5 Chernoff Bound

The Chernoff bound is an extremely useful bound on the tails of a sum of independent random variables. It is proved by bounding the moment generating function. This proof technique is interesting and important in its own right. We will see this proof method come up again later on when we prove martingale concentration inequalities. The method allows us to adapt the proof of the Chernoff bound to other distributions. Let us give the proof in the most basic case for simplicity and clarity.

Theorem 5.0.1 (Chernoff bound)
Let $S_n = X_1 + \cdots + X_n$ where $X_i \in \{-1, 1\}$ uniformly iid. Let $\lambda > 0$. Then

$$\mathbb{P}(S_n \geq \lambda\sqrt{n}) \leq e^{-\lambda^2/2}.$$

In contrast, Chebyshev's inequality gives a weaker bound $\mathbb{P}(S_n \geq \lambda\sqrt{n}) \leq 1/\lambda^2$. On the other hand, Chebyshev's inequality is application in wider settings as it only requires pairwise independence (for the second moment) as opposed to full independence.

Proof. Let $t \geq 0$. Consider the ***moment generating function***

$$\mathbb{E}\left[e^{tS_n}\right] = \mathbb{E}\left[e^{t\sum_i X_i}\right] = \mathbb{E}\left[\prod_i e^{tX_i}\right] = \prod_i \mathbb{E}\left[e^{tX_i}\right] = \left(\frac{e^{-t} + e^t}{2}\right)^n.$$

By comparing Taylor series, we have

$$\frac{e^{-t} + e^t}{2} = \sum_{k\geq 0} \frac{x^{2k}}{(2k)!} \leq \sum_{k\geq 0} \frac{x^{2k}}{k!2^k} = e^{t^2/2}.$$

By Markov's inequality,

$$\mathbb{P}(S_n \geq \lambda\sqrt{n}) \leq \frac{\mathbb{E}\left[e^{tS}\right]}{e^{t\lambda\sqrt{n}}} \leq e^{-t\lambda\sqrt{n}+t^2n/2}.$$

Setting $t = \lambda/\sqrt{n}$ gives the bound. □

***Remark* 5.0.2.** The technique of considering the moment generating function can be thought morally as taking an appropriately high moment. Indeed, $\mathbb{E}[e^{tS}] =$

$\sum_{n\geq 0} \mathbb{E}[S^n]t^n/n!$ contains all the moments data of the random variable.

The second moment method (Chebyshev + Markov) can be thought of as the first iteration of this idea. By taking fourth moments (now requiring 4-wise independence of the summands), we can obtain tail bounds of the form $\lesssim \lambda^{-4}$. And similarly with higher moments.

In some applications, where one cannot assume independence, but can estimate some high moments, the above philosophy can allow us to prove good tail bounds as well.

Also by symmetry, $\mathbb{P}(S_n \leq -\lambda\sqrt{n}) \leq e^{-\lambda^2/2}$. Thus we have the following two-sided tail bound.

Corollary 5.0.3

With S_n as before, for any $\lambda \geq 0$,

$$\mathbb{P}(|S_n| \geq \lambda\sqrt{n}) \leq 2e^{-\lambda^2/2}.$$

***Remark* 5.0.4.** It is easy to adapt the above proof so that each X_i is a mean-zero random variable taking $[-1, 1]$-values, and independent (but not necessarily identical) across all i. Indeed, by convexity, we have $e^{tx} \leq \frac{1-x}{2}e^{-t} + \frac{1+x}{2}e^{t}$ for all $x \in [-1, 1]$ by convexity, so that $\mathbb{E}[e^{tX}] \leq \frac{e^t+e^{-t}}{2}$. In particular, we obtain the following tail bounds on the binomial distribution.

Theorem 5.0.5 (Chernoff bound with bounded variables)

Let each X_i be an independent random variable taking values in $[-1, 1]$ and $\mathbb{E}X_i = 0$. Then $S_n = X_1 + \cdots + X_n$ satisfies

$$\mathbb{P}(S_n \geq \lambda\sqrt{n}) \leq e^{-\lambda^2/2}.$$

Corollary 5.0.6

Let X be a sum of n independent Bernoulli's (with not necessarily identitical probability). Let $\mu = \mathbb{E}X$ and $\lambda > 0$. Then

$$\mathbb{P}(X \geq \mu + \lambda\sqrt{n}) \leq e^{-\lambda^2/2} \quad \text{and} \quad \mathbb{P}(X \leq \mu - \lambda\sqrt{n}) \leq e^{-\lambda^2/2}$$

The Chernoff bound compares well to that of the normal distribution. For the standard normal $Z \sim N(0, 1)$, one has $\mathbb{E}[e^{tZ}] = e^{t^2/2}$ and so

$$\mathbb{P}(Z \geq \lambda) = \mathbb{P}(e^{tZ} \geq e^{t\lambda}) \leq e^{-t\lambda}\mathbb{E}[e^{tX}] = e^{-t\lambda+t^2/2}.$$

Set $t = \lambda$ and get

$$\mathbb{P}(Z \geq \lambda) \leq e^{-\lambda^2/2}.$$

And this is actually pretty tight, as, for $\lambda \to \infty$,

$$\mathbb{P}(Z \geq \lambda) = \frac{1}{\sqrt{2\pi}} \int_\lambda^\infty e^{-t^2/2}\, dt \sim \frac{e^{-\lambda^2/2}}{\sqrt{2\pi\lambda}}.$$

The same proof method allows you to prove bounds for other sums of random variables, which you can adjust based on the distributions. See the Alon–Spencer textbook, Appendix A, for examples of bounds and proofs.

For example, for a sum of independent Bernoulli random variables with small means, we can improve on the above estimates as follows.

Theorem 5.0.7

Let X be the sum of independent Bernoulli random variables (not necessarily the same probability). Let $\mu = \mathbb{E}X$. For all $\varepsilon > 0$,

$$\mathbb{P}(X \geq (1+\varepsilon)\mu) \leq e^{-((1+\varepsilon)\log(1+\varepsilon)-\varepsilon)\mu} \leq e^{-\frac{\varepsilon^2}{1+\varepsilon}\mu}$$

and

$$\mathbb{P}(X \leq (1-\varepsilon)\mu) \leq e^{-\varepsilon^2\mu/2}.$$

***Remark* 5.0.8.** The bounds for upper and lower tails are necessarily asymmetric when the probabilities are small. Why? Think about what happens when $X \sim \mathrm{Bin}(n, c/n)$, which, for a constant $c > 0$, converges as $n \to \infty$ to a Poisson distribution with mean c, whose value at k is $e^{-c}c^k/k! = e^{-\Theta(k \log k)}$ and not the sub-Gaussian decay $e^{-\Omega(k^2)}$ as one might naively predict by an incorrect application of the Chernoff bound formula.

Nonetheless, both formulas tell us that both tails exponentially decay like ε^2 for small values of $\varepsilon \in [0, 1]$.

5.1 Discrepancy

Given a hypergraph (i.e., set family), can we color the vertices red/blue so that every edge has roughly the same number of red versus blue vertices? (Contrast this problem to 2-coloring hypergraphs from Section 1.3.)

Theorem 5.1.1

Let $\mathcal{F}$ be a collection of m subsets of $[n]$. Then there exists some assignment $[n] \to \{-1, 1\}$ so that the sum on every set in $\mathcal{F}$ is $O(\sqrt{n \log m})$ in absolute value.

Proof. Put ± 1 iid uniformly at random on each vertex. On each edge, the probability that the sum exceeds $2\sqrt{n \log m}$ in absolute value is, by Chernoff bound, less than $2e^{-2 \log m} = 2/m^2$. By union bound over all m edges, with probability greater than $1 - 2/m \geq 0$, no edge has sum exceeding $2\sqrt{n \log m}$. □

***Remark* 5.1.2.** In a beautiful landmark paper titled *Six standard deviations suffice*, Spencer (1985) showed that one can remove the logarithmic term by a more sophisticated semirandom assignment algorithm.

Theorem 5.1.3 (Six standard deviations suffice: Spencer 1985)

Let $\mathcal{F}$ be a collection of n subsets of $[n]$. Then there exists some assignment $[n] \to \{-1, 1\}$ so that the sum on every set in $\mathcal{F}$ is at most $6\sqrt{n}$ in absolute value.

More generally, if $\mathcal{F}$ be a collection of $m \geq n$ subsets of $[n]$, then we can replace $6\sqrt{n}$ by $O(\sqrt{n \log(2m/n)})$.

***Remark* 5.1.4.** More generally, Spencer proves that the same holds if vertices have $[0, 1]$-valued weights.

The idea, very roughly speaking, is to first generalize from $\{-1, 1\}$-valued assignments to $[-1, 1]$-valued assignments. Then the all-zero vector is a trivially satisfying assignment. We then randomly, in iterations, alter the values from 0 to other values in $[-1, 1]$, while avoiding potential violations (e.g., edges with sum close to $6\sqrt{n}$ in absolute value), and finalizing a color of a color when its value moves to either -1 and 1.

Spencer's original proof was not algorithmic, and he suspected that it could not be made efficiently algorithmic. In a breakthrough result, Bansal (2010) gave an efficient algorithm for producing a coloring with small discrepancy. Lovett and Meka (2015) provided another element algorithm with a beautiful proof.

Here is a famous conjecture on discrepancy.

Conjecture 5.1.5 (Komlós)
There exists some absolute constant K so that for any $v_1, \dots, v_m \in \mathbb{R}^n$ all lying in the unit ball, there exist $\varepsilon_1, \dots, \varepsilon_m \in \{-1, 1\}$ such that

$$\varepsilon_1 v_1 + \cdots + \varepsilon_m v_m \in [-K, K]^n.$$

Banaszczyk (1998) proved the bound $K = O(\sqrt{\log n})$ in a beautiful paper using deep ideas from convex geometry.

Spencer's theorem's implies the special case of Komlós conjecture where all vectors v_i have the form $n^{-1/2}(\pm 1, \dots, \pm 1)$ (or more generally when all coordinates are $O(n^{-1/2})$). The deduction is easy when $m \le n$. When $m > n$, we use the following observation.

Lemma 5.1.6
Let $v_1, \dots, v_m \in \mathbb{R}^n$. Then there exists $a_1, \dots, a_m \in [-1, 1]^m$ with $|\{i : a_i \notin \{-1, 1\}\}| \le n$ such that

$$a_1 v_1 + \cdots + a_m v_m = 0$$

Proof. Find $(a_1, \dots, a_m) \in [-1, 1]^m$ satisfying and as many $a_i \in \{-1, 1\}$ as possible. Let $I = \{i : a_i \notin \{-1, 1\}\}$. If $|I| > n$, then we can find some nontrivial linear combination of the vectors $v_i, i \in I$, allowing us to to move $(a_i)_{i \in I}$'s to new values, while preserving $a_1 v_1 + \cdots + a_m v_m = 0$, and end up with at one additional a_i taking $\{-1, 1\}$-value. □

Let us explain how to deduce the special caes of Kómlos conjecture as stated earlier. Let $a_1, \dots, a_m$ and $I = \{i : a_i \notin \{-1, 1\}\}$ as in the Lemma. Take $\varepsilon_i = a_i$ for all $i \notin I$, and apply a corollary of Spencer's theorem to find $\varepsilon_i \in \{-1, 1\}^n$, $i \in I$ with

$$\sum_{i \in I} (\varepsilon_i - a_i) v_i \in [-K, K]^n,$$

which would yield the desired result. The above step can be deduced from Spencer's theorem by first assuming that each $a_i \in [-1, 1]$ has finite binary length (a compactness argument), and then rounding it off one digit at a time during Spencer's theorem, starting from the least significant bit (see Corollary 8 in Spencer's paper for details).

5.2 Nearly equiangular vectors

How many vectors can one place in $\mathbb{R}^d$ so that pairwise make equal angles?

5 *Chernoff Bound*

Let $S = \{v_1, \ldots, v_m\}$ be a set of unit vectors in $\mathbb{R}^n$ whose pairwise inner products all equal to some $\alpha \in [-1, 1)$. How large can S be?

The Gram matrix of S, defined as the matrix of pairwise inner products, has 1's on the diagonal and α off diagonal. So

$$\begin{pmatrix} | & \cdots & | \\ v_1 & \ddots & v_m \\ | & \cdots & | \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} | & \cdots & | \\ v_1 & \ddots & v_m \\ | & \cdots & | \end{pmatrix} = \begin{pmatrix} v_1 \cdot v_1 & \cdots & v_1 \cdot v_m \\ \vdots & \ddots & \vdots \\ v_m \cdot v_1 & \cdots & v_m \cdot v_m \end{pmatrix} = (1-\alpha) I_m + \alpha J_m$$

(here I_m and J_m are the $m \times m$ identity and all-ones matrix respectively). Since the eigenvalues of J_m are m (once) and 0 (repeated $m-1$ times), the eigenvalues of $I_m + (\alpha - 1)J_m$ are $(m-1)\alpha + 1$ (once) and $1 - \alpha$ ($m-1$ times). Since the Gram matrix is positive semidefinite, all its eigenvalues are nonnegative, and so $\alpha \geq -1/(m-1)$.

- If $\alpha \neq -1/(m-1)$, then this $m \times m$ matrix is non-singular, and since its rank is at most n (as $v_i \in \mathbb{R}^n$), we have $m \leq n$.
- If $\alpha = -1/(m-1)$, then this matrix has rank $m-1$, and we conclude that $m \leq n+1$.

It is left as an exercise to check all these bounds are tight.

Exercise: given m unit vectors in $\mathbb{R}^n$ whose pairwise inner products are all $\leq -\beta$, one has $m \leq 1 + \lfloor 1/\beta \rfloor$. (A bit more difficult: show that for $\beta = 0$, one has $m \leq 2n$).

What if instead of asking for exactly equal angles, we ask for approximately the same angle. It turns out that we can get many more vectors.

Theorem 5.2.1 (Exponentially many approximately equiangular vectors)
For every $\alpha \in (0, 1)$ and $\varepsilon > 0$, there exists $c > 0$ so that for every n, one can find at least 2^{cn} unit vectors in $\mathbb{R}^n$ whose pairwise inner products all lie in $[\alpha - \varepsilon, \alpha + \varepsilon]$.

***Remark* 5.2.2.** Such a collection of vectors is a type of "spherical code." Also, by examining the volume of spherical caps, one can prove an upper bound of the form $2^{C_{\alpha,\varepsilon} n}$.

Proof. Let $p = (1 + \sqrt{\alpha})/2$, and let $v_1, \ldots, v_m \in \{-1, 1\}^n$ be independent random vectors which each coordinate independently is $+1$ with probability p and -1 with probability $1 - p$. Then for $i \neq j$, the dot product $v_i \cdot v_j$ is a sum of n independent ± 1-valued random variables each with mean

$$p^2 + (1-p)^2 - 2p(1-p) = (p - (1-p))^2 = (2p-1)^2 = \alpha.$$

Applying Chernoff bound in the form of Theorem 5.0.5 (after linear transformation on each variable to make each term taking value in $[-1, 1]$ and mean centered at zero), we get

$$\mathbb{P}\left(|v_i \cdot v_j - n\alpha| \geq n\varepsilon\right) \leq 2e^{-\Omega(n\varepsilon^2)}.$$

By the union bound, the probability that $|v_i \cdot v_j - n\alpha| > n\varepsilon$ for some $i \neq j$ is $< m^2 e^{\Omega(n\varepsilon^2)}$, which is < 1 for some m at least 2^{cn}. So with positive probability, so such pair occurs, and then $v_1/\sqrt{n}, \ldots, v_m/\sqrt{n}$ is a collection of unit vectors in $\mathbb{R}^n$ whose pairwise inner products all lie in $[\alpha - \varepsilon, \alpha + \varepsilon]$. □

***Remark* 5.2.3** (Equiangular lines with a fixed angle)**.** Given a fixed angle θ, for large n, how many lines in $\mathbb{R}^n$ through the origin can one place whose pairwise angles are all exactly θ? This problem was solved by Jiang, Tidor, Yao, Zhang, Zhao (2021). This is the same as asking for a set of unit vectors in $\mathbb{R}^n$ whose pairwise inner products are $\pm\alpha$. It turns out that for fixed α, the maximum number of lines grows linearly with the dimension n, and the rate of growth depends on properties of α in relation to spectral graph theory. We refer to the cited paper for details.

5.3 Hajós conjecture counterexample

We begin by reviewing some classic result from graph theory. Recall some definitions:

- H is an ***induced subgraph*** of G if H can be obtained from G by removing vertices;
- H is a ***subgraph*** if G if H can be obtained from G by removing vertices and edges;
- H is a ***subdivision*** of G if H can be obtained from a subgraph of G by contracting induced paths to edges;
- H is a ***minor*** of G if H can be obtained from a subgraph of G by by contracting edges to vertices.

Kuratowski's theorem (1930). Every graph without $K_{3,3}$ and K_5 as subdivisions as subdivision is planar.

Wagner's theorem (1937). Every graph free of $K_{3,3}$ and K_5 as minors is planar.

(There is a short argument shows that Kuratowski and Wagner's theorems are equivalent.)

Four color theorem (Appel and Haken 1977) Every planar graph is 4-colorable.

Corollary: Every graph without $K_{3,3}$ and K_5 as minors is 4-colorable.

The condition on K_5 is clearly necessary, but what about $K_{3,3}$? What is the "real" reason for 4-colorability?

Hadwidger's conjecture, below, remains a major conjectures in graph theory.

Conjecture 5.3.1 (Hadwiger 1936)
For every $t \geq 1$, every graph without a K_{t+1} minor is t-colorable.

- $t = 1$ trivial
- $t = 2$ nearly trivial (if G is K_3-minor-free, then it's a tree)
- $t = 3$ elementary graph theoretic arguments
- $t = 4$ is equivalent to the 4-color theorem (Wagner 1937)
- $t = 5$ is equivalent to the 4-color theorem (Robertson–Seymour–Thomas 1994; this work won a Fulkerson Prize)
- $t \geq 6$ remains open

Let us explore a variation of Hadwiger's conjecture:

Hajós conjecture. (1961) Every graph without a K_{t+1}-subdivision is t-colorable.

Hajós conjecture is true for $t \leq 3$. However, it turns out to be false in general. Catlin (1979) constructed counterexamples for all $t \geq 6$ ($t = 4, 5$ are still open).

It turns out that Hajós conjecture is not just false, but very false.

Erdős–Fajtlowicz (1981) showed that almost every graph is a counterexample (it's a good idea to check for potential counterexamples among random graphs!)

Theorem 5.3.2
With probability $1 - o(1)$, $G(n, 1/2)$ has no K_t-subdivision with $t = \lceil 10\sqrt{n} \rceil$.

From Theorem 4.4.3 we know that, with high probability, $G(n, 1/2)$ has independence number $\sim 2 \log_2 n$ and hence chromatic number $\geq (1 + o(1)) \frac{n}{2 \log_2 n}$. Thus the above result shows that $G(n, 1/2)$ is whp a counterexample to Hajós conjecture.

Proof. If G had a K_t-subdivision, say with $S \subseteq V$, $|S| = t$. Each pair of vertices of S are connected via a path, whose intermediate vertices are outside S, and distinct for different pairs of vertices.

At most n of the $\binom{t}{2}$ pairs of vertices in S can be joined this way using a path of at least two edges, since each uses up a vertex outside S. Thus at $\geq \binom{t}{2} - n$ of the pairs of vertices of S form edges.

By Chernoff bound, for fixed t-vertex subset S

$$\mathbb{P}\left(e(S) \geq \binom{t}{2} - n\right) \leq \mathbb{P}\left(e(S) \geq \frac{3}{4}\binom{t}{2}\right) \leq e^{-t^2/10}.$$

Taking a union bound over all t-vertex subsets S, and noting that

$$\binom{n}{t} e^{-t^2/10} < n^t e^{-t^2/10} \leq e^{-10n + O(\sqrt{n}\log n)} = o(1)$$

we see that whp no such S exists, so that this $G(n, 1/2)$ whp has no K_t-subdivision □

***Remark* 5.3.3** (Quantitative question)**.** One can ask the following quantitative question regarding Hadwidger's conjecture:

Can every graph without a K_{t+1}-minor can be properly colored with a small number of colors?

Wagner (1964) showed that every graph without K_{t+1}-minor is 2^{t-1} colorable.

Here is the proof: assume that the graph is connected. Take a vertex v and let L_i be the set of vertices with distance exactly i from v. The subgraph induced on L_i has no K_t-minor, since otherwise such a K_t-minor would extend to a K_{t+1}-minor with v. Then by induction L_i is 2^{t-2}-colorable (check base cases), and using alternating colors for even and odd layers L_i yields a proper coloring of G.

This bound has been improved over time. Delcourt and Postle (2021+) showed that every graph with no K_t-minor is $O(t \log\log t)$-colorable.

For more on Hadwiger's conjecture, see Seymour's survey (2016).

Exercises

1. Prove that with probability $1 - o(1)$ as $n \to \infty$, every bipartite subgraph of $G(n, 1/2)$ has at most $n^2/8 + 10n^{3/2}$ edges.
2. *Unbalancing lights.* Prove that there is a constant C so that for every positive integer n, one can find an $n \times n$ matrix A with $\{-1, 1\}$ entries, so that for all vectors $x, y \in \{-1, 1\}^n$, $|y^\intercal A x| \leq Cn^{3/2}$.
3. Prove that there exists a constant $c > 1$ such that for every n, there are at least c^n points in $\mathbb{R}^n$ so that every triple of points form a triangle whose angles are all less than $61°$.

4. *Planted clique.* Give a deterministic polynomial-time algorithm for the following task so that it succeeds over the random input with probability approaching 1 as $n \to \infty$.

   Input: some unlabeled n-vertex G created as the union of $G(n, 1/2)$ and a clique on $t = \left\lfloor 100\sqrt{n \log n} \right\rfloor$ vertices.

   Output: a clique in G of size t.

5. *Weighing coins.* You are given n coins, each with one of two known weights, but otherwise indistinguishable. You can use a scale that outputs the combined weight of any subset of the coins. You must decide in advance which subsets $S_1, \ldots, S_k \subseteq [n]$ of the coins to weigh. We wish to determine the minimum number of weighings needed to identify the weight of every coin. (Below, X and Y represent two possibilities for which coins are of the first weight.)

   a) $\star$ Prove that if $k \le 1.99n/\log_2 n$ and n is sufficiently large, then for every $S_1, \ldots, S_k \subseteq [n]$, there are two distinct subsets $X, Y \subseteq [n]$ such that $|X \cap S_i| = |Y \cap S_i|$ for all $i \in [k]$.

      (There is a neat solution to part (a) using information theory, though here you are explicitly asked to solve it using the Chernoff bound.)

   b) $\star$ Show that there is some constant C such that (a) is false if 1.99 is replaced by C. (What is the best C you can get?)

18.226 Probabilistic Methods in Combinatorics

Fall 2022