Lecture 18: Course Summary

TTIC 31070 / CMSC 35470 / BUSF 36903 / STAT 31015

Convex Optimization

Prof. Zhiyuan Li

Spring 2026

What is this course about?

1. **Formalization**
2. **Analysis**
3. **Algorithms**

What is this course about?

1. Formalization 2. Analysis 3. Algorithms

Formalize concrete problems as convex optimization problems.

- Recognize convex objectives and feasible sets.
- Rewrite problems as LPs, QPs, SOCPs, SDPs, or conic programs.
- Specify the computational access model: first-order, projection, LMO, separation, stochastic, or barrier oracles.

What is this course about?

1. Formalization **2. Analysis** 3. Algorithms

Understand structural properties of convex sets, objectives, and programs.

- Separation, subgradients, conjugates, support functions, and perturbation functions.
- Weak duality, KKT certificates, strong duality, Slater's condition, and failure modes.
- Smoothness, strong convexity, relative smoothness, self-concordance, and their algorithmic implications.

What is this course about?

1. Formalization 2. Analysis **3. Algorithms**

Design and analyze first-order and second-order algorithms.

- First-order algorithms: cutting planes, steepest descent, mirror descent, subgradient descent, adaptive methods, Frank-Wolfe, and acceleration.
- Second-order algorithms: Newton's method and interior-point methods.
- Proof tools: descent lemmas, three-point inequalities, Bregman telescopes, Newton decrease, and central-path progress.
- Compare upper and lower bounds, and know which algorithm to use with which geometry and assumptions.

LP and Its Two Generalizations

Linear Programming

$$\min_x \quad c^\top x \quad \text{s.t.} \quad a_i^\top x - b_i \leq 0, \quad i = 1, \ldots, m.$$

Functional constraints

generalize the scalar function

$$a_i^\top x - b_i \quad \rightsquigarrow \quad f_i(x)$$

$$\min_x \quad f_0(x)$$
$$\text{s.t.} \quad f_i(x) \leq 0, \ i = 1, \ldots, m,$$
$$Ax = b.$$

e.g. $\|Bx + d\|_2 - r \leq 0.$

Conic form

generalize the cone

$$\mathbb{R}^n_+ \quad \rightsquigarrow \quad K \ \text{(closed convex cone)}$$

$$\min_x \quad \langle c, x \rangle$$
$$\text{s.t.} \quad Ax = b, \quad x \in K.$$

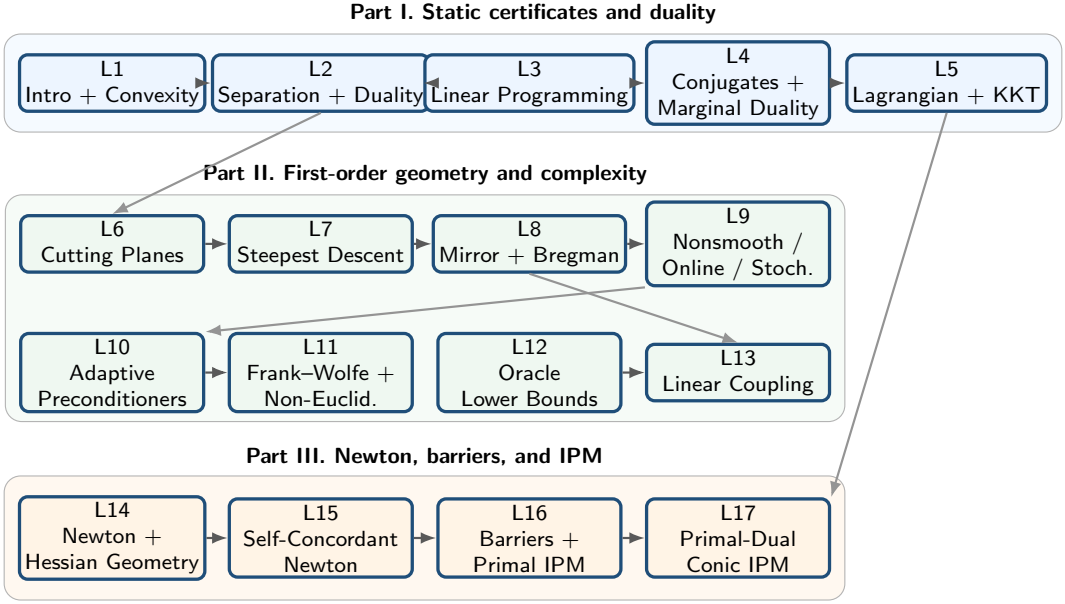
$\mathbb{R}^n_+$: LP — $\mathcal{Q}^n$: SOCP — $\mathbb{S}^n_+$: SDP.

Oracles: How do (computationally) we access the problem?

Oracle	Typical answer at query point
Zeroth-order	$x \mapsto f(x)$
First-order	$x \mapsto (f(x), \nabla f(x))$, or $g \in \partial f(x)$
Second-order	$x \mapsto (f(x), \nabla f(x), \nabla^2 f(x))$
Separation	either $x \in X$, or a separating hyperplane for X
Projection/prox	$y \mapsto \arg\min_{x \in X} \|x - y\|^2$, or $(y, g) \mapsto \arg\min_{x \in X} \{\langle g, x \rangle + D_\Phi(x, y)\}$
Linear minimization	$v \mapsto \arg\min_{x \in X} \langle v, x \rangle$
Stochastic	unbiased or controlled-noise gradient information
Barrier	second-order oracle for the self-concordant barrier of domain X

$$\text{overall runtime} = (\textbf{runtime of each iteration}) \times (\textbf{number of required iterations}),$$
$$\textbf{runtime of each iteration} = (\textbf{oracle runtime}) + (\textbf{other operations}).$$

Course Roadmap



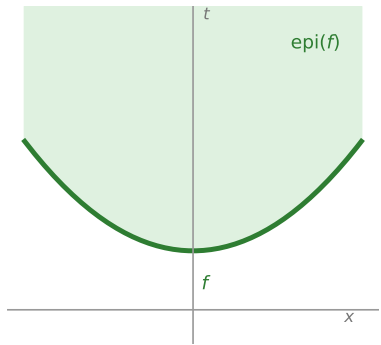
Convex functions $\Rightarrow$ Convex sets: Epigraph

Definition 2.4 (Epigraph). For $f : E \to \mathbb{R} \cup \{+\infty\}$:

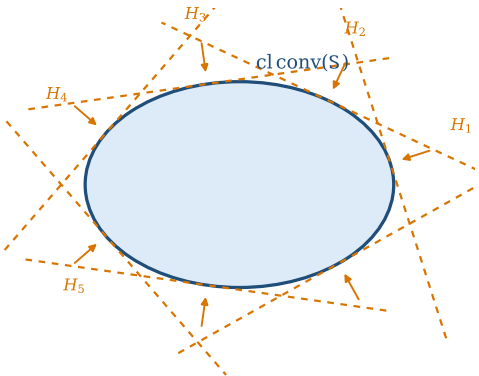
$$\text{epi}(f) := \{(x, t) \in E \times \mathbb{R} : \ t \geq f(x)\}.$$

Lemma 2.1 (Epigraph criterion).

f is convex $\iff$ epi(f) is a convex set.



Convex Sets Are Recovered From Halfspaces



$$\overline{\text{conv}}(S) = \bigcap_{\substack{\text{closed halfspaces } H \\ S \subseteq H}} H.$$

- A halfspace is a **linear certificate**.
- Separation says these certificates are complete for closed convex sets.
- Duality begins when this certificate system is applied to epigraphs, cones, and perturbation sets.

Subgradients Are Supporting Hyperplanes

$$g \in \partial f(x) \quad \iff \quad \forall (y,t) \in \operatorname{epi}(f),\ \big\langle (g,-1), (y,t) - (x, f(x)) \big\rangle \leq 0.$$

Theorem 2.10 (Subgradient existence on the relative interior). Let $f : E \to \mathbb{R} \cup \{+\infty\}$ be proper and convex. Then

$$\forall x \in \operatorname{ri}(\operatorname{dom} f), \qquad \partial f(x) \neq \varnothing.$$

Why relative interior? Apply supporting hyperplane to $\operatorname{epi}(f)$ inside $\operatorname{aff}(\operatorname{dom} f) \times \mathbb{R}$. The supporting covector has form (η, α). The condition $x \in \operatorname{ri}(\operatorname{dom} f)$ rules out $\alpha = 0$, hence $\alpha < 0$, so we can normalize it to $(g, -1)$.

Conjugates Are Affine Lower-Bound Certificates

$$f^*(y) = \sup_x \{\langle y, x\rangle - f(x)\} \quad \Longleftrightarrow \quad f(x) \geq \langle y, x\rangle - f^*(y).$$

- A conjugate records the best affine lower bound with slope y.
- Biconjugation says that a closed convex function is recovered from all affine lower bounds:

$$f(x) = \sup_y \{\langle y, x\rangle - f^*(y)\}.$$

- This is the halfspace-intersection theorem applied to $\mathrm{epi}(f)$.

Perturbation Function and Duality

For

$$\min_{x\in C} f_0(x) \quad \text{s.t.} \quad f_i(x) \le 0,\ i = 1,\dots,m, \qquad Ax = b,$$

the perturbation value function is

$$p(u,v) := \inf\{f_0(x) : x \in C,\ f_i(x) \le u_i,\ Ax - b = v\}.$$

Lagrangian.

$$L(x,\lambda,\nu) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \langle \nu, Ax - b\rangle, \qquad q(\lambda,\nu) := \inf_{x\in C} L(x,\lambda,\nu).$$

Duality is a supporting-hyperplane statement for p.

$$q(\lambda,\nu) = \inf_{u,v}\{p(u,v) + \lambda^\top u + \nu^\top v\} = -p^*(-\lambda,-\nu), \qquad \lambda \ge 0.$$

If $(-\lambda,-\nu) \in \partial p(0_m,0_p)$, then

$$p(u,v) \ge p(0_m,0_p) - \lambda^\top u - \nu^\top v,$$

so

$$q(\lambda,\nu) = \inf_{u,v}\{p(u,v) + \lambda^\top u + \nu^\top v\} \ge p(0_m,0_p).$$

Weak Duality: Every Multiplier Gives a Lower Bound

Theorem 5.1 (Weak duality). If $x \in C$ is primal feasible and $(\lambda, \nu) \in \mathbb{R}^m_+ \times \mathbb{R}^p$, then

$$q(\lambda, \nu) \le f_0(x).$$

Therefore value$(\mathrm{Dual}) \le$ value(Primal).

One-line proof idea.

$$\begin{aligned} q(\lambda,\nu) &= \inf_{z\in C} L(z,\lambda,\nu) \le L(x,\lambda,\nu) \\ &= f_0(x) + \underbrace{\sum_i \lambda_i f_i(x)}_{\le 0} + \underbrace{\langle \nu, Ax - b\rangle}_{=0} \le f_0(x). \end{aligned}$$

Weak duality is only a lower-bound statement; it does not say the best lower bound is attained.

$$\underbrace{\sup_{\lambda\ge 0,\nu} \inf_{x\in C} L(x,\lambda,\nu)}_{\text{dual value}} \le \underbrace{\inf_{x\in C} \sup_{\lambda\ge 0,\nu} L(x,\lambda,\nu)}_{\text{primal value}}.$$

KKT: When Weak Duality Becomes Equality

Definition 5.3 (KKT point of the program).

1. primal feasibility;
2. dual feasibility: $\lambda \geq 0$;
3. stationarity:

$$x^\star \in \operatorname*{argmin}_{x \in C} L(x, \lambda^\star, \nu^\star);$$

4. complementary slackness:

$$\lambda_i^\star f_i(x^\star) = 0 \quad \forall i.$$

Theorem 5.2 (KKT $\Rightarrow$ optimality). If $(x^\star, \lambda^\star, \nu^\star)$ is a KKT point, then

$x^\star$ is primal optimal, $(\lambda^\star, \nu^\star)$ is dual optimal,

and the duality gap is zero.

$$q(\lambda^\star, \nu^\star) = L(x^\star, \lambda^\star, \nu^\star) = f_0(x^\star).$$

KKT is an **optimality certificate**; Slater is about when such certificates must exist.

Slater: When the Certificate System Is Complete

Definition 5.4 (Slater's condition). There exists $\tilde{x} \in \operatorname{ri}(C)$ such that

$$f_i(\tilde{x}) < 0 \quad \forall i, \qquad A\tilde{x} = b.$$

Theorems 5.3 and 5.4. If the value is finite and Slater holds, then

no duality gap and dual optimum is attained.

If, in addition, the primal optimum is attained, then a KKT point exists.

Proof idea.

$$\text{Slater} \Rightarrow (0_m, 0_p) \in \operatorname{ri}(\operatorname{dom} p) \Rightarrow \partial p(0_m, 0_p) \neq \varnothing.$$

The last implication is Theorem 2.10 applied to the perturbation value function.

Slater is sufficient, but not necessary. See the counterexamples in the Lecture 5 slides.

LP Strong Duality: Farkas and Closed Cones

Farkas alternative. Exactly one of the following alternatives holds:

$$b \in A\mathbb{R}^n_+ \qquad \text{or} \qquad \exists y,\ A^\top y \geq 0,\ \langle y, b\rangle < 0.$$

Proof idea.

- The possible right-hand sides Ax with $x \geq 0$ form the cone

$$K_A := A\mathbb{R}^n_+.$$

- This cone is finitely generated, hence closed and convex.
- If $b \notin K_A$, closed convex separation gives a hyperplane y with

$$\langle y, b\rangle < 0 \leq \langle y, Ax\rangle \quad \forall x \geq 0.$$

- The right inequality is exactly $A^\top y \geq 0$, so y is the infeasibility certificate.

For LPs, strong duality ultimately comes from this exact separation certificate for polyhedral cones.

Summary of Dimension-free Optimization Algorithms

Setting	Algorithm	Key lemma / analysis term	Rate
smooth	steepest descent	descent lemma (Lem. 7.3)	LR^2/T
smooth + strongly convex	steepest descent	gap sandwich + contraction (Lem. 7.6, Thm. 7.7)	$\kappa \log(1/\epsilon)$
smooth over K	Frank–Wolfe	curvature + FW gap (Lem. 11.1, Thms. 11.2–11.3)	C_f/T
nonsmooth	subgrad / MD	regret telescope (Thm. 9.1, Lem. 9.3)	$GR/\sqrt{T}$
strongly convex	subgrad / MD	weighted telescope (Thm. 9.6)	$G^2/(\mu T)$
relative smooth	mirror descent	relative descent telescope (Thm. 8.6)	LD_Φ/T
unknown-norm smooth	AdaReg	adaptive regularizer comparison (Thm. 10.7)	$R_{\mathcal{H}}^2 S_{\mathcal{H}}/T$
accelerated smooth	linear coupling	primal + mirror progress (Thm. 13.3)	LR^2/T^2
self-concordant	Newton	Newton decrement: decrease + certificate + quadratic convergence (Thms. 15.7–15.9)	$(f(x_0) - f^\star)/\omega(1/4) + \log\log(1/\epsilon)$
barrier / IPM	Newton near path	Dikin + barrier gap (Lem. 16.1, Thm. 16.3)	$\sqrt{\nu}\log(\nu/\epsilon)$

Smooth vs. Nonsmooth: Descent or Averaging?

Smooth key estimate: descent lemma.

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2.$$

Plugging $y = x - \eta \nabla f(x)$ gives actual decrease when $\eta \le 1/L$.

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

Nonsmooth key estimate: linearization.

$$f(x_t) - f(u) \le \langle g_t, x_t - u \rangle, \qquad g_t \in \partial f(x_t).$$

Average the iterates:

$$f(\bar{x}_T) - f(u) \le \frac{1}{T}\sum_{t=1}^{T} \langle g_t, x_t - u \rangle.$$

Thus the problem reduces to controlling online-learning regret.

Mirror Descent: Three-Point Inequality and Telescope

Bregman step.

$$x_{t+1} = \operatorname*{argmin}_{x \in X} \{\eta_t \langle g_t, x \rangle + D_\Phi(x, x_t)\}.$$

Three-point inequality in dual-progress form. For every $u \in X$,

$$\eta_t \langle g_t, x_{t+1} - x_t \rangle + D_\Phi(x_{t+1}, x_t) \le \eta_t \langle g_t, u - x_t \rangle + D_\Phi(u, x_t) - D_\Phi(u, x_{t+1}).$$

Bounded gradients: regret bound. If Φ is α-strongly convex and $\|g_t\|_* \le G$, then

$$\sum_{t=1}^{T} \langle g_t, x_t - u \rangle \lesssim \frac{D_\Phi(u, x_1)}{\eta} + \frac{\eta G^2 T}{\alpha}.$$

Choose $\eta \asymp 1/\sqrt{T}$ to get the nonsmooth rate.

Smooth losses: descent telescope. If $g_t = \nabla f(x_t)$, f is L-smooth relative to Φ, and $\eta_t \le 1/L$, then

$$D_\Phi(x^\star, x_{t+1}) + \eta_t \big(f(x_{t+1}) - f^\star\big) \le D_\Phi(x^\star, x_t).$$

Summing gives the relative-smooth $1/T$ rate.

Noise: Average in Dual Space Before Nonlinear Maps

Mirror descent / SGD. Let $z_t := \nabla\Phi(x_t)$. With a constant stepsize,

$$z_{t+1} = z_t - \eta\widehat{g}_t, \qquad z_{T+1} = z_1 - \eta\sum_{t=1}^{T}\widehat{g}_t.$$

The noisy covectors are accumulated linearly in dual space. If $\mathbb{E}[\widehat{g}_t \mid \mathcal{F}_{t-1}] = g_t$, the martingale noise can average out before the geometry map

$$x_{T+1} = \nabla\Phi^*(z_{T+1})$$

is applied.

Naive stochastic Frank–Wolfe.

$$s_t = \operatorname*{argmin}_{s\in K}\langle\widehat{g}_t, s\rangle, \qquad x_{t+1} = (1-\gamma_t)x_t + \gamma_t s_t.$$

Here the noisy covector first passes through a nonlinear atom selector:

$$\mathbb{E}\left[\operatorname*{argmin}_{s\in K}\langle\widehat{g}, s\rangle\right] \neq \operatorname*{argmin}_{s\in K}\langle\mathbb{E}\widehat{g}, s\rangle.$$

Primal averaging of noisy atoms is not the same as dual averaging of noisy gradients.

Takeaway. Average noisy gradients before applying a nonlinear geometry map.

One Covector, Different Geometry Interfaces

At a point x, a first-order oracle returns a covector $g \in E^*$. The geometry specifies how this same g becomes the next point x^+.

Unnormalized steepest descent

$$v \in \operatorname*{argmin}_{\|z\| \le 1} \langle g, z \rangle$$

$$x^+ = x + \eta \|g\|_* v.$$

Interface: LMO over the unit ball; dual norm $\|g\|_*$.

Frank–Wolfe / normalized SD

$$v \in \operatorname*{argmin}_{z \in K} \langle g, z \rangle$$

$$x^+ = (1 - \gamma)x + \gamma v.$$

Interface: LMO over the feasible set K.

Mirror descent

$$x^+ = \operatorname*{argmin}_{u \in X} \Psi_x(u)$$

$\Psi_x(u) := \eta \langle g, u \rangle + D_\Phi(u, x)$.

Interface: Bregman prox generated by Φ.

Where Geometry Enters the Rate

The same proof template exposes different constants.

Method	Geometry-dependent quantity
Steepest descent	$L_{\|\cdot\|} R_{\|\cdot\|}^2$ for convex smooth rates; $\kappa_{\|\cdot\|} = L_{\|\cdot\|}/\mu_{\|\cdot\|}$ for linear rates.
Frank–Wolfe	Curvature $C_f(K)$, often $C_f(K) \lesssim L_{\|\cdot\|} \operatorname{diam}_{\|\cdot\|}(K)^2$.
Mirror descent	$\frac{D_\Phi(u, x_1)}{\eta} + \eta \sum_{t=1}^{T}$ local gradient cost.

Example: simplex. Let $X = \Delta_d$ and $\|g_t\|_\infty \le G$.

Entropy mirror geometry:

$$\mathrm{Reg}_T = O(G\sqrt{T \log d}).$$

Euclidean geometry:

$$\|g_t\|_2 \le G\sqrt{d} \Rightarrow \mathrm{Reg}_T = O(G\sqrt{dT}).$$

same template + different geometry = different rate

Oracle Lower Bounds: Quantifier Order

End-to-end statement. Let $\mathrm{err}_T(A; f, \mathsf{O}_f) := f(x_{T+1}) - \min_{x \in K} f(x)$.

$$\textbf{For each fixed } T: \qquad \inf_A \sup_{f, \mathsf{O}_f} \mathrm{err}_T(A; f, \mathsf{O}_f) \geq \Delta(T)$$

$$\iff \quad \forall T \quad \forall A \quad \exists (f, \mathsf{O}_f) : \mathrm{err}_T(A; f, \mathsf{O}_f) \geq \Delta(T).$$

Not claiming: $\forall A \, \exists (f, \mathsf{O}_f) \, \forall T : \mathrm{err}_T(A; f, \mathsf{O}_f) \geq \Delta(T)$.

Linear-span proof template. For this T, prove a hard oracle for span-respecting transcripts:

$$\exists (h, \mathsf{O}_h) \quad \forall \tau \in \mathrm{Span}(T), h(y_{T+1}) - \min_K h \geq \Delta(T).$$

Then lift to arbitrary deterministic methods:

$$\forall A \quad \exists U, \ f, \ \mathsf{O}_f, f(x_{T+1}) - \min_K f \geq \Delta(T).$$

Core insight.

$$y_{t+1} \in \mathrm{span}\{g_1, \ldots, g_t\}.$$

Only revealed directions can be used.

Two target lower bounds.

Nonsmooth G-Lipschitz convex:

$$\inf_A \sup_f f(x_{T+1}) - f^\star \gtrsim \frac{GR}{\sqrt{T+1}}.$$

***L*-smooth convex:**

$$\inf_A \sup_f f(x_{T+1}) - f^\star \gtrsim \frac{LR^2}{(T+1)^2}.$$

Algorithms with Dimension-Dependent Convergence Rates

Separation-oracle localization. Maintain a region known to contain the target set, query a center, and cut. The algorithms differ by the center they can compute and the representation they maintain.

Method	Oracle calls	Per iter.	Total runtime
Center of mass (Grünbaum 1960; Levin/Newman 1965)	$O\big(n \log \frac{1}{\epsilon}\big)$	$*$	$*$
Ellipsoid (Khachiyan 1979)	$O\big(n^2 \log \frac{1}{\epsilon}\big)$	$O(n^2)$	$O\big(n^4 \log \frac{1}{\epsilon}\big)$
Vaidya volumetric center (FOCS 1989)	$\widetilde{O}\big(n \log \frac{1}{\epsilon}\big)$	$O(n^3)$	$\widetilde{O}\big(n^4 \log \frac{1}{\epsilon}\big)$
Lee–Sidford–Wong (FOCS 2015)	$\widetilde{O}\big(n \log \frac{1}{\epsilon}\big)$	$\widetilde{O}(n^2)$	$\widetilde{O}\big(n^3 \log \frac{1}{\epsilon}\big)$
Lower bound	$\Omega\big(n \log \frac{1}{\epsilon}\big)$		

$*$: exact center of mass has the cleanest cut count but is not a cheap primitive. The table ignores separation-oracle cost and hides polylogarithmic factors in $\widetilde{O}(\cdot)$.

Volumetric barrier. For $P = \{x : s_i(x) = a_i^\top x - b_i > 0\}$, set $H(x) := \sum_i a_i a_i^\top / s_i(x)^2$, the log-barrier Hessian. Vaidya's volumetric center minimizes $V(x) := \frac{1}{2} \log \det H(x)$.

Self-Concordance: Local Norm and Dikin Ellipsoids

Definition. A convex C^3 function f on an open convex set is self-concordant if

$$|D^3 f(x)[h,h,h]| \le 2\big(D^2 f(x)[h,h]\big)^{3/2}.$$

Local norm.

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

Dikin ellipsoid for barriers. If Φ is a self-concordant barrier for K, then

$$\{x+u : \|u\|_x < 1\} \subseteq \mathrm{int}(K).$$

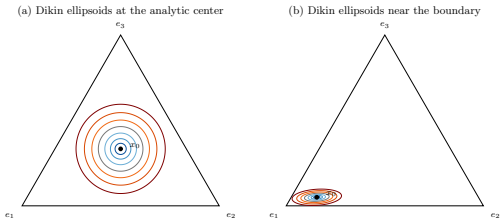
Multiplicative Hessian comparison. If $r := \|u\|_x < 1$, then for every direction v,

$$(1-r)^2\|v\|_x^2 \le \|v\|_{x+u}^2 \le \frac{1}{(1-r)^2}\|v\|_x^2.$$

Equivalently $\qquad (1-r)^2\nabla^2\Phi(x) \preceq \nabla^2\Phi(x+u) \preceq \frac{1}{(1-r)^2}\nabla^2\Phi(x).$

Two Local Geometries on the Simplex

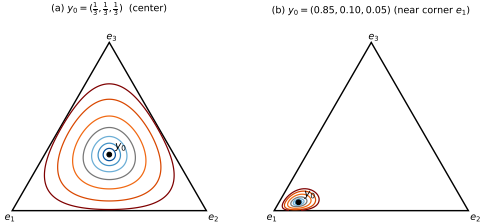
Log barrier: Dikin ellipsoids



$\Phi(x) = -\sum_i \log x_i$, Hessian metric $\|\cdot\|_x$.

Unit balls $r < 1$ stay inside (Lecture 16).

Entropy regularizer: KL Bregman balls



$h(x) = \sum_i x_i \log x_i$ (Lecture 8). Bregman ball $D_h(y, x_0) \le r$ around x_0.

Both geometries become anisotropic near the boundary. **Difference:** the log barrier's local unit ball is *guaranteed* to stay inside, while KL Bregman balls only do so for small enough radius.

Newton Decrement: Step Size and Certificate

Newton step and decrement.

$$d_x := -(\nabla^2 f(x))^{-1} \nabla f(x),$$

$$\lambda_f(x)^2 := \langle \nabla f(x), (\nabla^2 f(x))^{-1} \nabla f(x) \rangle.$$

$$\|d_x\|_x = \lambda_f(x), \qquad \langle \nabla f(x), d_x \rangle = -\lambda_f(x)^2.$$

Self-concordant key estimate. For $x^+ = x + d_x/(1+\lambda)$,

$$f(x^+) \le f(x) - \omega(\lambda), \qquad \omega(t) = t - \log(1+t).$$

If $\lambda < 1$, then

$$f(x) - f^\star \le \omega^*(\lambda), \qquad \omega^*(t) = -t - \log(1-t).$$

So $\lambda_f(x)$ is both a step-size quantity and an optimality certificate.

Central Path and Primal-Dual Complementarity

Barrier path.

$$F_t(x) := t\langle c, x\rangle + \Phi(x), \qquad \mu := 1/t.$$

$$x(t) \in \operatorname*{argmin}_{\substack{Ax=b \\ x\in \mathrm{int}(K)}} F_t(x).$$

Barrier gap.

$$\langle c, x(t)\rangle - p^\star \le \frac{\nu}{t}.$$

Central path as perturbed KKT.

$$\begin{aligned} Ax &= b, \\ A^* y + s &= c, \\ s &= -\mu \nabla \Phi(x). \end{aligned}$$

$$\langle x, s\rangle = \nu\mu \to 0.$$

LP: $x \odot s = \mu \mathbf{1}$.

SDP: $S = \mu X^{-1} \iff XS = \mu I$.

One IPM Step: Primal vs. Primal-Dual

Let

$$F_t(x) := t\langle c, x\rangle + \Phi(x), \qquad Ax = b, \quad x \in \operatorname{int}(K), \qquad \mu := 1/t.$$

Primal IPM step. Maintain $Ax = b$. Solve the equality-constrained Newton system

$$\begin{bmatrix} \nabla^2\Phi(x) & A^* \\ A & 0 \end{bmatrix} \begin{pmatrix} \Delta x \\ w \end{pmatrix} = -\begin{pmatrix} tc + \nabla\Phi(x) \\ 0 \end{pmatrix}.$$

Then take

$$x^+ = x + \alpha\Delta x,$$

with α chosen to stay in $\operatorname{int}(K)$ and reduce F_t.

Primal-dual IPM step. Use residuals

$$r_p = Ax - b, \quad r_d = A^*y + s - c, \quad r_c = s + \mu\nabla\Phi(x).$$

Linearize perturbed KKT:

$$\begin{aligned} A\Delta x &= -r_p, \\ A^*\Delta y + \Delta s &= -r_d, \\ \Delta s + \mu\nabla^2\Phi(x)\Delta x &= -r_c. \end{aligned}$$

Then take

$$(x^+, y^+, s^+) = (x, y, s) + \alpha(\Delta x, \Delta y, \Delta s).$$

Example: Large-Data ℓ_1 Loss

$$\min_{x \in \mathbb{R}^n} F(x) := \frac{1}{m} \sum_{i=1}^{m} |\langle a_i, x \rangle - b_i| + \frac{\lambda}{2} \|x\|_2^2.$$

Method	Formulation / update	Runtime scale
1. QP / IPM	$\min_{x,z} \ \frac{1}{m} \sum_i z_i + \frac{\lambda}{2} \|x\|_2^2$ s.t. $\ -z_i \le \langle a_i, x \rangle - b_i \le z_i$	$\widetilde{O}(\sqrt{m}\, n^3 \log(1/\epsilon))$. Newton steps are expensive.
2. Full subgradient	$x_{k+1} = x_k - \frac{1}{\lambda k m} \sum_{i=1}^{m} \operatorname{sign}(\langle a_i, x_k \rangle - b_i) a_i - \frac{1}{k} x_k$	$\widetilde{O}(\frac{nm}{\lambda \epsilon})$. Scan all data per step.
3. Stochastic subgrad.	sample $i_k \sim \mathrm{Unif}\{1, \dots, m\}$, $x_{k+1} = x_k - \frac{1}{\lambda k} \operatorname{sign}(\langle a_{i_k}, x_k \rangle - b_{i_k}) a_{i_k} - \frac{1}{k} x_k$.	$\widetilde{O}(\frac{n}{\lambda \epsilon})$. One datapoint per step.

IPM has better dependence on ϵ but worse dependence on dimension n; first-order methods have better dependence on n, especially with cheap stochastic gradients.

Some Things This Course Did Not Cover

Solver technology.

- quasi-Newton, conjugate gradient, trust-region methods
- line search, stopping rules, scaling, numerical stability
- production IPM: predictor-corrector, presolve, sparse factorization

Specialized problem classes.

- simplex, network-flow, and message-passing LP algorithms
- decomposition and problem-specific solvers
- nonlinear cone-valued constraints $F_i(x) \preceq_{K_i} 0$

Loewner-order example. $F(X) := -X^{1/2}$ on $\mathbb{S}^n_+$ satisfies

$$F(\theta X + (1-\theta)Y) \preceq \theta F(X) + (1-\theta)F(Y).$$

Other oracle models.

- zeroth-order and bandit optimization
- finite-sum variance reduction and modern stochastic optimizer engineering
- distributed, federated, and communication-limited optimization

Beyond convexity.

- nonconvex optimization and critical-point guarantees
- global, mixed-integer, and combinatorial optimization
- learning-theoretic and statistical limits beyond convex losses

Final Exam: Scope and Emphasis

Covers.

- **Modeling:** convex sets, convex functions, convex optimization programs, oracle models, and how to translate concrete problems into standard forms.
- **Certificates:** duality, separation, Slater condition, KKT-style reasoning, and when strong duality should or should not be expected.
- **Algorithms:** knowledge of optimization algorithms and convergence results. Basic convergence analysis through key lemmas. Be able to compare algorithms and choose a suitable algorithm and geometry for concrete problems.

What will not be included.

- Long convex-analysis proofs for their own sake.
- Detailed convergence-analysis proofs beyond the basic templates.
- Adaptive-algorithm analysis/proofs. You should know what AdaGrad-style adaptation is for, but not prove the adaptive regret theorem.
- Lower-bound reduction proofs.
- Technical results whose full proof was omitted in lecture because it was too complex.

Final Exam and Other Logistics

Final exam. Tuesday **May 26**, **1:00–4:00 pm**.

- One A4 cheat sheet, **double-sided**, allowed.
- **Must be handwritten** — no printed/typeset sheets.
- *Some HW problems will appear in the final.*

Practice exam. Released **May 22**, with solutions.

Bonus problems. Deadline **AoE, May 22**.

In-class bonus. If I promised you bonus points in lecture, please email me a reminder with the lecture, if you remember it, and the points.

Extra office hours. Beining will hold an additional OH on **Monday May 25**. Please drop by between **2–6pm**.

Thank you!