Lecture 7: Steepest Descent and Descent Lemmas

TTIC 31070 / CMSC 35470 / BUSF 36903 / STAT 31015

Convex Optimization

Prof. Zhiyuan Li

Spring 2026

From Oracles to Algorithms

Lecture 6: **cutting-plane methods** have iteration counts $O(n \log(\Delta/\delta))$ or $O(n^2 \log(\Delta/\delta))$ — the dimension n sits out front.

Today: **first-order methods**. Use a first-order oracle $x \mapsto (f(x), \nabla f(x))$ and the geometry of a norm.

Two new ingredients:

- L**-smoothness w.r.t. a norm** $\|\cdot\|$: gives a *quadratic upper model* of f at every point
- **Linear minimization oracle** on the unit ball $B_{\|\cdot\|}$: gives the *steepest-descent direction* for $\|\cdot\|$

Together they yield the **descent lemma**, which we convert to convergence rates by two routes: strong convexity (linear) and plain convexity with a radius bound ($O(1/T)$).

Smoothness

Definition 7.1 (L-smoothness w.r.t. $\|\cdot\|$). $f : E \to \mathbb{R}$ differentiable. We say f is *L-smooth w.r.t. $\|\cdot\|$* if

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \tfrac{L}{2}\|y - x\|^2 \qquad \forall\, x, y \in E.$$

Lemma 7.1 (Gradient Lipschitzness $\Rightarrow$ smoothness). If $\|\nabla f(y) - \nabla f(x)\|_* \leq L\|y - x\|$ for all x, y, then f is L-smooth.

Proof. $f(y) - f(x) - \langle \nabla f(x), y - x \rangle = \displaystyle\int_0^1 \langle \nabla f(\gamma(t)) - \nabla f(x), y - x \rangle \, dt$

where $\gamma(t) = x + t(y - x)$. By Hölder and Lipschitz:

$\leq \displaystyle\int_0^1 Lt\|y - x\|^2 \, dt = \tfrac{L}{2}\|y - x\|^2$. $\square$

Quadratic Upper Model and the Proxy Step

L-smoothness gives a **quadratic upper bound**:

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

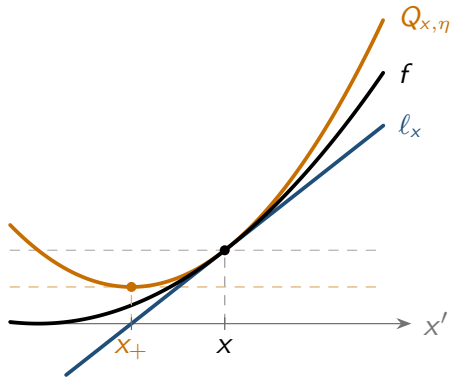
Define the model

$$Q_{x,\eta}(y) := f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2\eta}\|y - x\|^2.$$

For $\eta \le 1/L$: $Q_{x,\eta} \ge f$, tangent at x.

The **proxy step** is $x_+ := \operatorname{argmin}_y Q_{x,\eta}(y)$.

In ℓ_2: $x_+ = x - \eta \nabla f(x)$ (gradient step). For general norms: need a new primitive (next slide).



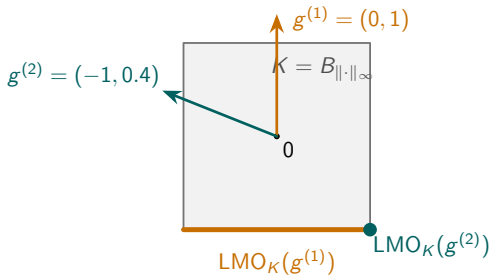
$f(y) = \frac{1}{2}y^2$, $x = 1$, $\eta = 1/L = 1/2$: affine lower bound ℓ_x, upper model $Q_{x,\eta}$, proxy step x_+.

Linear Minimization Oracle

Definition 7.2 (LMO). Let $K \subseteq E$ be nonempty, compact, convex. A *linear minimization oracle* is any map $\mathrm{LMO}_K : E^* \to E$ with $\mathrm{LMO}_K(g) \in \operatorname{argmin}_{s \in K} \langle g, s \rangle$.

Key fact. For the unit norm ball, $\min_{s \in B_{\|\cdot\|}} \langle g, s \rangle = -\|g\|_*$.

Examples (all return $-\|g\|_*$): ℓ_2 gives $-g/\|g\|_2$; ℓ_1 gives $-\operatorname{sign}(g_{i^*})\, e_{i^*}$ with $i^* = \operatorname{argmax}_i |g_i|$; ℓ_∞ gives $-\operatorname{sign}(g)$.



LMO returns a whole face (orange) or a unique corner (teal).

Steepest Descent via the Norm-Ball LMO

Proposition 7.2. Fix $x \in E$, $\eta > 0$. Then $x_+ = x + \eta \|\nabla f(x)\|_* \, \mathrm{LMO}_{B_{\|\cdot\|}}(\nabla f(x))$ is a minimizer of $Q_{x,\eta}$.

Proof. Write $h = ru$ with $r = \|h\| \geq 0$, $u \in B_{\|\cdot\|}$. Then

$$\langle g, h \rangle + \tfrac{1}{2\eta}\|h\|^2 = r\langle g, u \rangle + \tfrac{r^2}{2\eta}.$$

Minimize over $u \in B_{\|\cdot\|}$: $\langle g, u \rangle = -\|g\|_*$, direction $u = \mathrm{LMO}_{B_{\|\cdot\|}}(g)$.
Minimize over $r \geq 0$: $r = \eta\|g\|_*$. $\square$

A single choice of norm $\|\cdot\|$ fixes both pieces: the LMO on $B_{\|\cdot\|}$ picks the direction, and the dual-norm magnitude $\|g\|_*$ sets the optimal step length.

Steepest-Descent Updates

Definition 7.3 (Normalized and unnormalized steepest descent). Fix $\mathrm{LMO}_{B_{\|\cdot\|}}$ on the unit ball, $\eta > 0$.

$$\text{Unnormalized:} \quad x_{t+1} = x_t + \eta \, \|\nabla f(x_t)\|_* \, \mathrm{LMO}_{B_{\|\cdot\|}}(\nabla f(x_t))$$

$$\text{Normalized:} \quad x_{t+1} = x_t + \eta \, \mathrm{LMO}_{B_{\|\cdot\|}}(\nabla f(x_t))$$

- **Unnormalized**: step length scales with $\|\nabla f(x_t)\|_*$. Matches proxy-step minimizer $\Rightarrow$ convergence analysis via descent lemma.
- **Normalized**: step length is η independent of the gradient magnitude. Used for subgradient-style analysis later.

Euclidean case: unnormalized reduces to $x_{t+1} = x_t - \eta \nabla f(x_t)$, the classical gradient descent.

The Descent Lemma

Lemma 7.3 (One-step descent for unnormalized SD). Assume f is L-smooth w.r.t. $\|\cdot\|$. For $\eta \geq 0$ and $x_+ = x + \eta\|\nabla f(x)\|_* \,\mathrm{LMO}_{B_{\|\cdot\|}}(\nabla f(x))$:

$$f(x_+) \leq f(x) - \eta\Big(1 - \tfrac{L\eta}{2}\Big)\,\|\nabla f(x)\|_*^2.$$

Proof. Let $g = \nabla f(x)$, $s = \mathrm{LMO}(g)$, $u = \|g\|_* s$. Then $\langle g, u\rangle = -\|g\|_*^2$ and $\|u\| = \|g\|_*$.

By L-smoothness with $h = \eta u$:

$$f(x_+) \leq f(x) + \eta\langle g, u\rangle + \tfrac{L\eta^2}{2}\|u\|^2 = f(x) - \eta\left(1 - \tfrac{L\eta}{2}\right)\|g\|_*^2. \quad \square$$

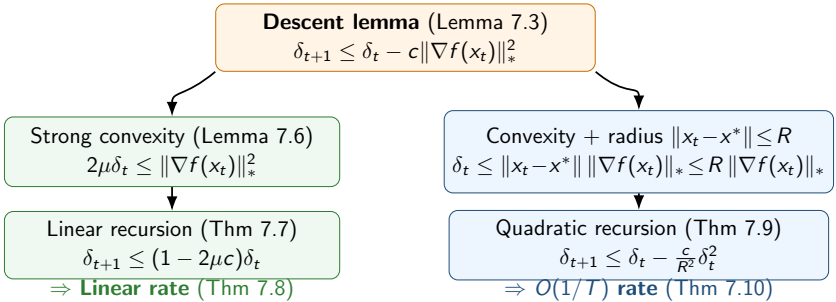
Best step: $\eta(1 - L\eta/2)$ maximized at $\eta = 1/L$, giving $f(x_+) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|_*^2$.

Note: this uses only smoothness — no convexity. The descent lemma already gives local improvement for nonconvex f.

Two Bridges from Descent Lemma to Convergence

Convexity **lower-bounds progress per step in terms of suboptimality**: it turns $\|\nabla f(x_t)\|_*^2 \geq \varphi(\delta_t)$, so the descent lemma becomes a recursion in $\delta_t = f(x_t) - f(x^*)$.

Without convexity this is impossible — SD can get trapped in a local minimum where $\nabla f = 0$ but $\delta_t > 0$.



Strong Convexity

Definition 7.4 (μ-strong convexity w.r.t. $\|\cdot\|$). For $\mu > 0$, $f : E \to (-\infty, +\infty]$ is *μ-strongly convex* if for all $x, y \in E$, $\theta \in [0, 1]$,

$$f\big((1-\theta)x + \theta y\big) \le (1-\theta)f(x) + \theta f(y) - \tfrac{\mu}{2}\theta(1-\theta)\|x - y\|^2.$$

Lemma 7.4 (Differentiable characterization). For differentiable $f : E \to \mathbb{R}$, Def 7.4 is equivalent to

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{\mu}{2}\|y - x\|^2 \qquad \forall\, x, y.$$

Proof idea. **($\Rightarrow$)** apply Def 7.4 along the segment $x + t(y - x)$ with $\theta = t$, rearrange, and let $t \downarrow 0$.
($\Leftarrow$) apply the first-order lower bound at $z = (1-\theta)x + \theta y$ separately to x and y, then take the convex combination. $\square$

Strong Convexity $\Rightarrow$ Unique Minimizer

Lemma 7.5. A proper, closed, μ-strongly convex $f : E \to (-\infty, +\infty]$ attains its minimum at a unique $x^* \in E$.

Proof idea. Pick $x_0 \in \operatorname{ri}(\operatorname{dom} f)$ and any subgradient $g_0 \in \partial f(x_0)$. Strong convexity gives the quadratic lower bound

$$f(y) \geq f(x_0) + \langle g_0, y - x_0 \rangle + \tfrac{\mu}{2}\|y - x_0\|^2 \to +\infty \quad \text{as} \quad \|y - x_0\| \to \infty,$$

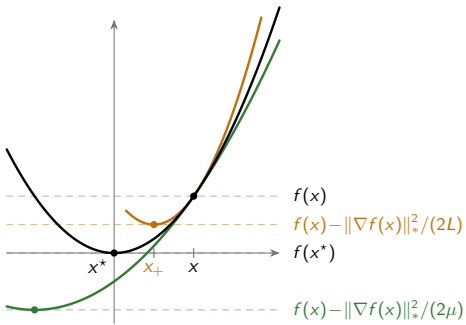
so every sublevel set of f is bounded.

A compact truncated slice of $\operatorname{epi} f$ then shows the infimum is attained (using closedness), and strict convexity from $\mu > 0$ gives uniqueness. $\square$

Gradient Sandwich

Lemma 7.6. If $f : E \to \mathbb{R}$ is L-smooth and μ-strongly convex with unique minimizer x^*, then, with $\delta(x) := f(x) - f(x^*)$,

$$2\mu\,\delta(x) \le \|\nabla f(x)\|_*^2 \le 2L\,\delta(x).$$



Proof. Smoothness $\Rightarrow f(x^*) \le \inf_h\{f(x) + \langle g, h\rangle + \frac{L}{2}\|h\|^2\}$. Lemma 7.4 $\Rightarrow$ the reverse with $\frac{\mu}{2}$. Use

$$\inf_{h\in E}\left\{\langle g, h\rangle + \tfrac{\alpha}{2}\|h\|^2\right\} = -\tfrac{1}{2\alpha}\|g\|_*^2,$$

with $\alpha = L$ and $\alpha = \mu$. $\square$

Linear Rate under Strong Convexity

Theorem 7.7 (One-step recursion). Under L-smooth and μ-strongly convex, $\eta \in (0, 2/L]$, $\delta_t = f(x_t) - f(x^*)$:

$$\delta_{t+1} \le \big(1 - 2\mu\eta(1 - L\eta/2)\big)\delta_t.$$

Proof. Descent lemma (Lemma 7.3): $\delta_{t+1} \le \delta_t - \eta(1 - L\eta/2)\|\nabla f(x_t)\|_*^2$. Lower bound in Lemma 7.6: $\|\nabla f(x_t)\|_*^2 \ge 2\mu\delta_t$. Substitute. $\square$

Theorem 7.8 (Linear rate). $\delta_T \le \big(1 - 2\mu\eta(1 - L\eta/2)\big)^T \delta_0$. With $\eta = 1/L$: $f(x_T) - f(x^*) \le (1 - \mu/L)^T \delta_0$.

Proof. Iterate Thm 7.7; Lemma 7.6 gives $\mu \le L$, so the contraction factor lies in $[0, 1]$.

Condition number $\kappa := L/\mu \ge 1$: iteration count $T = O(\kappa \log(1/\varepsilon))$.

Convex Case: No Strong Convexity

Theorem 7.9 (One-step recursion, convex case). Assume f convex, L-smooth, with minimizer x^*. For $\eta \in (0, 2/L]$ and $x_t \neq x^*$:

$$\delta_{t+1} \leq \delta_t - \frac{\eta(1 - L\eta/2)}{\|x_t - x^*\|^2}\, \delta_t^2.$$

Proof. Convexity gives the **gap estimate**
$\delta_t \leq \langle \nabla f(x_t), x_t - x^* \rangle \leq \|x_t - x^*\| \|\nabla f(x_t)\|_*$, so

$$\|\nabla f(x_t)\|_*^2 \geq \delta_t^2 / \|x_t - x^*\|^2.$$

Combine with the descent lemma. □

Problem: the bound depends on $\|x_t - x^*\|$, which changes over time ⇒ need a uniform radius.

$O(1/T)$ Rate with Radius Bound

Theorem 7.10. Setting of Theorem 7.9, plus $\|x_t - x^*\| \le R$ for all t. With $c = \eta(1 - L\eta/2)/R^2$:

$$\delta_T \le \frac{1}{\delta_0^{-1} + cT}, \qquad \text{so with } \eta = 1/L: \quad f(x_T) - f(x^*) \le \frac{2LR^2}{T}.$$

Proof sketch. Theorem 7.9 gives $\delta_{t+1} \le \delta_t - c\delta_t^2$. Using $1/(1-z) \ge 1 + z$ for $z \in [0, 1]$:

$$\frac{1}{\delta_{t+1}} \ge \frac{1}{\delta_t(1 - c\delta_t)} \ge \frac{1}{\delta_t} + c \;\Rightarrow\; \frac{1}{\delta_T} \ge \frac{1}{\delta_0} + cT. \;\square$$

Telescoping trick: pass to reciprocals to linearize the quadratic recursion.

Where Does the Radius Bound Come From?

Corollary 7.11 (Bounded sublevel set). If $S_0 = \{x : f(x) \le f(x_0)\}$ is bounded, let $R := \sup\{\|x - x^*\| : x \in S_0\}$. Then Theorem 7.10 holds with this R.

Why: descent lemma gives $f(x_{t+1}) \le f(x_t)$ for $\eta \le 2/L$, so all iterates stay in S_0.

Theorem 7.12 (Unconditional Euclidean rate). For f convex, L-smooth in $\|\cdot\|_2$, gradient descent $x_{t+1} = x_t - \eta \nabla f(x_t)$ with $\eta \le 1/L$ is *non-expansive*: $\|x_{t+1} - x^*\|_2 \le \|x_t - x^*\|_2$. Hence $R = \|x_0 - x^*\|_2$ works and

$$f(x_T) - f(x^*) \le 2L\|x_0 - x^*\|_2^2 / T.$$

Why non-expansive. Expand $\|x_{t+1} - x^*\|_2^2 = \|x_t - x^*\|_2^2 - 2\eta\langle g_t, x_t - x^*\rangle + \eta^2\|g_t\|_2^2$. Use convexity $f(x_t) - f(x^*) \le \langle g_t, x_t - x^*\rangle$ and descent lemma $f(x_{t+1}) \le f(x_t) - \frac{\eta}{2}\|g_t\|_2^2$ (for $\eta \le 1/L$):

$$\|x_{t+1} - x^*\|_2^2 \le \|x_t - x^*\|_2^2 - 2\eta(f(x_{t+1}) - f(x^*)) \le \|x_t - x^*\|_2^2.$$

Summary of Rates

	Convex + L-smooth	**μ-strong convex + L-smooth**
Rate	$\frac{2LR^2}{T}$	$\left(1-\frac{\mu}{L}\right)^T \delta_0$
Iterations for ε	$O(LR^2/\varepsilon)$	$O(\kappa \log(1/\varepsilon))$
Needs radius bound?	yes (general) / no (ℓ_2)	no
Key theorem	Thm 7.10 / Thm 7.12	Thm 7.8

Conceptual takeaway.

- Smoothness $\Rightarrow$ descent lemma with factor $\eta(1 - L\eta/2)$
- Two bridges: strong convexity $\rightarrow$ linear; convexity+radius $\rightarrow O(1/T)$
- Norm choice enters via LMO and the dual norm $\|\nabla f\|_*$

Summary & What's Next

Today:

- L-smoothness w.r.t. a norm $\Rightarrow$ quadratic upper model
- Linear minimization oracle on the unit ball $\Rightarrow$ steepest-descent direction
- Descent lemma: $f(x_+) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|_*^2$
- Two bridges to convergence:
  - Strong convexity + gradient sandwich $\Rightarrow$ linear rate
  - Convexity + radius bound $\Rightarrow$ $O(1/T)$ rate

Next lecture:

- Mirror descent and Bregman divergences
- Non-Euclidean geometry: entropy on the simplex, matrix entropy, ...