to customer 1's system time, so $w_2 = w_1 + y_1 - x_2$. Customer 3 arrives when the system is empty, so it enters service immediately with no wait in queue, *i.e.*, $w_3 = 0$.

Since Y_{n-1} and X_n are coupled together in this equation for each n, it is convenient to define $U_n = Y_{n-1} - X_n$. Note that $\{U_n;\, n \geq 1\}$ is a sequence of IID random variables. From (7.4), $W_n = \max[W_{n-1} + U_n,\, 0]$, and iterating on this equation,

$$\begin{aligned} W_n &= \max[\max[W_{n-2}+U_{n-1},\, 0]+U_n,\, 0] \\ &= \max[(W_{n-2} + U_{n-1} + U_n),\, U_n,\, 0] \\ &= \max[(W_{n-3}+U_{n-2}+U_{n-1}+U_n),\ (U_{n-1}+U_n),\ U_n,\ 0] \\ &= \cdots \quad \cdots \\ &= \max[(U_1+U_2+\ldots+U_n),\ (U_2+U_3+\ldots+U_n),\ldots,(U_{n-1}+U_n),\ U_n,\ 0]. \end{aligned} \tag{7.5}$$

It is not necessary for the theorem below, but we can understand this maximization better by realizing that if the maximization is achieved at $U_i + U_{i+1} + \cdots + U_n$, then a busy period must start with the arrival of customer $i-1$ and continue at least through the service of customer n. To see this intuitively, note that the analysis above starts with the arrival of customer 0 to an empty system at time 0, but the choice of 0 time and customer number 0 has nothing to do with the analysis, and thus the analysis is valid for any arrival to an empty system. Choosing the largest customer number before n that starts a busy period must then give the correct queueing delay, and thus maximizes (7.5). Exercise 7.2 provides further insight into this maximization.

Define $Z_1^n = U_n$, define $Z_2^n = U_n + U_{n-1}$, and in general, for $i \leq n$, define $Z_i^n = U_n + U_{n-1} + \cdots + U_{n-i+1}$. Thus $Z_n^n = U_n + \cdots + U_1$. With these definitions, (7.5) becomes

$$W_n = \max[0, Z_1^n, Z_2^n, \ldots, Z_n^n]. \tag{7.6}$$

Note that the terms in $\{Z_i^n;\, 1 \leq i \leq n\}$ are the first n terms of a random walk, but it is not the random walk based on $U_1, U_2, \ldots$, but rather the random walk going backward,

starting with U_n. Note also that W_{n+1}, for example, is the maximum of a different set of variables, *i.e.*, it is the walk going backward from U_{n+1}. Fortunately, this doesn't matter for the analysis since the ordered variables $(U_n, U_{n-1} \dots, U_1)$ are statistically identical to $(U_1, \dots, U_n)$. The probability that the wait is greater than or equal to a given value α is

$$\Pr\{W_n \geq \alpha\} = \Pr\{\max[0, Z_1^n, Z_2^n, \dots, Z_n^n] \geq \alpha\}. \tag{7.7}$$

This says that, for the n^{th} customer, $\Pr\{W_n \geq \alpha\}$ is equal to the probability that the random walk $\{Z_i^n; 1 \leq i \leq n\}$ crosses a threshold at α by the n^{th} trial. Because of the initialization used in the analysis, we see that W_n is the queueing delay of the n^{th} arrival after the beginning of a busy period (although this n^{th} arrival might belong to a later busy period than that initial busy period).

As noted above, $(U_n, U_{n-1}, \dots, U_1)$ is statistically identical to $(U_1, \dots, U_n)$. This means that $\Pr\{W_n \geq \alpha\}$ is the same as the probability that the first n terms of a random walk based on $\{U_i; i \geq 1\}$ crosses a threshold at α. Since the first $n+1$ terms of this random walk provide one more opportunity to cross α than the first n terms, we see that

$$\cdots \leq \Pr\{W_n \geq \alpha\} \leq \Pr\{W_{n+1} \geq \alpha\} \leq \cdots \leq 1. \tag{7.8}$$

Since this sequence of probabilities is non-decreasing, it must have a limit as $n \to \infty$, and this limit is denoted $\Pr\{W \geq \alpha\}$. Mathematically,[4] this limit is the probability that a random walk based on $\{U_i; i \geq 1\}$ ever crosses a threshold at α. Physically, this limit is the probability that the queueing delay is at least α for any given very large-numbered customer (*i.e.*, for customer n when the influence of a busy period starting n customers earlier has died out). These results are summarized in the following theorem.

Theorem 7.2.1. *Let $\{X_i; i \geq 1\}$ be the IID interarrival intervals of a G/G/1 queue, let $\{Y_i; i \geq 0\}$ be the IID service times, and assume that the system is empty at time 0 when customer 0 arrives. Let W_n be the queueing delay for the n^{th} customer. Let $U_n = Y_{n-1} - X_n$ for $n \geq 1$ and let $Z_i^n = U_n + U_{n-1} + \cdots + U_{n-i+1}$ for $1 \leq i \leq n$. Then for every $\alpha > 0$, and $n \geq 1$, $W_n = \max[0, Z_1^n, Z_2^n, \dots, Z_n^n]$. Also, $\Pr\{W_n \geq \alpha\}$ is the probability that the random walk based on $\{U_i; i \geq 1\}$ crosses a threshold at α on or before the n^{th} trial. Finally, $\Pr\{W \geq \alpha\} = \lim_{n\to\infty} \Pr\{W_n \geq \alpha\}$ is equal to the probability that the random walk based on $\{U_i; i \geq 1\}$ ever crosses a threshold at α.*

Note that the theorem specifies the distribution function of W_n for each n, but says nothing about the joint distribution of successive queueing delays. These are not the same as the distribution of successive terms in a random walk because of the reversal of terms above.

We shall find a relatively simple upper bound and approximation to the probability that a random walk crosses a positive threshold in Section 7.4. From Theorem 7.2.1, this can be applied to the distribution of queueing delay for the G/G/1 queue (and thus also to the M/G/1 and M/M/1 queues).

[4] More precisely, the sequence of queueing delays $W_1, W_2 \dots$, converge in distribution to W, *i.e.*, $\lim_n \mathsf{F}_{W_n}(w) = \mathsf{F}_W(w)$ for each w. We refer to W as the queueing delay in steady state.

7.3 Detection, decisions, and hypothesis testing

Consider a situation in which we make n noisy observations of the outcome of a single discrete random variable H and then guess, on the basis of the observations alone, which sample value of H occurred. In communication technology, this is called a *detection problem*. It models, for example, the situation in which a symbol is transmitted over a communication channel and n noisy observations are received. It similarly models the problem of detecting whether or not a target is present in a radar observation. In control theory, such situations are usually referred to as *decision problems*, whereas in statistics, they are referred to both as *hypothesis testing* and *statistical inference problems*.

The above communication, control, and statistical problems are basically the same, so we discuss all of them using the terminology of hypothesis testing. Thus the sample values of the rv H are called hypotheses. We assume throughout this section that H has only two possible values. Situations where H has more than 2 values can usually be viewed in terms of multiple binary decisions.

It makes no difference in the analysis of binary hypothesis testing what the two hypotheses happen to be called, so we simply denote them as hypothesis 0 and hypothesis 1. Thus H is a binary rv and we abbreviate its PMF as $\mathsf{p}_H(0) = p_0$ and $\mathsf{p}_H(1) = p_1$. Thus p_0 and $p_1 = 1 - p_0$ are the *a priori* probabilities[5] for the two sample values (hypotheses) of the random variable H.

Let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ be the n observations. Suppose, for initial simplicity, that these variables, conditional on $H = 0$ or $H = 1$, have joint conditional densities, $\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0)$ and $\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 1)$, that are strictly positive over their common region of definition. The case of primary importance in this section (and the case where random walks are useful) is that in which, conditional on a given sample value of H, the rv's $Y_1, \ldots, Y_n$ are IID. Still assuming for simplicity that the observations have densities, say $\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid \ell)$ for $\ell = 0$ or $\ell = 1$, the joint density, conditional on $H = \ell$, is given by

$$\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid \ell) = \prod_{i=1}^{n} \mathsf{f}_{Y|H}(y_i \mid \ell). \tag{7.9}$$

Note that all components of $\boldsymbol{Y}$ are conditioned on the same H, *i.e.*, for a single sample value of H, there are n corresponding sample values of Y.

[5]Statisticians have argued since the earliest days of statistics about the 'validity' of choosing *a priori* probabilities for the hypotheses to be tested. Bayesian statisticians are comfortable with this practice and non-Bayesians are not. Both are comfortable with choosing a probability model for the observations conditional on each hypothesis. We take a Bayesian approach here, partly to take advantage of the power of a complete probability model, and partly because non-Bayesian results, *i.e.*, results that do not depend on the *a priori* probabilies, are often easier to derive and interpret within a collection of probability models using different choices for the *a priori* probabilities. As will be seen, the Bayesian approach also makes it natural to incorporate the results of earlier observations into updated *a priori* probabilities for analyzing later observations. In defense of non-Bayesians, note that the results of statistical tests are often used in areas of significant public policy disagreement, and it is important to give the appearance of a lack of bias. Statistical results can be biased in many more subtle ways than the use of *a priori* probabilities, but the use of an *a priori* distribution is particularly easy to attack.

Assume that it is necessary[6] to select a sample value for H on the basis of the sample value $\boldsymbol{y}$ of the n observations. It is usually possible for the selection (decision) to be incorrect, so we distinguish the sample value h of the actual experimental outcome from the sample value $\hat{h}$ of our selection. That is, the outcome of the experiment specifies h and $\boldsymbol{y}$, but only $\boldsymbol{y}$ is observed. The selection $\hat{h}$, based only on $\boldsymbol{y}$, might be unequal to h. We say that an error has occurred if $\hat{h} \neq h$

We now analyze both how to make decisions and how to evaluate the resulting probability of error. Given a particular sample of n observations $\boldsymbol{y} = y_1, y_2, \ldots, y_n$, we can evaluate $\Pr\{H{=}0 \mid \boldsymbol{y}\}$ as

$$\Pr\{H{=}0 \mid \boldsymbol{y}\} = \frac{p_0 f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0)}{p_0 f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0) + p_1 f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 1)}. \tag{7.10}$$

We can evaluate $\Pr\{H{=}1 \mid \boldsymbol{y}\}$ in the same way The ratio of these quantities is given by

$$\frac{\Pr\{H{=}0 \mid \boldsymbol{y}\}}{\Pr\{H{=}1 \mid \boldsymbol{y}\}} = \frac{p_0 f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0)}{p_1 f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 1)}. \tag{7.11}$$

For a given $\boldsymbol{y}$, if we select $\hat{h} = 0$, then an error is made if $h = 1$. The probability of this event is $\Pr\{H = 1 \mid \boldsymbol{y}\}$. Similarly, if we select $\hat{h} = 1$, the probability of error is $\Pr\{H = 0 \mid \boldsymbol{y}\}$. Thus the probability of error is minimized, for a given $\boldsymbol{y}$, by selecting $\hat{h}{=}0$ if the ratio in (7.11) is greater than 1 and selecting $\hat{h}{=}1$ otherwise. If the ratio is equal to 1, the error probability is the same whether $\hat{h}{=}0$ or $\hat{h}{=}1$ is selected, so we arbitrarily select $\hat{h} = 1$ to make the rule deterministic.

The above rule for choosing $\hat{H}{=}0$ or $\hat{H}{=}1$ is called the *Maximum a posteriori probability* detection rule, abbreviated as the MAP rule. Since it minimizes the error probability for each $\boldsymbol{y}$, it also minimizes the overall error probability as an expectation over $\boldsymbol{Y}$ and H.

The structure of the MAP test becomes clearer if we define the *likelihood ratio* $\Lambda(\boldsymbol{y})$ for the binary decision as

$$\Lambda(\boldsymbol{y}) = \frac{f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0)}{f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 1)}.$$

The likelihood ratio is a function only of the observation $\boldsymbol{y}$ and does not depend on the *a priori* probabilities. There is a corresponding rv $\Lambda(\boldsymbol{Y})$ that is a function[7] of $\boldsymbol{Y}$, but we will tend to deal with the sample values. The MAP test can now be compactly stated as

$$\Lambda(\boldsymbol{y}) \begin{cases} > p_1/p_0 \quad ; & \text{select } \hat{h}{=}0 \\ \leq p_1/p_0 \quad ; & \text{select } \hat{h}{=}1. \end{cases} \tag{7.12}$$

[6] In a sense, it is in making decisions where the rubber meets the road. For real-world situations, a probabilistic analysis is often connected with studying the problem, but eventually one technology or another must be chosen for a system, one person or another must be hired for a job, one law or another must be passed, etc.

[7] Note that in a non-Bayesian formulation, $\boldsymbol{Y}$ is a different random vector, in a different probability space, for $H = 0$ than for $H = 1$. However, the corresponding conditional probabilities for each exist in each probability space, and thus $\Lambda(\boldsymbol{Y})$ exists (as a different rv) in each space.

For the primary case of interest, where the n observations are IID conditional on H, the likelihood ratio is given by

$$\Lambda(\boldsymbol{y}) = \prod_{i=1}^{n} \frac{f_{Y|H}(y_i \mid 0)}{f_{Y|H}(y_i \mid 1)}.$$

The MAP test then takes on a more attractive form if we take the logarithm of each side in (7.12). The logarithm of $\Lambda(\boldsymbol{y})$ is then a sum of n terms, $\sum_{i=1}^{n} z_i$, where z_i is given by

$$z_i = \ln \frac{f_{Y|H}(y_i \mid 0)}{f_{Y|H}(y_i \mid 1)}.$$

The sample value z_i is called the log-likelihood ratio of y_i for each i, and $\ln \Lambda(\boldsymbol{y})$ is called the log-likelihood ratio of $\boldsymbol{y}$. The test in (7.12) is then expressed as

$$\sum_{i=1}^{n} z_i \quad \begin{cases} > \ln(p_1/p_0) \quad ; & \text{select } \hat{h}{=}0 \\ \leq \ln(p_1/p_0) \quad ; & \text{select } \hat{h}{=}1. \end{cases} \tag{7.13}$$

As one might imagine, log-likelihood ratios are far more convenient to work with than likelihood ratios when dealing with conditionally IID rv's.

Example 7.3.1. Consider a Poisson process for which the arrival rate λ might be λ_0 or might be λ_1. Assume, for $i \in \{0, 1\}$, that p_i is the *a priori* probability that the rate is λ_i. Suppose we observe the first n arrivals, $Y_1, \ldots Y_n$ and make a MAP decision about the arrival rate from the sample values $y_1, \ldots, y_n$.

The conditional probability densities for the observations $Y_1, \ldots, Y_n$ are given by

$$f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid i) = \prod_{i=1}^{n} \lambda_i e^{-\lambda_i y_i}.$$

The test in (7.13) then becomes

$$n \ln(\lambda_0/\lambda_1) + \sum_{i=1}^{n} (\lambda_1 - \lambda_0) y_i \quad \begin{cases} > \ln(p_1/p_0) \quad ; & \text{select } \hat{h}{=}0 \\ \leq \ln(p_1/p_0) \quad ; & \text{select } \hat{h}{=}1. \end{cases}$$

Note that the test depends on $\boldsymbol{y}$ only through the time of the nth arrival. This should not be surprising, since we know that, under each hypothesis, the first $n-1$ arrivals are uniformly distributed conditional on the nth arrival time.

The MAP test in (7.12), and the special case in (7.13), are examples of *threshold tests*. That is, in (7.12), a decision is made by calculating the likelihood ratio and comparing it to a threshold $\eta = p_1/p_0$. In (7.13), the log-likelihood ratio $\ln(\Lambda(\boldsymbol{y}))$ is compared with $\ln \eta = \ln(p_1/p_0)$.

There are a number of other formulations of hypothesis testing that also lead to threshold tests, although in these alternative formulations, the threshold $\eta > 0$ need not be equal to

p_1/p_0, and indeed *a priori* probabilities need not even be defined. In particular, then, the threshold test at η is defined by

$$\Lambda(\boldsymbol{y}) \begin{cases} > \eta \quad ; & \text{select } \hat{h}{=}0 \\ \leq \eta \quad ; & \text{select } \hat{h}{=}1. \end{cases} \tag{7.14}$$

For example, *maximum likelihood* (ML) detection selects hypothesis $\hat{h} = 0$ if $f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0) > f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 1)$, and selects $\hat{h} = 1$ otherwise. Thus the ML test is a threshold test with $\eta = 1$. Note that ML detection is equivalent to MAP detection with equal *a priori* probabilities, but it can be used in many other cases, including those with undefined *a priori* probabilities.

In many detection situations there are unequal costs, say C_0 and C_1, associated with the two kinds of errors. For example one kind of error in a medical prognosis could lead to serious illness and the other to an unneeded operation. A *minimum cost* decision could then minimize the expected cost over the two types of errors. As shown in Exercise 7.5, this is also a threshold test with the threshold $\eta = (C_1 p_1)/(C_0 p_0)$. This example also illustrates that, although assigning costs to errors provides a rational approach to decision making, there might be no widely agreeable way to assign costs.

Finally, consider the situation in which one kind of error, say $\Pr\{\mathsf{e} \mid H{=}1\}$ is upper bounded by some tolerable limit α and $\Pr\{\mathsf{e} \mid H{=}0\}$ is minimized subject to this constraint. The solution to this is called the *Neyman-Pearson rule*. The Neyman-Pearson rule is of particular interest since it does not require any assumptions about *a priori* probabilities. The next subsection shows that the Neyman-Pearson rule is essentially a threshold test, and explains why one rarely looks at tests other than threshold tests.

7.3.1 The error curve and the Neyman-Pearson rule

Any test, *i.e.*, any deterministic rule for selecting a binary hypothesis from an observation $\boldsymbol{y}$, can be viewed as a function[8] mapping each possible observation $\boldsymbol{y}$ to 0 or 1. If we define A as the set of n-vectors $\boldsymbol{y}$ that are mapped to hypothesis 1 for a given test, then the test can be identified by its corresponding set A.

Given the test A, the error probabilities, given $H = 0$ and $H = 1$ respectively, are given by

$$\Pr\{\boldsymbol{Y} \in A \mid H = 0\}; \qquad \Pr\{\boldsymbol{Y} \in A^c \mid H = 1\}.$$

Note that these conditional error probabilities depend only on the test A and not on the *a priori* probabilities. We will abbreviate these error probabilities as

$$q_0(A) = \Pr\{\boldsymbol{Y} \in A \mid H = 0\}; \qquad q_1(A) = \Pr\{\boldsymbol{Y} \in A^c \mid H = 1\}.$$

For given *a priori* probabilities, p_0 and p_1, the overall error probability is

$$\Pr\{\mathsf{e}(A)\} = p_0 q_0(A) + p_1 q_1(A).$$

[8]By assumption, the decision must be made on the basis of the observation $\boldsymbol{y}$, so a deterministic rule is based solely on $\boldsymbol{y}$, *i.e.*, is a function of $\boldsymbol{y}$. We will see later that randomized rather than deterministic rules are required for some purposes, but such rules are called randomized rules rather than tests.

If A is a threshold test, with threshold η, the set A is given by

$$A = \left\{ \boldsymbol{y} : \frac{\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0)}{\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 1)} \leq \eta \right\}.$$

Since threshold tests play a very special role here, we abuse the notation by using η in place of A to refer to a threshold test at η. We can now characterize the relationship between threshold tests and other tests. The following lemma is illustrated in Figure 7.3.

Lemma 7.3.1. *Consider a two dimensional plot in which the error probabilities for each test A are plotted as $(q_1(A), q_0(A))$. Then for any theshold test η, $0 < \eta < \infty$, and any A, the point $(q_1(A), q_0(A))$ is on the closed half plane above and to the right of a straight line of slope $-\eta$ passing through the point $(q_1(\eta), q_0(\eta))$.*

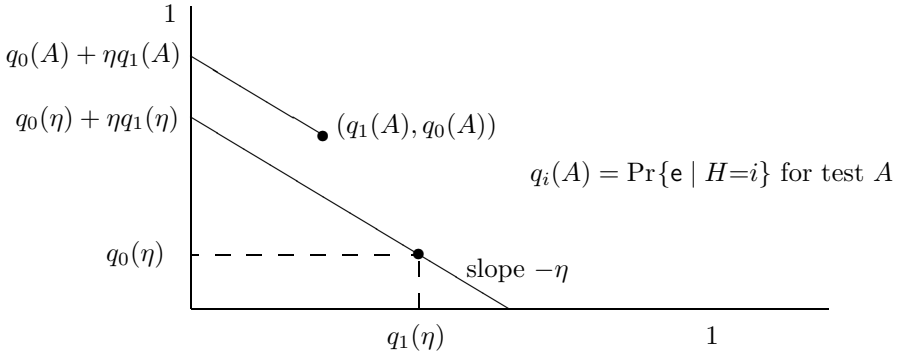


Figure 7.3: Illustration of Lemma 7.3.1

Proof: For any given η, consider the *a priori* probabilities[9] (p_0, p_1) for which $\eta = p_1/p_0$. The overall error probability for test A using these *a priori* probabilities is then

$$\Pr\{\mathsf{e}(A)\} = p_0 q_0(A) + p_1 q_1(A) = p_0 \big[q_0(A) + \eta q_1(A) \big].$$

Similarly, the overall error probability for the threshold test η using the same *a priori* probabilities is

$$\Pr\{\mathsf{e}(\eta)\} = p_0 q_0(\eta) + p_1 q_1(\eta) = p_0 \left[q_0(\eta) + \eta q_1(\eta) \right].$$

This latter error probability is the MAP error probability for the given p_0, p_1, and is thus the minimum overall error probability (for the given p_0, p_1) over all tests. Thus

$$q_0(\eta) + \eta q_1(\eta) \leq q_0(A) + \eta q_1(A).$$

As shown in the figure, these are the points at which the lines of slope $-\eta$ from $(q_1(A), q_0(A))$ and $(q_1(\eta), q_0(\eta))$ respectively cross the ordinate axis. Thus all points on the first line

[9]Note that the lemma does not assume any *a priori* probabilities. The MAP test for each *a priori* choice (p_0, p_1) determines $q_0(\eta)$ and $q_1(\eta)$ for $\eta = p_1/p_0$. Its interpretation as a MAP test depends on a model with *a priori* probabilities, but its calculation depends only on p_0, p_1 viewed as parameters.

(including $(q_1(A), q_0(A))$) lie in the closed half plane above and to the right of all points on the second, completing the proof. □

The straight line of slope $-\eta$ through the point $(q_1(\eta), q_0(\eta))$ has the equation $f_\eta(\alpha) = q_0(\eta) + \eta(q_1(\eta) - \alpha)$. Since the lemma is valid for all η, $0 < \eta < \infty$, the point $(q_1(A), q_0(A))$ for an arbitrary test lies above and to the right of the entire family of straight lines that, for each η, pass through $(q_1(\eta), q_0\eta))$ with slope $-\eta$. This family of straight lines has an upper envelope called the error curve, $u(\alpha)$, defined by

$$u(\alpha) = \sup_{0 \le \eta < \infty} q_0(\eta) + \eta(q_1(\eta) - \alpha). \tag{7.15}$$

The lemma then asserts that for every test A (including threshold tests), we have $u(q_1(A)) \le q_0(A)$. Also, since every threshold test lies on one of these straight lines, and therefore on or below the curve $u(\alpha)$, we see that the error probabilities $(q_1(\eta), q_0(\eta)$ for each threshold test must lie on the curve $u(\alpha)$. Finally, since each straight line defining $u(\alpha)$ forms a tangent of $u(\alpha)$ and lies on or below $u(\alpha)$, the function $u(\alpha)$ is convex.[10] Figure 7.4 illustrates the error curve.

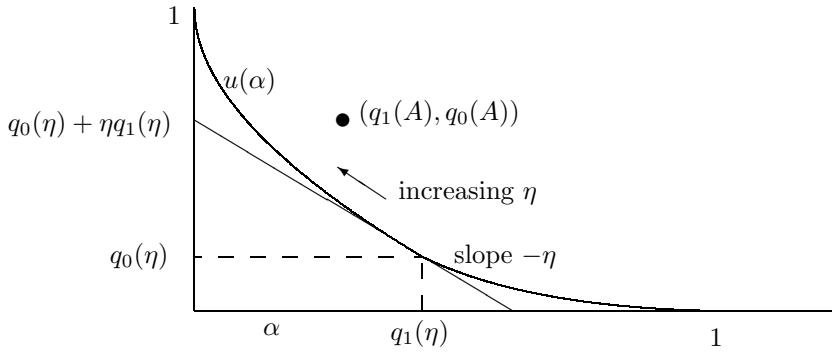


Figure 7.4: Illustration of the error curve $u(\alpha)$ (see (7.15)). Note that $u(\alpha)$ is convex, lies on or above its tangents, and on or below all tests. It can also be seen, either directly from the curve above or from the definition of a threshold test, that $q_1(\eta)$ is non-increasing in η and $q_0(\eta)$ is non-decreasing.

The error curve essentially gives us a tradeoff between the probability of error given $H = 0$ and that given $H = 1$. Threshold tests, since they lie on the error curve, provide optimal points for this tradeoff.

The argument above lacks one important feature. Although all threshold tests lie on the error curve, we have not shown that all points on the error curve correspond to threshold tests. In fact, as the following example shows, it is sometimes necessary to generalize threshold tests by randomization in order to reach all points on the error curve.

Before proceeding to this example, note that Lemma 7.3.1 and the definition of the error curve apply to a broader set of models than discussed so far. First, the lemma still holds if

[10] A convex function of a real variable is sometimes defined as a function with a nonnegative second derivative, but defining it as a function lying on or above all its tangents is more general, allowing for step discontinuities in the first derivative of f. We see the need for this generality shortly.

$\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid \ell)$ is zero over an arbitrary set of $\boldsymbol{y}$ for one or both hypotheses ℓ. The likelihood ratio $\Lambda(\boldsymbol{y})$ is infinite where $\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0) > 0$ and $\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 1) = 0$, but this does not affect the proof of the lemma. Exercise 7.6 helps explain how this situation can affect the error curve.

In addition, it can be seen that the lemma also holds if $\boldsymbol{Y}$ is an n-tuple of discrete rv's. The following example further explains the discrete case and also shows why not all points on the error curve correspond to threshold tests. What we assume throughout is that $\Lambda(\boldsymbol{Y})$ is a rv conditional on $H = 1$ and is a possibly-defective[11] rv conditional on $H = 0$.

Example 7.3.2. About the simplest example of a detection problem is that with a one-dimensional binary observation Y. Assume then that

$$\mathsf{p}_{Y|H}(0 \mid 0) = \mathsf{p}_{Y|H}(1 \mid 1) = \frac{2}{3}; \qquad \mathsf{p}_{Y|H}(0 \mid 1) = \mathsf{p}_{Y|H}(1 \mid 0) = \frac{1}{3}.$$

Thus the observation Y equals H with probability 2/3. The 'natural' decision rule would be to select $\hat{h}$ to agree with the observation y, thus making an error with probability 1/3, both conditional on $H = 1$ and $H = 0$. It is easy to verify from (7.12) (using PMF's in place of densities) that this 'natural' rule is the MAP rule if the *a priori* probabilities are in the range $1/3 \leq p_1 < 2/3$.

For $p_1 < 1/3$, the MAP rule can be seen to be $\hat{h} = 0$, no matter what the observation is. Intuitively, hypothesis 1 is too unlikely *a priori* to be overcome by the evidence when $Y = 1$. Similarly, if $p_1 \geq 2/3$, then the MAP rule selects $\hat{h} = 1$, independent of the observation.

The corresponding threshold test (see (7.14)) selects $\hat{h} = y$ for $1/2 \leq \eta < 2$. It selects $\hat{h} = 0$ for $\eta < 1/2$ and $\hat{h} = 1$ for $\eta \geq 2$. This means that, although there is a threshold test for each η, there are only 3 resulting error probability pairs, *i.e.*, $(q_1(\eta), q_0(\eta)$ can be (0, 1) or (1/3, 1/3), or (1, 0). The first pair holds for $\eta \geq 2$, the second for $1/2 \leq \eta < 2$, and the third for $\eta < 1/2$. This is illustrated in Figure 7.5.

We have just seen that there is a threshold test for each η, $0 < \eta < \infty$, but those threshold tests map to only 3 distinct points $(q_1(\eta), q_0(\eta))$. As can be seen, the error curve joins these 3 points by straight lines.

Let us look more carefully at the tangent of slope -1/2 through the point (1/3, 1/3. This corresponds to the MAP test at $\mathsf{p}_1 = 1/3$. As seen in (7.12), this MAP test selects $\hat{h} = 1$ for $y = 1$ and $\hat{h} = 0$ for $y = 0$. The selection of $\hat{h} = 1$ when $y = 1$ is a don't-care choice in which selecting $\hat{h} = 0$ would yield the same overall error probability, but would change $(q_1(\eta), q_0(\eta)$ from (1/3, 1/3) to (1, 0).

It is not hard to verify (since there are only 4 tests, *i.e.*, deterministic rules, for mapping a binary variable to another binary variable) that no test can achieve the tradeoff between $q_1(A)$ and $q_0(A)$ indicated by the interior points on the straight line between (1/3, 1/3) and (1, 0). However, if we use a randomized rule, mapping 1 to 1 with probability θ and into 0 with probability $1 - \theta$ (along with always mappng 0 to 0), then all points on the straight

[11] If $\mathsf{p}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 1) = 0$ and $\mathsf{p}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0) > 0$ for some $\boldsymbol{y}$, then $\Lambda(\boldsymbol{y})$ is infinite, and thus defective, conditional on $H = 0$. Since the given $\boldsymbol{y}$ has zero probability conditional on $H = 0$, we see that $\Lambda(\boldsymbol{Y})$ is not defective conditional on $H = 1$.

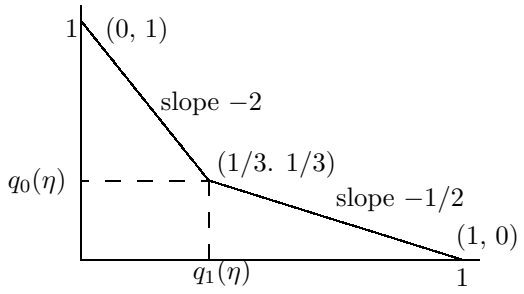


Figure 7.5: Illustration of the error curve $u(\alpha)$ for Example 7.3.2. For all η in the range $1/2 \leq \eta < 2$, the threshold test selects $\hat{h} = y$. The corresponding error probabilities are then $q_1(\eta) = q_0(\eta) = 1/3$. For $\eta < 1/2$, the threshold test selects $\hat{h} = 0$ for all y, and for $\eta > 2$, it selects $\hat{h} = 1$ for all y. The error curve (see (7.15)) for points to the right of (1/3, 1/3) is maximized by the straight line of slope -1/2 through (1/3, 1/3). Similarly, the error curve for points to the left of (1/3, 1/3) is maximized by the straight line of slope -2 through (1/3, 1/3). One can visualize the tangent lines as an inverted see-saw, first see-sawing around (0,1), then around (1/3, 1/3), and finally around (1, 0).

line from (1/3, 1/3) to (1, 0) are achieved as θ goes from 0 to 1. In other words, a don't-care choice for MAP becomes an important choice in the tradeoff between $q_1(A)$ and $q_0(A)$.

In the same way, all points on the straight line from (0, 1) to (1/3, 1/3) can be achieved by a randomized rule that maps 0 to 0 with given probability θ (along with always mapping 1 to 1).

In the general case, the error curve can contain straight line segments whenever the distribution function of the likelihood ratio, conditional on $H = 1$, is discontinuous. This is always the case for discrete observations, and, as illustrated in Exercise 7.6, might also occur with continuous observations. To understand this, suppose that $\Pr\{\Lambda(\boldsymbol{Y}) = \eta \mid H{=}1\} = \beta > 0$ for some $\eta,\ 0 < \eta < \infty$. Then the MAP test at η has a don't-care region of probability β given $H = 1$. This means that if the MAP test is changed to resolve the don't-care case in favor of $H = 0$, then the error probability q_1 is increased by β and the error probability q_0 is decreased by $\eta\beta$. Lemma 7.3.1 is easily seen to be valid whichever way the don't-care cases are resolved, and thus both (q_1, q_0) and $(q_1{+}\beta,\ q_0{-}\eta\beta)$ lie on the error curve. Since all tests lie above and to the right of the straight line of slope $-\eta$ through these points, the error curve has a straight line segment between these points. As mentioned before, any pair of error probabilities on this straight line segment can be realized by using a randomized threshold test at η.

The Neyman-Pearson rule is the rule (randomized where needed) that realizes any desired error probability pair (q_1, q_0) on the error curve, *i.e.*, where $q_0 = u(q_1)$. To be specific, assume the constraint that $q_1 = \alpha$ for any given α, $0 < \alpha \leq 1$. Since $\Pr\{\Lambda(\boldsymbol{Y}) > \eta \mid H = 1\}$ is a complementary distribution function, it is non-increasing in η, perhaps with step discontinuities. At $\eta = 0$, it has the value $1 - \Pr\{\Lambda(\boldsymbol{Y}) = 0\}$. As η increases, it decreases to 0, either for some finite η or as $\eta \to \infty$. Thus if $0 < \alpha \leq 1 - \Pr\{\Lambda(\boldsymbol{Y}) = 0\}$, we see that α is either equal to $\Pr\{\Lambda(\boldsymbol{y}) > \eta \mid H = 1\}$ for some η or α is on a step discontinuity at some

η. Defining $\eta(\alpha)$ as an η where one of these alternatives occur,[12] there is a solution to the following:

$$\Pr\{\Lambda(\boldsymbol{y})>\eta(\alpha) \mid H=1\} \leq \alpha \leq \Pr\{\Lambda(\boldsymbol{y})\geq\eta(\alpha) \mid H=1\}. \tag{7.16}$$

Note that (7.16) also holds, with $\eta(\alpha) = 0$ if $1 - \Pr\{\Lambda(\boldsymbol{Y})=0\} < \alpha \leq 1$.

The Neyman-Pearson rule, given α, is to use a threshold test at $\eta(\alpha)$ if $\Pr\{\Lambda(\boldsymbol{y})=\eta(\alpha) \mid H=1\} = 0$. If $\Pr\{\Lambda(\boldsymbol{y})=\eta(\alpha) \mid H=1\} > 0$, then a randomized test is used at $\eta(\alpha)$. When $\Lambda(\boldsymbol{y}) = \eta(\alpha)$, $\hat{h}$ is chosen to be 1 with probability θ where

$$\theta = \frac{\alpha - \Pr\{\Lambda(\boldsymbol{y}) > \eta(\alpha) \mid H = 1\}}{\Pr\{\Lambda(\boldsymbol{y}) = \eta(\alpha) \mid H = 1\}}. \tag{7.17}$$

This is summarized in the following theorem:

Theorem 7.3.1. *Assume that the likelihood ratio $\Lambda(\boldsymbol{Y})$ is a rv under $H = 1$. For any α, $0 < \alpha \leq 1$, the constraint that $\Pr\{\mathsf{e} \mid H = 1\} \leq \alpha$ implies that $\Pr\{\mathsf{e} \mid H = 0\} \geq u(\alpha)$ where $u(\alpha)$ is the error curve. Furthermore, $\Pr\{\mathsf{e} \mid H = 0\} = u(\alpha)$ if the Neyman-Pearson rule, specified in (7.16) and (7.17), is used.*

Proof: We have shown that the Neyman-Pearson rule has the stipulated error probabilities. An arbitrary decision rule, randomized or not, can be specified by a finite set $A_1, \ldots, A_k$ of deterministic decision rules along with a rv V, independent of H and $\boldsymbol{Y}$, with sample values $1, \ldots, k$. Letting the PMF of V be $(\theta_1, \ldots, \theta_k)$, an arbitrary decision rule, given $\boldsymbol{Y} = \boldsymbol{y}$ and $V = i$ is to use decision rule A_i if $V = i$. The error probabilities for this rule are $\Pr\{\mathsf{e} \mid H=1\} = \sum_{i=1}^{k} \theta_i q_1(A_i)$ and $\Pr\{\mathsf{e} \mid H=0\} = \sum_{i=1}^{k} \theta_i q_0(A_i)$. It is easy to see that Lemma 7.3.1 applies to these randomized rules in the same way as to deterministic rules. □

All of the decision rules we have discussed are threshold rules, and in all such rules, the first part of the decision rule is to find the likelihood ratio $\Lambda(\boldsymbol{y})$ from the observation $\boldsymbol{y}$. This simplifies the hypothesis testing problem from processing an n-dimensional vector $\boldsymbol{y}$ to processing a single number $\Lambda(\boldsymbol{y})$, and this is typically the major component of making a threshold decision. Any intermediate calculation from $\boldsymbol{y}$ that allows $\Lambda(\boldsymbol{y})$ to be calculated is called a *sufficient statistic*. These usually play an important role in detection, especially for detection in noisy environments.

There are some hypothesis testing problems in which threshold rules are in a sense inappropriate. In particular, the cost of an error under one or the other hypothesis could be highly dependent on the observation $\boldsymbol{y}$. A minimum cost threshold test for each $\boldsymbol{y}$ would still be appropriate, but a threshold test based on the likelihood ratio might be inappropriate since different observations with very different cost structures could have the same likelihood ratio. In other words, in such situations, the likelihood ratio is no longer sufficient to make a minimum cost decision.

[12] In the discrete case, there can be multiple solutions to (7.16) for some values of α, but this does not affect the decision rule. One can choose the smallest value of η to satisfy (7.16) if one wants to eliminate this ambiguity.

So far we have assumed that a decision is made after a predetermined number n of observations. In many situations, observations are costly and introduce an unwanted delay in decisions. One would prefer, after a given number of observations, to make a decision if the resulting probability of error is small enough, and to continue with more observations otherwise. Common sense dictates such a strategy, and the branch of probability theory analyzing such strategies is called *sequential analysis*. In a sense, this is a generalization of Neyman-Pearson tests, which employed a tradeoff between the two types of errors. Here we will have a three-way tradeoff between the two types of errors and the time required to make a decision.

We now return to the hypothesis testing problem of major interest where the observations, conditional on the hypothesis, are IID. In this case the log-likelihood ratio after n observations is the sum $S_n = Z_1 + \cdots Z_n$ of the n IID individual log likelihood ratios. Thus, aside from the question of making a decision after some number of observations, the sequence of log-likelihood ratios $S_1, S_2, \ldots$ is a random walk.

Essentially, we will see that an appropriate way to choose how many observations to make, based on the result of the earlier observations, is as follows: The probability of error under either hypothesis is based on $S_n = Z_1 + \cdots + Z_n$. Thus we will see that an appropriate rule is to choose H=0 if the sample value s_n of S_n is more than some positive threshold α, to choose H_1 if $s_n \leq \beta$ for some negative threshold β, and to continue testing if the sample value has not exceeded either threshold. In other words, the decision time for these sequential decision problems is the time at which the corresponding random walk crosses a threshold. An error is made when the wrong threshold is crossed.

We have now seen that both the queueing delay in a G/G/1 queue and the above sequential hypothesis testing problem can be viewed as threshold crossing problems for random walks. The following section begins a systemic study of such problems.

7.4 Threshold crossing probabilities in random walks

Let $\{X_i;\, i \geq 1\}$ be a sequence of IID random variables (rv's), each with the distribution function $\mathsf{F}_X(x)$, and let $\{S_n;\, n \geq 1\}$ be a random walk with $S_n = X_1 + \cdots + X_n$. Let $g_X(r) = \mathsf{E}\left[e^{rX}\right]$ be the MGF of X and let r_- and r_+ be the upper and lower ends of the interval over which $g(r)$ is finite. We assume throughout that $r_- < 0 < r_+$. Each of r_- and r_+ might be contained in the interval where $g_X(r)$ is finite, and each of r_- and r_+ might be infinite.

The major objective of this section is to develop results about the probability that the sequence $\{S_n;\ n \geq 1\}$ ever crosses a given threshold $\alpha > 0$, *i.e.*, $\Pr\{\bigcup_{n=1}^{\infty}\{S_n \geq \alpha\}\}$. We assume throughout this section that $\overline{X} < 0$ and that X takes on both positive and negative values with positive probability. Under these conditions, we will see that $\Pr\{\bigcup_{n=1}^{\infty}\{S_n \geq \alpha\}\}$ is essentially bounded by an exponentially decreasing function of α. In this section and the next, this exponent is evaluated and shown to be exponentially tight with increasing α. The next section also finds an exponentially tight bound on the probability that a threshold at α is crossed before another threshold at $\beta < 0$ is crossed.

7.4.1 The Chernoff bound

The Chernoff bound was derived and discussed in Section 1.4.3. It was shown in (1.62) that for any $r \in (0, r_+)$ and any $a > \overline{X}$, that

$$\Pr\{S_n \geq na\} \leq \exp\left(n[\gamma_X(r) - ra]\right). \tag{7.18}$$

where $\gamma_X(r) = \ln g_X(r)$ is the semi-invariant MGF of X. The tightest bound of this form is given by

$$\Pr\{S_n \geq na\} \leq \exp[n\mu_X(a)] \qquad \text{where } \mu_X(a) = \inf_{r\in(0,r_+)} \gamma_X(r) - ra.$$

This optimization is illustrated in Figure 7.6 for the case where $\gamma_X(r)$ is finite for all real r. The optimization is simplified by the fact (shown in Exercise 1.24) that $\gamma''(r) > 0$ for $r \in (0, r_+)$. Lemma 1.4.1 (which is almost obvious from the figure) shows that $\mu_X(a) < 0$ for $a > \overline{X}$. This implies that for fixed a, this bound decreases exponentially in n.

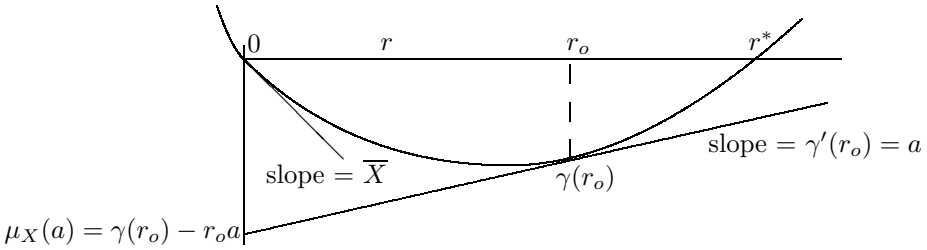


Figure 7.6: Graphical minimization of $\gamma(r) - ar$. For any r, $\gamma(r) - ar$ is the vertical axis intercept of a line of slope a through the point $(r, \gamma(r))$. The minimum occurs when the line of slope a is tangent to the curve, *i.e.*, for the r such that $\gamma'(r) = a$.

Expressing the optimization in the figure algebraically, we try to minimize $\gamma(r) - ar$ by setting the derivitive with respect to r equal to 0. Assuming for the moment that this has a solution and the solution is at some value r_o, we have

$$\Pr\{S_n \geq na\} \leq \exp\left\{n\left[\gamma(r_o) - r_o\gamma'(r_o)\right]\right\} \qquad \text{where } \gamma'(r_o) = a. \tag{7.19}$$

This can be expressed somewhat more compactly by substituting $\gamma'(r_o)$ for a in the left side of (7.19). Thus, the Chernoff bound says that for each $r \in (0, r_+)$,

$$\Pr\left\{S_n \geq n\gamma'(r)\right\} \leq \exp\left\{n\left[\gamma(r) - r\gamma'(r)\right]\right\}. \tag{7.20}$$

In principle, we could now bound the probability of threshold crossing, $\Pr\{\bigcup_{n=1}^{\infty}\{S_n \geq \alpha\}\}$, by using the union bound over n and then bounding each term by (7.19). This would be quite tedious and would also give us little insight. Instead, we pause to develop the concept of tilted probabilities. We will use these tilted probabilities in three ways, first to get a better understanding of the Chernoff bound, second to prove that the Chernoff bound is exponentially tight in the limit of large n, and third to prove the Wald identity in the next section. The Wald identity in turn will provide an upper bound on $\Pr\{\bigcup_{n=1}^{\infty}\{S_n \geq \alpha\}\}$ that is essentially exponentially tight in the limit of large α.

7.4.2 Tilted probabilities

As above, let $\{X_n;\, n \geq 1\}$ be a sequence of IID rv's and assume that $g_X(r)$ is finite for $r \in (r_-, r_+)$. Initially assume that X is discrete with the PMF $\mathsf{p}_X(x)$ for each sample value x of X. For any given $r \in (r_-,\, r_+)$, define a new PMF (called a tilted PMF) on X by

$$\mathsf{q}_{X,r}(x) = \mathsf{p}_X(x) \exp[rx - \gamma(r)]. \tag{7.21}$$

Note that $\mathsf{q}_{X,r}(x) \geq 0$ for all sample values x and $\sum_x \mathsf{q}_{X,r}(x) = \sum_x \mathsf{p}_X(x) e^{rx} / \mathsf{E}\,[e^{rx}] = 1$, so this is a valid probability assignment.

Imagine a random walk, summing IID rv's X_i, using this new probability assignment rather than the old. We then have the same mapping from sample points of the underlying sample space to sample values of rv's, but we are dealing with a new probability space, *i.e.*, we have changed the probability model, and thus we have changed the probabilities in the random walk. We will further define this new probability measure so that the rv's $X_1, X_2, \ldots$ in this new probability space are IID.[13] For every event in the old walk, the same event exists in the new walk, but its probability has changed.

Using (7.21), along with the assumption that $X_1, X_2, \ldots$ are independent in the tilted probability assignment, the tilted PMF of an n tuple $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is given by

$$\mathsf{q}_{\boldsymbol{X}^n,r}(x_1, \ldots, x_n) = \mathsf{p}_{\boldsymbol{X}^n}(x_1, \ldots, x_n) \exp(\sum_{i=1}^{n} [rx_i - \gamma(r)]. \tag{7.22}$$

Next we must related the PMF of the sum, $\sum_{i=1}^n X_i$, in the original probability measure to that in the tilted probability measure. From (7.22), note that for every n-tuple $(x_1, \ldots, x_n)$ for which $\sum_{i=1}^n x_i = s_n$ (for any given s_n) we have

$$\mathsf{q}_{\boldsymbol{X}^n,r}(x_1, \ldots, x_n) = \mathsf{p}_{\boldsymbol{X}^n}(x_1, \ldots, x_n) \exp[rs_n - n\gamma(r)].$$

Summing over all $\boldsymbol{x}^n$ for which $\sum_{i=1}^n x_i = s_n$, we then get a remarkable relation between the PMF for S_n in the original and the tilted probability measures:

$$\mathsf{q}_{\boldsymbol{S}_n,r}(s_n) = \mathsf{p}_{S_n}(s_n) \exp[rs_n - n\gamma(r)]. \tag{7.23}$$

This equation is the key to large deviation theory applied to sums of IID rv's . The tilted measure of S_n, for positive r, increases the probability of large values of S_n and decreases that of small values. Since S_n is an IID sum under the tilted measure, however, we can use the laws of large numbers and the CLT around the tilted mean to get good estimates and bounds on the behavior of S_n far from the mean for the original measure.

[13] One might ask whether $X_1, X_2, \ldots$ are the same rv's in this new probability measure as in the old. It is usually convenient to view them as being the same, since they correspond to the same mapping from sample points to sample values in both probability measures. However, since the probability space has changed, one can equally well view them as different rv's. It doesn't make any difference which viewpoint is adopted, since all we use the relationship in (7.21), and other similar relationships, to calculate probabilities in the original system.

We now denote the mean of X in the tilted measure as $\mathsf{E}_r[X]$. Using (7.21),

$$\begin{aligned}\mathsf{E}_r[X] &= \sum_x x\mathsf{q}_{X,r}(x) = \sum_x x\mathsf{p}_X(x)\exp[rx-\gamma(r)] \\ &= \frac{1}{g_X(r)}\sum_x \frac{d}{dr}\mathsf{p}_X(x)\exp[rx] \\ &= \frac{g'_X(r)}{g_X(r)} = \gamma'(r). \end{aligned} \tag{7.24}$$

Higher moments of X under the tilted measure can be calculated in the same way, but, more elegantly, the MGF of X under the tilted measure can be seen to be $\mathsf{E}_r[\exp(tX)] = g_X(t+r)/g_X(r)$.

The following theorem uses (7.23) and (7.24) to show that the Chernoff bound is exponentially tight.

Theorem 7.4.1. *Let $\{X_n; n \geq 1\}$ be a discrete IID sequence with a finite MGF for $r \in (r_-, r_+)$ where $r_- < 0 < r_+$. Let $S_n = \sum_{i=1}^n x_i$ for each $n \geq 1$. Then for any $r \in (0, r_+)$, and any $\epsilon > 0$ and $\delta > 0$ there is an m such that for all $n \geq m$,*

$$\Pr\{S_n \geq n(\gamma'(r)-\epsilon)\} \;\geq\; (1-\delta)\exp[n(\gamma(r)-r\gamma'(r)-r\epsilon)]. \tag{7.25}$$

Proof: The weak law of large numbers, in the form of (1.76), applies to the tilted measure on each S_n. Writing out the limit on n there, we see that for any ϵ, δ, there is an m such that for all $n \geq m$,

$$\begin{aligned} 1-\delta &\leq \sum_{(\gamma'(r)-\epsilon)n \,\leq\, s_n \,\leq\, (\gamma'(r)+\epsilon)n} \mathsf{q}_{S_n,r}(s_n) \\ &= \sum_{(\gamma'(r)-\epsilon)n \,\leq\, s_n \,\leq\, (\gamma'(r)+\epsilon)n} \mathsf{p}_{S_n}(s_n)\exp[rs_n - n\gamma(r)] && (7.26) \\ &\leq \sum_{(\gamma'(r)-\epsilon)n \,\leq\, s_n \,\leq\, (\gamma'(r)+\epsilon)n} \mathsf{p}_{S_n}(s_n)\exp[n(r\gamma'(r)+r\epsilon-\gamma(r))] && (7.27) \\ &\leq \sum_{(\gamma'(r)-\epsilon)n \,\leq\, s_n} \mathsf{p}_{S_n}(s_n)\exp[n(r\gamma'(r)+r\epsilon-\gamma(r))] && (7.28) \\ &= \exp[n(r\gamma'(r)+r\epsilon-\gamma(r))]\,\Pr\{S_n \geq n(\gamma'(r)-\epsilon)\}. && (7.29) \end{aligned}$$

The equality (7.26) follows from (7.23) and the inequality (7.27) follows because $s_n \leq \gamma'(r)+\epsilon$ in the sum. The next inequality is the result of adding additional positive terms into the sum, and (7.29) simply replaces the sum over a PMF with the probability of the given event. Since (7.29) is equivalent to (7.25), the proof is complete. □

The structure of the above proof can be used in many situations. A tilted probability measure is used to focus on the tail of a distribution, and then some known result about the tilted distribution is used to derive the desired result about the given distribution.

7.4.3 Back to threshold crossings

The Chernoff bound is convenient (and exponentially tight) for understanding $\Pr\{S_n \geq na\}$ as a function of n for fixed a. It does not give us as much direct insight into $\Pr\{S_n \geq \alpha\}$ as a function of n for fixed α, which tells us something about *when* a threshold at α is most likely to be crossed. Additional insight can be achieved by substituing $\alpha/\gamma'(r_0)$ for n in (7.19), getting

$$\Pr\{S_n \geq \alpha\} \leq \exp\left\{\alpha\left[\frac{\gamma(r_o)}{\gamma'(r_o)} - r_o\right]\right\} \qquad \text{where } \gamma'(r_o) = \alpha/n. \tag{7.30}$$

A nice graphic interpretation of this equation is given in Figure 7.7. Note that the exponent in α, namely $\gamma(r_0)/\gamma'(r_0) - r_0$ is the negative of the horizontal axis intercept of the line of slope $\gamma'(r_0) = \alpha/n$ in Figure 7.7.

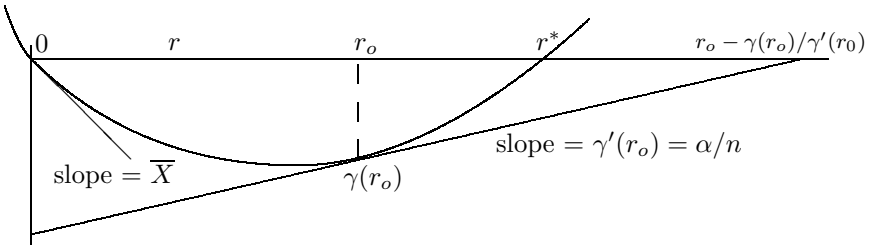


Figure 7.7: The exponent in α for $\Pr\{S_n \geq \alpha\}$, minimized over r. The minimization is the same as that in Figure 7.6, but $\gamma(r_o)/\gamma'(r_o) - r_o$ is the negative of the horizontal axis intercept of the line tangent to $\gamma(r)$ at $r = r_0$.

For fixed α, then, we see that for very large n, the slope α/n is very small and this horizontal intercept is very large. As n is decreased, the slope increases, r_0 increases, and the horizontal intercept decreases. When r_0 increases to the point labeled r^* in the figure, namely the $r > 0$ at which $\gamma(r) = 0$, then the exponent decreases to r^*. When n decreases even further, r_0 becomes larger than r^* and the horizontal intercept starts to increase again.

Since r^* is the minimum horizontal axis intercept of this class of straight lines, we see that the following bound holds for all n,

$$\Pr\{S_n \geq \alpha\} \leq \exp(-r^*\alpha) \qquad \text{for arbitrary } \alpha > 0,\ n \geq 1, \tag{7.31}$$

where r^* is the positive root of $\gamma(r) = 0$.

The graphical argument above assumed that there is some $r^* > 0$ such that $\gamma_X(r^*) = 0$. However, if $r_+ = \infty$, then (since X takes on positive values with positive probability by assumption) $\gamma(r)$ can be seen to approach ∞ as $r \to \infty$. Thus r^* must exist because of the continuity of $\gamma(r)$ in (r_-, r_+). This is summarized in the following lemma.

Lemma 7.4.1. *Let $\{S_n = X_1 + \cdots + X_n;\ n \geq 1\}$ be a random walk where $\{X_i;\ i \geq 1\}$ is an IID sequence where $g_X(r)$ exists for all $r \geq 0$, where $\overline{X} < 0$, and where X takes on both positive and negative values with positive probability. Then for all integer $n > 0$ and*

all $\alpha > 0$

$$\Pr\{S_n \geq \alpha\} \;\leq\; \exp\{\alpha[\gamma(r_0)n/\alpha - r_0]\} \;\leq\; \exp(-r^*\alpha). \tag{7.32}$$

where $\gamma'(r_o) = \alpha/n$.

Next we briefly consider the situation where $r_+ < \infty$. There are two cases to consider, first where $\gamma(r_+)$ is infinite and second where it is finite. Examples of these cases are given in Exercise 1.22, parts b) and c) respectively. For the case $\gamma(r_+) = \infty$, Exercise 1.23 shows that $\gamma(r)$ goes to ∞ as r increases toward r_+. Thus r^* must exist in this case 1 by continuity.

The case $\gamma(r_+) < \infty$ is best explained by Figure 7.8. As explained there, if $\gamma'(r) = \alpha/n$ for some $r_0 < r_+$, then (7.19) and (7.30) hold as before. Alternatively, if $\alpha/n > \gamma'(r)$ for all $r < r_+$, then the Chernoff bound is optimized at $r = r_+$, and we have

$$\Pr\{S_n \geq \alpha\} \leq \exp\{n[\gamma(r_+) - r_+\alpha/n]\} = \exp\{\alpha[\gamma(r_+)n/\alpha - r_+]\}. \tag{7.33}$$

If we modify the definition of r^* to be the supremum of $r > 0$ such that $\gamma(r) < 0$, then $r^* = r_+$ in this case, and (7.32) is valid in general with the obvious modification of r_o.

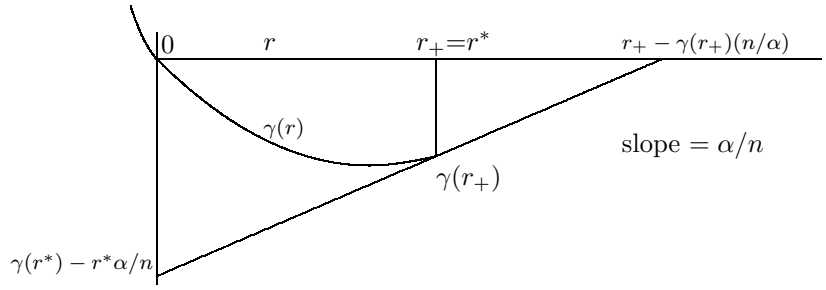


Figure 7.8: Graphical minimization of $\gamma(r) - (\alpha/n)r$ for the case where $\gamma(r^+) < \infty$. As before, for any $r < r_+$, we can find $\gamma(r) - r\alpha/n$ by drawing a line of slope (α/n) from the point $(r, \gamma(r))$ to the vertical axis. If $\gamma'(r) = \alpha/n$ for some $r < r_+$, the minimum occurs at that r. Otherwise, as shown in the figure, it occurs at $r = r_+$.

We could now use this result, plus the union bound, to find an upper bound on the probabiity of threshold crossing, *i.e.*, $\Pr\{\bigcup_{n=1}^{\infty}\{S_n \geq \alpha\}\}$. The coefficients in this are somewhat messy and change according to the special cases discussed above. It is far more instructive and elegant to develop Wald's identity, which shows that $\Pr\{\bigcup_n\{S_n \geq \alpha\}\} \leq \exp[-\alpha r^*]$. This is slightly stronger than the Chernoff bound approach in that the bound on the probability of the union is the same as that on a single term in the Chernoff bound approach. The main value of the Chernoff bound approach, then, is to provide assurance that the bound is exponentially tight.

7.5 Thresholds, stopping rules, and Wald's identity

The following lemma shows that a random walk with both a positive and negative threshold, say $\alpha > 0$ and $\beta < 0$, eventually crosses one of the thresholds. Figure 7.9 illustrates two

sample paths and how they cross thresholds. More specifically, the random walk first crosses a threshold at trial n if $\beta < S_i < \alpha$ for $1 \leq i < n$ and either $S_n \geq \alpha$ or $S_n \leq \beta$. For now we make no assumptions about the mean or MGF of S_n.

The trial at which a threshold is first crossed is a possibly-defective rv J. The event $J = n$ (*i.e.*, the event that a threshold is first crossed at trial n), is a function of $S_1, S_2, \ldots, S_n$, and thus, in the notation of Section 4.5, J is a possibly-defective stopping trial. The following lemma shows that J is actually a stopping trial, namely that stopping (threshold crossing) eventually happens with probability 1.

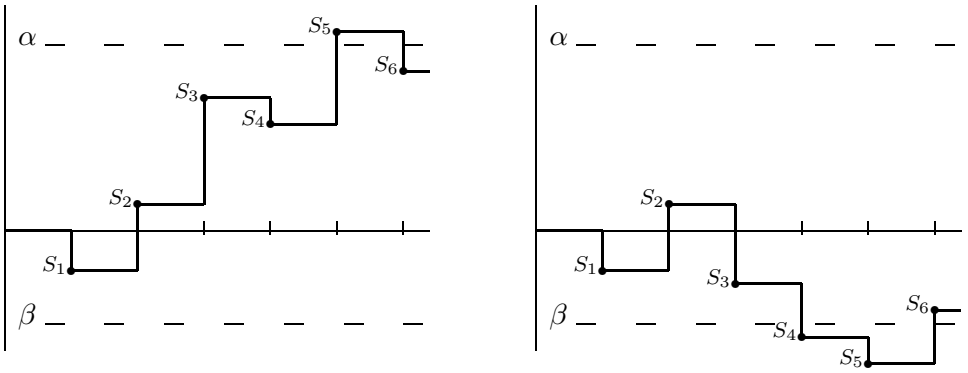


Figure 7.9: Two sample paths of a random walk with two thresholds. In the first, the threshold at α is crossed at $J = 5$. In the second, the threshold at β is crossed at $J = 4$

Lemma 7.5.1. *Let $\{X_i; i \geq 1\}$ be IID rv's, not identically 0. For each $n \geq 1$, let $S_n = X_1 + \cdots + X_n$. Let $\alpha > 0$ and $\beta < 0$ be arbitrary, and let J be the smallest n for which either $S_n \geq \alpha$ or $S_n \leq \beta$. Then J is a random variable (*i.e., $\lim_{m\to\infty} \Pr\{J \geq m\} = 0$*) and has finite moments of all orders.*

Proof: Since X is not identically 0, there is some n for which either $\Pr\{S_n \leq -\alpha + \beta\} > 0$ or for which $\Pr\{S_n \geq \alpha - \beta\} > 0$. For any such n, define ϵ by

$$\epsilon = \max[\Pr\{S_n \leq -\alpha + \beta\},\ \Pr\{S_n \geq \alpha - \beta\}].$$

For any integer $k \geq 1$, given that $J > n(k-1)$, and given any value of $S_{n(k-1)}$ in (β, α), a threshold will be crossed by time nk with probability at least ϵ. Thus,

$$\Pr\{J > nk \mid J > n(k-1)\} \leq 1 - \epsilon,$$

Iterating on k,

$$\Pr\{J > nk\} \leq (1-\epsilon)^k.$$

This shows that J is finite with probability 1 and that $\Pr\{J \geq j\}$ goes to 0 at least geometrically in j. It follows that the moment generating function $\mathsf{g}_J(r)$ of J is finite in a region around $r = 0$, and that J has moments of all orders. □

7.5.1 Wald's identity for two thresholds

Theorem 7.5.1 (Wald's identity). *Let $\{X_i;\, i \geq 1\}$ be IID and let $\gamma(r) = \ln\{\mathsf{E}\left[e^{rX}\right]\}$ be the semi-invariant MGF of each X_i. Assume $\gamma(r)$ is finite in the open interval (r_-, r_+) with $r_- < 0 < r_+$. For each $n \geq 1$, let $S_n = X_1 + \cdots + X_n$. Let $\alpha > 0$ and $\beta < 0$ be arbitrary real numbers, and let J be the smallest n for which either $S_n \geq \alpha$ or $S_n \leq \beta$. Then for all $r \in (r_-, r_+)$,*

$$\mathsf{E}\left[\exp(rS_J - J\gamma(r))\right] = 1. \tag{7.34}$$

The following proof is a simple application of the tilted probability distributions discussed in the previous section..

Proof: We assume that X_i is discrete for each i with the PMF $\mathsf{p}_X(x)$. For the arbitrary case, the PMF's must be replaced by distribution functions and the sums by Stieltjes integrals, thus complicating the technical details but not introducing any new ideas.

For any given $r \in (r_-, r_+)$, we use the tilted PMF $\mathsf{q}_{X,r}(x)$ given in (7.21) as

$$\mathsf{q}_{X,r}(x) = \mathsf{p}_X(x) \exp[rx - \gamma(r)].$$

Taking the X_i to be independent in the tilted probability measure, the tilted PMF for the n-tuple $\boldsymbol{X}^n = (X_1, X_2, \ldots, X_n)$ is given by

$$\mathsf{q}_{\boldsymbol{X}^n,r}(\boldsymbol{x}^n) = \mathsf{p}_{\boldsymbol{X}^n}(\boldsymbol{x}^n) \exp[rs_n - n\gamma(r)] \qquad \text{where } s_n = \sum_{i=1}^{n} x_i.$$

Now let $\mathcal{T}_n$ be the set of n-tuples $X_1, \ldots, X_n$ such that $\beta < S_i < \alpha$ for $1 \leq i < n$ and either $S_n \geq \alpha$ or $S_n \leq \beta$. That is, $\mathcal{T}_n$ is the set of $\boldsymbol{x}^n$ for which the stopping trial J has the sample value n. The PMF for the stopping trial J in the tilted measure is then

$$\begin{aligned} \mathsf{q}_{J,r}(n) &= \sum_{\boldsymbol{x}^n \in \mathcal{T}_n} \mathsf{q}_{\boldsymbol{X}^n,r}(\boldsymbol{x}^n) = \sum_{\boldsymbol{x}^n \in \mathcal{T}_n} \mathsf{p}_{\boldsymbol{X}^n}(\boldsymbol{x}^n) \exp[rs_n - n\gamma(r)] \\ &= \mathsf{E}\left[\exp[rS_n - n\gamma(r) \mid J{=}n\right] \Pr\{J = n\}. \end{aligned} \tag{7.35}$$

Lemma 7.5.1 applies to the tilted PMF on this random walk as well as to the original PMF, and thus the sum of $\mathsf{q}_J(n)$ over n is 1. Also, the sum over n of the expression on the right is $\mathsf{E}\left[\exp[rS_J - J\gamma(r)]\right]$, thus completing the proof. □

After understanding the details of this proof, one sees that it is essentially just the statement that J is a non-defective stopping rule in the tilted probability space.

We next give a number of examples of how Wald's identity can be used.

7.5.2 The relationship of Wald's identity to Wald's equality

The trial J at which a threshold is crossed in Wald's identity is a stopping trial in the terminology of Chapter 4. If we take the derivative with respect to r of both sides of (7.34), we get

$$\mathsf{E}\left[\left[S_J - J\gamma'(r)\right] \exp\{rS_J - J\gamma(r)\}\right. = 0.$$

Setting $r = 0$ and recalling that $\gamma(0) = 0$ and $\gamma'(0) = \overline{X}$, this becomes Wald's equality as established in Theorem 4.5.1,

$$\mathsf{E}[S_J] = \mathsf{E}[J]\,\overline{X}. \tag{7.36}$$

Note that this derivation of Wald's equality is restricted[14] to a random walk with two thresholds (and this automatically satisfies the constraint in Wald's equality that $\mathsf{E}[J] < \infty$). The result in Chapter 4 is more general, applying to any stopping trial such that $\mathsf{E}[J] < \infty$.

The second derivative of (7.34) with respect to r is

$$\mathsf{E}\left[\left[(S_J - J\gamma'(r))^2 - J\gamma''(r)\right]\right] \exp\{rS_J - J\gamma(r)\} = 0.$$

At $r = 0$, this is

$$\mathsf{E}\left[S_J^2 - 2JS_J\overline{X} + J^2\overline{X}^2\right] = \sigma_X^2\,\mathsf{E}[J]. \tag{7.37}$$

This equation is often difficult to use because of the cross term between S_J and J, but its main application comes in the case where $\overline{X} = 0$. In this case, Wald's equality provides no information about $\mathsf{E}[J]$, but (7.37) simplifies to

$$\mathsf{E}\left[S_J^2\right] = \sigma_X^2\,\mathsf{E}[J]. \tag{7.38}$$

7.5.3 Zero-mean simple random walks

As an example of (7.38), consider the simple random walk of Section 7.1.1 with $\Pr\{X{=}1\} = \Pr\{X{=}-1\} = 1/2$, and assume that $\alpha > 0$ and $\beta < 0$ are integers. Since S_n takes on only integer values and changes only by ± 1, it takes on the value α or β before exceeding either of these values. Thus $S_J = \alpha$ or $S_J = \beta$. Let q_α denote $\Pr\{S_J = \alpha\}$. The expected value of S_J is then $\alpha q_\alpha + \beta(1 - q_\alpha)$. From Wald's equality, $\mathsf{E}[S_J] = 0$, so

$$q_\alpha = \frac{-\beta}{\alpha - \beta}; \qquad 1 - q_\alpha = \frac{\alpha}{\alpha - \beta}. \tag{7.39}$$

From (7.38),

$$\sigma_X^2\,\mathsf{E}[J] = \mathsf{E}\left[S_J^2\right] = \alpha^2 q_\alpha + \beta^2(1 - q_\alpha). \tag{7.40}$$

Using the value of q_α from (7.39) and recognizing that $\sigma_X^2 = 1$,

$$\mathsf{E}[J] = -\beta\alpha/\sigma_X^2 = -\beta\alpha. \tag{7.41}$$

As a sanity check, note that if α and β are each multiplied by some large constant k, then $\mathsf{E}[J]$ increases by k^2. Since $\sigma_{S_n}^2 = n$, we would expect S_n to fluctuate with increasing n, with typical values growing as $\sqrt{n}$, and thus it is reasonable that the expected time to reach a threshold increases with the product of the distances to the thresholds.

[14] This restriction is quite artificial and made simply to postpone any consideration of generalizations until discussing martingales.

We also notice that if β is decreased toward $-\infty$, while holding α constant, then $q_\alpha \to 1$ and $\mathsf{E}[J] \to \infty$. This helps explain Example 4.5.2 where one plays a coin tossing game, stopping when finally ahead. This shows that if the coin tosser has a finite capital b, *i.e.*, stops either on crossing a positive threshold at 1 or a negative threshold at $-b$, then the coin tosser wins a small amount with high probability and loses a large amount with small probability.

For more general random walks with $\overline{X} = 0$, there is usually an overshoot when the threshold is crossed. If the magnitudes of α and β are large relative to the range of X, however, it is often reasonable to ignore the overshoots, and $-\beta\alpha/\sigma_X^2$ is then often a good approximation to $\mathsf{E}[J]$. If one wants to include the overshoot, then the effect of the overshoot must be taken into account both in (7.39) and (7.40).

7.5.4 Exponential bounds on the probability of threshold crossing

We next apply Wald's identity to complete the analysis of crossing a threshold at $\alpha > 0$ when $\overline{X} < 0$.

Corollary 7.5.1. *Under the conditions of Theorem 7.5.1, assume that $\overline{X} < 0$ and that $r^* > 0$ exists such that $\gamma(r^*) = 0$. Then*

$$\Pr\{S_J \geq \alpha\} \leq \exp(-r^*\alpha). \tag{7.42}$$

Proof: Wald's identity, with $r = r^*$, reduces to $\mathsf{E}[\exp(r^*S_J)] = 1$. We can express this as

$$\Pr\{S_J \geq \alpha\}\,\mathsf{E}[\exp(r^*S_J) \mid S_J \geq \alpha] + \Pr\{S_J \leq \beta\}\,\mathsf{E}[\exp(r^*S_J) \mid S_J \leq \beta] = 1. \tag{7.43}$$

Since the second term on the left is nonnegative,

$$\Pr\{S_J \geq \alpha\}\,\mathsf{E}[\exp(r^*S_J) \mid S_J \geq \alpha] \leq 1. \tag{7.44}$$

Given that $S_J \geq \alpha$, we see that $\exp(r^*S_J) \geq \exp(r^*\alpha)$. Thus

$$\Pr\{S_J \geq \alpha\}\exp(r^*\alpha) \leq 1. \tag{7.45}$$

which is equivalent to (7.42). □

This bound is valid for all $\beta < 0$, and thus is also valid in the limit $\beta \to -\infty$ (see Exercise 7.10 for a more detailed demonstration that (7.42) is valid without a lower threshold). We see from this that the case of a single threshold is really a special case of the two threshold problem, but as seen in the zero-mean simple random walk, having a second threshold is often valuable in further understanding the single threshold case. Equation (7.42) is also valid for the case of Figure 7.8, where $\gamma(r) < 0$ for all $r \in (0, r_+]$ and $r^* = r_+$.

The exponential bound in (7.31) shows that $\Pr\{S_n \geq \alpha\} \leq \exp(-r^*\alpha)$ for each n; (7.42) is stronger than this. It shows that $\Pr\{\bigcup_n \{S_n \geq \alpha\}\} \leq \exp(-r^*\alpha)$. This also holds in the limit $\beta \to -\infty$.

When Corollary 7.5.1 is applied to the G/G/1 queue in Theorem 7.2.1, (7.42) is referred to as the *Kingman Bound*.

Corollary 7.5.2 (Kingman Bound). *Let $\{X_i;\, i \geq 1\}$ and $\{Y_i;\, i \geq 0\}$ be the interarrival intervals and service times of a G/G/1 queue that is empty at time 0 when customer 0 arrives. Let $\{U_i = Y_{i-1} - X_i;\, i \geq 1\}$, and let $\gamma(r) = \ln\{\mathsf{E}\left[e^{Ur}\right]\}$ be the semi-invariant moment generating function of each U_i. Assume that $\gamma(r)$ has a root at $r^* > 0$. Then W_n, the queueing delay of the nth arrival, and W, the steady state queueing delay, satisfy*

$$\Pr\{W_n \geq \alpha\} \leq \Pr\{W \geq \alpha\} \leq \exp(-r^*\alpha) \quad ; \qquad \text{for all } \alpha > 0. \tag{7.46}$$

In most applications, a positive threshold crossing for a random walk with a negative drift corresponds to some exceptional, and often undesirable, circumstance (for example an error in the hypothesis testing problem or an overflow in the G/G/1 queue). Thus an upper bound such as (7.42) provides an assurance of a certain level of performance and is often more useful than either an approximation or an exact expression that is very difficult to evaluate. Since these bounds are exponentially tight, they also serve as rough approximations.

For a random walk with $\overline{X} > 0$, the exceptional circumstance is $\Pr\{S_J \leq \beta\}$. This can be analyzed by changing the sign of X and β and using the results for a negative expected value. These exponential bounds do not work for $\overline{X} = 0$, and we will not analyze that case here other than the result in (7.38).

Note that the simple bound on the probability of crossing the upper threshold in (7.42) (and thus also the Kingman bound) is an upper bound (rather than an equality) because, first, the effect of the lower threshold was eliminated (see (7.44)), and, second, the overshoot was bounded by 0 (see (7.45). The effect of the second threshold can be taken into account by recognizing that $\Pr\{S_J \leq \beta\} = 1 - \Pr\{S_J \geq \alpha\}$. Then (7.43) can be solved, getting

$$\Pr\{S_J \geq \alpha\} = \frac{1 - \mathsf{E}\left[\exp(r^*S_J) \mid S_J \leq \beta\right]}{\mathsf{E}\left[\exp(r^*S_J) \mid S_J \geq \alpha\right] - \mathsf{E}\left[\exp(r^*S_J) \mid S_J \leq \beta\right]}. \tag{7.47}$$

Solving for the terms on the right side of (7.47) usually requires analyzing the overshoot upon crossing a barrier, and this is often difficult. For the case of the simple random walk, overshoots don't occur, since the random walk changes only in unit steps. Thus, for α and β integers, we have $\mathsf{E}\left[\exp(r^*S_J) \mid S_J \leq \beta\right] = \exp(r^*\beta)$ and $\mathsf{E}\left[\exp(r^*S_J) \mid S_J \geq \alpha\right] = \exp(r^*\alpha)$. Substituting this in (7.47) yields the exact solution

$$\Pr\{S_J \geq \alpha\} = \frac{\exp(-r^*\alpha)[1 - \exp(r^*\beta)]}{1 - \exp[-r^*(\alpha - \beta)]}. \tag{7.48}$$

Solving the equation $\gamma(r^*) = 0$ for the simple random walk with probabilities p and q yields $r^* = \ln(q/p)$. This is also valid if X takes on the three values -1, 0, and $+1$ with $p = \Pr\{X = 1\}$, $q = \Pr\{X = -1\}$, and $1 - p - q = \Pr\{X = 0\}$. It can be seen that if α and $-\beta$ are large positive integers, then the simple bound of (7.42) is almost exact for this example.

Equation (7.48) is sometimes used as an approximation for (7.47). Unfortunately, for many applications, the overshoots are more significant than the effect of the opposite threshold. Thus (7.48) is only negligibly better than (7.42) as an approximation, and has the further disadvantage of not being a bound.

If $\Pr\{S_J \geq \alpha\}$ must actually be calculated, then the overshoots in (7.47) must be taken into account. See Chapter 12 of [8] for a treatment of overshoots.

7.5.5 Binary hypotheses testing with IID observations

The objective of this subsection and the next is to understand how to make binary decisions on the basis of a variable number of observations, choosing to stop observing additional data when a sufficiently good decision can be made. This initial subsection lays the groundwork for this by analyzing the large deviation aspects of binary detection with a large but fixed number of IID observations.

Consider the binary hypothesis testing problem of Section 7.3 in which H is a binary hypothesis with apriori probabilities $\mathsf{p}_H(0) = p_0$ and $\mathsf{p}_H(1) = p_1$. The observation $Y_1, Y_2, \ldots$, conditional on $H = 0$, is a sequence of IID rv's with the probability density $\mathsf{f}_{Y|H}(y \mid 0)$. Conditional on $H = 1$, the observations are IID with density $\mathsf{f}_{Y|H}(y \mid 1)$. For any given number n of sample observations, $y_1, \ldots, y_n$, the likelihood ratio was defined as

$$\Lambda_n(\boldsymbol{y}) = \prod_{i=1}^{n} \frac{\mathsf{f}_{Y|H}(y_n \mid 0)}{\mathsf{f}_{Y|H}(y_n \mid 1)}$$

and the log-likelihood-ratio was defined as

$$s_n = \sum_{i=1}^{n} z_n; \qquad \text{where } z_n = \ln \frac{\mathsf{f}_{Y|H}(y_n \mid 0)}{\mathsf{f}_{Y|H}(y_n \mid 1)}. \tag{7.49}$$

The MAP test gives the maximum aposteriori probability of correct decision based on the n observations $(y_1, \ldots, y_n)$. It is defined as the following threshold test:

$$s_n \begin{cases} > \ln(p_1/p_0) \quad ; & \text{select } \hat{h}{=}0 \\ \leq \ln(p_1/p_0) \quad ; & \text{select } \hat{h}{=}1. \end{cases}$$

We can use the Chernoff bound to get an exponentially tight bound on the probability of error using the the MAP test and given $H = 1$ (and similarly given $H = 0$). That is, $\Pr\{e \mid H = 1\}$ is simply $\Pr\{S_n \geq \ln(p_1/p_0) \mid H{=}1\}$ where S_n is the rv whose sample value is given by (7.49), *i.e.*, $S_n = \sum_{i=1}^{n} Z_n$ where $Z_n = \ln[f_{Y|H}(Y \mid 0)/f_{Y|H}(Y \mid 1)]$. The semi-invariant MGF $\gamma_1(r)$ of Z given $H = 1$ is

$$\begin{aligned} \gamma_1(r) &= \ln \int_y \mathsf{f}_{Y|H}(y \mid 1) \exp\left\{ r \left[\ln \frac{\mathsf{f}_{Y|H}(y \mid 0)}{\mathsf{f}_{Y|H}(y \mid 1)} \right] \right\} dy \\ &= \ln \int_y [\mathsf{f}_{Y|H}(y \mid 1)]^{1-r} [\mathsf{f}_{Y|H}(y \mid 0)]^r \, dy \end{aligned} \tag{7.50}$$

Surprisingly, we see that $\gamma_1(1) = \ln \int_y \mathsf{f}_{Y|H}(y \mid 0)\, dy = 0$, so that $r^* = 1$ for $\gamma_1(r)$. Since $\gamma_1(r)$ has a positive second derivative, this shows that $\gamma_1'(0) < 0$ and $\gamma_1'(1) > 0$. Figure 7.10 illustrates the optimized exponent in the Chernoff bound,

$$\Pr\{e \mid H{=}1\} \leq \exp\left\{ n \left[[\min_r \gamma_1(r) - ra \right] \right\} \qquad \text{where } a = \frac{1}{n} \ln(p_1/p_0)$$

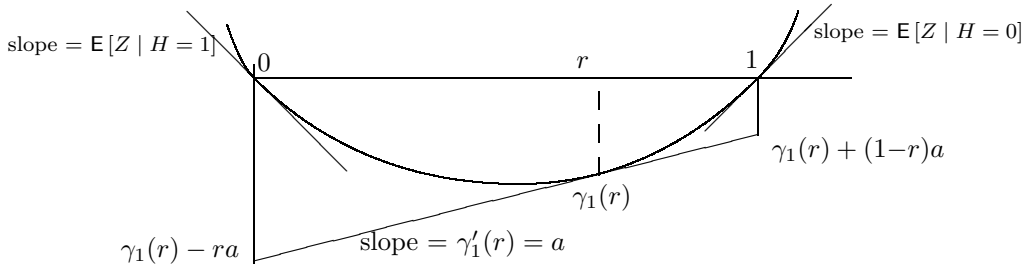


Figure 7.10: Graphical description of the exponents in the optimized Chernoff bounds for $\Pr\{e \mid H{=}1\} \leq \exp\{n[\gamma_1(r) - ra]\}$ and $\Pr\{e \mid H{=}0\} \leq \exp\{n[\gamma_1(r) + (1-r)a]\}$. The slope a is equal to $\frac{1}{n}\ln(p_1/p_0)$. For an arbitrary threshold test with threshold η, the slope a is $\frac{1}{n}\ln\eta$.

We can find the Chernoff bound for $\Pr\{e \mid H{=}0\}$ in the same way. The semi-invariant MGF for Z conditional on $H = 0$ is given by

$$\gamma_0(r) = \ln \int_y [\mathsf{f}_{Y|H}(y \mid 1)]^{-r} [\mathsf{f}_{Y|H}(y \mid 0)]^{1+r}\, dy = \gamma_1(r{-}1) \tag{7.51}$$

We can see from this that the distribution of Z conditional on $H = 0$ is the tilted distribution (at $r = 1$) of Z conditional on $H = 1$. Since $\gamma_0(r) = \gamma_1(r{-}1)$, the optimized Chernoff bound for $\Pr\{e \mid H{=}0\}$ is given by

$$\Pr\{e \mid H{=}0\} \leq \exp\left\{n\left[[\min_r \gamma_1(r) + (1{-}r)a\right]\right\} \qquad \text{where } a = \frac{1}{n}\ln(p_1/p_0)$$

In general, the threshold p_1/p_0 can be replaced by an arbitrary threshold η. As illustrated in Figure 7.10, the slope $\gamma_1'(r) = \frac{1}{n}\ln(\eta)$ at which the optimized bound occurs increases with η, varying from $\gamma'(0) = \mathsf{E}[Z \mid H{=}1]$ at $r = 0$ to $\gamma'(1) = \mathsf{E}[Z \mid H{=}0]$ at $r = 1$. The tradeoff between the two error exponents are seen to vary as the two ends of an inverted see-saw. One could in principle achieve a still larger magnitude of exponent for $\Pr\{(\}\, e \mid H{=}1)$ by using $r > 1$, but this would be somewhat silly since $\Pr\{e \mid H{=}0\}$ would then be very close to 1 and it would usually make more sense to simply decide on $H = 1$ without looking at the observed data at all.

We can view the tradeoff of exponents above as a large deviation form of the Neyman-Pearson criterion. That is, rather than fixing the allowable value of $\Pr\{e \mid H{=}1\}$ and choosing a test to minimize $\Pr\{e \mid H{=}0\}$, we fix the allowable exponent for $\Pr\{e \mid H{=}1\}$ and then optimize the exponent for $\Pr\{e \mid H{=}1\}$. This is essentially the same as the Neyman-Pearson test, and is a threshold test as before, but it allows us to ignore the exact solution of the Neyman-Pearson problem and focus on the exponentially tight exponents.

7.5.6 Sequential decisions for binary hypotheses

Common sense tells us that it should be valuable to have the ability to make additional observations when the current observations do not lead to a clear choice. With such an

ability, we have 3 possible choices at the end of each observation: choose $\hat{H} = 1$, choose $\hat{H} = 0$, or continue with additional observations. It is evident that we need a stopping time J to decide when to stop. We assume for the time being that stopping occurs with probability 1, both under hypothesis $H = 1$ and $H = 0$, and the indicator rv $\mathbb{I}_{J=n}$ for each $n \geq 1$ is a function of $Z_1, \ldots, Z_n$.

Given $H = 1$, the sequence of LLR's $\{S_n; n \geq 1\}$ (where $S_n = \sum_{i=1}^{n} Z_i$) is a random walk. Assume for now that the stopping rule is to stop when the random walk crosses either a threshold[15] at $\alpha > 0$ or a threshold at $\beta < 0$. Given $H = 0$, $S_n = \sum_{i=1}^{n} Z_n$ is also a random walk (with a different probability measure). The same stopping rule must be used, since the decision to stop at n can be based only on $Z_1, \ldots, Z_n$ and not on knowledge of H. We saw that the random walk based on $H = 0$ uses probabilities tilted from those based on $H = 1$, a fact that is interesting but not necessary here.

We assume that when stopping occurs, the decision $\hat{h} = 0$ is made if $S_J \geq \alpha$ and $\hat{h} = 1$ is made if $S_J \leq \beta$. Given that $H = 1$, an error is then made if $S_J \geq \alpha$. Thus, using the probability distribution for $H = 1$, we apply (7.42), along with $r^* = 1$, to get

$$\Pr\{e \mid H{=}1\} \;=\; \Pr\{S_J \geq \alpha \mid H{=}1\} \;\leq\; \exp -\alpha \tag{7.52}$$

Similarly, conditional on $H = 0$, an error is made if the threshold at β is crossed before that at α and the error probability is

$$\Pr\{e \mid H{=}0\} \;=\; \Pr\{S_J \leq \beta \mid H{=}0\} \;\leq\; \exp \beta \tag{7.53}$$

The error probabilities can be made as small as desired by increasing the magnitudes of α and β, but the cost of increasing α is to increase the number of observations required when $H = 0$. From Wald's equality,

$$\mathsf{E}\left[J \mid H{=}0\right] \;=\; \frac{\mathsf{E}\left[S_J \mid H{=}0\right]}{\mathsf{E}\left[Z \mid H{=}0\right]} \;\approx\; \frac{\alpha + \mathsf{E}\left[\text{overshoot}\right]}{\mathsf{E}\left[Z \mid H{=}0\right]}$$

In the approximation, we have ignored the possibility of S_J crossing the threshold at β conditional on $H = 0$ since this is a very small-probability event when α and β have large magnitudes. Thus we see that the expected number of observations (given $H = 0$) is essentially linear in α.

We next ask what has been gained quantitatively by using the sequential decision procedure here. Suppose we use a fixed-length test with $n = \alpha/\mathsf{E}\left[Z \mid H{=}0\right]$. Referring to Figure 7.10, we see that if we choose the slope $a = \gamma'(1)$, then the (exponentially tight) Chernoff bound on $\Pr\{e \mid H{=}1\}$ is given by $e^{-\alpha}$, but the exponent on $\Pr\{e \mid H{=}0\}$ is 0. In other words, by using a sequential test as described here, we simultaneously get the error exponent for

[15] It is unfortunate that the word 'threshold' has a universally accepted meaning for random walks (*i.e.*, the meaning we are using here), and the word 'threshold test' has a universally accepted meaning for hypothesis testing. Stopping when S_n crosses either α or β can be viewed as an extension of an hypothesis testing threshold test in the sense that both the stopping trial and the decision is based only on the LLR, and is based on the LLR crossing either a positive or negative threshold, but the results about threshold tests in Section 7.3 are not necessarily valid for this extension.

$H = 1$ that a fixed test would provide if we gave up entirely on an error exponent for $H = 0$, and vice versa.[16]

A final question to be asked is whether any substantial improvement on this sequential decision procedure would result from letting the thresholds at α and β vary with the number of observations. Assuming that we are concerned only with the expected number of observations, the answer is no. We will not carry this argument out here, but it consists of using the Chernoff bound as a function of the number of observations. This shows that there is a typical number of observations at which most errors occur, and changes in the thresholds elsewhere can increase the error probability, but not substantially decrease it.

7.5.7 Joint distribution of crossing time and barrier

Next we look at $\Pr\{J \geq n,\ S_J \geq \alpha\}$, where again we assume that $\overline{X} < 0$ and that $\gamma(r^*) = 0$ for some $r^* > 0$. For any r in the region where $\gamma(r) \leq 0$ (*i.e.*, $0 \leq r \leq r^*$), we have $-J\gamma(r) \geq -n\gamma(r)$ for $J \geq n$. Thus, from the Wald identity, we have

$$\begin{aligned} 1 &\geq \mathsf{E}\left[\exp[rS_J - J\gamma(r)] \mid J \geq n,\ S_J \geq \alpha\right] \Pr\{J \geq n,\ S_J \geq \alpha\} \\ &\geq \exp[r\alpha - n\gamma(r)] \Pr\{J \geq n, S_J \geq \alpha\} \\ \Pr\{J \geq n, S_J \geq \alpha\} &\leq \exp[-r\alpha + n\gamma(r)]\ ; \quad \text{for } \ r \in [0, r^*]. \end{aligned} \tag{7.54}$$

Since this is vaid for all $r \in (0, r^*]$, we can obtain the tightest bound of this form by minimizing the right hand side of (7.54). This is the same minimization (except for the constraint $r \leq r^*$) as in Figures 7.7. Assume that $r^* < r_+$ (*i.e.*, that the exceptional situation[17] of Figure 7.8) does not occur. Define n^* as $\alpha/\gamma'(r^*)$. The result is then

$$\Pr\{J \geq n, S_J \geq \alpha\} \leq \begin{cases} \exp\left[n\gamma(r_o) - r_o\alpha\right] & \text{for } n > n^*,\ \alpha/n = \gamma'(r_o) \\ \\ \exp\left[-r^*\alpha\right] & n \leq n^*. \end{cases} \tag{7.55}$$

The interpretation of (7.55) is that n^* is an estimate of the typical value of J given that the threshold at α is crossed. For n greater than this typical value, (7.55) provides a tighter bound on $\Pr\{J \geq n, S_J \geq \alpha\}$ than the bound on $\Pr\{S_J \geq \alpha\}$ in (7.42), whereas (7.55) provides nothing new for $n \leq n^*$. In Section 7.8, we shall derive the slightly stronger result that $\Pr\left\{\bigcup_{i \geq n}[S_i \geq \alpha]\right\}$ is also upper bounded by the right hand side of (7.55).

[16] In the communication context, decision rules are used to detect sequentially transmitted data. The use of a sequential decision rule usually requires feedback from receiver to transmitter, and also requires a variable rate of transmission. Thus the substantial reductions in error probability are accompanied by substantial system complexity.

[17] The exceptional situation where $\gamma(r)$ is negative for $r \leq r_+$ and discontinuous at $r^* = r_+$, *i.e.*, the situation in Figure 7.8), will be treated in the exercises, and is quite different from the case here. In particular, as can be seen from the figure, the optimized Chernoff bound on $\Pr\{S_n \geq \alpha\}$ is optimized at $n = 1$.

An almost identical upper bound to $\Pr\{J \leq n,\ S_J \geq \alpha\}$ can be found (again assuming that $r^* < r_+$) by using the Wald identity for $r > r^*$. Here $\gamma(r) > 0$, so $-J\gamma(r) \geq -n\gamma(r)$ for $J \leq n$. The result is

$$\Pr\{J \leq n, S_J \geq \alpha\} \leq \begin{cases} \exp\left[n\gamma(r_o) - r_o\alpha\right] & \text{for } n < n^*,\ \alpha/n = \gamma'(r_o) \\ \\ \exp\left[-r^*\alpha\right] & n \geq n^*. \end{cases} \tag{7.56}$$

This strengthens the interpretation of n^* as the typical value of J conditional on crossing the threshold at α. That is, (7.56) provides information on the lower tail of the distribution of J (conditional on $S_J \geq \alpha$), whereas (7.55) provides information on the upper tail.

7.6 Martingales

A martingale is defined as an integer-time stochastic process $\{Z_n;\ n \geq 1\}$ with the properties that $\mathsf{E}\left[|Z_n|\right] < \infty$ for all $n \geq 1$ and

$$\mathsf{E}\left[Z_n \mid Z_{n-1}, Z_{n-2}, \ldots, Z_1\right] = Z_{n-1}; \qquad \text{for all } n \geq 2. \tag{7.57}$$

The name martingale comes from gambling terminology where martingales refer to gambling strategies in which the amount to be bet is determined by the past history of winning or losing. If one visualizes Z_n as representing the gambler's fortune at the end of the n^{th} play, the definition above means, first, that the game is fair (in the sense that the expected increase in fortune from play $n-1$ to n is zero), and, second, that the expected fortune on the n^{th} play depends on the past only through the fortune on play $n-1$.

The important part of the definition of a martingale, and what distinguishes martingales from other kinds of proceesses, is the form of dependence in (7.57). However, the restriction that $\mathsf{E}\left[|Z_n|\right] < \infty$ is also important, particularly since martingales are so abstract and general that one often loses the insight to understand intuitively when this restriction is important. Students are advised to ignore this restriction when first looking at something that might be a martingale, and to check later after acquiring some understanding.

There are two interpretations of (7.57); the first and most straightforward is to view it as shorthand for $\mathsf{E}\left[Z_n \mid Z_{n-1}{=}z_{n-1},\ Z_{n-2}{=}z_{n-2},\ \ldots,\ Z_1{=}z_1\right] = z_{n-1}$ for all possible sample values $z_1, z_2, \ldots, z_{n-1}$. The second is that $\mathsf{E}\left[Z_n \mid Z_{n-1}{=}z_{n-1},\ \ldots,\ Z_1{=}z_1\right]$ is a function of the sample values $z_1, \ldots, z_{n-1}$ and thus $\mathsf{E}\left[Z_n \mid Z_{n-1}, \ldots, Z_1\right]$ is a random variable which is a function of the random variables $Z_1, \ldots, Z_{n-1}$ (and, for a martingale, a function only of Z_{n-1}). Students are encouraged to take the first viewpoint initially and to write out the expanded type of expression in cases of confusion. The second viewpoint, however, is very powerful, and, with experience, is the more useful viewpoint.

It is important to understand the difference between martingales and Markov chains. For the Markov chain $\{X_n;\ n \geq 1\}$, each rv X_n is conditioned on the past only through X_{n-1},

whereas for the martingale $\{Z_n;\, n \geq 1\}$, it is only the expected value of Z_n that is conditioned on the past only through Z_{n-1}. The rv Z_n itself, conditioned on Z_{n-1}, can also be dependent on all the earlier Z_i's. It is very surprising that so many results can be developed using such a weak form of conditioning.

In what follows, we give a number of important examples of martingales, then develop some results about martingales, and then discuss those results in the context of the examples.

7.6.1 Simple examples of martingales

Example 7.6.1 (Random walks). One example of a martingale is a zero-mean random walk, since if $Z_n = X_1 + X_2 + \cdots + X_n$, where the X_i are IID and zero mean, then

$$\begin{aligned} \mathsf{E}\left[Z_n \mid Z_{n-1}, \ldots, Z_1\right] &= \mathsf{E}\left[X_n + Z_{n-1} \mid Z_{n-1}, \ldots, Z_1\right] && (7.58)\\ &= \mathsf{E}\left[X_n\right] + Z_{n-1} \;=\; Z_{n-1}. && (7.59) \end{aligned}$$

Extending this example, suppose that $\{X_i;\, i \geq 1\}$ is an arbitrary sequence of IID random variables with mean $\overline{X}$ and let $\widetilde{X}_i = X_i - \overline{X}$. Then $\{S_n;\, n \geq 1\}$ is a random walk with $S_n = X_1 + \cdots + X_n$ and $\{Z_n;\, n \geq 1\}$ is a martingale with $Z_n = \widetilde{X}_1 + \cdots + \widetilde{X}_n$. The random walk and the martingale are simply related by $Z_n = S_n - n\overline{X}$, and thus general results about martingales can easily be applied to arbitrary random walks.

Example 7.6.2 (Sums of dependent zero-mean variables). Let $\{X_i;\, i \geq 1\}$ be a set of dependent random variables satisfying $\mathsf{E}\left[X_i \mid X_{i-1}, \ldots, X_1\right] = 0$. Then $\{Z_n;\, n \geq 1\}$, where $Z_n = X_1 + \cdots + X_n$, is a zero mean martingale. To see this, note that

$$\begin{aligned} \mathsf{E}\left[Z_n \mid Z_{n-1}, \ldots, Z_1\right] &= \mathsf{E}\left[X_n + Z_{n-1} \mid Z_{n-1}, \ldots, Z_1\right] \\ &= \mathsf{E}\left[X_n \mid X_{n-1}, \ldots, X_1\right] + \mathsf{E}\left[Z_{n-1} \mid Z_{n-1}, \ldots, Z_1\right] = Z_{n-1}. \end{aligned}$$

This is a more general example than it appears, since given any martingale $\{Z_n;\, n \geq 1\}$, we can define $X_n = Z_n - Z_{n-1}$ for $n \geq 2$ and define $X_1 = Z_1$. Then $\mathsf{E}\left[X_n \mid X_{n-1}, \ldots, X_1\right] = 0$ for $n \geq 2$. If the martingale is zero mean (*i.e.*, if $\mathsf{E}\left[Z_1\right] = 0$), then $\mathsf{E}\left[X_1\right] = 0$ also.

Example 7.6.3 (Product-form martingales). Another example is a product of unit mean IID random variables. Thus if $Z_n = X_1 X_2 \ldots X_n$, we have

$$\begin{aligned} \mathsf{E}\left[Z_n \mid Z_{n-1}, \ldots, Z_1\right] &= \mathsf{E}\left[X_n, Z_{n-1} \mid Z_{n-1}, \ldots, Z_1\right] \\ &= \mathsf{E}\left[X_n\right] \mathsf{E}\left[Z_{n-1} \mid Z_{n-1}, \ldots, Z_1\right] && (7.60)\\ &= \mathsf{E}\left[X_n\right] \mathsf{E}\left[Z_{n-1} \mid Z_{n-1}\right] = Z_{n-1}. && (7.61) \end{aligned}$$

A particularly simple case of this product example is where $X_n = 2$ with probability $1/2$ and $X_n = 0$ with probability $1/2$. Then

$$\Pr\{Z_n = 2^n\} = 2^{-n}; \qquad \Pr\{Z_n = 0\} = 1 - 2^{-n}; \qquad \mathsf{E}\left[Z_n\right] = 1. \tag{7.62}$$

Thus $\lim_{n\to\infty} Z_n = 0$ with probability 1, but $\mathsf{E}[Z_n] = 1$ for all n and $\lim_{n\to\infty} \mathsf{E}[Z_n] = 1$. This is an important example to keep in mind when trying to understand why proofs about martingales are necessary and non-trivial. This type of phenomenon will be clarified somewhat by Lemma 7.8.1 when we discussed stopped martingales in Section 7.8.

An important example of a product-form martingale is as follows: let $\{X_i;\, i \geq 1\}$ be an IID sequence, and let $\{S_n = X_1 + \cdots + X_n;\, n \geq 1\}$ be a random walk. Assume that the semi-invariant generating function $\gamma(r) = \ln\{\mathsf{E}[\exp(rX)]\}$ exists in some region of r around 0. For each $n \geq 1$, let Z_n be defined as

$$\begin{aligned} Z_n &= \exp\{rS_n - n\gamma(r)\} && (7.63)\\ &= \exp\{rX_n - \gamma(r)\}\exp\{rS_{n-1} - (n-1)\gamma(r)\} \\ &= \exp\{rX_n - \gamma(r)\}Z_{n-1}. && (7.64)\end{aligned}$$

Taking the conditional expectation of this,

$$\begin{aligned} \mathsf{E}[Z_n \mid Z_{n-1}, \ldots, Z_1] &= \mathsf{E}[\exp(rX_n - \gamma(r))]\ \mathsf{E}[Z_{n-1} \mid Z_{n-1}, \ldots, Z_1] \\ &= Z_{n-1}. && (7.65)\end{aligned}$$

where we have used the fact that $\mathsf{E}[\exp(rX_n)] = \exp(\gamma(r))$. Thus we see that $\{Z_n;\ n \geq 1\}$ is a martingale of the product-form.

7.6.2 Scaled branching processes

A final simple example of a martingale is a "scaled down" version of a branching process $\{X_n;\, n \geq 0\}$. Recall from Section 5.5 that, for each n, X_n is defined as the aggregate number of elements in generation n. Each element i of generation n, $1 \leq i \leq X_n$ has a number $Y_{i,n}$ of offspring which collectively constitute generation $n+1$, *i.e.*, $X_{n+1} = \sum_{i=1}^{X_n} Y_{i,n}$. The rv's $Y_{i,n}$ are IID over both i and n.

Let $\overline{Y} = \mathsf{E}[Y_{i,n}]$ be the mean number of offspring of each element of the population. Then $\mathsf{E}[X_n \mid X_{n-1}] = \overline{Y}X_{n-1}$, which resembles a martingale except for the factor of $\overline{Y}$. We can convert this branching process into a martingale by scaling it, however. That is, define $Z_n = X_n/\overline{Y}^n$. It follows that

$$\mathsf{E}[Z_n \mid Z_{n-1}, \ldots, Z_1] = \mathsf{E}\left[\frac{X_n}{\overline{Y}^n} \mid X_{n-1}, \ldots, X_1\right] = \frac{\overline{Y}X_{n-1}}{\overline{Y}^n} = Z_{n-1}. \qquad (7.66)$$

Thus $\{Z_n;\, n \geq 1\}$ is a martingale. We will see the surprising result later that this implies that Z_n converges with probability 1 to a limiting rv as $n \to \infty$.

7.6.3 Partial isolation of past and future in martingales

Recall that for a Markov chain, the states at all times greater than a given n are independent of the states at all times less than n, conditional on the state at time n. The following lemma shows that at least a small part of this independence of past and future applies to martingales.

Lemma 7.6.1. *Let $\{Z_n;\, n \geq 1\}$ be a martingale. Then for any $n > i \geq 1$,*

$$\mathsf{E}\left[Z_n \mid Z_i, Z_{i-1}, \ldots, Z_1\right] = Z_i. \tag{7.67}$$

Proof: For $n = i+1$, $\mathsf{E}\left[Z_{i+1} \mid Z_i, \ldots, Z_1\right] = Z_i$ by the definition of a martingale. Now consider $n = i+2$. Then $\mathsf{E}\left[Z_{i+2} \mid Z_{i+1}, \ldots, Z_1\right]$ is a rv equal to Z_{i+1}. We can view $\mathsf{E}\left[Z_{i+2} \mid Z_i, \ldots, Z_1\right]$ as the conditional expectation of the rv $\mathsf{E}\left[Z_{i+2} \mid Z_{i+1}, \ldots, Z_1\right]$ over Z_{i+1}, conditional on $Z_i, \ldots, Z_1$.

$$\begin{aligned}\mathsf{E}\left[Z_{i+2} | Z_i, \ldots, Z_1\right] &= \mathsf{E}\left[\mathsf{E}\left[Z_{i+2} \mid Z_{i+1}, Z_i, \ldots, Z_1\right] \mid Z_i, \ldots, Z_1\right] \\ &= \mathsf{E}\left[Z_{i+1} \mid Z_i, \ldots, Z_1\right] = Z_i. \end{aligned} \tag{7.68}$$

For $n = i+3$, (7.68), with i incremented, shows us that the rv $\mathsf{E}\left[Z_{i+3} \mid Z_{i+1}, \ldots, Z_1\right] = Z_{i+1}$. Taking the conditional expectation of this rv over Z_{i+1} conditional on $Z_i, \ldots, Z_1$, we get

$$\mathsf{E}\left[Z_{i+3} \mid Z_i, \ldots, Z_1\right] = Z_i.$$

This argument can be applied successively to any $n > i$. □

This lemma is particularly important for $i = 1$, where it says that $\mathsf{E}\left[Z_n \mid Z_{n-1}, \ldots, Z_1\right] = Z_1$. The left side of this is a rv which is a function (in fact the identity function) of Z_1. Thus, by taking the expected value of each side, we see that

$$\mathsf{E}\left[Z_n\right] = \mathsf{E}\left[Z_1\right] \quad \text{for all } n > 1. \tag{7.69}$$

An important application of this is to the product form martingale in (7.65). This says that

$$\begin{aligned}\mathsf{E}\left[\exp(rS_n - n\gamma(r)\right] &= \mathsf{E}\left[\exp(rX - \gamma(r))\right] \\ &= \mathsf{E}\left[\exp(rX)\right]/g(r) \;=\; 1. \end{aligned} \tag{7.70}$$

We will come back later to relate this to Wald's identity.

7.7 Submartingales and supermartingales

Submartingales and supermartingales are simple generalizations of martingales that provide many useful results for very little additional work. We will subsequently derive the Kolmogorov submartingale inequality, which is a powerful generalization of the Markov inequality. We use this both to give a simple proof of the strong law of large numbers and also to better understand threshold crossing problems for random walks.

Definition 7.7.1. *A submartingale is an integer-time stochastic process $\{Z_n;\, n \geq 1\}$ that satisfies the relations*

$$\mathsf{E}\left[|Z_n|\right] < \infty \quad ; \quad \mathsf{E}\left[Z_n \mid Z_{n-1}, Z_{n-2}, \ldots, Z_1\right] \geq Z_{n-1} \quad ; \quad n \geq 1. \tag{7.71}$$

A supermartingale is an integer-time stochastic process $\{Z_n;\, n \geq 1\}$ that satisfies the relations

$$\mathsf{E}\left[|Z_n|\right] < \infty \quad ; \quad \mathsf{E}\left[Z_n \mid Z_{n-1}, Z_{n-2}, \ldots, Z_1\right] \leq Z_{n-1} \quad ; \quad n \geq 1. \tag{7.72}$$

In terms of our gambling analogy, a submartingale corresponds to a game that is at least fair, *i.e.*, where the expected fortune of the gambler either increases or remains the same. A *supermartingale* is a process with the opposite type of inequality.[18]

Since a martingale satisfies both (7.71) and (7.72) with equality, a martingale is both a submartingale and a supermartingale. Note that if $\{Z_n;\, n \geq 1\}$ is a submartingale, then $\{-Z_n;\, n \geq 1\}$ is a supermartingale, and conversely. Thus, some of the results to follow are stated only for submartingales, with the understanding that they can be applied to supermartingales by changing signs as above.

Lemma 7.6.1, with the equality replaced by inequality, also applies to submartingales and supermartingales. That is, if $\{Z_n;\, n \geq 1\}$ is a submartingale, then

$$\mathsf{E}\left[Z_n \mid Z_i, Z_{i-1}, \ldots, Z_1\right] \geq Z_i \quad ; \quad 1 \leq i < n, \tag{7.73}$$

and if $\{Z_n;\, n \geq 1\}$ is a supermartingale, then

$$\mathsf{E}\left[Z_n \mid Z_i, Z_{i-1}, \ldots, Z_1\right] \leq Z_i \quad ; \quad 1 \leq i < n. \tag{7.74}$$

Equations (7.73) and (7.74) are verified in the same way as Lemma 7.6.1 (see Exercise 7.18). Similarly, the appropriate generalization of (7.69) is that if $\{Z_n;\, n \geq 1\}$ is a submartingale, then

$$\mathsf{E}\left[Z_n\right] \geq \mathsf{E}\left[Z_i\right] \quad ; \quad \text{for all } i,\ 1 \leq i < n. \tag{7.75}$$

and if $\{Z_n;\, n \geq 1\}$ is a supermartingale, then

$$\mathsf{E}\left[Z_n\right] \leq \mathsf{E}\left[Z_i\right] \quad ; \quad \text{for all } i,\ 1 \leq i < n. \tag{7.76}$$

A random walk $\{S_n;\, n \geq 1\}$ with $S_n = X_1 + \cdots + X_n$ is a submartingale, martingale, or supermartingale respectively for $\overline{X} \geq 0$, $\overline{X} = 0$, or $\overline{X} \leq 0$. Also, if X has a semi-invariant moment generating function $\gamma(r)$ for some given r, and if Z_n is defined as $Z_n = \exp(rS_n)$, then the process $\{Z_n;\, n \geq 1\}$ is a submartingale, martingale, or supermartingale respectively for $\gamma(r) \geq 0$, $\gamma(r) = 0$, or $\gamma(r) \leq 0$. The next example gives an important way in which martingales and submartingales are related.

Example 7.7.1 (Convex functions of martingales). Figure 7.11 illustrates the graph of a convex function h of a real variable x. A function h of a real variable is defined to be *convex* if, for each point x_1, there is a real number c with the property that $h(x_1) + c(x - x_1) \leq h(x)$ for all x.

Geometrically, this says that every tangent to $h(x)$ lies on or below $h(x)$. If $h(x)$ has a derivative at x_1, then c is the value of that derivative and $h(x_1) + c(x - x_1)$ is the tangent line at x_1. If $h(x)$ has a discontinuous slope at x_1, then there might be many choices for c; for example, $h(x) = |x|$ is convex, and for $x_1 = 0$, one could choose any c in the range -1 to $+1$.

A simple condition that implies convexity is a nonnegative second derivative everywhere. This is not necessary, however, and functions can be convex even when the first derivative

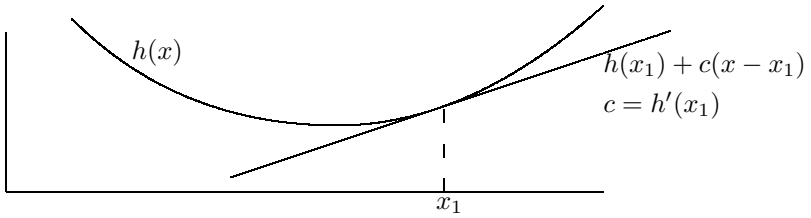


Figure 7.11: Convex functions: For each x_1, there is a value of c such that, for all x, $h(x_1) + c(x - x_1) \leq h(x)$. If h is differentiable at x_1, then c is the derivative of h at x_1.

does not exist everywhere. For example, the function $\gamma(r)$ in Figure 7.8 is convex even though it it is finite at $r = r_+$ and infinite for all $r > r_+$.

Jensen's inequality states that if h is convex and X is a random variable with an expectation, then $h(\mathsf{E}[X]) \leq \mathsf{E}[h(X)]$. To prove this, let $x_1 = \mathsf{E}[X]$ and choose c so that $h(x_1) + c(x - x_1) \leq h(x)$. Using the random variable X in place of x and taking expected values of both sides, we get Jensen's inequality. Note that for any particular event A, this same argument applies to X conditional on A, so that $h(\mathsf{E}[X \mid A]) \leq \mathsf{E}[h(X) \mid A]$. Jensen's inequality is very widely used; it is a minor miracle that we have not required it previously.

Theorem 7.7.1. *Assume that h is a convex function of a real variable, that $\{Z_n; n \geq 1\}$ is a martingale or submartingale, and that $\mathsf{E}[|h(Z_n)|] < \infty$ for all n. Then $\{h(Z_n); n \geq 1\}$ is a submartingale.*

Proof: For any choice of $z_1, \ldots, z_{n-1}$, we can use Jensen's inequality with the conditioning probabilities to get

$$\mathsf{E}[h(Z_n)|Z_{n-1}{=}z_{n-1}, \ldots, Z_1{=}z_1] \geq h(\mathsf{E}[Z_n \mid Z_{n-1}{=}z_{n-1}, \ldots, Z_1{=}z_1]) = h(z_{n-1}). \quad (7.77)$$

For any choice of numbers $h_1, \ldots, h_{n-1}$ in the range of the function h, let $z_1, \ldots, z_{n-1}$ be arbitrary numbers satisfying $h(z_1){=}h_1, \ldots, h(z_{n-1}){=}h_{n-1}$. For each such choice, (7.77) holds, so that

$$\begin{aligned} \mathsf{E}[h(Z_n) \mid h(Z_{n-1}){=}h_{n-1}, \ldots, h(Z_1){=}h_1] &\geq h(\mathsf{E}[Z_n \mid h(Z_{n-1}){=}h_{n-1}, \ldots, h(Z_1){=}h_1]) \\ &= h(z_{n-1}) = h_{n-1}. \end{aligned} \quad (7.78)$$

completing the proof. □

Some examples of this result, applied to a martingale $\{Z_n; n \geq 1\}$, are as follows:

$$\{|Z_n|; n \geq 1\} \text{ is a submartingale} \quad (7.79)$$

$$\{Z_n^2; n \geq 1\} \text{ is a submartingale if } \mathsf{E}\left[Z_n^2\right] < \infty \quad (7.80)$$

$$\{\exp(rZ_n); n \geq 1\} \text{ is a submartingale for } r \text{ such that } \mathsf{E}[\exp(rZ_n)] < \infty. \quad (7.81)$$

[18]The easiest way to remember the difference between a submartingale and a supermartingale is to remember that it is contrary to what common sense would dictate.

A function of a real variable $h(x)$ is defined to be concave if $-h(x)$ is convex. It then follows from Theorem 7.7.1 that if h is concave and $\{Z_n; n \geq 1\}$ is a martingale, then $\{h(Z_n); n \geq 1\}$ is a supermartingale (assuming that $\mathsf{E}[|h(Z_n)|] < \infty$). For example, if $\{Z_n; n \geq 1\}$ is a positive martingale and $\mathsf{E}[|\ln(Z_n)|] < \infty$, then $\{\ln(Z_n); n \geq 1\}$ is a supermartingale.

7.8 Stopped processes and stopping trials

The definition of stopping trials in Section 4.5 applies to arbitrary integer-time processes $\{Z_n; n \geq 1\}$ as well as to IID sequences. Recall that J is a stopping trial for a sequence $\{Z_n; n \geq 1\}$ of rv's if $\mathbb{I}_{J=n}$ is a function of $Z_1, \ldots, Z_n$ and if J is a rv.

If $\mathbb{I}_{J=n}$ is a function of $Z_1, \ldots, Z_n$ and J is a defective rv, then J is called a defective stopping trial. For some of the results to follow, it is unimportant whether J is a random variable or a defective random variable (*i.e.*, whether or not the process stops with probability 1). If it is not specified whether J is a random variable or a defective random variable, we refer to the stopping trial as a *possibly-defective stopping trial*; we consider J to take on the value ∞ if the process does not stop.

Definition 7.8.1. *A stopped process $\{Z_n^*; n \geq 1\}$ for a possibly-defective stopping trial J relative to a process $\{Z_n; n \geq 1\}$ is the process for which $Z_n^* = Z_n$ for $n \leq J$ and $Z_n^* = Z_J$ for $n > J$.*

As an example, suppose Z_n models the fortune of a gambler at the completion of the nth trial of some game, and suppose the gambler then modifies the game by deciding to stop gambling under some given circumstances (*i.e.*, at the stopping trial). Thus, after stopping, the fortune remains constant, so the stopped process models the gambler's fortune in time, including the effect of the stopping trial.

As another example, consider a random walk with a positive and negative threshold, and consider the process to stop after reaching or crossing a threshold. The stopped process then stays at that point beyond the threshold as an artifice to simplify analysis. The use of stopped processes is similar to the artifice that we employed in Section 3.5 for first-passage times in Markov chains; recall that we added an artificial trapping state after the desired passage to simplify analysis.

We next show that the possibly-defective stopped process of a martingale is itself a martingale; the intuitive reason is that, before stopping, the stopped process is the same as the martingale, and, after stopping, $Z_n^* = Z_{n-1}^*$. The following theorem establishes this and the corresponding results for submartingales and supermartingales.

Theorem 7.8.1. *Given a stochastic process $\{Z_n; n \geq 1\}$ and a possibly-defective stopping trial J for the process, the stopped process $\{Z_n^*; n \geq 1\}$ is a submartingale if $\{Z_n; n \geq 1\}$ is a submartingale, is a martingale if $\{Z_n; n \geq 1\}$ is a martingale, and is a supermartingale if $\{Z_n; n \geq 1\}$ is a supermartingale.*

Proof: First we show that, for all three cases, the stopped process satisfies $\mathsf{E}[|Z_n^*|] < \infty$ for any given $n \geq 1$. Conditional on $J = i$ for some $i < n$, we have $Z_n^* = Z_i$, so

$$\mathsf{E}[|Z_n^*| \mid J = i] = \mathsf{E}[|Z_i| \mid J = i] < \infty \text{ for each } i < n \text{ such that } \Pr\{J = i\} > 0.$$

The reason for this is that if $\mathsf{E}[|Z_i| \mid J = i] = \infty$ and $\Pr\{J = i\} > 0$, then $\mathsf{E}[|Z_i|] = \infty$, contrary to the assumption that $\{Z_n; n \geq 1\}$ is a martingale, submartingale, or supermartingale. Similarly, for $J \geq n$, we have $Z_n^* = Z_n$ so

$$\mathsf{E}[|Z_n^*| \mid J \geq n] = \mathsf{E}[|Z_n| \mid J \geq n] < \infty \quad \text{if } \Pr\{J \geq n\} > 0.$$

Averaging,

$$\mathsf{E}[|Z_n^*|] = \sum_{i=1}^{n-1} \mathsf{E}[|Z_n^*| \mid J{=}i] \Pr\{J{=}i\} + \mathsf{E}[|Z_n^*| \mid J \geq n] \Pr\{J \geq n\} < \infty.$$

Next assume that $\{Z_n; n \geq 1\}$ is a submartingale. For any given $n > 1$, consider an arbitrary initial sample sequence $(Z_1 = z_1, Z_2 = z_2, \ldots, Z_{n-1} = z_{n-1})$. Note that z_1 specifies whether or not $J = 1$. Similarly, (z_1, z_2) specifies whether or not $J = 2$, and so forth up to $(z_1, \ldots, z_{n-1})$, which specifies whether or not $J = n - 1$. Thus $(z_1, \ldots, z_{n-1})$ specifies the sample value of J for $J \leq n - 1$ and specifies that $J \geq n$ otherwise.

For $(z_1, \ldots, z_{n-1})$ such that $\mathbb{I}_{J \geq n} < n$, we have $z_n^* = z_{n-1}^*$. For all such sample values,

$$\mathsf{E}\left[Z_n^* | Z_{n-1}^* {=} z_{n-1}^*, \ldots, Z_1^* {=} z_1^*\right] = z_{n-1}^*. \tag{7.82}$$

For the remaining case, where $(z_1, \ldots, z_{n-1})$ is such that $\mathbb{I}_{J \geq n} \geq n$, we have $z_n^* = z_n$. Thus

$$\mathsf{E}\left[Z_n^* | Z_{n-1}^* {=} z_{n-1}^*, \ldots, Z_1^* {=} z_1^*\right] \geq z_{n-1}^*. \tag{7.83}$$

The same argument works for martingales and supermartingales by replacing the inequality in (7.83) by equality for the martingale case and the opposite inequality for the supermartingale case. □

Theorem 7.8.2. Given a stochastic process $\{Z_n; n \geq 1\}$ and a possibly-defective stopping trial J for the process, the stopped process $\{Z_n^*; n \geq 1\}$ satisfies the following conditions for all $n \geq 1$ if $\{Z_n; n \geq 1\}$ is a submartingale, martingale, or supermartingale respectively:

$$\mathsf{E}[Z_1] \leq \mathsf{E}[Z_n^*] \leq \mathsf{E}[Z_n] \qquad \text{(submartingale)} \tag{7.84}$$

$$\mathsf{E}[Z_1] = \mathsf{E}[Z_n^*] = \mathsf{E}[Z_n] \qquad \text{(martingale)} \tag{7.85}$$

$$\mathsf{E}[Z_1] \geq \mathsf{E}[Z_n^*] \geq \mathsf{E}[Z_n] \qquad \text{(supermartingale).} \tag{7.86}$$

Proof: Since a process cannot stop before epoch 1, $Z_1 = Z_1^*$ in all cases. First consider the case in which $\{Z_n; n \geq 1\}$ is a submartingale. Theorem 7.8.1 shows that $\{Z_n^*; n \geq 1\}$ is a submartingale, and from (7.75), $\mathsf{E}[Z_1] \leq \mathsf{E}[Z_n^*]$ for all $n \geq 1$. This establishes the first half of (7.84) and we next prove the second half. First condition on the set of sequences for which the initial segment $(z_1, \ldots, z_i)$ specifies that $\mathbb{I}_{J \geq n} = i$ for any given $i < n$. Then $\mathsf{E}[Z_n^*] = z_i$. From (7.73), $\mathsf{E}[Z_n] \geq z_i$, proving (7.84) for this case. For those sequences not

having such an initial segment, $Z_n^* = Z_n$, establishing (7.84) in that case. Averaging over these two cases gives (7.84) in general.

Finally, if $\{Z_n;\, n \geq 1\}$ is a supermartingale, then $\{-Z_n;\, n \geq 1\}$ is a submartingale, verifying (7.86). Since a martingale is both a submartingale and supermartingale, (7.85) follows and the proof is complete. □

Consider a (non-defective) stopping trial J for a martingale $\{Z_n;\, n \geq 1\}$. Since the stopped process is also a martingale, we have

$$\mathsf{E}\left[Z_n^*\right] = \mathsf{E}\left[Z_1^*\right] = \mathsf{E}\left[Z_1\right];\ n \geq 1. \tag{7.87}$$

Since $Z_n^* = Z_J$ for all $n \geq J$ and since J is finite with probability 1, we see that $\lim_{n\to\infty} Z_n^* = Z_J$ with probability 1. Unfortunately, in general, $\mathsf{E}[Z_J]$ is unequal to $\lim_{n\to\infty} \mathsf{E}[Z_n^*] = \mathsf{E}[Z_1]$. An example in which this occurs is the binary product martingale in (7.62). Taking the stopping trial J to be the smallest n for which $Z_n = 0$, we have $Z_J = 0$ with probability 1, and thus $\mathsf{E}[Z_J] = 0$. But $Z_n^* = Z_n$ for all n, and $\mathsf{E}[Z_n^*] = 1$ for all n. The problem here is that, given that the process has not stopped by time n, Z_n and Z_n^* each have the value 2^n. Fortunately, in most situations, this type of bizarre behavior does not occur and $\mathsf{E}[Z_J] = \mathsf{E}[Z_1]$. To get a better understanding of when $\mathsf{E}[Z_J] = \mathsf{E}[Z_1]$, note that for any n, we have

$$\mathsf{E}\left[Z_n^*\right] = \sum_{i=1}^{n} \mathsf{E}\left[Z_n^* \mid J=i\right] \Pr\{J=i\} + \mathsf{E}\left[Z_n^* \mid J>n\right] \Pr\{J>n\} \tag{7.88}$$

$$= \sum_{i=1}^{n} \mathsf{E}\left[Z_J \mid J=i\right] \Pr\{J=i\} + \mathsf{E}\left[Z_n \mid J>n\right] \Pr\{J>n\}. \tag{7.89}$$

The left side of this equation is $\mathsf{E}[Z_1]$ for all n. If the final term on the right converges to 0 as $n \to \infty$, then the sum must converge to $\mathsf{E}[Z_1]$. If $\mathsf{E}[|Z_J|] < \infty$, then the sum also converges to $\mathsf{E}[Z_J]$. Without the condition $\mathsf{E}[|Z_J|] < \infty$, the sum might consist of alternating terms which converge, but whose absolute values do not converge, in which case $\mathsf{E}[Z_J]$ does not exist (see Exercise 7.21 for an example). Thus we have established the following lemma.

Lemma 7.8.1. Let J be a stopping trial for a martingale $\{Z_n;\, n \geq 1\}$. Then $\mathsf{E}[Z_J] = \mathsf{E}[Z_1]$ if and only if

$$\lim_{n\to\infty} \mathsf{E}\left[Z_n \mid J>n\right] \Pr\{J>n\} = 0 \qquad \text{and} \quad \mathsf{E}\left[|Z_J|\right] < \infty. \tag{7.90}$$

Example 7.8.1 (Random walks with thresholds). Recall the generating function product martingale of (7.63) in which $\{Z_n = \exp[rS_n - n\gamma(r)];\, n \geq 1\}$ is a martingale defined in terms of the random walk $\{S_n = X_1 + \cdots + X_n;\, n \geq 1\}$. From (7.85), we have $\mathsf{E}[Z_n] = \mathsf{E}[Z_1]$, and since $\mathsf{E}[Z_1] = \mathsf{E}[\exp\{rX_1 - \gamma(r)\}] = 1$, we have $\mathsf{E}[Z_n] = 1$ for all n. Also, for any possibly-defective stopping trial J, we have $\mathsf{E}[Z_n^*] = \mathsf{E}[Z_1] = 1$. If J is a non-defective stopping trial, and if (7.90) holds, then

$$\mathsf{E}\left[Z_J\right] = \mathsf{E}\left[\exp\{rS_J - J\gamma(r)\}\right] = 1. \tag{7.91}$$

If there are two thresholds, one at $\alpha > 0$, and the other at $\beta < 0$, and the stopping rule is to stop when either threshold is crossed, then (7.91) is just the Wald identity, (7.34).

The nice part about the approach here is that it also applies naturally to other stopping rules. For example, for some given integer n, let J_{n+} be the smallest integer $i \geq n$ for which $S_i \geq \alpha$ or $S_i \leq \beta$. Then, in the limit $\beta \to -\infty$, $\Pr\{S_{J_{n+}} \geq \alpha\} = \Pr\{\cup_{i=n}^{\infty}(S_i \geq \alpha)\}$. Assuming $\overline{X} < 0$, we can find an upper bound to $\Pr\{S_{J_{n+}} \geq \alpha\}$ for any $r > 0$ and $\gamma(r) \leq 0$ (*i.e.*, for $0 < r \leq r^*$) by the following steps

$$1 = \mathsf{E}\left[\exp\{rS_{J_{n+}} - J_{n+}\gamma(r)\}\right] \geq \Pr\{S_{J_{n+}} \geq \alpha\} \exp[r\alpha - n\gamma(r)]$$
$$\Pr\{S_{J_{n+}} \geq \alpha\} \leq \exp[-r\alpha + n\gamma(r)]; \qquad 0 \leq r \leq r^*. \tag{7.92}$$

This is almost the same result as (7.54), except that it is slightly stronger; (7.54) bounded the probability that the *first* threshold crossing crossed α at some epoch $i \geq n$, whereas this includes the possibility that $S_m \geq \alpha$ and $S_i \geq \alpha$ for some $m < n \leq i$.

7.9 The Kolmogorov inequalities

We now use the previous theorems to establish Kolmogorov's submartingale inequality, which is a major strengthening of the Markov inequality. Just as the Markov inequality in Section 1.7 was used to derive the Chebychev inequality and then the weak law of large numbers, the Kolmogorov submartingale inequality will be used to strengthen the Chebychev inequality, from which the strong law of large numbers will follow.

Theorem 7.9.1 (Kolmogorov's submartingale inequality). *Let $\{Z_n;\, n \geq 1\}$ be a nonnegative submartingale. Then for any positive integer m and any $a > 0$,*

$$\Pr\left\{\max_{1\leq i\leq m} Z_i \geq a\right\} \leq \frac{\mathsf{E}[Z_m]}{a}. \tag{7.93}$$

Proof: Given a nonnegative submartingale $\{Z_n;\, n \geq 1\}$, given $a > 0$, and given a positive integer m, let J be the stopping trial defined as the smallest $n \leq m$ such that $Z_n \geq a$. If $Z_n < a$ for all $n \leq m$, then $J = m$. Thus the process must stop by time m, and $Z_J \geq a$ if and only if $Z_n \geq a$ for some $n \leq m$. Thus

$$\Pr\left\{\max_{1\leq n\leq m} Z_n \geq a\right\} = \Pr\{Z_J \geq a\} \leq \frac{\mathsf{E}[Z_J]}{a}. \tag{7.94}$$

where we have used the Markov inequality. Finally, since the process must be stopped by time m, we have $Z_J = Z_m^*$. From (7.84), $\mathsf{E}[Z_m^*] \leq \mathsf{E}[Z_m]$, so the right hand side of (7.94) is less than or equal to $\mathsf{E}[Z_m]/a$, completing the proof. □

The following simple corollary shows that (7.93) takes a simpler form for nonnegative martingales.

Corollary 7.9.1 (Nonnegative martingale inequality). *Let $\{Z_n;\, n \geq 1\}$ be a nonnegative martingale. Then*

$$\Pr\left\{\sup_{n\geq 1} Z_n \geq a\right\} \leq \frac{\mathsf{E}[Z_1]}{a}; \qquad \textit{for all } a > 0. \tag{7.95}$$

Proof For a martingale, $\mathsf{E}[Z_m] = \mathsf{E}[Z_1]$. Thus, from (7.93), $\Pr\{\max_{1\le i\le m} Z_i \ge a\} \le \frac{\mathsf{E}[Z_1]}{a}$ for all $m > 1$. Passing to the limit $m \to \infty$ essentially yields (7.95). Exercise 7.22 illustrates why the limiting operation here is a little tricky, and shows that it is valid. □

The following corollary bears the same relationship to the submartingale inequality as the Chebychev inequality does to the Markov inequality.

Corollary 7.9.2 (Kolmogorov's martingale inequality). *Let $\{Z_n; n \ge 1\}$ be a martingale with $\mathsf{E}\left[Z_n^2\right] < \infty$ for all $n \ge 1$. Then*

$$\Pr\left\{\max_{1\le n\le m} |Z_n| \ge b\right\} \le \frac{\mathsf{E}\left[Z_m^2\right]}{b^2}; \quad \text{for all integer } m \ge 2, \text{all } b > 0. \tag{7.96}$$

Proof: Since $\{Z_n; n \ge 1\}$ is a martingale and Z_n^2 is a convex function of Z_n, it follows from Theorem 7.7.1 that $\{Z_n^2; n \ge 1\}$ is a submartingale. Since Z_n^2 is nonnegative, we can use the Kolmogorov submartingale inequality to see that

$$\Pr\left\{\max_{n\le m} Z_n^2 \ge a\right\} \le \mathsf{E}\left[Z_m^2\right]/a \quad \text{for any } a > 0.$$

Substituting b^2 for a, we get (7.96). □

Corollary 7.9.3 (Kolmogorov's random walk inequality). *Let $\{S_n; n \ge 1\}$ be a random walk with $S_n = X_1 + \cdots + X_n$ where $\{X_i; i \ge i\}$ is a set of IID random variables with mean $\overline{X}$ and variance σ^2. Then for any positive integer m and any $\epsilon > 0$,*

$$\Pr\left\{\max_{1\le n\le m} |S_n - n\overline{X}| \ge m\epsilon\right\} \le \frac{\sigma^2}{m\epsilon^2}. \tag{7.97}$$

Proof: $\{Z_n = S_n - n\overline{X}; n \ge 1\}$ is a zero mean random walk, and thus a martingale. Since $\mathsf{E}\left[Z_m^2\right] = m\sigma^2$, (7.97) follows by substituting $m\epsilon$ for b in 7.96). □

Recall that the simplest form of the weak law of large numbers was given in (1.75) as $\Pr\{|S_m/m - \overline{X}| \ge \epsilon\} \le \sigma^2/(m\epsilon^2)$. This is strengthened in (7.97) to upper bound the probability that any of the first m terms deviate from the mean by more than $m\epsilon$. It is this strengthening that will allow us to prove the strong law of large numbers assuming only a finite variance.

The following corollary yields essentially the same result as (7.56), but is included here as another example of the use of the Kolmogorov submartingale inequality.

Corollary 7.9.4. *Let $\{S_n; n \ge 1\}$ be a random walk, $S_n = X_1 + \cdots + X_n$ where each X_i has mean $\overline{X} < 0$ and semi-invariant moment generating function $\gamma(r)$. For any $r > 0$ such that $0 < \gamma(r) < \infty$ (*i.e.*, for $r > r^*$), and for any $a > 0$.*

$$\Pr\left\{\max_{1\le i\le n} S_i \ge \alpha\right\} \le \exp\{-r\alpha + n\gamma(r)\}. \tag{7.98}$$

Proof: For $r > r^*$, $\{\exp(rS_n);\, n \geq 1\}$ is a submartingale. Taking a $= \exp(r\alpha)$ in (7.93), we get (7.98). □

The following theorem about supermartingales is, in a sense, the dual of the Kolmogorov submartingale inequality. Note, however, that it applies to the terms $n \geq m$ in the supermartingale rather than $n \leq m$.

Theorem 7.9.2. *Let $\{Z_n;\, n \geq 1\}$ be a nonnegative supermartingale. Then for any positive integer m and any $a > 0$,*

$$\Pr\left\{\bigcup_{i\geq m} \{Z_i \geq a\}\right\} \leq \frac{\mathsf{E}[Z_m]}{a}. \tag{7.99}$$

Proof: For given $m \geq 1$ and $a > 0$, let J be a possibly-defective stopping trial defined as the smallest $i \geq m$ for which $Z_i \geq a$. Let $\{Z_n^*;\, n \geq 1\}$ be the corresponding stopped process, which is also nonnegative and is a supermartingale from Theorem 7.8.1. For any $k > m$, note that $Z_k^* \geq a$ iff $\max_{m\leq i\leq k} Z_i \geq a$. Thus

$$\Pr\left\{\max_{m\leq i\leq k} Z_i \geq a\right\} = \Pr\{Z_k^* \geq a\} \leq \frac{\mathsf{E}[Z_k^*]}{a}.$$

Since $\{Z_n^*;\, n \geq 1\}$ is a supermartingale, (7.76) shows that $\mathsf{E}[Z_k^*] \leq \mathsf{E}[Z_m^*]$. On the other hand, $Z_m^* = Z_m$ since the process can not stop before epoch m. Thus $\Pr\{\max_{m\leq i\leq k} Z_i \geq a\}$ is at most $\mathsf{E}[Z_m]/a$. Since k is arbitrary, we can pass to the limit, getting (7.99) and completing the proof. □

7.9.1 The strong law of large numbers (SLLN)

We now proceed to prove the strong law of large numbers assuming only a second moment. Recall that we proved the SLLN under the assumption of a finite fourth moment in Section 4.2.1. Here we use the Kolmogorov martingale inequality to show that only a second moment is required. The theorem is also true assuming only a first absolute moment, but the truncation argument we used for the weak law in Theorem 1.5.3 does not carry over simply here.

Theorem 7.9.3 (SLLN). *Let $\{X_i;\, i \geq 1\}$ be a sequence of IID random variables with mean $\overline{X}$ and standard deviation $\sigma < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then for any $\epsilon > 0$,*

$$\Pr\left\{\lim_{n\to\infty} \frac{S_n}{n} = \overline{X}\right\} = 1 \qquad \text{and} \tag{7.100}$$

$$\lim_{n\to\infty} \Pr\left\{\bigcup_{m>n} \left|\frac{S_m}{m} - \overline{X}\right| > \epsilon\right\} = 0. \tag{7.101}$$

Proof: From Section 4.2.2, (7.100) and (7.101) are equivalent, so we establish (7.101). As n increases, successive terms are droppped out of the union above, so the probability is

non-increasing with n. Thus we can restrict attention to n of the form 2^k for integer k. For any given k, the union above can be separated into blocks as follows:

$$\begin{aligned}
\Pr\left\{\bigcup_{m>2^k}\left\{\left|\frac{S_m}{m}-\overline{X}\right|>\epsilon\right\}\right\} &= \Pr\left\{\bigcup_{j=k}^{\infty}\bigcup_{m=2^j+1}^{2^{j+1}}\left\{\left|\frac{S_m}{m}-\overline{X}\right|>\epsilon\right\}\right\} \\
&\le \sum_{j=k}^{\infty}\Pr\left\{\bigcup_{m=2^j+1}^{2^{j+1}}\left\{\left|\frac{S_m}{m}-\overline{X}\right|>\epsilon\right\}\right\} \qquad (7.102)\\
&= \sum_{j=k}^{\infty}\Pr\left\{\bigcup_{m=2^j+1}^{2^{j+1}}\left\{|S_m-m\overline{X}|>\epsilon m\right\}\right\} \\
&\le \sum_{j=k}^{\infty}\Pr\left\{\bigcup_{m=2^j+1}^{2^{j+1}}\left\{|S_m-m\overline{X}|>\epsilon 2^j\right\}\right\} \qquad (7.103)\\
&= \sum_{j=k}^{\infty}\Pr\left\{\max_{2^j+1\le m\le 2^{j+1}}|S_m-m\overline{X}|>\epsilon 2^j\right\} \\
&\le \sum_{j=k}^{\infty}\Pr\left\{\max_{1\le m\le 2^{j+1}}|S_m-m\overline{X}|>\epsilon 2^j\right\} \qquad (7.104)\\
&\le \sum_{j=k}^{\infty}\frac{2^{j+1}\sigma^2}{\epsilon^2 2^{2j}} = \frac{2^{-k+2}\sigma^2}{\epsilon^2}. \qquad (7.105)
\end{aligned}$$

In (7.102), we used the union bound on the union over j. In (7.103), we used the fact that $m \ge 2^j$ to increase the size of the sets in the union. In (7.104), we upper bounded by adding additional terms into the maximization, and in (7.105) we used the Kolmogorov martingale inequality. The proof is completed by noting that the upper bound in (7.105) goes to 0 with increasing k. □

It should be noted that, although the proof consists of a large number of steps, the steps are all quite small and familiar. Basically the proof is a slick way of upper bounding the probability of the high-order terms in a sequence by using the Kolmogorov martingale inequality, which bounds the low-order terms.

7.9.2 The martingale convergence theorem

Another famous result that follows from the Kolmogorov submartingale inequality is the martingale convergence theorem. This states that if a martingale $\{Z_n; n \ge 1\}$ has the property that there is some finite M such that $\mathsf{E}[|Z_n|] \le M$ for all n, then $\lim_{n\to\infty} Z_n$ exists (and is finite) with probability 1. This is a powerful theorem in more advanced work, but it is not quite as useful as it appears, since the restriction $\mathsf{E}[|Z_n|] \le M$ is more than a technical restriction; for example it is not satisfied by a zero-mean random walk. We prove the theorem with the additional restriction that there is some finite M such that $\mathsf{E}\left[Z_n^2\right] \le M$ for all n.

Theorem 7.9.4 (Martingale convergence theorem). *Let $\{Z_n; n \geq 1\}$ be a martingale and assume that there is some finite M such that $\mathsf{E}\left[Z_n^2\right] \leq M$ for all n. Then there is a random variable Z such that, for all sample sequences except a set of probability 0, $\lim_{n\to\infty} Z_n = Z$.*

Proof*: From Theorem 7.7.1 and the assumption that $\mathsf{E}\left[Z_n^2\right] \leq M$, $\{Z_n^2; n \geq 1\}$ is a submartingale. Thus, from (7.75), $\mathsf{E}\left[Z_n^2\right]$ is nondecreasing in n, and since $\mathsf{E}\left[Z_n^2\right]$ is bounded, $\lim_{n\to\infty} \mathsf{E}\left[Z_n^2\right] = M'$ for some $M' \leq M$. For any integer k, the process $\{Y_n = Z_{k+n} - Z_k; n \geq 1\}$ is a zero mean martingale (see Exercise 7.29). Thus from Kolmogorov's martingale inequality,

$$\Pr\left\{\max_{1\leq n\leq m} |Z_{k+n} - Z_k| \geq b\right\} \leq \mathsf{E}\left[(Z_{k+m} - Z_k)^2\right]/b^2. \tag{7.106}$$

Next, observe that $\mathsf{E}\left[Z_{k+m}Z_k \mid Z_k = z_k, Z_{k-1} = z_{k-1}, \ldots, Z_1 = z_1\right] = z_k^2$, and therefore, $\mathsf{E}\left[Z_{k+m}Z_k\right] = \mathsf{E}\left[Z_k^2\right]$. Thus $\mathsf{E}\left[(Z_{k+m} - Z_k)^2\right] = \mathsf{E}\left[Z_{k+m}^2\right] - \mathsf{E}\left[Z_k^2\right] \leq M' - \mathsf{E}\left[Z_k^2\right]$. Since this is independent of m, we can pass to the limit, obtaining

$$\Pr\left\{\sup_{n\geq 1} |Z_{k+n} - Z_k| \geq b\right\} \leq \frac{M' - \mathsf{E}\left[Z_k^2\right]}{b^2}. \tag{7.107}$$

Since $\lim_{k\to\infty} \mathsf{E}\left[Z_k^2\right] = M'$, we then have, for all $b > 0$,

$$\lim_{k\to\infty} \Pr\left\{\sup_{n\geq 1} |Z_{k+n} - Z_k| \geq b\right\} = 0. \tag{7.108}$$

This means that with probability 1, a sample sequence of $\{Z_n; n \geq 1\}$ is a Cauchy sequence, and thus approaches a limit, concluding the proof. □

This result can be relatively easily interpreted for branching processes. For a branching process $\{X_n; n \geq 1\}$ where $\overline{Y}$ is the expected number of offspring of an individual, $\{X_n/\overline{Y}^n; n \geq 1\}$ is a martingale that satisfies the above conditions. If $\overline{Y} \leq 1$, the branching process dies out with probability 1, so $X_n/\overline{Y}^n$ approaches 0 with probability 1. For $\overline{Y} > 1$, however, the branching process dies out with some probability less than 1 and approaches ∞ otherwise. Thus, the limiting random variable Z is 0 with the probability that the process ultimately dies out, and is positive otherwise. In the latter case, for large n, the interpretation is that when the population is very large, a law of large numbers effect controls its growth in each successive generation, so that $X_n/\overline{Y}^n$ tends to change in a random way for small n, and then changes increasingly little as n increases.

7.10 Markov modulated random walks

Frequently it is useful to generalize random walks to allow some dependence between the variables being summed. The particular form of dependence here is the same as the Markov reward processes of Section 3.5. The treatment in Section 3.5 discussed only expected

rewards, whereas the treatment here focuses on the random variables themselves. Let $\{Y_m;\, m \geq 0\}$ be a sequence of (possibly dependent) rv's, and let

$$\{S_n;\, n \geq 1\} \quad \text{where } S_n = \sum_{m=0}^{n-1} Y_m. \tag{7.109}$$

be the process of successive sums of these random variables. Let $\{X_n;\, n \geq 0\}$ be a Markov chain, and assume that each Y_n can depend on X_n and X_{n+1}. Conditional on X_n and X_{n+1}, however, Y_n is independent of $Y_{n-1}, \ldots, Y_1$, and of X_i for all $i \neq n, n-1$. Assume that Y_n, conditional on X_n and X_{n+1} has a distribution function $\mathsf{F}_{ij}(y) = \Pr\{Y_n \leq y \mid X_n = i, X_{n+1} = j\}$. Thus each rv Y_n depends only on the associated transition in the Markov chain, and this dependence is the same for all n.

The process $\{S_n;\, n \geq 1\}$ is called a *Markov modulated random walk*. If each Y_m is positive, it is also the sequence of epochs in a semi-Markov process. For each m, Y_m is associated with the transition in the Markov chain from time m to $m+1$, and S_n is the aggregate reward up to but not including time n. Let $\overline{Y}_{ij}$ denote $\mathsf{E}[Y_n \mid X_n = i, X_{n+1} = j]$ and $\overline{Y}_i$ denote $\mathsf{E}[Y_n \mid X_n = i]$. Let $\{P_{ij}\}$ be the set of transition probabilities for the Markov chain, so $\overline{Y}_i = \sum_j P_{ij}\overline{Y}_{ij}$. We may think of the process $\{Y_n;\, n \geq 0\}$ as evolving along with the Markov chain. The distributions of the variables Y_n are associated with the transitions from X_n to X_{n+1}, but the Y_n are otherwise independent random variables.

In order to define a martingale related to the process $\{S_n;\, n \geq 1\}$, we must subtract the mean reward from $\{S_n\}$ and must also compensate for the effect of the state of the Markov chain. The appropriate compensation factor turns out to be the relative-gain vector defined in Section 3.5.

For simplicity, consider only finite-state irreducible Markov chains with M states. Let $\pi = (\pi_{\mathbf{1}}, \ldots, \pi_{\mathsf{M}})$ be the steady-state probability vector for the chain, let $\overline{\boldsymbol{Y}} = (\overline{Y}_1, \ldots, \overline{Y}_{\mathsf{M}})^{\mathsf{T}}$ be the vector of expected rewards, let $g = \pi\overline{Y}$ be the steady-state gain per unit time, and let $\boldsymbol{w} = (w_1, \ldots, w_{\mathsf{M}})^{\mathsf{T}}$ be the relative-gain vector. From Theorem 3.5.1, $\boldsymbol{w}$ is the unique solution to

$$\boldsymbol{w} + g\boldsymbol{e} = \overline{\boldsymbol{Y}} + [P]\boldsymbol{w} \quad ; \quad w_1 = 0. \tag{7.110}$$

We assume a fixed starting state $X_0 = k$. As we now show, the process $Z_n;\, n \geq 1$ given by

$$Z_n = S_n - ng + w_{X_n} - w_k \quad ; \quad n \geq 1 \tag{7.111}$$

is a martingale. First condition on a given state, $X_{n-1} = i$.

$$\mathsf{E}[Z_n \mid Z_{n-1}, Z_{n-2}, \ldots, Z_1, X_{n-1} = i]. \tag{7.112}$$

Since $S_n = S_{n-1} + Y_{n-1}$, we can express Z_n as

$$Z_n = Z_{n-1} + Y_{n-1} - g + w_{X_n} - w_{X_{n-1}}. \tag{7.113}$$

Since $\mathsf{E}[Y_{n-1} \mid X_{n-1} = i] = \overline{Y}_i$ and $\mathsf{E}[w_{X_n} \mid X_{n-1} = i] = \sum_j P_{ij}w_j$, we have

$$\mathsf{E}[Z_n \mid Z_{n-1}, Z_{n-2}, \ldots, Z_1, X_{n-1} = i] = Z_{n-1} + \overline{Y}_i - g + \sum_j P_{ij}w_j - w_i. \tag{7.114}$$

From (7.110) the final four terms in (7.114) sum to 0, so

$$\mathsf{E}\left[Z_n \mid Z_{n-1}, \ldots, Z_1, X_{n-1}=i\right] = Z_{n-1}. \tag{7.115}$$

Since this is valid for all choices of X_{n-1}, we have $\mathsf{E}\left[Z_n \mid Z_{n-1}, \ldots, Z_1\right] = Z_{n-1}$. Since the expected values of all the reward variables $\overline{Y}_i$ exist, we see that $\mathsf{E}\left[|Y_n|\right] < \infty$, so that $\mathsf{E}\left[|Z_n|\right] < \infty$ also. This verifies that $\{Z_n;\ n \geq 1\}$ is a martingale. It can be verified similarly that $\mathsf{E}\left[Z_1\right] = 0$, so $\mathsf{E}\left[Z_n\right] = 0$ for all $n \geq 1$.

In showing that $\{Z_n;\ n \geq 1\}$ is a martingale, we actually showed something a little stronger. That is, (7.115) is conditioned on X_{n-1} as well as $Z_{n-1}, \ldots, Z_1$. In the same way, it follows that for all $n > 1$,

$$\mathsf{E}\left[Z_n \mid Z_{n-1}, X_{n-1}, Z_{n-2}, X_{n-2}, \ldots, Z_1, X_1\right] = Z_{n-1}. \tag{7.116}$$

In terms of the gambling analogy, this says that $\{Z_n;\ n \geq 1\}$ is fair for each possible past sequence of states. A martingale $\{Z_n;\ n \geq 1\}$ with this property (*i.e.*, satisfying (7.116)) is said to be a *martingale relative to the joint process* $\{Z_n, X_n;\ n \geq 1\}$. We will use this martingale later to discuss threshold crossing problems for Markov modulated random walks. We shall see that the added property of being a martingale relative to $\{Z_n, X_n\}$ gives us added flexibility in defining stopping times.

As an added bonus to this example, note that if $\{X_n;\ n \geq 0\}$ is taken as the embedded chain of a Markov process (or semi-Markov process), and if Y_n is taken as the time interval from transition n to $n+1$, then S_n becomes the epoch of the nth transition in the process.

7.10.1 Generating functions for Markov random walks

Consider the same Markov chain and reward variables as in the previous example, and assume that for each pair of states, i, j, the moment generating function

$$\mathsf{g}_{ij}(r) = \mathsf{E}\left[\exp(rY_n) \mid X_n = i, X_{n+1} = j\right]. \tag{7.117}$$

exists over some open interval (r_-, r_+) containing 0. Let $[\Gamma(r)]$ be the matrix with terms $P_{ij}\mathsf{g}_{ij}(r)$. Since $[\Gamma(r)]$ is an irreducible nonnegative matrix, Theorem 3.4.1 shows that $[\Gamma(r)]$ has a largest real eigenvalue, $\rho(r) > 0$, and an associated positive right eigenvector, $\nu(r) = (\nu_1(r), \ldots, \nu_{\mathsf{M}}(r))^T$ that is unique within a scale factor. We now show that the process $\{M_n(r);\ n \geq 1\}$ defined by

$$M_n(r) = \frac{\exp(rS_n)\nu_{X_n}(r)}{\rho(r)^n \nu_k(r)}. \tag{7.118}$$

is a product type Martingale for each $r \in (r_-, r_+)$. Since $S_n = S_{n-1} + Y_{n-1}$, we can express $M_n(r)$ as

$$M_n(r) = M_{n-1}(r)\ \frac{\exp(rY_{n-1})\,\nu_{X_n}(r)}{\rho(r)\nu_{X_{n-1}}(r)}. \tag{7.119}$$

The expected value of the ratio in (7.119), conditional on $X_{n-1} = i$, is

$$E\left[\frac{\exp(rY_{n-1})\nu_{X_n}(r)}{\rho(r)\nu_i(r)} \mid X_{n-1}{=}i\right] = \frac{\sum_j P_{ij}\mathsf{g}_{ij}(r)\nu_j(r)}{\rho(r)\nu_i(r)} = 1. \tag{7.120}$$

where, in the last step, we have used the fact that $\nu(r)$ is an eigenvector of $[\Gamma(r)]$. Thus, $\mathsf{E}[M_n(r) \mid M_{n-1}(r), \ldots, M_1(r), X_{n-1} = i] = M_{n-1}(r)$. Since this is true for all choices of i, the condition on $X_{n-1} = i$ can be removed and $\{M_n(r);\, n \geq 1\}$ is a martingale. Also, for $n > 1$,

$$\mathsf{E}[M_n(r) \mid M_{n-1}(r), X_{n-1}, \ldots, M_1(r), X_1] = M_{n-1}(r). \tag{7.121}$$

so that $\{M_n(r);\, n \geq 1\}$ is also a martingale relative to the joint process $\{M_n(r), X_n;\, n \geq 1\}$.

It can be verified by the same argument as in (7.120) that $\mathsf{E}[M_1(r)] = 1$. It then follows that $\mathsf{E}[M_n(r)] = 1$ for all $n \geq 1$.

One of the uses of this martingale is to provide exponential upper bounds, similar to (7.18), to the probabilities of threshold crossings for Markov modulated random walks. Define

$$\widetilde{M}_n(r) = \frac{\exp(rS_n)\min_j(\nu_j(r))}{\rho(r)^n\nu_k(r)}. \tag{7.122}$$

Then $\widetilde{M}_n(r) \leq M_n(r)$, so $\mathsf{E}\left[\widetilde{M}_n(r)\right] \leq 1$. For any $\mu > 0$, the Markov inequality can be applied to $\widetilde{M}_n(r)$ to get

$$\Pr\left\{\widetilde{M}_n(r) \geq \mu\right\} \leq \frac{1}{\mu}\mathsf{E}\left[\widetilde{M}_n(r)\right] \leq \frac{1}{\mu}. \tag{7.123}$$

For any given α, and any given r, $0 \leq r < r_+$, we can choose $\mu = \exp(r\alpha)\rho(r)^{-n}\min_j(\nu_j(r))/\nu_k(r)$, and for $r > 0$. Combining (7.122) and (7.123),

$$\Pr\{S_n \geq \alpha\} \leq \rho(r)^n \exp(-r\alpha)\nu_k(r)/\min_j(\nu_j(r)). \tag{7.124}$$

This can be optimized over r to get the tightest bound in the same way as (7.18).

7.10.2 stopping trials for martingales relative to a process

A martingale $\{Z_n;\, n \geq 1\}$ relative to a joint process $\{Z_n, X_n;\, n \geq 1\}$ was defined as a martingale for which (7.116) is satisfied, *i.e.*, $\mathsf{E}[Z_n \mid Z_{n-1}, X_{n-1}, \ldots, Z_1, X_1] = Z_{n-1}$. In the same way, we can define a *submartingale or supermartingale* $\{Z_n;\, n \geq 1\}$ *relative to a joint process* $\{Z_n, X_n;\, n \geq 1\}$ as a submartingale or supermartingale satisfying (7.116) with the $=$ sign replaced by $\geq$ or $\leq$ respectively. The purpose of this added complication is to make it easier to define useful stopping rules.

As generalized in Definition 4.5.2, a generalized stopping trial $\mathbb{I}_{J\geq n}$ for a sequence of pairs of rv's $(Z_1, X_1), (Z_2, X_2) \ldots$, is a positive integer-valued rv such that, for each $n \geq 1$, $\mathbb{I}_{\{J=n\}}$ is a function of $Z_1, X_1, Z_2, X_2, \ldots, Z_n, X_n$.

Theorems 7.8.1, 7.8.2 and Lemma 7.8.1 all carry over to martingales (submartingales or supermartingales) relative to a joint process. These theorems are stated more precisely in Exercises 7.23 to 7.26. To summarize them here, assume that $\{Z_n;\, n \geq 1\}$ is a martingale (submartingale or supermartingale) relative to a joint process $\{Z_n, X_n;\, n \geq 1\}$ and assume that J is a stopping trial for $\{Z_n;\, n \geq 1\}$ relative to $\{Z_n, X_n;\, n \leq 1\}$. Then the stopped process is a martingale (submartingale or supermartingale) respectively, (7.84 — 7.86) are satisfied, and, for a martingale, $\mathsf{E}[Z_J] = \mathsf{E}[Z_1]$ is satisfied iff (7.90) is satisfied.

7.10.3 Markov modulated random walks with thresholds

We have now developed two martingales for Markov modulated random walks, both conditioned on a fixed initial state $X_0 = k$. The first, given in (7.111), is $\{Z_n = S_n - ng + w_{X_n} - w_k;\, n \geq 1\}$. Recall that $\mathsf{E}[Z_n] = 0$ for all $n \geq 1$ for this martingale. Given two thresholds, $\alpha > 0$ and $\beta < 0$, define J as the smallest n for which $S_n \geq \alpha$ or $S_n \leq \beta$. The indicator function $\mathbb{I}_{J=n}$ of $\{J = n\}$, is 1 iff $\beta < S_i < \alpha$ for $1 \leq i \leq n-1$ and either $S_n \geq \alpha$ or $S_n \leq \beta$. Since $S_i = Z_i + ig - w_{X_i} + w_k$, S_i is a function of Z_i and X_i, so the stopping trial is a function of both Z_i and X_i for $1 \leq i \leq n$. It follows that J is a stopping trial for $\{Z_n;\, n \geq 1\}$ relative to $\{Z_n, X_n;\, n \geq 1\}$. From Lemma 7.8.1, we can assert that $\mathsf{E}[Z_J] = \mathsf{E}[Z_1] = 0$ if (7.90) is satisfied, *i.e.*, if $\lim_{n\to\infty} \mathsf{E}[Z_n \mid J > n] \Pr\{J > n\} = 0$ is satisfied. Using the same argument as in Lemma 7.5.1, we can see that $\Pr\{J > n\}$ goes to 0 at least geometrically in n. Conditional on $J > n$, $\beta < S_n < \alpha$, so S_n is bounded independent of n. Also w_{X_n} is bounded, since the chain is finite state, and ng is linear in n. Thus $\mathsf{E}[Z_n \mid J > n]$ varies at most linearly with n, so (7.90) is satisfied, and

$$0 = \mathsf{E}[Z_J] = \mathsf{E}[S_J] - \mathsf{E}[J]\, g + \mathsf{E}[w_{X_n}] - w_k. \tag{7.125}$$

Recall that Wald's equality for random walks is $\mathsf{E}[S_J] = \mathsf{E}[J]\,\overline{X}$. For Markov modulated random walks, this is modified, as shown in (7.125), by the relative-gain vector terms.

The same arguments can be applied to the generating function martingale of (7.118). Again, let J be the smallest n for which $S_n \geq \alpha$ or $S_n \leq \beta$. As before, S_i is a function of $M_i(r)$ and X_i, so $\mathbb{I}_n$ is a function of $M_i(r)$ and X_i for $1 \leq i \leq n-1$. It follows that J is a stopping trial for $\{M_n(r);\, n \geq 1\}$ relative to $\{M_n(r), X_n;\, n \geq 1\}$ Next we need the following lemma:

Lemma 7.10.1. *For the martingale $\{M_n(r);\, n \geq 1\}$ relative to $\{M_n(r), X_n;\, n \geq 1\}$ defined in (7.118), where $\{X_n;\, n \geq 0\}$ is a finite-state Markov chain, and for the above stopping trial J,*

$$\lim_{n\to\infty} \mathsf{E}[M_n(r) \mid J > n] \Pr\{J > n\} = 0. \tag{7.126}$$

Proof: From lemma 4, slightly modified for the case here, there is a $\delta > 0$ such that for all states i, j, and all $n > 1$ such that $\Pr\{J = n, X_{n-1} = i, X_n = j\} > 0$,

$$\mathsf{E}[\exp(rS_n \mid J = n, X_{n-1} = i, X_n = j] \geq \delta. \tag{7.127}$$

Since the stopped process, $\{M_n^*(r);\, n \geq 1\}$, is a martingale, we have for each m,

$$1 = \mathsf{E}[M_m^*(r)] \geq \sum_{n=1}^{m} \frac{\mathsf{E}[\exp(rS_n)\nu_{X_n}(r) \mid J = n]}{\rho(r)^n \nu_k(r)}. \tag{7.128}$$

From (7.127), we see that there is some $\delta' > 0$ such that

$$\mathsf{E}\left[\exp(rS_n)\nu_{X_n}(r)\right]/\nu_k(r) \mid J = n] \geq \delta'$$

for all n such that $\Pr\{J = n\} > 0$. Thus (7.128) is bounded by

$$1 \geq \delta' \sum_{n \leq m} \rho(r)^n \Pr\{J = n\}.$$

Since this is valid for all m, it follows by the argument in the proof of theorem 7.5.1 that $\lim_{n\to\infty} \rho(r)^n \Pr\{J > n\} = 0$. This, along with (7.127), establishes (7.126), completing the proof. □

From Lemma 7.8.1, we have the desired result:

$$\mathsf{E}\left[M_J(r)\right] = E\left[\frac{\exp(rS_J)\nu_{X_J}(r)}{[\rho(r)]^J \nu_k(r)}\right] = 1; \qquad r_- < r < r_+. \tag{7.129}$$

This is the extension of the Wald identity to Markov modulated random walks, and is used in the same way as the Wald identity. As shown in Exercise 7.28, the derivative of (7.129), evaluated at $r = 0$, is the same as (7.125).

7.11 Summary

Each term in a random walk $\{S_n; n \geq 1\}$ is a sum of IID random variables, and thus the study of random walks is closely related to that of sums of IID variables. The focus in random walks, however, as in most of the processes we have studied, is more in the relationship between the terms (such as which term first crosses a threshold) than in the individual terms. We started by showing that random walks are a generalization of renewal processes, are central to studying the queueing delay for G/G/1 queues, and to sequential analysis for hypothesis testing.

A major focus of the chapter was on estimating the probabilities of very unlikely events, a topic known as large deviation theory. We started by studying the Chernoff bound to $\Pr\{S_n \geq \alpha\}$ for $\alpha > 0$ and $\mathsf{E}[X] < 0$. We then developed the Wald identity, which can be used to find tight upper bounds to the probability that a threshold is ever crossed by a random walk. One of the insights gained here was that if a threshold at α is crossed, it is likely to be crossed at a time close to $n^* = \alpha/\gamma'(r^*)$, where r^* is the positive root of $\gamma(r)$. We also found that r^* plays a fundamental role in the probability of threshold crossings. For questions of typical behavior, the mean and variance of a random variable are the major quantities of interest, but when interested in atypically large deviations, r^* is the major parameter of interest.

We next introduced martingales, submartingales and supermartingales. These are sometimes regarded as somewhat exotic topics in mathematics, but in fact they are very useful in a large variety of relatively simple processes. For example, we showed that all of the random walk issues of earlier sections can be treated as a special case of martingales, and

that martingales can be used to model both sums and products of random variables. We also showed how Markov modulated random walks can be treated as martingales.

Stopping trials, as first introduced in chapter 4, were then applied to martingales. We defined a stopped process $\{Z_n^*;\, n \geq 1\}$ to be the same as the original process $\{Z_n;\, n \geq 1\}$ up to the stopping point, and then constant thereafter. Theorems 7.8.1 and 7.8.2 showed that the stopped process has the same form (martingale, submartingale, or supermartingale) as the original process, and that the expected values $\mathsf{E}[Z_n^*]$ are between $\mathsf{E}[Z_1]$ and $\mathsf{E}[Z_n]$. We also looked at $\mathsf{E}[Z_J]$ and found that it is equal to $\mathsf{E}[Z_1]$ iff (7.90) is satisfied. The Wald identity can be viewed as $\mathsf{E}[Z_J] = \mathsf{E}[Z_1] = 1$ for the Wald martingale, $Z_n = \exp\{rS_n - n\gamma(r)\}$. We then found a similar identity for Markov modulated random walks. In deriving results for Markov modulated random walks, it was necessary to define martingales relative to other processes in order to find suitable stopping trials, also defined on martingales relative to other processes. This added restriction on martingales is useful in other contexts.

The Kolmogorov inequalities were next developed. They are analogs of the Markov inequality and Chebyshev inequality, except they bound initial segments of submartingales and martingales rather than single rv's. They were used, first, to prove the SLLN with only a second moment and, second,the martingale convergence theorem.

A standard reference on random walks, and particularly on the analysis of overshoots is [Fel66]. Dembo and Zeitouni, [5] develop large deviation theory in a much more general and detailed way than the introduction here. The classic reference on martingales is [6], but [4] and [16] are more accessible.

7.12 Exercises

Exercise 7.1. Consider the simple random walk $\{S_n;\, n \geq 1\}$ of Section 7.1.1 with $S_n = X_1 + \cdots + X_n$ and $\Pr\{X_i = 1\} = p$; $\Pr\{X_i = -1\} = 1 - p$; assume that $p \leq 1/2$.

a) Show that $\Pr\Big\{\bigcup_{n\geq 1} \{S_n \geq k\}\Big\} = \Big[\Pr\Big\{\bigcup_{n\geq 1} \{S_n \geq 1\}\Big\}\Big]^k$ for any positive integer k. Hint: Given that the random walk ever reaches the value 1, consider a new random walk starting at that time and explore the probability that the new walk ever reaches a value 1 greater than its starting point.

b) Find a quadratic equation for $y = \Pr\Big\{\bigcup_{n\geq 1} \{S_n \geq 1\}\Big\}$. Hint: explore each of the two possibilities immediately after the first trial.

c) For $p < 1/2$, show that the two roots of this quadratic equation are $p/(1-p)$ and 1. Argue that $\Pr\Big\{\bigcup_{n\geq 1} \{S_n \geq 1\}\Big\}$ cannot be 1 and thus must be $p/(1-p)$.

d) For $p = 1/2$, show that the quadratic equation in part c) has a double root at 1, and thus $\Pr\Big\{\bigcup_{n\geq 1} \{S_n \geq 1\}\Big\} = 1$. Note: this is the very peculiar case explained in the section on Wald's equality.

e) For $p < 1/2$, show that $p/(1-p) = \exp(-r^*)$ where r^* is the unique positive root of $g(r) = 1$ where $g(r) = \mathsf{E}\left[e^{rX}\right]$.

Exercise 7.2. Consider a G/G/1 queue with IID arrivals $\{X_i;\, i \geq 1\}$, IID FCFS service times $\{Y_i;\, i \geq 0\}$, and an initial arrival to an empty system at time 0. Define $U_i = X_i - Y_{i-1}$ for $i \geq 1$. Consider a sample path where $(u_1, \ldots, u_6) = (1, -2, 2, -1, 3, -2)$.

a) Let $Z_i^6 = U_6 + U_{6-1} + \ldots + U_{6-i+1}$. Find the queueing delay for customer 6 as the maximum of the 'backward' random walk with elements $0, Z_1^6, Z_2^6, \ldots, Z_6^6$; sketch this random walk.

b) Find the queueing delay for customers 1 to 5.

c) Which customers start a busy period (*i.e.*, arrive when the queue and server are both empty)? Verify that if Z_i^6 maximizes the random walk in part a), then a busy period starts with arrival $6 - i$.

d) Now consider a forward random walk $V_n = U_1 + \cdots + U_n$. Sketch this walk for the sample path above and show that the queueing delay for each customer is the difference between two appropriately chosen values of this walk.

Exercise 7.3. A G/G/1 queue has a deterministic service time of 2 and interarrival times that are 3 with probability p and 1 with probability $1 - p$.

a) Find the distribution of W_1, the wait in queue of the first arrival after the beginning of a busy period.

b) Find the distribution of W_∞, the steady-state wait in queue.

c) Repeat parts a) and b) assuming the service times and interarrival times are exponentially distributed with rates μ and λ respectively.

Exercise 7.4. A sales executive hears that one of his sales people is routing half of his incoming sales to a competitor. In particular, arriving sales are known to be Poisson at rate one per hour. According to the report (which we view as hypothesis 1), each second arrival is routed to the competition; thus under hypothesis 1 the interarrival density for successful sales is $f(y|\mathsf{H}_1) = ye^{-y}$; $y \geq 0$. The alternate hypothesis (H_0) is that the rumor is false and the interarrival density for successful sales is $f(y|\mathsf{H}_0) = e^{-y}$; $y \geq 0$. Assume that, *a priori*, the hypotheses are equally likely. The executive, a recent student of stochastic processes, explores various alternatives for choosing between the hypotheses; he can only observe the times of successful sales however.

a) Starting with a successful sale at time 0, let S_i be the arrival time of the i^{th} subsequent successful sale. The executive observes $S_1, S_2, \ldots, S_n (n \geq 1)$ and chooses the maximum aposteriori probability hypothesis given this data. Find the joint probability density $f(S_1, S_2, \ldots, S_n|\mathsf{H}_1)$ and $f(S_1, \ldots, S_n|\mathsf{H}_0)$ and give the decision rule.

b) This is the same as part a) except that the system is in steady state at time 0 (rather than starting with a successful sale). Find the density of S_1 (the time of the first arrival after time 0) conditional on H_0 and on H_1. What is the decision rule now after observing $S_1, \ldots, S_n$.

c) This is the same as part b), except rather than observing n successful sales, the successful sales up to some given time t are observed. Find the probability, under each hypothesis, that the first successful sale occurs in $(s_1, s_1 + \Delta]$, the second in $(s_2, s_2 + \Delta], \ldots$, and the last in $(s_{N(t)}, s_{N(t)} + \Delta]$ (assume Δ very small). What is the decision rule now?

Exercise 7.5. For the hypothesis testing problem of Section 7.3, assume that there is a cost C_0 of choosing H_1 when H_0 is correct, and a cost C_1 of choosing H_0 when H_1 is correct. Show that a threshold test minimizes the expected cost using the threshold $\eta = (C_1 p_1)/(C_0 p_0)$.

Exercise 7.6. Consider a binary hypothesis testing problem where H is 0 or 1 and a one dimensional observation Y is given by $Y = H + U$ where U is uniformly distributed over [-1, 1] and is independent of H.

a) Find $\mathsf{f}_{Y|H}(y \mid 0)$, $\mathsf{f}_{Y|H}(y \mid 1)$ and the likelihood ratio $\Lambda(y)$.

b) Find the threshold test at η for each η, $0 < \eta < \infty$ and evaluate the conditional error probabilities, $q_0(\eta)$ and $q_1(\eta)$.

c) Find the error curve $u(\alpha)$ and explain carefully how $u(0)$ and $u(1/2)$ are found (hint: $u(0) = 1/2$).

d) Describe a decision rule for which the error probability under each hypothesis is 1/4. You need not use a randomized rule, but you need to handle the don't-care cases under the threshold test carefully.

Exercise 7.7. a) For given α, $0 < \alpha \leq 1$, let η^* achieve the supremum $\sup_{0\leq\eta<\infty} q_1(\eta) + \eta(q_0(\eta) - \alpha)$. Show that $\eta^* \leq 1/\alpha$. Hint: Think in terms of Lemma 7.3.1 applied to a very simple test.

b) Show that the magnitude of the slope of the error curve $u(\alpha)$ at α is at most $1/\alpha$.

Exercise 7.8. Define $\gamma(r)$ as $\ln[\mathsf{g}(r)]$ where $\mathsf{g}(r) = \mathsf{E}[\exp(rX)]$. Assume that X is discrete with possible outcomes $\{a_i;\, i \geq 1\}$, let p_i denote $\Pr\{X = a_i\}$, and assume that $\mathsf{g}(r)$ exists in some open interval (r_-, r_+) containing $r = 0$. For any given r, $r_- < r < r_+$, define a random variable X_r with the same set of possible outcomes $\{a_i;\, i \geq 1\}$ as X, but with a probability mass function $q_i = \Pr\{X_r = a_i\} = p_i \exp[a_i r - \gamma(r)]$. X_r is not a function of X, and is not even to be viewed as in the same probability space as X; it is of interest simply because of the behavior of its defined probability mass function. It is called a tilted random variable relative to X, and this exercise, along with Exercise 7.9 will justify our interest in it.

a) Verify that $\sum_i q_i = 1$.

b) Verify that $\mathsf{E}[X_r] = \sum_i a_i q_i$ is equal to $\gamma'(r)$.

c) Verify that $\mathrm{Var}[X_r] = \sum_i a_i^2 q_i - (\mathsf{E}[X_r])^2$ is equal to $\gamma''(r)$.

d) Argue that $\gamma''(r) \geq 0$ for all r such that $\mathsf{g}(r)$ exists, and that $\gamma''(r) > 0$ if $\gamma''(0) > 0$.

e) Give a similar definition of X_r for a random variable X with a density, and modify parts a) to d) accordingly.

Exercise 7.9. Assume that X is discrete, with possible values $\{a_i;\, i \geq 1)$ and probabilities $\Pr\{X = a_i\} = p_i$. Let X_r be the corresponding tilted random variable as defined in Exercise 7.8. Let $S_n = X_1 + \cdots + X_n$ be the sum of n IID rv's with the distribution of X, and let $S_{n,r} = X_{1,r} + \cdots + X_{n,r}$ be the sum of n IID tilted rv's with the distribution of X_r. Assume that $\overline{X} < 0$ and that $r > 0$ is such that $\gamma(r)$ exists.

a) Show that $\Pr\{S_{n,r}{=}s\} = \Pr\{S_n{=}s\} \exp[sr - n\gamma(r)]$. Hint: first show that

$$\Pr\{X_{1,r}{=}v_1, \ldots, X_{n,r}{=}v_n\} = \Pr\{X_1{=}v_1, \ldots, X_n{=}v_n\} \exp[sr - n\gamma(r)]$$

where $s = v_1 + \cdots + v_n$.

b) Find the mean and variance of $S_{n,r}$ in terms of $\gamma(r)$.

c) Define $a = \gamma'(r)$ and $\sigma_r^2 = \gamma''(r)$. Show that $\Pr\{|S_{n,r} - na| \leq \sqrt{2n}\sigma_r\} > 1/2$. Use this to show that

$$\Pr\Big\{|\, S_n - na \,| \leq \sqrt{2n}\, \sigma_r \,\Big\} > (1/2) \exp[-r(an + \sqrt{2n}\, \sigma_r) + n\gamma(r)].$$

d) Use this to show that for any ϵ and for all sufficiently large n,

$$\Pr\{S_n \geq n(\gamma'(r) - \epsilon)\} > \frac{1}{2} \exp[-rn(\gamma'(r) + \epsilon) + n\gamma(r)].$$

Exercise 7.10. Consider a random walk with thresholds $\alpha > 0$, $\beta < 0$. We wish to find $\Pr\{S_J \geq \alpha\}$ in the absence of a lower threshold. Use the upper bound in (7.42) for the probability that the random walk crosses α before β.

a) Given that the random walk crosses β first, find an upper bound to the probability that α is now crossed before a yet lower threshold at 2β is crossed.

b) Given that 2β is crossed before α, upperbound the probability that α is crossed before a threshold at 3β. Extending this argument to successively lower thresholds, find an upper bound to each successive term, and find an upper bound on the overall probability that α is crossed. By observing that β is arbitrary, show that (7.42) is valid with no lower threshold.

Exercise 7.11. This exercise verifies that Corollary 7.5.1 holds in the situation where $\gamma(r) < 0$ for all $r \in (r_-, r_+)$ and where r^* is taken to be r_+ (see Figure 7.7).

a) Show that for the situation above, $\exp(rS_J) \leq \exp(rS_J - J\gamma(r))$ for all $r \in (0, r^*)$.

b) Show that $\mathsf{E}[\exp(rS_J)] \leq 1$ for all $r \in (0, r^*)$.

c) Show that $\Pr\{S_J \geq \alpha\} \leq \exp(-r\alpha)$ for all $r \in (0, r^*)$. Hint: Follow the steps of the proof of Corollary 7.5.1.

d) Show that $\Pr\{S_J \geq \alpha\} \leq \exp(-r^*\alpha)$.

Exercise 7.12. a) Use Wald's equality to show that if $\overline{X} = 0$, then $\mathsf{E}[S_J] = 0$ where J is the time of threshold crossing with one threshold at $\alpha > 0$ and another at $\beta < 0$.

b) Obtain an expression for $\Pr\{S_J \geq \alpha\}$. Your expression should involve the expected value of S_J conditional on crossing the individual thresholds (you need not try to calculate these expected values).

c) Evaluate your expression for the case of a simple random walk.

d) Evaluate your expression when X has an exponential density, $\mathsf{f}_X(x) = a_1 e^{-\lambda x}$ for $x \geq 0$ and $\mathsf{f}_X(x) = a_2 e^{\mu x}$ for $x < 0$ and where a_1 and a_2 are chosen so that $\overline{X} = 0$.

Exercise 7.13. A random walk $\{S_n;\, n \geq 1\}$, with $S_n = \sum_{i=1}^{n} X_i$, has the following probability density for X_i

$$\mathsf{f}_X(x) = \begin{cases} \frac{e^{-x}}{e-e^{-1}}\ ; & -1 \leq x \leq 1 \\ \\ = 0\ ; & \text{elsewhere.} \end{cases}$$

a) Find the values of r for which $\mathsf{g}(r) = \mathsf{E}[\exp(rX)] = 1$.

b) Let P_α be the probability that the random walk ever crosses a threshold at α for some $\alpha > 0$. Find an upper bound to P_α of the form $P_\alpha \leq e^{-\alpha A}$ where A is a constant that does not depend on α; evaluate A.

c) Find a lower bound to P_α of the form $P_\alpha \geq Be^{-\alpha A}$ where A is the same as in part b) and B is a constant that does not depend on α. Hint: keep it simple — you are not expected to find an elaborate bound. Also recall that $\mathsf{E}\left[e^{r^* S_J}\right] = 1$ where J is a stopping trial for the random walk and $\mathsf{g}(r^*) = 1$.

Exercise 7.14. Let $\{X_n;\, n \geq 1\}$ be a sequence of IID integer valued random variables with the probability mass function $P_X(k) = Q_k$. Assume that $Q_k > 0$ for $|k| \leq 10$ and $Q_k = 0$ for $|k| > 10$. Let $\{S_n;\, n \geq 1\}$ be a random walk with $S_n = X_1 + \cdots + X_n$. Let $\alpha > 0$ and $\beta < 0$ be integer valued thresholds, let J be the smallest value of n for which either $S_n \geq \alpha$ or $S_n \leq \beta$. Let $\{S_n^*;\, n \geq 1\}$ be the stopped random walk; *i.e.*, $S_n^* = S_n$ for $n \leq J$ and $S_n^* = S_J$ for $n > J$. Let $\pi_i^* = \Pr\{S_J = i\}$.

a) Consider a Markov chain in which this stopped random walk is run repeatedly until the point of stopping. That is, the Markov chain transition probabilities are given by $P_{ij} = Q_{j-i}$ for $\beta < i < \alpha$ and $P_{i0} = 1$ for $i \leq \beta$ and $i \geq \alpha$. All other transition probabilities are 0 and the set of states is the set of integers $[-9 + \beta, 9 + \alpha]$. Show that this Markov chain is ergodic.

b) Let $\{\pi_i\}$ be the set of steady-state probabilities for this Markov chain. Find the set of probabilities $\{\pi_i^*\}$ for the stopping states of the stopped random walk in terms of $\{\pi_i\}$.

c) Find $\mathsf{E}[S_J]$ and $\mathsf{E}[J]$ in terms of $\{\pi_i\}$.

Exercise 7.15. a) Conditional on H_0 for the hypothesis testing problem, consider the random variables $Z_i = \ln[f(Y_i|\mathsf{H}_1)/f(Y_i|\mathsf{H}_0)]$. Show that r^*, the positive solution to $\mathsf{g}(r) = 1$, where $g(r) = \mathsf{E}[\exp(rZi)]$, is given by $r^* = 1$.

b) Assuming that Y is a discrete random variable (under each hypothesis), show that the tilted random variable Z_r with $r = 1$ has the PMF $P_Y(y|\mathsf{H}_1)$.

Exercise 7.16. a) Suppose $\{Z_n;\, n \geq 1\}$ is a martingale. Verify (7.69); *i.e.*, $\mathsf{E}[Z_n] = \mathsf{E}[Z_1]$ for $n > 1$.

b) If $\{Z_n;\, n \geq 1\}$ is a submartingale, verify (7.75), and if a supermartingale, verify (7.76).

Exercise 7.17. Suppose $\{Z_n;\, n \geq 1\}$ is a martingale. Show that

$$\mathsf{E}\left[Z_m \mid Z_{n_i}, Z_{n_{i-1}}, \ldots, Z_{n_1}\right] = Z_{n_i} \quad \text{for all } 0 < n_1 < n_2 < \ldots < n_i < m.$$

Exercise 7.18. a) Assume that $\{Z_n;\, n \geq 1\}$ is a submartingale. Show that

$$\mathsf{E}[Z_m \mid Z_n, Z_{n-1}, \ldots, Z_1] \geq Z_n \text{ for all } n < m.$$

b) Show that

$$\mathsf{E}\left[Z_m \mid Z_{n_i}, Z_{n_{i-1}}, \ldots, Z_{n_1}\right] \geq Z_{n_i} \text{ for all } \; 0 < n_1 < n_2 < \ldots < n_i < m.$$

c) Assume now that $\{Z_n;\, n \geq 1\}$ is a supermartingale. Show that parts a) and b) still hold with $\geq$ replaced by $\leq$.

Exercise 7.19. Let $\{Z_n = \exp[rS_n - n\gamma(r)];\ n \geq 1\}$ be the generating function martingale of (7.63) where $S_n = X_1 + \cdots + X_n$ and $X_1, \ldots X_n$ are IID with mean $\overline{X} < 0$. Let J be the possibly-defective stopping trial for which the process stops after crossing a threshold at $\alpha > 0$ (there is no negative threshold). Show that $\exp[r^*\alpha]$ is an upper bound to the probability of threshold crossing by considering the stopped process $\{Z_n^*;\ n \geq 1\}$.

The purpose of this exercise is to illustrate that the stopped process can yield useful upper bounds even when the stopping trial is defective.

Exercise 7.20. This problem uses martingales to find the expected number of trials $\mathsf{E}[J]$ before a fixed pattern, $a_1, a_2, \ldots, a_k$, of binary digits occurs within a sequence of IID binary random variables $X_1, X_2, \ldots$ (see Exercises 4.32 and 3.28 for alternate approaches). A mythical casino and set of gamblers who follow a prescribed strategy will be used to determine $\mathsf{E}[J]$. The casino has a game where, on the ith trial, gamblers bet money on either 1 or 0. After bets are placed, X_i above is used to select the outcome 0 or 1. Let $p(1) = \mathsf{p}_X(1)$ and $p(0) = 1 - p(1) = \mathsf{p}_X(0)$. If an amount s is bet on 1, the casino receives s if $X_i = 0$, and pays out $s/p(1) - s$ (plus returning the bet s) if $X_i = 1$. If s is bet on 0, the casino receives s if $X_i = 1$, and pays out $s/p(0) - s$ (plus the bet s) if $X_i = 0$.

a) Assume an arbitrary pattern of bets by various gamblers on various trials (some gamblers might bet arbitrary amounts on 0 and some on 1 at any given trial). Let Y_i be the net gain of the casino on trial i. Show that $\mathsf{E}[Y_i] = 0$ (*i.e.*, show that the game is fair). Let $Z_n = Y_1 + Y_2 + \cdots + Y_n$ be the aggregate gain of the casino over n trials. Show that for the given pattern of bets, $\{Z_n;\ n \geq 1\}$ is a martingale.

b) In order to determine $\mathsf{E}[J]$ for a given pattern $a_1, a_2, \ldots, a_k$, we program our gamblers to bet as follows:

i) Gambler 1 has an initial capital of 1 which is bet on a_1 at trial 1. If he wins, his capital grows to $1/p(a_1)$, which is bet on a_2 at trial 2. If he wins again, he bets his entire capital, $1/[p(a_1)p(a_2)]$, on a_3 at trial 3. He continues, at each trial i, to bet his entire capital on a_i until he loses at some trial (in which case he leaves with no money) or he wins on k successive trials (in which case he leaves with $1/[p(a_1)\ldots p(a_k)]$.

ii) Gambler j, $j > 1$, follows exactly the same strategy but starts at trial j. Note that if the pattern $a_1, \ldots, a_k$ appears for the first time at $J = n$, then gambler $n - k + 1$ leaves at time n with capital $1/[p(a_1)\ldots p(a_k)]$ and gamblers $j < n - k + 1$ all lose their capital.

Suppose the string $(a_1, \ldots, a_k)$ is $(0, 1)$. Let for the above gambling strategy. Given that $J = 3$ (*i.e.*, that $X_2 = 0$ and $X_3 = 1$), note that gambler 1 loses his money at either trial 1 or 2, gambler 2 leaves at time 3 with $1/[p(0)p(1)]$ and gambler 3 loses his money at time 3. Show that $Z_J = 3 - 1/[p(0)p(1)]$ given $J = 3$. Find Z_J given $J = n$ for arbitrary $n \geq 2$ (note that the condition $J = n$ uniquely specifies Z_J).

c) Find $\mathsf{E}[Z_J]$ from part a). Use this plus part b) to find $\mathsf{E}[J]$.

d) Repeat parts b) and c) using the string $(a_1, \ldots, a_k) = (1, 1)$. Be careful about gambler 3 for $J = 3$. Show that $\mathsf{E}[J] = 1/[p(1)p(1)] + 1/p(1)$

e) Repeat parts b) and c) for $(a_1, \ldots, a_k) = (1, 1, 1, 0, 1, 1)$.

Exercise 7.21. a) This exercise shows why the condition $\mathsf{E}[|Z_J|] < \infty$ is required in Lemma 7.8.1. Let $Z_1 = -2$ and, for $n \geq 1$, let $Z_{n+1} = Z_n[1 + X_n(3n+1)/(n+1)]$ where $X_1, X_2, \ldots$ are IID and take on the values $+1$ and -1 with probability $1/2$ each. Show that $\{Z_n; n \geq 1\}$ is a martingale.

b) Consider the stopping trial J such that J is the smallest value of $n > 1$ for which Z_n and Z_{n-1} have the same sign. Show that, conditional on $n < J$, $Z_n = (-2)^n/n$ and, conditional on $n = J$, $Z_J = -(-2)^n(n-2)/(n^2-n)$.

c) Show that $\mathsf{E}[|Z_J|]$ is infinite, so that $\mathsf{E}[Z_J]$ does not exist according to the definition of expectation, and show that $\lim_{n\to\infty} \mathsf{E}[Z_n | J > n] \Pr\{J > n\} = 0$.

Exercise 7.22. This exercise shows why the sup of a martingale can behave markedly differently from the maximum of an arbitrarily large number of the variables. More precisly, it shows that $\Pr\{\sup_{n\geq 1} Z_n \geq a\}$ can be unequal to $\Pr\left\{\bigcup_{n\geq 1}\{Z_n \geq a\}\right\}$.

a) Consider a martingale where Z_n can take on only the values 2^{-n-1} and $1 - 2^{-n-1}$, each with probability $1/2$. Given that Z_n, conditional on Z_{n-1}, is independent of $Z_1, \ldots Z_{n-2}$, find $\Pr\{Z_n | Z_{n-1}\}$ for each n so that the martingale condition is satisfied.

b) Show that $\Pr\{\sup_{n\geq 1} Z_n \geq 1\} = 1/2$ and show that $\Pr\{\bigcup_n \{Z_n \geq 1\}\} = 0$.

c Show that for every $\epsilon > 0$, $\Pr\{\sup_{n\geq 1} Z_n \geq a\} \leq \frac{\overline{Z}_1}{a-\epsilon}$. Hint: Use the relationship between $\Pr\{\sup_{n\geq 1} Z_n \geq a\}$ and $\Pr\{\bigcup_n \{Z_n \geq a\}\}$ while getting around the issue in part b).

d Use part c) to establish (7.95).

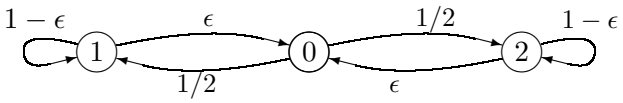
Exercise 7.23. Show that Theorem 7.7.1 is also valid for martingales relative to a joint process. That is, show that if h is a convex function of a real variable and if $\{Z_n; n \geq 1\}$ is a martingale relative to a joint process $\{Z_n, X_n; n \geq 1\}$, then $\{h(Z_n); n \geq 1\}$ is a submartingale relative to $\{h(Z_n), X_n; n \geq 1\}$.

Exercise 7.24. Show that if $\{Z_n; n \geq 1\}$ is a martingale (submartingale or supermartingale) relative to a joint process $\{Z_n, X_n; n \geq 1\}$ and if J is a stopping trial for $\{Z_n; n \geq 1\}$ relative to $\{Z_n, X_n; n \geq 1\}$, then the stopped process is a martingale (submartingale or supermartingale) respectively relative to the joint process.

Exercise 7.25. Show that if $\{Z_n; n \geq 1\}$ is a martingale (submartingale or supermartingale) relative to a joint process $\{Z_n, X_n; n \geq 1\}$ and if J is a stopping trial for $\{Z_n; n \geq 1\}$ relative to $\{Z_n, X_n; n \geq 1\}$, then the stopped process satisfies (7.84), (7.85), or (7.86) respectively.

Exercise 7.26. Show that if $\{Z_n; n1\}$ is a martingale relative to a joint process $\{Z_n, X_n; n \geq 1\}$ and if J is a stopping trial for $\{Z_n; n \geq 1\}$ relative to $\{Z_n, X_n; n \geq 1\}$, then $\mathsf{E}[Z_J] = \mathsf{E}[Z_1]$ if and only f (7.90) is satisfied.

Exercise 7.27. Consider the Markov modulated random walk in the figure below. The random variables Y_n in this example take on only a single value for each transition, that value being 1 for all transitions from state 1, 10 for all transitions from state 2, and 0 otherwise. $\epsilon > 0$ is a very small number, say $\epsilon < 10^{-6}$.



a) Show that the steady-state gain per transition is $5.5/(1+\epsilon)$. Show that the relative-gain vector is $\boldsymbol{w} = (0, (\epsilon - 4.5)/[\epsilon(1+\epsilon)], (10\epsilon + 4.5)/[\epsilon(1+\epsilon)])$.

b) Let $S_n = Y_0 + Y_1 + \cdots + Y_{n-1}$ and take the starting state X_0 to be 0. Let J be the smallest value of n for which $S_n \geq 100$. Find $\Pr\{J = 11\}$ and $\Pr\{J = 101\}$. Find an estimate of $\mathsf{E}[J]$ that is exact in the limit $\epsilon \to 0$.

c) Show that $\Pr\{X_J = 1\} = (1 - 45\epsilon + o(\epsilon))/2$ and that $\Pr\{X_J = 2\} = (1 + 45\epsilon + o(\epsilon))/2$. Verify, to first order in ϵ that (7.125) is satisfied.

Exercise 7.28. Show that (7.125) results from taking the derivative of (7.129) and evaluating it at $r = 0$.

Exercise 7.29. Let $\{Z_n;\ n \geq 1\}$ be a martingale, and for some integer m, let $Y_n = Z_{n+m} - Z_m$.

a) Show that $\mathsf{E}[Y_n \mid Z_{n+m-1} = z_{n+m-1}, Z_{n+m-2} = z_{n+m-2}, \ldots, Z_m = z_m, \ldots, Z_1 = z_1] = z_{n+m-1} - z_m$.

b) Show that $\mathsf{E}[Y_n \mid Y_{n-1} = y_{n-1}, \ldots, Y_1 = y_1] = y_{n-1}$

c) Show that $\mathsf{E}[|Y_n|] < \infty$. Note that **b)** and **c)** show that $\{Y_n;\ n \geq 1\}$ is a martingale.

Exercise 7.30. a) Show that Theorem 7.7.1 is valid if $\{Z_n;\ n \geq 1\}$ is a submartingale rather than a martingale. Hint: Simply follow the proof of Theorem 7.7.1 in the text.

b) Show that the Kolmogorov martingale inequality also holds if $\{Z_n;\ n \geq 1\}$ is a submartingale rather than a martingale.

Exercise 7.31 (Continuation of continuous-time branching). This exercse views the continuous-time branching process of Exercise 6.15 as a stopped random walk. Recall that the process was specified there as a Markov process such that for each state j, $j \geq 0$, the transition rate to $j+1$ is $j\lambda$ and to $j-1$ is $j\mu$. There are no other transitions, and in particular, there are no transitions out of state 0, so that the Markov process is reducible. Recall that the embedded Markov chain is the same as the embedded chain of an M/M/1 queue except that there is no transition from state 0 to state 1.

a) To model the possible extinction of the population, convert the embedded Markov chain abve to a stopped random walk, $\{S_n;\ n \geq 0$. The stopped random walk starts at $S_0 =$

0 and stops on reaching a threshold at -1. Before stopping, it moves up by one with probability $\frac{\lambda}{\lambda+\mu}$ and downward by 1 with probability $\frac{\mu}{\lambda+\mu}$ at each step. Give the (very simple) relationship between the state X_n of the Markov chain and the state S_n of the stopped random walk for each $n \geq 0$.

b) Find the probability that the population eventually dies out as a function of λ and μ. Be sure to consider all three cases $\lambda > \mu$, $\lambda < \mu$, and $\lambda = \mu$.

Appendix A

Table of standard random variables

The following tables summarize the properties of some common random variables. If a density or PMF is specified only in a given region, it is assumed to be zero elsewhere. The parameters λ, σ, and a are assumed to be positive, $p \in (0,1)$, and n is a positive integer.

Name	**Density** $\mathsf{f}_X(x)$	**Mean**	**Variance**	**MGF**
Exponential:	$\lambda \exp(-\lambda x);\ x{\geq}0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda - r};\quad \text{for } r < \lambda$
Erlang:	$\frac{\lambda^n x^{n-1} \exp(-\lambda x)}{(n-1)!};\quad x{\geq}0$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$	$\left(\frac{\lambda}{\lambda - r}\right)^n;\quad \text{for } r < \lambda$
Gaussian:	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-a)^2}{2\sigma^2}\right)$	a	σ^2	$\exp(ra + r^2\sigma^2/2)$
Uniform:	$\frac{1}{a};\ 0{\leq}x{\leq}a$	$\frac{a}{2}$	$\frac{a^2}{12}$	$\frac{\exp(ra)-1}{ra}$

Name	**PMF** $\mathsf{p}_M(m)$	**Mean**	**Variance**	**MGF**
Binary:	$\mathsf{p}_M(1) = p;\ \mathsf{p}_M(0) = 1 - p$	p	$p(1-p)$	$1 - p + pe^r$
Binomial:	$\binom{n}{m} p^m (1-p)^{n-m};\ 0{\leq}m{\leq}n$	np	$np(1-p)$	$[1 - p + pe^r]^n$
Geometric:	$p(1{-}p)^{m-1};\ m{\geq}1$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^r}{1-(1-p)e^r};\ \text{for } r < \ln\frac{1}{1-p}$
Poisson:	$\frac{\lambda^n \exp(-\lambda)}{n!};\ n{\geq}0$	λ	λ	$\exp[\lambda(e^r - 1)]$

Bibliography

[1] Bellman, R., Dynamic Programming, Princeton University Press, Princeton, NJ., 1957.

[2] D. Bertsekas and J. Tsitsiklis, *An Introduction to Probability Theory*, 2nd Ed. Athena Scientific, Belmont, MA, 2008.

[3] Bertsekas, D. P., Dynamic Programming - Deterministic and Stochastic Models, Prentice Hall, Englewood Cliffs, NJ, 1987.

[4] Bhattacharya, R.N. & E.C. Waymire, *Stochastic Processes with Applications*, Wiley, NY, 1990.

[5] Dembo, A. and O. Zeitouni, *Large Deviation Techniques and Applications*, Jones & Bartlett Publishers, 1993.

[6] Doob, J. L., *Stochastic Processes*, Wiley, NY, 1953.

[7] Feller, W., *An Introduction to Probability Theory and its Applications*, vol 1, 3rd Ed., Wiley, NY, 1968.

[8] Feller, W., *An Introduction to Probability Theory and its Applications*, vol 2, Wiley, NY, 1966.

[9] Gallager, R.G., *Discrete Stochastic Processes*, Kluwer, Norwell MA 1996.

[10] Gantmacher, F. R., *The Theory of Matrices*, (English translation), Chelsea, NY, 1959 (2 volumes).

[11] Grimmett, G. and D. Starzaker, *Probability and Random Processes*, 3rd ed., Oxford University Press, Oxford G.B., 2001.

[12] Harris, T. E., *The Theory of Branching Processes*, Springer Verlag, Berlin, and Prentice Hall, Englewood Cliffs, NJ, 1963.

[13] Kelly, F.P. *Reversibility and Stochastic Networks*, Wiley, NY, 1979.

[14] Kolmogorov, A. N., *Foundations of the Theory of Probability*, Chelsea NY 1950 (Translation of *Grundbegriffe der Wahrscheinlchksrechnung*, Springer Berlin, 1933).

[15] Ross, S., *A First Course in Probability*, 4th Ed., McMillan & Co., 1994.

[16] Ross, S., *Stochastic Processes*, 2nd Ed., Wiley, NY, 1983.

[17] Rudin, W. *Real and Complex Analysis*, McGraw Hill, NY, 1966.

[18] Rudin, W. *Principles of Mathematical Analysis* 3rd Ed., McGraw Hill, NY, 1975.

[19] Strang, G. *Linear Algebra*, 4th Edition, Wellesley-Cambridge Press, Wellesley, MA 2009.

[20] Schweitzer, P. J. & A. Federgruen, "The Asymptotic Behavior of Undiscounted Value Iteration in Markov Decision Problems," *Math. of Op. Res.*, 2, Nov.1967, pp 360-381.

[21] Wald, A. *Sequential Analysis*, Wiley, NY, 1947.

[22] Wolff, R. W. *Stochastic Modeling and the Theory of Queues*, Prentice Hall, Englewood Cliffs, NJ, 1989.

[23] Yates, R. D. & D.J. Goodman, *Probability and Stochastic Processes*, Wiley, NY, 1999.

[24] Yates, R. D., *High Speed Round Robin Queueing Networks*, LIDS-TH-1983, M.I.T., Cambridge MA, 1983.

MIT OpenCourseWare
https://ocw.mit.edu/

6.262 Discrete Stochastic Processes
Spring 2011

For information about citing these materials or our Terms of Use, visit: https://ocw.mit.edu/terms.