6.7220 | 15.084 — Final Review

Spring 2025

Final administrivia

Policies

- **Cheatsheet**: two sheets (front and back), i.e., 4 pages. No other written or digital materials are permitted.
- **Attendance**: only students registered for *credit* (no listeners)
- **Scope**: Everything seen so far, with a focus on the second half
- **Collaboration**: individual final (no collaboration)

Organization for today

1. Final's goals
2. Review of the important concepts and tools
3. Your questions

Goal of final

- The goals of the final:
  - Test your understanding of all the important ideas covered in the course
  - To some extent, give you another possibility to show your understanding of the material covered in the midterm
  - Test your ability to implement (simple) optimization algorithms
- The final will **not** require *complex calculations*
- The final will **not** require long, difficult proofs
- Questions will generally require providing arguments in favor or against a statement (*i.e.*, either a proof or a counterexample)
- Much **shorter** (and easier) than problem sets
- Similar length as the midterm

Examples of questions (1/2)

- Show that a set is convex
- Show that a function is convex
- Construct a separation oracle
- Determine the convex cone at a point
- Prove the problem admits a solution
- Compute the solution to an optimization problem
- Write KKT conditions for a problem
- Determine whether constraint qualifications hold
- Prove separating hyperplane theorem
- Strong / strict convexity proofs

Examples of questions (2/2)

- Work out the update of gradient descent, mirror descent, ...
- Prove convergence of an optimization algorithm
- The PŁ condition: definition, relationship with strong convexity, consequences with respect to convergence of gradient descent
- Prove the gradient descent lemma / Euclidean mirror descent lemma / full mirror desent lemma
- Prove a function is self-concordant
- Work out the Newton step for a function
- Determine the second-order direction of descent
- Prove the Newton's method lemma
- Show existence of solutions for lower bounded PŁ functions and self-concordant functions

Any questions?

Gradient descent

- Definition of gradient descent update:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

- Definition of projected gradient descent update:

$$x_{t+1} = \Pi_{\Omega}(x_t - \eta \nabla f(x_t))$$

- Definition of mirror descent update:

$$x_{t+1} = \arg\min_{x \in \Omega} \eta \langle \nabla f(x_t), x - x_t \rangle + D_{\varphi}(x \parallel x_t)$$

L-smoothness and PŁ condition

Definition

(**L-smoothness**) A differentiable function $f : \Omega \to \mathbb{R}$ is L-smooth if its gradient is L-Lipschitz continuous, that is,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \qquad \forall x, y \in \Omega.$$

Definition

(**PŁ condition**) A lower-bounded function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the *Polyak-Łojasiewicz (PŁ) condition* if there exists $\mu > 0$ such that

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f_\star) \qquad \forall x \in \mathbb{R}^n,$$

where $f_\star := \inf f$.

Descent lemmas

Theorem

(**Gradient descent lemma**) Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth. Then, for any $0 < \eta \leq \frac{1}{L}$, each step of gradient descent guarantees

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|_2^2.$$

Theorem

(**Euclidean mirror descent lemma**) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex and differentiable function. Then, for any choice of stepsize η, any two consecutive points (x_t, x_{t+1}) produced by the gradient descent algorithm satisfy

$$f(x_t) \leq f(y) + \frac{1}{2\eta}\left(\|y - x_t\|_2^2 - \|y - x_{t+1}\|_2^2 + \|x_{t+1} - x_t\|_2^2\right) \qquad \forall y \in \mathbb{R}^n$$

Descent lemmas

- You should know how to prove the descent lemmas
- It should be clear to you the difference in spirit between the two lemmas
- You should be able to analyze (unconstrained) gradient descent pretty easily
- No need to know the details of the proofs of linear coupling

Distance-generating function and Bregman divergence

Definition

Let $\varphi : \Omega \to \mathbb{R}$ be a differentiable and μ-strongly convex ($\mu > 0$) function with respect to a norm $\|\cdot\|$, that is, satisfy

$$\varphi(x) \geq \varphi(y) + \langle \nabla\varphi(y), x - y\rangle + \frac{\mu}{2}\|x - y\|^2 \qquad \forall x, y \in \Omega.$$

The Bregman divergence *centered in* $y \in \Omega$ is the function $\mathrm{D}_\varphi(x \,\|\, y)$ defined as

$$\mathrm{D}_\varphi(x \,\|\, y) := \varphi(x) - \varphi(y) - \langle \nabla\varphi(y), x - y\rangle.$$

- Basic properties
  - Positive definiteness
  - Strong convexity given a fixed center
  - ...

Proximal steps

- Definition
- Special cases:
  - Squared Euclidean norm

Theorem

Consider the squared Euclidean norm distance-generating function $\varphi(x) = \frac{1}{2}\|x\|_2^2$. Then, proximal steps and projected gradient steps are equivalent:

$$\operatorname{Prox}_{\varphi}(\eta \nabla f(x), x) = \Pi_{\Omega}\Big(x - \eta \nabla f(x)\Big) \qquad \forall x \in \Omega.$$

  - Negative entropy
- You should be able to prove that proximal steps exist and are unique

Stochastic gradient descent

- Empirical risk minimization problems

$$J_{\text{emp}}(\theta) := \frac{1}{k} \sum_{i=1}^{k} \ell(g_\theta(z_i), y_i),$$

where $\ell : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ is a loss function that measures the discrepancy between the model's prediction $g_\theta(z_i)$ and the true label y_i.

Stochastic descent lemmas

Theorem

(**Stochastic gradient descent lemma**) Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth and $\eta > 0$ be an arbitrary stepsize. Two consecutive iterates (x_t, x_{t+1}) produced by the stochastic gradient descent algorithm satisfy

$$\mathbb{E}_t\left[f(x_{t+1})\right] \leq f(x_t) - \eta\|\nabla f(x_t)\|_2^2 + \frac{L}{2}\eta^2\ \mathbb{E}_t\left[\left\|\tilde{\nabla} f(x_t)\right\|_2^2\right].$$

Theorem

(**Stochastic Euclidean mirror descent lemma**) Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Then, for any stepsize $\eta > 0$ and $y \in \mathbb{R}^n$, two consecutive iterates (x_t, x_{t+1}) produced by the SGD algorithm satisfy

$$f(x_t) \leq f(y) + \frac{1}{2\eta}\mathbb{E}_t\left[\|x_t - y\|_2^2 - \left\|x_{t+1} - y\right\|_2^2 + \left\|x_t - x_{t+1}\right\|_2^2\right].$$

- Final guarantee:

Theorem

Consider the SGD algorithm run on a convex function. By picking $\eta = \frac{1}{\sqrt{GT}}$, this immediately implies that at least one of the iterates $x_t, t \in \{0, \ldots, T-1\}$, satisfies

$$\mathbb{E}[f(x_t) - f(x_\star)] \leq \frac{\sqrt{G}}{2}\left(1 + \|x_0 - x_\star\|_2^2\right)\frac{1}{\sqrt{T}}.$$

By convexity, the same bound holds also for $\mathbb{E}\left[f\left(\overline{x}^T\right) - f(x_\star)\right]$, where $\overline{x}^T$ is the average of the iterates $x_0, \ldots, x_{T-1}$.

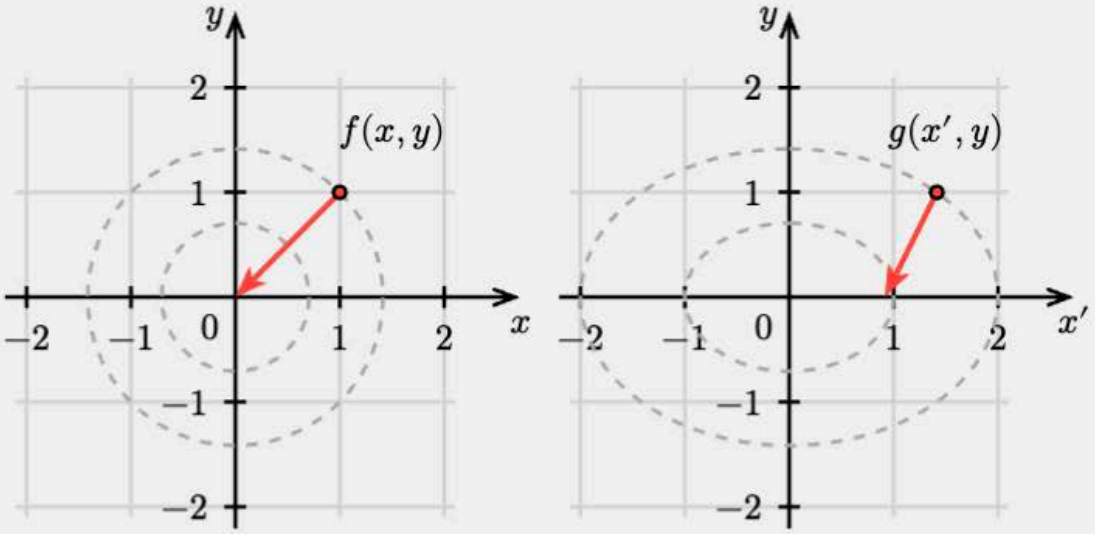
Preconditioning

- The first-order methods we have seen so far are sensitive to rescaling / reparameterizations / linear transformations of variables.

Example. As a small numerical example, consider the objective function (say, parameterized in meters) $f(x, y) = \frac{1}{2}x^2 + \frac{1}{2}y^2$. After a change of units of x, consider now the reparameterized objective

$$g(x', y) = f\left(\frac{x'}{\sqrt{2}}, y\right) = \frac{1}{4}x'^2 + \frac{1}{2}y^2,$$

plotted on the right.



The two plots show contour lines for the respective objectives, together with an identical initial point (up to reparameterization). The red arrows show the gradient descent direction at the initial point. It is clear that after one step of gradient descent, the two points will no longer be equivalent.

AdaGrad & ADAM

At each iteration t, AdaGrad keeps a tally of the sum of the squared gradients up to time t for each variable. This is done by maintaining a vector s_t of components

$$[s_t]_i := \sqrt{\sum_{\tau=0}^{t} [\nabla f(x_\tau)]_i^2},$$

where $[\nabla f(x_t)]_i$ is the i-th component of the gradient at time t. The update rule for AdaGrad is then

$$x_{t+1} = x_t - \eta M_t^{-1} \nabla f(x_t),$$

where

$$M_t := \operatorname{diag}\left([s_t]_i := \sqrt{\sum_{\tau=0}^{t} [\nabla f(x_\tau)]_i^2} : i = 1, ..., n \right).$$

AdaGrad & Adam

- The same algorithm can be used in the stochastic setting, where as usual the gradient is replaced by a stochastic gradient
- It can also be used in the projected setting, where the update is projected onto a feasible set
- **ADAM**: AdaGrad with Momentum
  - At each iteration t, ADAM keeps track of the momentum (discounted sum of past gradients)

$$g_t = \gamma g_{t-1} + (1-\gamma)\nabla f(x_t); \qquad g_{-1} := 0.$$

  - The scaling factors s_t are also accumulated with a discount rate β as

$$[s_t]_i^2 = \beta[s_{t-1}]_i^2 + (1-\beta)[\nabla f(x_t)]_i^2 \qquad i = 1, ..., n; \qquad s_{-1} := 0.$$

  - Finally, ADAM updates the iterate as follows:

$$x_{t+1} = x_t - \eta\ M_t^{-1}\ g_t, \quad \text{where } M_t := \text{diag}(s_t)$$

Equivalence between AdaGrad and mirror descent

Theorem

The AdaGrad update rule is equivalent to the mirror descent update rule with the distance-generating function $\varphi_t(x) = \frac{1}{2} x^\top M_t x$, where $M_t := \operatorname{diag}(s_t)$.

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex and differentiable function. AdaGrad is competitive with the best preconditioner in hindsight: for any choice of coefficients $\lambda_i \geq 0, i = 1, \ldots, n$, AdaGrad with stepsize $\eta = D/\sqrt{2}$ satisfies

$$\frac{1}{T} \sum_{t=0}^{T-1} f(x_t) \leq f(x_\star) + \frac{\sqrt{2n} D}{T} \sqrt{\min_{\lambda \in \mathbb{R}^n_{\geq 0}, \|\lambda\|_1 = n} \sum_{t=0}^{T-1} \nabla f(x_t)^\top \operatorname{diag}(\lambda)^{-1} \nabla f(x_t)},$$

where $D := \max_{t=0}^{T} \|x_t - x_\star\|_\infty$.

Hessian preconditioning & Newton's method

- Ultimate form of preconditioning: Hessian preconditioning
  - Leads to affine invariance
- It also matches the second-order direction of descent we would get from the Taylor expansion:

$$\underbrace{d = -\nabla f(x_t)}_{\text{using first-order Taylor approximation}} \quad \text{to} \quad \underbrace{d = -\left[\nabla^2 f(x_t)\right]^{-1} \nabla f(x_t)}_{\text{using second-order Taylor approximation}}$$

- Damped Newton's method:

$$x_{t+1} = x_t - \eta \left[\nabla^2 f(x_t)\right]^{-1} \nabla f(x_t)$$

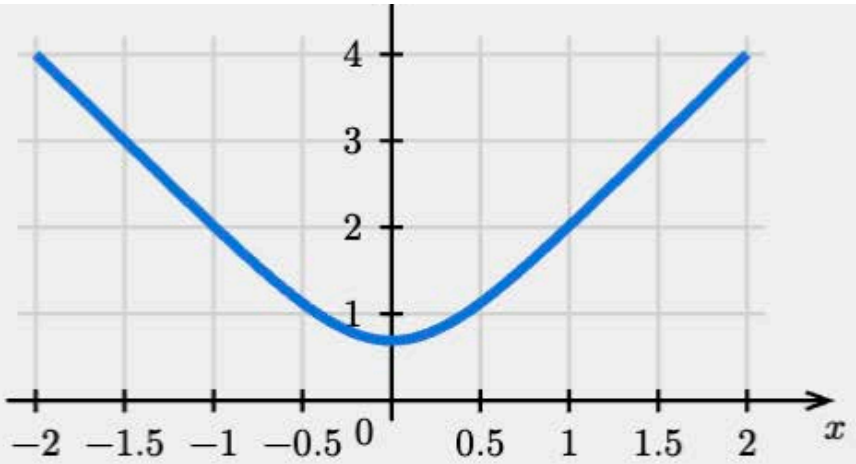
- "Newton's method": special case of $\eta = 1$

$$f(x) = \log(e^{2x} + e^{-2x}),$$

plotted on the right, whose gradient and Hessian are respectively computed as

$$\nabla f(x) = 2 \cdot \frac{e^{4x} - 1}{e^{4x} + 1},$$

$$\nabla^2 f(x) = 16 \cdot \frac{e^{4x}}{(e^{4x} + 1)^2}.$$



The two tables below show the first 10 iterates of Newton's method and gradient descent when applied to $f(x)$ starting at two close initial points: $x_0 = 0.5$ on the left, and $x_0 = 0.7$ on the right. As you can see, the behavior of Newton's method is very different: while it converges extremely quickly to the minimum when starting at $x_0 = 0.5$, it diverges when starting at $x_0 = 0.7$.

	Newton's method	GD ($\eta = 0.1$)
$t = 0$	**0.5000**	**0.5000**
$t = 1$	−0.4067	0.3477
$t = 2$	0.2047	0.2274
$t = 3$	−0.0237	0.1422
$t = 4$	3.53×10^{-5}	0.0868
$t = 5$	-1.17×10^{-13}	0.0524
$t = 6$	-1.14×10^{-17}	0.0315
$t = 7$	0.0000	0.0189
$t = 8$	0.0000	0.0114

	Newton's method	GD ($\eta = 0.1$)
$t = 0$	**0.7000**	**0.7000**
$t = 1$	−1.3480	0.5229
$t = 2$	26.1045	0.3669
$t = 3$	-2.79×10^{44}	0.2418
$t = 4$	diverged	0.1520
$t = 5$	diverged	0.0930
$t = 6$	diverged	0.0562
$t = 7$	diverged	0.0338
$t = 8$	diverged	0.0203

Classical analysis

Theorem

(**Local quadratic convergence**) Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable with M-Lipschitz continuous Hessian, and $x_\star$ be a local minimum such that

$$\nabla f(x_\star) = 0, \qquad \text{and} \qquad \nabla^2 f(x_\star) \succeq \mu I$$

for some $\mu > 0$. Then, as long as we start Newton's method from x_0 s.t.

$$\|x_0 - x_\star\|_2 \leq \frac{\mu}{2M}$$

from the local minimum, the distance to optimality of the iterates x_t generated by Newton's method decays as

$$\frac{\|x_{t+1} - x_\star\|_2}{\mu/M} \leq \left(\frac{\|x_t - x_\star\|_2}{\mu/M} \right)^2.$$

Classical analysis

- Idea of the proof:

Theorem

The distance to optimality of the iterates x_t generated by the damped Newton's method with stepsize $\eta > 0$ satisfy

$$x_{t+1} - x_\star = (I - \eta H_t)(x_t - x_\star),$$

where

$$H_t := \left[\nabla^2 f(x_t)\right]^{-1} \int_0^1 \nabla^2 f(x_\star + \lambda(x_t - x_\star))\, \mathrm{d}\lambda.$$

- Then, bounding the spectral norm of $I - \eta H_t$ is a worthwhile goal.

Classical analysis

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable, μ-strongly convex and L-smooth. Then, the distance to optimality of the iterates x_t generated by damped Newton's method with stepsize $\eta \leq \frac{\mu}{L}$ decays exponentially fast at the rate

$$\left\|x_{t+1} - x_\star\right\|_2 \leq \frac{L}{\mu}\left(1 - \eta\frac{\mu}{L}\right)^t \|x_1 - x_\star\|_2.$$

- We already know that gradient descent can reach a similar convergence rate for the same class of smooth and strongly convex function, without even needing to invert the Hessian (see Lecture 7, section on convergence rate for smooth PŁ functions).

Self-concordance

- Two shortcomings of the standard analysis
  - Lack of affine invariance: relies on extrinsic norms (e.g., Euclidean norm)
  - Fails to capture important cases of interest

Example 1.1.

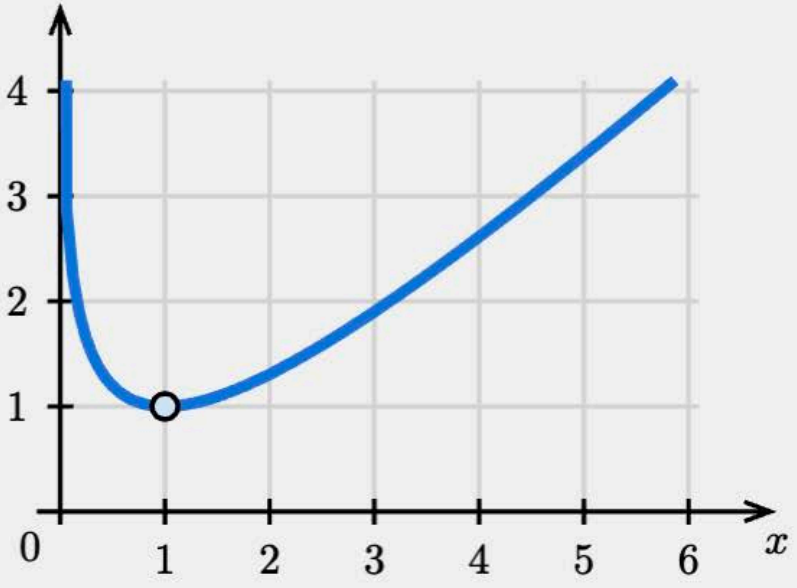
Consider the function $f : (0, \infty) \to \mathbb{R}$

$$f(x) = x - \log(x),$$

whose minimum is at $x = 1$. The update of Newton's method is

$$\begin{aligned} x_{t+1} &= x_t - [f''(x_t)]^{-1} f'(x_t) \\ &= x_t - x_t^2 \cdot \left(1 - \frac{1}{x_t}\right) = 2x_t - x_t^2. \end{aligned}$$

So, we have $1 - x_{t+1} = 1 + x_t^2 - 2x_t = (1 - x_t)^2$.



Self-concordance

- We need to move away from Lipschitz continuity of the Hessian
- Conceptually we want this condition:
  - Sufficiently close to each point…
  - …the quadratic approximation of the function is good
- Questions:
  - **Q1**: How to measure "sufficiently close" in an intrinsic way?
  - **Q2**: How to define "good" approximation?
- Answers:
  - **A1**: Use the *intrinsic norm* induced by the Hessian

$$\|v\|_x := \sqrt{\langle v, \nabla^2 f(x) v \rangle}$$

  - **A2**: Require that the Hessian change is bounded by intrinsic distance

Self-concordance

Definition

Let $\Omega \subseteq \mathbb{R}^n$ be open and convex. A twice-differentiable function $f : \Omega \to \mathbb{R}$ is said to be *strongly nondegenerate self-concordant* if:

1. The Hessian of f is positive definite everywhere on Ω;
2. The ellipsoid $W(x) := \left\{y \in \mathbb{R}^n : \|y - x\|_x^2 < 1\right\} \subseteq \Omega$ for all $x \in \Omega$; and
3. Inside of the ellipsoid $W(x)$, the function f is *almost quadratic*: for all $x \in \Omega, y \in W(x)$,

$$(1 - \|y - x\|_x)^2 \nabla^2 f(x) \preceq \nabla^2 f(y) \preceq \frac{1}{(1 - \|y - x\|_x)^2} \nabla^2 f(x)$$

which is equivalent to the statement

$$(1 - \|y - x\|_x)\|v\|_x \leq \|v\|_y \leq \frac{\|v\|_x}{1 - \|y - x\|_x} \qquad \forall x \in \Omega, y \in W(x), v \in \mathbb{R}^n.$$

Self-concordance

Equivalent characterization:

Theorem

If $f : \Omega \to \mathbb{R}$ is three-time differentiable with positive definite Hessian everywhere on Ω, then strong nondegenerate self-concordance is equivalent to f satisfying the following two properties:

1. for any $x_0 \in \Omega$ and $d \in \mathbb{R}^n$, the restriction $\varphi(\gamma) := f(x_0 + \gamma d)$ of f to the segment $\{\gamma : x_0 + \gamma d \in \Omega\}$ satisfies

$$\varphi'''(\gamma) \leq 2\varphi''(\gamma)^{3/2}; \qquad \text{and}$$

2. any sequence $\{x_k\}$ converging to a point on the boundary of Ω is such that $f(x_k) \to +\infty$.

Notable self-concordant functions

- $f(x) = -\sum_{i=1}^{n} \log(x_i)$ on the domain $\Omega = \mathbb{R}^n_{>0}$. We saw this in class together.
- $f(x) = -\log\det(x)$ on the set of positive definite matrices. This was part of HW5.
- $f(x) = -\log(1 - \|x\|_2^2)$ on the open unit ball $B_2(0,1) \subseteq \mathbb{R}^n$. This was also part of HW5.
- $f(x) = -\sum_i \log(b_i - a_i^\top x)$ for the open polytope $\{a_i^\top x < b_i\}$.

Properties of self-concordant functions

Theorem

Let $f : \Omega \to \mathbb{R}$ be self-concordant and lower bounded. Then, f attains a unique minimum.

Newton's method on self-concordant functions works especially well. The quantity

$$n(x) := -\left[\nabla^2 f(x)\right]^{-1} \nabla f(x).$$

is especially important for capturing the convergence.

Newton's method on self-concordant functions

Theorem

Let $f : \Omega \to \mathbb{R}$ be self-concordant. If a point $x \in \Omega$ is such that $\|n(x)\|_x \leq 1/9$, then there exists a minimum z of f within distance

$$\|z - x\|_x \leq 3 \cdot \|n(x)\|_x.$$

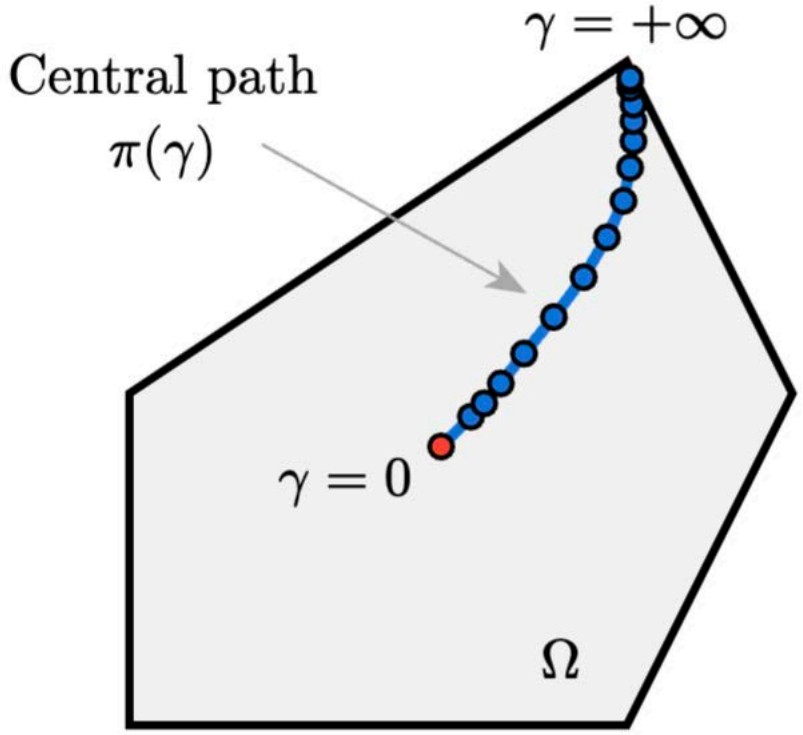
Theorem

Let $f : \Omega \to \mathbb{R}$ be self-concordant. If a point $x_t \in \Omega$ is such that $\|n(x_t)\|_{x_t} < 1$, then

$$\|n(x_{t+1})\|_{x_{t+1}} \leq \left(\frac{\|n(x_t)\|_{x_t}}{1 - \|n(x_t)\|_{x_t}} \right)^2.$$

Interior-point methods

- Idea: chase the **central path** to find the minimum of $\langle c, x \rangle$ on $\overline{\Omega}$:



$$\pi(\gamma) := \arg\min_{x} \ \gamma\langle c, x \rangle + f(x)$$
$$\text{s.t.} \quad x \in \Omega.$$

Interior-point methods

- The function $f(x)$ should be a **barrier** function

Definition

(**Barrier function**) A *strongly nondegenerate self-concordant barrier* is a strongly nondegenerate self-concordant function f whose complexity parameter

$$\theta_f := \sup_{x \in \Omega} \|n(x)\|_x^2$$

is *finite*.

Example

The *logarithmic barrier* for the positive orthant $\mathbb{R}^n_{>0}$, defined as

$$f : \mathbb{R}^n_{>0} \to \mathbb{R} \qquad \text{where} \qquad f(x) = -\sum_{i=1}^{n} \log(x_i)$$

has complexity parameter $\theta_f = n$.

Short-step barrier

- **Idea**: follow the central path *closely* (step norm $\leq \frac{1}{9}$) at every iteration

$$\gamma_{t+1} := \beta\gamma_t, \qquad x_{t+1} := x_t - [\nabla^2 f(x_t)]^{-1}(\gamma_{t+1}c + \nabla f(x_t))$$

Theorem

If x_t is close to the central path, in the sense that $\left\|n_{\gamma_t}(x_t)\right\|_{x_t} \leq \frac{1}{9}$, then by setting

$$\gamma_{t+1} := \beta\gamma_t \qquad \text{with} \quad \beta := \left(1 + \frac{1}{8\sqrt{\theta_f}}\right),$$

the same proximity is guaranteed at time $t+1$, that is, $\left\|n_{\gamma_{t+1}}(x_{t+1})\right\|_{x_{t+1}} \leq \frac{1}{9}$.

- How to find the starting point? Path switching and auxiliary central path

Any questions?

How to practice

- Look back at homework
- Look back at the midterm [+ make-up midterm + 2024 final]
- Look back at lecture notes, including all the examples
- Look back at exercises in recitation. New concepts introduced in recitation (for example, subdifferentials) will **not** appear in the final
- Piazza also has some posts you might find interesting
- Several comments in grey left in the lecture notes pointing to opportunities to try to complete arguments orthink about extensions

Office hours

We are here to **help**:

- Office hours will continue as usual this week

Beyond this course

- I hope you all enjoyed this course and that you got to appreciate the beauty of optimization theory and algorithms
- We got to see a lot of deep and beautiful results
  1. Normal cones $\rightarrow$ Lagrange multipliers $\rightarrow$ duality
  2. Separation as the driver of duality. NP $\cap$ coNP. Certificates of infeasibility (Farkas lemma)
  3. Separation is an algorithmic tool (Ellipsoid method)
  4. Convexity and sufficiency of first-order conditions
  5. Effect of curvature (strong convexity, self-concordance, PŁ functions)
  6. Equivalence between projections, linear optimization, convex optimization, separation. Polarity as a form of duality.
  7. Conic optimization and modern applications in SOS / polynomial opt.
  8. First-order methods as minimization of first-order approximations
  9. Second-order methods as minimization of second-order approximations
  10. Newton and interior-point methods. Chasing the central path efficiently

Beyond this course

- This was a pretty ambitious course. We jumped straight into duality ideas and were not shy away from a lot of modern developments
- Optimization is a very vibrant field. We had to make choices for this course, and while I think we made the right ones, there are a lot of other topics that were could not cover
  - Connections between optimization and sampling (e.g., diffusion, log-concave sampling, etc.)
  - Calculus of variations
  - Optimal control
  - Online optimization / game theory
  - Zeroth-order optimization / Bayesian optimization
    (I uploaded some (very introductory) notes I wrote on Bayesian opt. to Canvas)
  - Semi-algebraic methods in optimization
  - Optimization for deep learning (NTK / edge of stability / etc.)

6.7220 Nonlinear Optimization

Spring 2025