Lecture 9: Nonsmooth, Online, and Stochastic Mirror Descent

TTIC 31070 / CMSC 35470 / BUSF 36903 / STAT 31015

Convex Optimization

Prof. Zhiyuan Li

Spring 2026

What Changes Without Smoothness?

Lecture 8 one-step telescope (constrained, Thm 8.9):

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

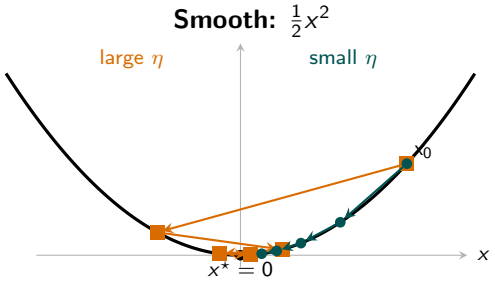
Smooth case (L8): $g_t = \nabla f(x_t)$, relative smoothness $\Rightarrow$ local decrement $\le \eta_t \big(f(x_t) - f(x_{t+1})\big) \Rightarrow$ *last-iterate* rate on one fixed objective.

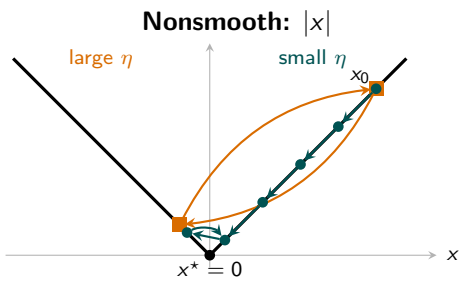
Today. Drop smoothness. Now g_t is an arbitrary subgradient, possibly of a *different* loss at each step. The same inequality becomes a **pathwise linearized comparison bound**.

Three readings of that one bound:

- **Online** (adversarial): regret against a fixed comparator
- **Offline** nonsmooth: convergence of the *average* iterate
- **Stochastic**: expected suboptimality via online-to-batch

Constant Stepsize: Smooth vs Nonsmooth





Online Convex Optimization

Definition 9.1 (OCO protocol). E f.d. normed space, $X \subseteq E$ nonempty closed convex. An *online algorithm* is $A : \mathcal{F}_X^* \to X$ (maps past losses to next iterate). The induced play: for $t = 1, 2, \ldots$

(i) Learner plays $x_t \leftarrow A(f_1, \ldots, f_{t-1})$.

(ii) Environment reveals convex $f_t : X \to \mathbb{R}$; learner incurs $f_t(x_t)$.

Definition 9.2 (Regret). Against comparator $u \in X$, after T rounds,

$$\mathrm{Reg}_T(u) := \sum_{t=1}^{T} \big(f_t(x_t) - f_t(u)\big).$$

From OCO to offline. If $f_t \equiv f$ and $\bar{x}_T = \frac{1}{T}\sum x_t$, Jensen gives $T\big(f(\bar{x}_T) - f(u)\big) \le \mathrm{Reg}_T(u)$.

Linear Losses Are the Hard Core of OCO

Online Linear Optimization (OLO). The special case of OCO (Def 9.1) where the round-t loss is linear: the environment reveals $g_t \in E^*$ and the loss is $u \mapsto \langle g_t, u \rangle$.

Remark 9.1 (Linear losses as the hard core). For general convex f_t, pick $g_t \in \partial f_t(x_t)$ (or $g_t = \nabla f_t(x_t)$ if differentiable). Convexity gives

$$f_t(x_t) - f_t(u) \le \langle g_t, x_t - u \rangle \qquad \forall u \in X,$$

hence $\mathrm{Reg}_T(u) \le \sum_{t=1}^{T} \langle g_t, x_t - u \rangle$.

Consequence. Controlling the linear losses on the right $\Rightarrow$ controlling OCO regret on the left. Today's pathwise quantity is the stepsize-weighted version $\sum_{t=1}^{T} \eta_t \langle g_t, x_t - u \rangle$.

Master: Bregman-Form Weighted Linearized Regret

Theorem 9.1 (Master pathwise bound). Φ mirror map, $X \subseteq \operatorname{dom}\Phi$ closed convex with $X \cap \operatorname{int}(\operatorname{dom}\Phi) \neq \emptyset$. For $t = 1, \dots, T$: $x_t \in X \cap \operatorname{int}(\operatorname{dom}\Phi)$, $g_t \in E^*$, $\eta_t > 0$, and

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

Then for every $u \in X$,

$$\sum_{t=1}^{T} \eta_t \langle g_t, x_t - u \rangle \leq \underbrace{D_\Phi(u, x_1) - D_\Phi(u, x_{T+1})}_{\text{telescope}} + \sum_{t=1}^{T} \underbrace{\left(\eta_t \langle g_t, x_t - x_{t+1} \rangle - D_\Phi(x_{t+1}, x_t)\right)}_{\substack{\text{maximal decrement of} \\ \text{a Bregman-type upper bound}}}.$$

Proof. Apply Thm 8.9 at each t with comparator u; the $D_\Phi(u, x_t) - D_\Phi(u, x_{t+1})$ terms telescope. $\square$

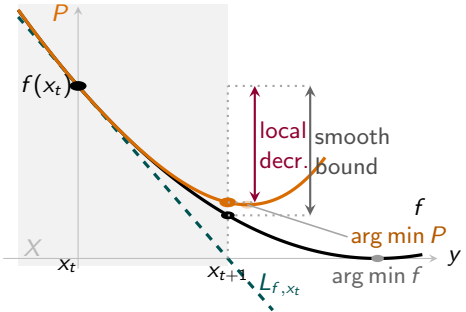
Visualizing the Local Decrement

$L_{f,x_t}(y) := f(x_t) + \langle g_t, y - x_t \rangle$ (linearization of f at x_t); $P := L_{f,x_t} + \frac{1}{\eta_t} D_\Phi(\cdot, x_t)$ (*Bregman proxy*);

$Q := L_{f,x_t} + \frac{\alpha}{2\eta_t} \|\cdot - x_t\|^2$ (*Quadratic proxy*).

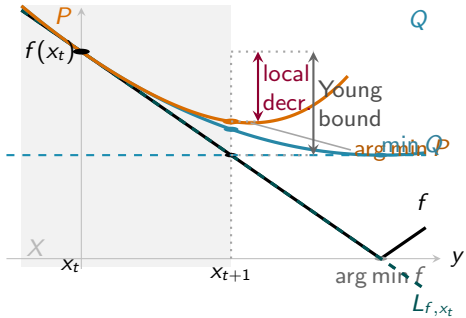
Local decrement $:= \eta_t \langle g_t, x_t - x_{t+1} \rangle - D_\Phi(x_{t+1}, x_t) = \eta_t \big[f(x_t) - P(x_{t+1}) \big] \geq 0.$

Smooth: $f = \frac{1}{2}(y-2)^2$, L-smooth rel. Φ



$f \leq P \Rightarrow$ local decr. $\leq \eta_t \big(f(x_t) - f(x_{t+1}) \big)$

Nonsmooth: $f = 2|y-2|$, use α-SC of Φ



$P \geq Q$, complete sq. on Q:

local decr. $\leq \frac{\eta_t^2}{2\alpha} \|g_t\|_*^2$

Sanity Check: Lecture 8 Is the Smooth Specialization

Let $g_t = \nabla f(x_t)$, f convex and L-smooth rel. Φ, $\eta_t \le 1/L$. Relative smoothness gives

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

Using convexity at $u = x^*$ turns Thm 9.1 into

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

the last-iterate telescope of Lecture 8.

Two regimes share one master:

- *Smooth* (Lecture 8): local decr. $\to$ **descent** $\Rightarrow$ last-iterate rate.
- *Nonsmooth* (today): local decr. bounded by $\frac{\eta_t^2}{2\alpha} \|g_t\|_*^2 \Rightarrow$ **averaged** rate.

From Pathwise Bound to Offline Averaging

Theorem 9.2 (Offline nonsmooth master). Under Thm 9.1's hypotheses, if $f_t \equiv f$ convex and $g_t \in \partial f(x_t)$, define $A_T := \sum_{t=1}^{T} \eta_t$ and the weighted average $\bar{x}_T^{(\eta)} := \frac{1}{A_T} \sum_{t=1}^{T} \eta_t x_t$. Then for every $u \in X$:

$$A_T\big(f(\bar{x}_T^{(\eta)}) - f(u)\big) \le D_\Phi(u, x_1) + \sum_{t=1}^{T} \big(\eta_t \langle g_t, x_t - x_{t+1}\rangle - D_\Phi(x_{t+1}, x_t)\big).$$

Proof. (1) Subgrad. ineq.: $f(x_t) - f(u) \le \langle g_t, x_t - u\rangle$. Multiply by η_t and sum. (2) Jensen: $f(\bar{x}_T^{(\eta)}) \le \frac{1}{A_T} \sum_t \eta_t f(x_t) \Rightarrow A_T(f(\bar{x}_T^{(\eta)}) - f(u)) \le \sum_t \eta_t (f(x_t) - f(u))$. (3) Combine (1),(2) $\Rightarrow A_T(f(\bar{x}_T^{(\eta)}) - f(u)) \le \sum_t \eta_t \langle g_t, x_t - u\rangle$. (4) Apply Thm 9.1 to RHS. $\square$

Dual-Norm Bound on the Local Decrement

Lemma 9.3. If Φ is α-strongly convex w.r.t. $\|\cdot\|$ on $X \cap \operatorname{int}(\operatorname{dom}\Phi)$, i.e., $D_\Phi(y,x) \geq \frac{\alpha}{2}\|y-x\|^2$, then for any $x, y \in X \cap \operatorname{int}(\operatorname{dom}\Phi)$ and $g \in E^*$:

$$\langle g, x-y\rangle - D_\Phi(y,x) \;\leq\; \frac{1}{2\alpha}\|g\|_*^2.$$

Proof.

1. Duality: $\langle g, x-y\rangle \leq \|g\|_*\|x-y\|$.
2. Young ($ab \leq \frac{a^2}{2\alpha} + \frac{\alpha b^2}{2}$) with $a = \|g\|_*$, $b = \|x-y\|$:
$$\|g\|_*\|x-y\| \leq \tfrac{1}{2\alpha}\|g\|_*^2 + \tfrac{\alpha}{2}\|x-y\|^2.$$
3. Strong convexity: $D_\Phi(y,x) \geq \frac{\alpha}{2}\|y-x\|^2$, so $-D_\Phi(y,x) \leq -\frac{\alpha}{2}\|x-y\|^2$.
4. Add: $\langle g, x-y\rangle - D_\Phi(y,x) \leq \frac{1}{2\alpha}\|g\|_*^2$. $\square$

Takeaway. α-SC of Φ turns the local decrement into a **dual-norm budget**. Applying it to $g \to \eta_t g_t$ gives the $\frac{\eta_t^2}{2\alpha}\|g_t\|_*^2$ budget we need.

Norm-Based Regret Bound

Corollary 9.4. Under Thm 9.1 with Φ α-strongly convex w.r.t. $\|\cdot\|$:

$$\sum_{t=1}^{T} \eta_t \langle g_t, x_t - u \rangle \le D_\Phi(u, x_1) - D_\Phi(u, x_{T+1}) + \sum_{t=1}^{T} \frac{\eta_t^2}{2\alpha} \|g_t\|_*^2.$$

With constant $\eta_t \equiv \eta$ and dropping $D_\Phi(u, x_{T+1}) \ge 0$:

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

Proof. Apply Lem 9.3 with $x = x_t$, $y = x_{t+1}$, $g = \eta_t g_t$:

$$\langle \eta_t g_t, x_t - x_{t+1} \rangle - D_\Phi(x_{t+1}, x_t) \le \frac{\eta_t^2}{2\alpha} \|g_t\|_*^2.$$

Substitute into the pathwise master bound (Thm 9.1). $\square$

Offline Nonsmooth: $O(1/\sqrt{T})$ Rate

Corollary 9.5 (Offline subgradient rate). f convex, $g_t \in \partial f(x_t)$, $\|g_t\|_* \le G$, constant η, $D_\Phi(x^\star, x_1) \le R^2$. Then

$$f(\bar{x}_T) - f(x^\star) \;\le\; \frac{D_\Phi(x^\star, x_1)}{\eta T} + \frac{\eta G^2}{2\alpha}.$$

Optimal $\eta = \dfrac{R\sqrt{2\alpha}}{G\sqrt{T}}$ gives $\boxed{f(\bar{x}_T) - f(x^\star) \le RG\sqrt{\tfrac{2}{\alpha T}}.}$

Proof. (1) Cor 9.4 (const. η): $\sum_t \langle g_t, x_t - u\rangle \le \frac{D_\Phi(u,x_1)}{\eta} + \frac{\eta T G^2}{2\alpha}$. (2) Thm 9.2 + const. $\eta \Rightarrow \bar{x}_T^{(\eta)} = \bar{x}_T$: $T(f(\bar{x}_T) - f(u)) \le \sum_t \langle g_t, x_t - u\rangle$. (3) Divide by $T \Rightarrow$ first display. (4) Min over η: $\eta^* = \frac{R\sqrt{2\alpha}}{G\sqrt{T}}$ balances both terms at $\frac{RG}{\sqrt{2\alpha T}}$ each. $\square$

Compare: smooth (L8) $O(LR^2/T)$ vs. nonsmooth $O(GR/\sqrt{T})$ — $\sqrt{T}$ is the price of nonsmoothness.

Strongly Convex Nonsmooth: μ-SC Relative to Φ

Definition of relative strong convexity (used inside Thm 9.6): f is μ-strongly convex relative to Φ on X if

$$f - \mu\Phi \text{ is convex on } X.$$

Consequence. For $x \in X \cap \operatorname{int}(\operatorname{dom}\Phi)$, $y \in X$, $g_x \in \partial f(x)$:

$$f(y) \geq f(x) + \langle g_x, y - x\rangle + \mu D_\Phi(y, x).$$

(Derived in the proof of Thm 9.6 from $f - \mu\Phi$ convex + Φ differentiable at x.)

Why weighted averaging? Equal-weight average + SC only proves $(\log T)/T$. Weighting iterates by t (later iterates count more) restores $\boxed{O(1/T)}$. Intuition: later iterates are closer to $x^\star$.

SC Nonsmooth: Weighted Template

Theorem 9.6. Φ α-SC w.r.t. $\|\cdot\|$; f μ-SC rel. Φ ($f-\mu\Phi$ convex); $g_t \in \partial f(x_t)$, $\|g_t\|_* \le G$; $x^\star \in \operatorname{argmin}_X f$. Fix weights $\lambda_1, \dots, \lambda_T > 0$; set $\Lambda_t := \sum_{s=1}^{t} \lambda_s$, stepsize $\eta_t := \dfrac{\lambda_t}{\mu \Lambda_t}$, weighted average $\widetilde{x}_T^{(\lambda)} := \dfrac{1}{\Lambda_T} \sum_{t=1}^{T} \lambda_t x_t$. Then

$$\boxed{f\big(\widetilde{x}_T^{(\lambda)}\big) - f(x^\star) \le \frac{G^2}{2\alpha\mu\Lambda_T} \sum_{t=1}^{T} \frac{\lambda_t^2}{\Lambda_t}.}$$

How to pick η_t. The one-step recursion (next slide) has form $f(x_t) - f(x^\star) \le \big(\frac{1}{\eta_t} - \mu\big) D_t - \frac{1}{\eta_t} D_{t+1} + \cdots$. Multiplying by λ_t and asking D_t-coefficients to telescope forces $\lambda_t\big(\frac{1}{\eta_t} - \mu\big) = \frac{\lambda_{t-1}}{\eta_{t-1}}$, solved by $\frac{\lambda_t}{\eta_t} = \mu\Lambda_t$.

Thm 9.6: Proof ($D_t := D_\Phi(x^\star, x_t)$, $\Lambda_0 := 0$)

Step 1. μ-subgradient inequality. $f - \mu\Phi$ convex + $\theta \downarrow 0$:
$f(y) \geq f(x_t) + \langle g_t, y - x_t\rangle + \mu D_\Phi(y, x_t)$. At $y = x^\star$: $f(x_t) - f(x^\star) \leq \langle g_t, x_t - x^\star\rangle - \mu D_t$.

Step 2. One-step + Young. Thm 8.9 ($u = x^\star$) + Lem 9.3 ($g = \eta_t g_t$):
$\eta_t\langle g_t, x_t - x^\star\rangle \leq D_t - D_{t+1} + \frac{\eta_t^2 G^2}{2\alpha}$. Combine with Step 1; divide by η_t:

$$f(x_t) - f(x^\star) \leq \big(\tfrac{1}{\eta_t} - \mu\big) D_t - \tfrac{1}{\eta_t} D_{t+1} + \tfrac{\eta_t G^2}{2\alpha}.$$

Step 3. Weighted recursion. With $\eta_t = \lambda_t/(\mu\Lambda_t)$:

$$\lambda_t\big(\tfrac{1}{\eta_t} - \mu\big) = \mu\Lambda_{t-1}, \quad \lambda_t/\eta_t = \mu\Lambda_t, \quad \lambda_t\eta_t = \tfrac{\lambda_t^2}{\mu\Lambda_t}.$$

Multiply Step 2 by λ_t: $\lambda_t(f(x_t) - f(x^\star)) \leq \mu\Lambda_{t-1}D_t - \mu\Lambda_t D_{t+1} + \frac{G^2}{2\alpha\mu}\frac{\lambda_t^2}{\Lambda_t}$.

Step 4. Sum + Jensen. D_t-terms telescope; $-\mu\Lambda_T D_{T+1} \leq 0$:
$\sum \lambda_t(f(x_t) - f(x^\star)) \leq \frac{G^2}{2\alpha\mu}\sum \frac{\lambda_t^2}{\Lambda_t}$. Jensen on $\widetilde{x}_T^{(\lambda)}$; divide by Λ_T. $\square$

Cor 9.7: Linear Weights $\lambda_t = t$ Give $O(1/T)$

Corollary 9.7. Under Thm 9.6 with $\lambda_t := t$:

$$\Lambda_t = \tfrac{t(t+1)}{2}, \quad \eta_t = \tfrac{2}{\mu(t+1)}, \quad \widetilde{x}_T^{(\lambda)} = \tfrac{2}{T(T+1)} \sum_{t=1}^{T} t\, x_t,$$

$$\boxed{f(\widetilde{x}_T^{(\lambda)}) - f(x^\star) \le \tfrac{2G^2}{\alpha\mu(T+1)}.}$$

Proof.

1. $\Lambda_t = \sum_{s=1}^{t} s = \frac{t(t+1)}{2} \Rightarrow \eta_t = \frac{t}{\mu\, t(t+1)/2} = \frac{2}{\mu(t+1)}$.
2. $\frac{\lambda_t^2}{\Lambda_t} = \frac{t^2}{t(t+1)/2} = \frac{2t}{t+1} \le 2$, so $\sum_{t=1}^{T} \frac{\lambda_t^2}{\Lambda_t} \le 2T$.
3. Thm 9.6: bound $\le \frac{G^2}{2\alpha\mu \cdot T(T+1)/2} \cdot 2T = \frac{2G^2}{\alpha\mu(T+1)}$. $\square$

Nonsmooth analogue of L8's $(1-\mu/L)^T$ linear rate — polynomial $1/T$ only.

Online-to-Stochastic Reduction

Setup (inline; no separate definition). $(\Xi, \mathcal{A}, \mathbb{P})$ probability space; $f : X \times \Xi \to \mathbb{R}$ convex in x; population risk $F(x) := \mathbb{E}_\xi[f(x, \xi)]$; $(\xi_t)_{t \geq 1}$ i.i.d.; x_t measurable w.r.t. $\sigma(\xi_1, \ldots, \xi_{t-1})$; $\hat{g}_t \in \partial_x f(x_t, \xi_t)$; $A_T = \sum \eta_t$, $\bar{x}_T^{(\eta)} = \frac{1}{A_T} \sum \eta_t x_t$.

Theorem 9.8. If a.s. $\sum_{t=1}^{T} \eta_t \langle \hat{g}_t, x_t - u \rangle \leq \mathcal{R}_T(u)$ for some integrable $\mathcal{R}_T(u)$, then

$$\mathbb{E}\big[F(\bar{x}_T^{(\eta)}) - F(u)\big] \;\leq\; \frac{\mathbb{E}[\mathcal{R}_T(u)]}{A_T}.$$

This is a wrapper. Any pathwise bound $\mathcal{R}_T$ plugs in (smooth, nonsmooth, SC) — same reduction.

Thm 9.8: Proof

Step 1. Jensen. Convexity of F gives $F(\bar{x}_T^{(\eta)}) \le \frac{1}{A_T}\sum \eta_t F(x_t)$, so

$$A_T(F(\bar{x}_T^{(\eta)}) - F(u)) \le \sum_t \eta_t (F(x_t) - F(u)).$$

Step 2. Conditional expectation. x_t is $\sigma(\xi_1, \dots, \xi_{t-1})$-measurable, ξ_t independent $\Rightarrow$ $\mathbb{E}[f(x_t, \xi_t) - f(u, \xi_t) \mid \xi_1, \dots, \xi_{t-1}] = F(x_t) - F(u)$. Hence

$$A_T \mathbb{E}[F(\bar{x}_T^{(\eta)}) - F(u)] \le \mathbb{E}\sum_t \eta_t (f(x_t, \xi_t) - f(u, \xi_t)).$$

Step 3. Convexity of $f(\cdot, \xi_t)$. $f(x_t, \xi_t) - f(u, \xi_t) \le \langle \hat{g}_t, x_t - u \rangle$. So

$$A_T \mathbb{E}[F(\bar{x}_T^{(\eta)}) - F(u)] \le \mathbb{E}\sum_t \eta_t \langle \hat{g}_t, x_t - u \rangle \le \mathbb{E}\mathcal{R}_T(u).$$

Divide by A_T. $\square$

Bounded-Noise Stochastic Oracle

Definition 9.3 (Unbiased oracle with bounded pop. subgradient and bounded noise). Filtration $(\mathcal{F}_t)_{t\geq 0}$. A stochastic subgradient sequence $(\hat{g}_t)$ is unbiased with bounded population subgradient and noise if $\exists$ predictable (g_t) and constants $G, \sigma \geq 0$ with

$$\mathbb{E}[\hat{g}_t \mid \mathcal{F}_{t-1}] = g_t, \quad g_t \in \partial F(x_t), \quad \|g_t\|_* \leq G,$$

$$\mathbb{E}[\|\hat{g}_t - g_t\|_*^2 \mid \mathcal{F}_{t-1}] \leq \sigma^2.$$

Reading. Two separate budgets:

- G: size of the *population* subgradient (deterministic part).
- σ: size of the *noise* $\hat{g}_t - g_t$ (martingale difference).

Together: $\mathbb{E}\|\hat{g}_t\|_*^2 \leq 2G^2 + 2\sigma^2$ (via triangle + $(a+b)^2 \leq 2a^2 + 2b^2$).

Nonsmooth SMD: $O(1/\sqrt{T})$

Corollary 9.9. Φ α-SC; oracle as in Def 9.3; constant η; $D_\Phi(x^\star, x_1) \le R^2$.

$$\mathbb{E}\big[F(\bar{x}_T) - F(x^\star)\big] \;\le\; \frac{R^2}{\eta T} + \frac{\eta}{\alpha}(G^2 + \sigma^2).$$

Optimal $\eta = \sqrt{\frac{\alpha R^2}{T(G^2+\sigma^2)}}$ gives $\boxed{\mathbb{E}[F(\bar{x}_T) - F(x^\star)] \le 2R\sqrt{\frac{G^2+\sigma^2}{\alpha T}}.}$

Reading. G^2 (deterministic subgradient size) and σ^2 (noise variance) combine *additively* under $\sqrt{\cdot}$. Rate $O(1/\sqrt{T})$ matches offline nonsmooth (Cor 9.5) — noise is free to leading order.

Proof map (next slide): well-posedness $\to$ Cor 9.4 pathwise $\to$ Thm 9.8 reduction $\to$ bound $\mathbb{E}\|\hat{g}_t\|_*^2$ $\to$ optimize η.

Cor 9.9: Proof

Step 0. Well-posedness. Lem 8.8 recursively $\Rightarrow x_t \in X \cap \operatorname{int}(\operatorname{dom}\Phi)$, so Cor 9.4 applies.

Step 1. Pathwise (Cor 9.4). $\sum_{t=1}^{T} \eta\langle \hat{g}_t, x_t - u\rangle \le D_\Phi(u, x_1) + \frac{\eta^2}{2\alpha}\sum_{t=1}^{T}\|\hat{g}_t\|_*^2$.

Step 2. Online-to-stochastic (Thm 9.8). Take $\mathcal{R}_T$ = RHS:

$$\mathbb{E}[F(\bar{x}_T) - F(u)] \le \frac{D_\Phi(u,x_1)}{\eta T} + \frac{\eta}{2\alpha T}\sum_{t}\mathbb{E}\|\hat{g}_t\|_*^2.$$

Step 3. Bound $\mathbb{E}\|\hat{g}_t\|_*^2$. Triangle: $\|\hat{g}_t\|_* \le \|g_t\|_* + \|\hat{g}_t - g_t\|_*$. $(a+b)^2 \le 2a^2+2b^2$, then $\mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ and Def 9.3: $\mathbb{E}\|\hat{g}_t\|_*^2 \le 2G^2 + 2\sigma^2$.

Step 4. Assemble. Sub. Step 3 into Step 2, set $u = x^\star$:
$\mathbb{E}[F(\bar{x}_T) - F(x^\star)] \le \frac{R^2}{\eta T} + \frac{\eta}{\alpha}(G^2+\sigma^2)$.

Step 5. Optimize. Min of $\frac{a}{\eta} + b\eta$ is $2\sqrt{ab}$ at $\eta = \sqrt{a/b}$. With $a = R^2/T$, $b = (G^2+\sigma^2)/\alpha$:
$\eta^* = \sqrt{\frac{\alpha R^2}{T(G^2+\sigma^2)}}$, bound $= 2R\sqrt{\frac{G^2+\sigma^2}{\alpha T}}$. $\square$

Smooth SMD under Relative Smoothness

Theorem 9.10. F convex, L-smooth rel. Φ (i.e., $F(y) \le F(x) + \langle \nabla F(x), y - x\rangle + L D_\Phi(y, x)$); Φ α-SC; unbiased $\hat{g}_t$ with $\mathbb{E}[\|\hat{g}_t - \nabla F(x_t)\|_*^2 \mid \mathcal{F}_{t-1}] \le \sigma^2$; $\eta \in (0, 1/L)$; post-update avg $\bar{x}_T^+ := \frac{1}{T}\sum_{t=1}^{T} x_{t+1}$. Then

$$\mathbb{E}\big[F(\bar{x}_T^+) - F(x^\star)\big] \;\le\; \frac{D_\Phi(x^\star, x_1)}{\eta T} + \frac{\eta\,\sigma^2}{2\alpha(1 - \eta L)}.$$

Choosing $\eta = \min\{\frac{1}{2L}, \frac{R\sqrt{\alpha}}{\sigma\sqrt{T}}\}$ (with $D_\Phi(x^\star, x_1) \le R^2$, $\sigma > 0$):

$$\boxed{\mathbb{E}[F(\bar{x}_T^+) - F(x^\star)] \le \tfrac{4LR^2}{T} + \tfrac{2\sigma R}{\sqrt{\alpha T}}.}$$

Two regimes in the rate: deterministic $O(LR^2/T)$ (smooth-descent part) + **noise** $O(\sigma R/\sqrt{\alpha T})$ (stochastic floor). Small noise $\to$ smooth GD $(1/T)$; large noise $\to$ like nonsmooth SMD $(1/\sqrt{T})$.

Thm 9.10: Proof ($s_t := x_t - x_{t+1}$, $\xi_t := \hat{g}_t - \nabla F(x_t)$)

Step 0. Well-posedness. Lem 8.8 applied recursively keeps $x_t \in X \cap \mathrm{int}(\mathrm{dom}\,\Phi)$, so Thm 9.1 applies to the run.

Step 1. Split. $\eta\langle \hat{g}_t, s_t\rangle = \eta\langle \nabla F(x_t), s_t\rangle + \eta\langle \xi_t, s_t\rangle$.

Step 2. Smooth part. Rel. smoothness with $y = x_{t+1}$:
$F(x_{t+1}) \le F(x_t) - \langle \nabla F(x_t), s_t\rangle + L D_\Phi(x_{t+1}, x_t)$. Hence

$$\eta\langle \nabla F(x_t), s_t\rangle - D_\Phi(x_{t+1}, x_t) \le \eta[F(x_t) - F(x_{t+1})] - (1-\eta L) D_\Phi(x_{t+1}, x_t).$$

Step 3. Noise part. Young ($\alpha(1-\eta L) > 0$): $\eta\langle \xi_t, s_t\rangle \le \frac{\eta^2}{2\alpha(1-\eta L)}\|\xi_t\|_*^2 + \frac{\alpha(1-\eta L)}{2}\|s_t\|^2$.
$D_\Phi(x_{t+1}, x_t) \ge \frac{\alpha}{2}\|s_t\|^2 \Rightarrow$ last term $\le (1-\eta L) D_\Phi(x_{t+1}, x_t)$.

Step 4. Combine. Add Steps 2+3, cancel $(1-\eta L) D_\Phi$:

$$\eta\langle \hat{g}_t, s_t\rangle - D_\Phi(x_{t+1}, x_t) \le \eta[F(x_t) - F(x_{t+1})] + \frac{\eta^2}{2\alpha(1-\eta L)}\|\xi_t\|_*^2.$$

Thm 9.10: Proof (continued)

Step 5. Apply Thm 9.1 at $u = x^\star$. Summing Step 4:

$$\sum_t \eta\langle \hat{g}_t, x_t - x^\star\rangle \le D_\Phi(x^\star, x_1) + \eta\sum_t [F(x_t) - F(x_{t+1})] + \tfrac{\eta^2}{2\alpha(1-\eta L)}\sum_t \|\xi_t\|_*^2.$$

Step 6. Convexity of F. $F(x_t) - F(x^\star) \le \langle \hat{g}_t, x_t - x^\star\rangle - \langle \xi_t, x_t - x^\star\rangle$. Multiply by η, sum; F-telescoping:

$$\eta\sum_t [F(x_{t+1}) - F(x^\star)] \le D_\Phi(x^\star, x_1) + \tfrac{\eta^2}{2\alpha(1-\eta L)}\sum_t \|\xi_t\|_*^2 - \eta\sum_t \langle \xi_t, x_t - x^\star\rangle.$$

Step 7. Expectation. $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0 \Rightarrow$ last sum vanishes; $\mathbb{E}\|\xi_t\|_*^2 \le \sigma^2$; Jensen on $\bar{x}_T^+$:

$$\mathbb{E}[F(\bar{x}_T^+) - F(x^\star)] \le \tfrac{D_\Phi(x^\star, x_1)}{\eta T} + \tfrac{\eta\sigma^2}{2\alpha(1-\eta L)}.$$

Step 8. Optimize η. $\eta \le 1/(2L) \Rightarrow 1 - \eta L \ge 1/2$, so $\mathbb{E}[\cdot] \le \frac{R^2}{\eta T} + \frac{\eta\sigma^2}{\alpha}$. Min over $(0, 1/(2L)]$: $\eta = R\sqrt{\alpha}/(\sigma\sqrt{T})$ gives $2\sigma R/\sqrt{\alpha T}$; $\eta = 1/(2L)$ gives $4LR^2/T$. Take max: boxed bound. $\square$

Summary: One Master, Many Rates

Thm 9.1 (pathwise) + Young + Jensen + conditional expectation.

Setting	Rate	Ref
Offline nonsmooth (convex)	$O\big(GR/\sqrt{T}\big)$	Cor 9.5
Offline nonsmooth (μ-SC rel. Φ)	$O\big(G^2/(\alpha\mu T)\big)$	Cor 9.7
Stochastic nonsmooth	$O\big(R\sqrt{(G^2+\sigma^2)/(\alpha T)}\big)$	Cor 9.9
Stochastic smooth (rel. Φ)	$O(LR^2/T) + O\big(\sigma R/\sqrt{\alpha T}\big)$	Thm 9.10

Next lecture (L10). Unfreeze the geometry: $\Phi_H(x) = \frac{1}{2}x^\top Hx$ with H chosen adaptively from observed gradients (AdaGrad, Adam).

Symbol Review

Geometry		Iterates & outputs	
Φ	mirror map	x_t	iterate at round t
$D_\Phi(y, x)$	Bregman divergence	x_{t+1}	mirror-descent step
α	SC const. of Φ w.r.t. $\|\cdot\|$	$\bar{x}_T$	uniform average $\frac{1}{T}\sum x_t$
$\|\cdot\|, \|\cdot\|_*$	primal / dual norm	$\bar{x}_T^{(\eta)}$	η-weighted avg
R	$\sqrt{D_\Phi(x^\star, x_1)}$ bound	$\bar{x}_T^{(\lambda)}$	λ-weighted avg (Thm 9.6)
		$\bar{x}_T^{+}$	post-update avg (Thm 9.10)
Loss / objective		**Oracles & noise**	
f, f_t	convex loss / per-round loss	g_t	(sub)gradient $\in \partial f_t(x_t)$
F	population risk $\mathbb{E}_\xi f(x, \xi)$	$\hat{g}_t$	stochastic subgrad. (noisy)
$f(x, \xi)$	sample loss	ξ_t	noise $\hat{g}_t - \nabla F(x_t)$
L	smoothness const. rel. Φ	G	pop. subgrad. bound
μ	SC const. of f rel. Φ	σ	noise L^2-bound
Game		**Regret**	
u	comparator $\in X$	$\mathrm{Reg}_T(u)$	$\sum(f_t(x_t) - f_t(u))$
η_t	stepsize at round t	$\mathrm{Reg}_T^{\mathrm{OLO}}(u)$	linearized: $\sum\langle g_t, x_t - u\rangle$
A_T	$\sum \eta_t$		