Lecture 9: Nonsmooth, Online, and Stochastic Mirror Descent

Lecture 8 analyzed mirror descent in the relative-smooth setting. Lecture 9 drops smoothness, allows arbitrary covectors or subgradients, and asks what remains of the same one-step geometry. The answer is a pathwise Bregman inequality whose three readings are an online comparison bound, offline nonsmooth convex optimization, and stochastic optimization after taking expectations.

9.1 Motivation and Online Convex Optimization

Lecture 8 analyzed mirror descent in the smooth setting. The geometric input was the constrained one-step mirror inequality from Theorem 8.9. Relative smoothness was then used to control the local term and turn that one-step inequality into a descent estimate for one fixed objective.

This raises the next question: what remains true when the objective is convex but not smooth, so that each step sees only a subgradient rather than a gradient? Without smoothness, the one-step mirror inequality no longer turns into a descent estimate for one fixed objective. In particular, even with an arbitrarily small constant stepsize, mirror descent need not make $f(x_t)$ decrease at every step, and the last iterate need not converge. The natural output is therefore often not the final iterate but an average of the whole trajectory; for strongly convex objectives, a different weighted average leads to a faster rate.

Example 9.1 (Constant stepsizes in smooth and nonsmooth one-dimensional models)**.** Consider descent on $\mathbb{R}$ with a fixed stepsize $\eta > 0$. For the smooth quadratic

$$f_{\mathrm{sm}}(x) := \frac{1}{2}x^2,$$

gradient descent satisfies

$$x_{t+1} = x_t - \eta \nabla f_{\mathrm{sm}}(x_t) = (1-\eta)x_t,$$

so $x_t \to 0$ whenever $0 < \eta < 2$.
In contrast, for the nonsmooth objective

$$f_{\mathrm{ns}}(x) := |x|,$$

every nonzero iterate of subgradient descent satisfies

$$x_{t+1} = x_t - \eta \operatorname{sign}(x_t).$$

Thus, as long as $x_t \neq 0$, one has

$$|x_{t+1} - x_t| = \eta.$$

Unless an iterate lands exactly at 0, the step size never goes to 0, so (x_t) cannot converge.

The point is that the same nonsmooth mirror-descent analysis extends, with no new geometry, to a

Figure 1: Constant-stepsize gradient and subgradient descent started from a shared x_0. **Left (smooth $\frac{1}{2}x^2$):** a large η alternates sign across the minimizer but still contracts geometrically (orange), while a small η simply contracts monotonically (teal) — both converge. **Right (nonsmooth $|x|$):** a large η overshoots the kink already on the first step and gets trapped in a two-point oscillation (orange); and even a small η, which descends monotonically for several steps, eventually overshoots as well and bounces across the kink forever (teal) — neither converges.

more general setting in which the loss may change from round to round. This more general problem is online convex optimization (OCO). We therefore formulate OCO first, and then recover offline nonsmooth convex optimization as the repeated-objective special case. Stochastic optimization will later follow by taking expectations.

Definition 9.1 (Online convex optimization protocol)**.** Let E be a finite-dimensional normed vector space, and let $X \subseteq E$ be nonempty, closed, and convex. Let $\mathcal{F}_X$ denote the set of convex functions $f : X \to \mathbb{R}$. An online optimization algorithm on X is a map

$$A : \mathcal{F}_X^* \to X,$$

where $\mathcal{F}_X^*$ denotes the set of finite sequences of elements of $\mathcal{F}_X$. Given such an algorithm A and a loss sequence $(f_t)_{t\geq 1} \subseteq \mathcal{F}_X$, the induced online play is the following protocol:

Algorithm 1 Online convex optimization on X

Require: A nonempty closed convex set $X \subseteq E$, a loss sequence $(f_t)_{t\geq 1} \subseteq \mathcal{F}_X$, and an online optimization algorithm $A : \mathcal{F}_X^* \to X$.
Ensure: A play consisting of iterates $x_t \in X$ and incurred losses $f_t(x_t)$.

1: **for** $t = 1, 2, \ldots$ **do**
2: Set $x_t \leftarrow A(f_1, \ldots, f_{t-1})$, with the convention $x_1 = A(\emptyset)$.
3: The learner incurs the loss $f_t(x_t)$.
4: **end for**

Definition 9.2 (Regret)**.** Given an online play induced by A and $f_1, \ldots, f_T$, for any fixed comparator $u \in X$, the regret after $T \in \mathbb{N}$ rounds is

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

If the same convex objective $f : X \to \mathbb{R}$ is used at every round, so that $f_t \equiv f$, then the usual regret

already has an immediate offline meaning. Indeed, if one sets

$$\bar{x}_T := \frac{1}{T}\sum_{t=1}^{T} x_t,$$

then $\bar{x}_T$ is an offline candidate solution, and Jensen's inequality gives

$$f(\bar{x}_T) \le \frac{1}{T}\sum_{t=1}^{T} f(x_t),$$

so for every $u \in X$,

$$T\big(f(\bar{x}_T) - f(u)\big) \le \mathrm{Reg}_T(u).$$

Thus regret bounds in OCO immediately imply convergence guarantees for the uniform average in offline convex optimization. Later, when stepsizes enter the mirror-descent analysis, the same idea will reappear in a weighted form.

Remark 9.1 (Linear losses as the hard core of OCO). The theorems below are written in terms of covectors $g_t \in E^*$ rather than the losses f_t themselves. Informally, this means that once one can solve online linear optimization, meaning the special case in which the loss at round t is the linear form

$$u \mapsto \langle g_t, u\rangle,$$

one can also solve online convex optimization (OCO): at round t, first-order convexity gives a covector $g_t \in E^*$ such that

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

If f_t is differentiable, one takes $g_t = \nabla f_t(x_t)$; if f_t is merely convex, one takes any subgradient at x_t. Consequently,

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

Thus any control of the linear losses on the right-hand side immediately gives a control of the regret on the left-hand side. In this intuitive sense, the linear case is the hardest part of OCO: once the linear case is understood, the general convex case follows by linearization.

9.2 Mirror Descent and Linearized Regret

The geometric input of this section is still Theorem 8.9. What changes is the interpretation of the covectors g_t. In Lecture 8 they were $\nabla f(x_t)$, and relative smoothness turned the local term into a descent quantity. Here they are arbitrary revealed covectors or subgradients, so the same one-step inequality becomes a pathwise estimate for a weighted linearized comparison bound. Since mirror descent is run with stepsizes $\eta_t > 0$, the natural quantities are

$$\mathrm{Reg}_T^{(\eta)}(u) := \sum_{t=1}^{T} \eta_t (f_t(x_t) - f_t(u)), \qquad \sum_{t=1}^{T} \eta_t \langle g_t, x_t - u\rangle.$$

Equivalently, if one rescales the losses by $\tilde{f}_t := \eta_t f_t$, then $\mathrm{Reg}_T^{(\eta)}(u)$ is just the ordinary regret for the sequence $(\tilde{f}_t)$. In the repeated-objective setting, this weighted form will correspond to a weighted average of the iterates.

Theorem 9.1 (Bregman-form weighted linearized regret decomposition)**.** *Let E be a finite-dimensional normed vector space. Let $\Phi : E \to (-\infty, +\infty]$ be a mirror map, and let $X \subseteq \operatorname{dom}\Phi$ be nonempty, closed, and convex with $X \cap \operatorname{int}(\operatorname{dom}\Phi) \neq \emptyset$. Let $T \in \mathbb{N}$. For each $t \in \{1, \dots, T\}$, let*

$$x_t \in X \cap \operatorname{int}(\operatorname{dom}\Phi), \qquad g_t \in E^*, \qquad \eta_t > 0,$$

and let

$$x_{t+1} \in \arg\min_{x \in X} \Big\{ \eta_t \langle g_t, x - x_t \rangle + D_\Phi(x, x_t) \Big\}.$$

Then, for every $u \in X$,

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

Proof of Theorem 9.1. Apply Theorem 8.9 at each time t with the same comparator u. Summing over $t \in \{1, \dots, T\}$ makes the terms $D_\Phi(u, x_t) - D_\Phi(u, x_{t+1})$ telescope. □

The local decrement $\eta_t \langle g_t, x_t - x_{t+1} \rangle - D_\Phi(x_{t+1}, x_t)$ appearing on the right-hand side of Theorem 9.1 can be read geometrically. Introduce the linearization of f at x_t and two local upper proxies,

$$L_{f,x_t}(y) := f(x_t) + \langle g_t, y - x_t \rangle, \qquad P(y) := L_{f,x_t}(y) + \tfrac{1}{\eta_t} D_\Phi(y, x_t), \qquad Q(y) := L_{f,x_t}(y) + \tfrac{\alpha}{2\eta_t} \|y - x_t\|^2,$$

where α is the strong-convexity constant of Φ. We refer to P as the *Bregman proxy* and Q as the *quadratic proxy*. The mirror step is $x_{t+1} \in \arg\min_{x \in X} P(x)$, and since $P(x_t) = f(x_t)$, the local decrement equals $\eta_t [f(x_t) - P(x_{t+1})] \geq 0$. Figure 2 shows both readings in a 1D example where the minimizer of f lies outside the feasible interval X, so the mirror step is pulled to the boundary.

Remark 9.2 (How the master theorem recovers Lecture 8). Suppose now that $f : X \to \mathbb{R}$ is convex and differentiable on $X \cap \operatorname{int}(\operatorname{dom}\Phi)$, and that we return to the smooth setting of Lecture 8 by taking

$$g_t = \nabla f(x_t).$$

If f is L-smooth relative to Φ and $\eta_t \leq 1/L$, then relative smoothness gives

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + L D_\Phi(x_{t+1}, x_t),$$

hence

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

Applying Theorem 9.1 with $u = x^\star \in \arg\min_X f$ and using convexity,

$$f(x_t) - f(x^\star) \leq \langle \nabla f(x_t), x_t - x^\star \rangle,$$

we obtain

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

This is exactly the one-step telescope behind the last-iterate theorem of Lecture 8. If one further assumes objective relative strong convexity, then the same telescope is also the input to the linear-convergence theorem there. In other words, Lecture 8 is the special case of the present master theorem in which relative smoothness turns the local term into a descent quantity for one fixed objective.

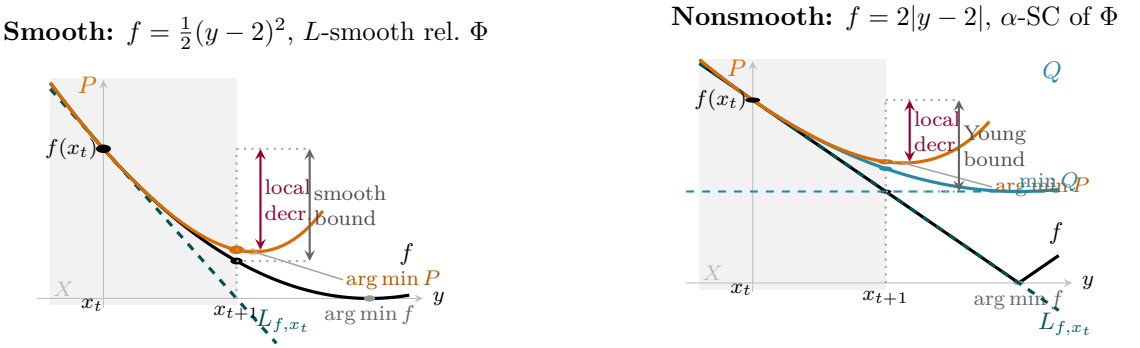

Figure 2: Visualizing the local decrement from Theorem 9.1. **Left (smooth):** f is L-smooth rel. Φ, so $f \le P$, and the vertical gap $f(x_t) - P(x_{t+1})$ (purple) is bounded by $f(x_t) - f(x_{t+1})$ (the "smooth bound"). **Right (nonsmooth):** f is no longer below P, but α-strong convexity of Φ gives $P \ge Q$, and completing the square on Q gives the "Young floor" at height $f(x_t) - \frac{\eta_t}{2\alpha}\|g_t\|_*^2$; the gap to this floor (black) bounds the local decrement, giving Lemma 9.3.

9.3 From OCO to Offline Convex, Nonsmooth Optimization

We now specialize to the case where the same convex objective is used at every iteration. In that repeated-objective setting, the weighted linearized regret bound becomes a convergence theorem for the weighted average of the iterates.

Theorem 9.2 (Offline nonsmooth mirror theorem)**.** *Under the hypotheses of Theorem 9.1, assume moreover that $f : X \to \mathbb{R}$ is convex and that*

$$g_t \in \partial f(x_t) \qquad \forall t \in \{1, \dots, T\}.$$

Define

$$A_T := \sum_{t=1}^{T} \eta_t, \qquad \bar{x}_T^{(\eta)} := \frac{1}{A_T} \sum_{t=1}^{T} \eta_t x_t.$$

Then, for every $u \in X$,

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

Proof of Theorem 9.2. By convexity of f,

$$f(x_t) - f(u) \le \langle g_t, x_t - u \rangle \qquad \forall t \in \{1, \dots, T\}.$$

Multiplying by η_t and summing gives

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

By Jensen's inequality applied to the convex function f,

$$f(\bar{x}_T^{(\eta)}) \le \frac{1}{A_T} \sum_{t=1}^{T} \eta_t f(x_t).$$

Combining the last two displays yields

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

Now apply Theorem 9.1. □

The next lemma records the standard bound on the local term when the mirror map is strongly convex with respect to a norm.

Lemma 9.3 (Local-term bound under strong convexity of the mirror map)**.** *Let E be a finite-dimensional normed vector space. Let $\Phi : E \to (-\infty, +\infty]$ be a mirror map, let $X \subseteq \operatorname{dom}\Phi$ be nonempty and convex, and assume that*

$$\forall x, y \in X \cap \operatorname{int}(\operatorname{dom}\Phi), \qquad D_\Phi(x, y) \ge \frac{\alpha}{2} \|x - y\|^2$$

for some norm $\|\cdot\|$ on E and some $\alpha > 0$. Let

$$x, y \in X \cap \operatorname{int}(\operatorname{dom}\Phi), \qquad g \in E^*.$$

Then

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

Proof of Lemma 9.3. By duality and Young's inequality,

$$\langle g, x - y \rangle \le \|g\|_* \|x - y\| \le \frac{1}{2\alpha} \|g\|_*^2 + \frac{\alpha}{2} \|x - y\|^2 .$$

By the assumed strong convexity of Φ,

$$D_\Phi(y, x) \ge \frac{\alpha}{2} \|y - x\|^2 .$$

Subtracting this lower bound proves the claim. □

Corollary 9.4 (Norm-based regret bound under strongly convex mirror map)**.** *Let E be a finite-dimensional normed vector space. Let $\Phi : E \to (-\infty, +\infty]$ be a mirror map, let $X \subseteq \operatorname{dom}\Phi$ be closed and convex with $X \cap \operatorname{int}(\operatorname{dom}\Phi) \neq \emptyset$, and assume that Φ is α-strongly convex on $X \cap \operatorname{int}(\operatorname{dom}\Phi)$ with respect to a norm $\|\cdot\|$ on E for some $\alpha > 0$. Under the hypotheses of Theorem 9.1, we have, for every $u \in X$,*

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

In particular, if $\eta_t \equiv \eta > 0$, then

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

Proof of Corollary 9.4. Apply Theorem 9.1. For each $t \in \{1, \dots, T\}$, Lemma 9.3 with $x = x_t$, $y = x_{t+1}$, and $g = \eta_t g_t$ gives

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

Substituting these bounds into Theorem 9.1 proves the first display. The constant-stepsize statement follows after dropping the nonnegative term $D_\Phi(u, x_{T+1})$. □

Corollary 9.5 (Offline subgradient rate via averaging)**.** *Assume the hypotheses of Corollary 9.4. Let $f : X \to \mathbb{R}$ be convex, and suppose that for each $t \in \{1, \dots, T\}$ one chooses a subgradient*

$$g_t \in \partial f(x_t),$$

and that $\|g_t\|_* \le G$ for all $t \in \{1, \dots, T\}$. Then, for every constant stepsize $\eta > 0$ and every $u \in X$,

$$f(\bar{x}_T) - f(u) \le \frac{D_\Phi(u, x_1)}{\eta T} + \frac{\eta G^2}{2\alpha}, \qquad \bar{x}_T := \frac{1}{T} \sum_{t=1}^{T} x_t.$$

In particular, if $x^\star \in \arg\min_X f$ exists and $D_\Phi(x^\star, x_1) \le R^2$, then choosing

$$\eta = \frac{R\sqrt{2\alpha}}{G\sqrt{T}}$$

yields

$$f(\bar{x}_T) - f(x^\star) \le RG\sqrt{\frac{2}{\alpha T}}.$$

Proof of Corollary 9.5. The constant-stepsize version of Corollary 9.4 gives

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

Apply Theorem 9.2 and use that with constant stepsizes the weighted average $\bar{x}_T^{(\eta)}$ is the simple average $\bar{x}_T$. The optimized choice of η balances the two terms on the right-hand side. □

9.4 Strongly Convex, Non-smooth Optimization

The previous subsection assumed that the geometry Φ was strongly convex. We now separate that from a different condition: strong convexity of the objective itself. This is the mechanism behind the $1/T$ rate for nonsmooth strongly convex optimization. The faster rate does not come from the standard equal-weight regret bound. The theorem below records the general weighted template. The $1/T$ rate comes from the linear choice $\lambda_t = t$, stated afterward as a corollary.

Theorem 9.6 (Weighted template for strongly convex nonsmooth optimization)**.** *Let $\Phi : E \to (-\infty, +\infty]$ be a mirror map, and let $X \subseteq \operatorname{dom} \Phi$ be nonempty, closed, and convex with $X \cap \operatorname{int}(\operatorname{dom} \Phi) \neq \emptyset$. Assume that Φ is α-strongly convex on $X \cap \operatorname{int}(\operatorname{dom} \Phi)$ with respect to a norm $\|\cdot\|$ on E for some $\alpha > 0$. Let $f : X \to \mathbb{R}$ be μ-strongly convex relative to Φ, i.e.,*

assume that $f - \mu\Phi$ is convex on X for some $\mu > 0$. Assume also that $x^\star \in \arg\min_X f$. Let $\lambda_1, \dots, \lambda_T > 0$, and define

$$\Lambda_t := \sum_{s=1}^{t} \lambda_s \qquad \forall t \in \{1, \dots, T\}.$$

For each $t \in \{1, \dots, T\}$, let

$$x_t \in X \cap \operatorname{int}(\operatorname{dom}\Phi), \qquad g_t \in \partial f(x_t), \qquad \|g_t\|_* \le G, \qquad \eta_t := \frac{\lambda_t}{\mu\Lambda_t},$$

and let

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

Define the weighted average

$$\widetilde{x}_T^{(\lambda)} := \frac{1}{\Lambda_T} \sum_{t=1}^{T} \lambda_t x_t.$$

Then

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

Proof of Theorem 9.6. Write

$$D_t := D_\Phi(x^\star, x_t) \qquad \forall t \in \{1, \dots, T+1\}.$$

Fix $t \in \{1, \dots, T\}$, and let $y \in X$. For $\theta \in (0, 1)$, write

$$z_\theta := (1 - \theta)x_t + \theta y \in X.$$

Since $f - \mu\Phi$ is convex on X,

$$(f - \mu\Phi)(z_\theta) \le (1 - \theta)(f - \mu\Phi)(x_t) + \theta(f - \mu\Phi)(y).$$

Rearranging gives

$$f(y) - f(x_t) \ge \frac{f(z_\theta) - f(x_t)}{\theta} - \mu\frac{\Phi(z_\theta) - \Phi(x_t)}{\theta} + \mu(\Phi(y) - \Phi(x_t)).$$

Since $g_t \in \partial f(x_t)$ and $z_\theta - x_t = \theta(y - x_t)$, the subgradient inequality yields

$$\frac{f(z_\theta) - f(x_t)}{\theta} \ge \langle g_t, y - x_t \rangle .$$

Letting $\theta \downarrow 0$ and using differentiability of Φ at $x_t \in \operatorname{int}(\operatorname{dom}\Phi)$, we obtain

$$f(y) \ge f(x_t) + \langle g_t, y - x_t \rangle + \mu D_\Phi(y, x_t).$$

Applying this with $y = x^\star$ yields

$$f(x_t) - f(x^\star) \le \langle g_t, x_t - x^\star \rangle - \mu D_t.$$

Applying Theorem 8.9 with $u = x^\star$ gives

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

By Lemma 9.3 with $x = x_t$, $y = x_{t+1}$, and $g = \eta_t g_t$,

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

Therefore

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

Equivalently,

$$f(x_t) - f(x^\star) \le \left(\frac{1}{\eta_t} - \mu\right) D_t - \frac{1}{\eta_t} D_{t+1} + \frac{\eta_t G^2}{2\alpha}. \tag{1}$$

Now let $\Lambda_0 := 0$. Multiplying Equation (1) by λ_t and using

$$\eta_t = \frac{\lambda_t}{\mu\Lambda_t}$$

gives

$$\lambda_t \left(\frac{1}{\eta_t} - \mu\right) = \mu\Lambda_{t-1}, \qquad \frac{\lambda_t}{\eta_t} = \mu\Lambda_t, \qquad \lambda_t\eta_t = \frac{\lambda_t^2}{\mu\Lambda_t}.$$

Therefore

$$\lambda_t(f(x_t) - f(x^\star)) \le \mu\Lambda_{t-1}D_t - \mu\Lambda_t D_{t+1} + \frac{G^2}{2\alpha\mu}\frac{\lambda_t^2}{\Lambda_t}.$$

Summing over $t \in \{1, \dots, T\}$, the Bregman terms telescope:

$$\sum_{t=1}^{T} \lambda_t(f(x_t) - f(x^\star)) \le -\mu\Lambda_T D_{T+1} + \frac{G^2}{2\alpha\mu}\sum_{t=1}^{T}\frac{\lambda_t^2}{\Lambda_t} \le \frac{G^2}{2\alpha\mu}\sum_{t=1}^{T}\frac{\lambda_t^2}{\Lambda_t}.$$

Finally, convexity of f gives

$$f(\widetilde{x}_T^{(\lambda)}) \le \frac{1}{\Lambda_T}\sum_{t=1}^{T}\lambda_t f(x_t),$$

so

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

Dividing by Λ_T proves the claim. □

Corollary 9.7 (Linear weights give the $1/T$ rate)**.** *Under the hypotheses of Theorem 9.6, take $\lambda_t := t$. Then*

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

and

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

Proof of Corollary 9.7. Here

$$\Lambda_t = \sum_{s=1}^{t} s = \frac{t(t+1)}{2},$$

so

$$\eta_t = \frac{\lambda_t}{\mu \Lambda_t} = \frac{t}{\mu\, t(t+1)/2} = \frac{2}{\mu(t+1)}.$$

Also

$$\widetilde{x}_T^{(\lambda)} = \frac{1}{\Lambda_T} \sum_{t=1}^{T} \lambda_t x_t = \frac{2}{T(T+1)} \sum_{t=1}^{T} t\, x_t.$$

Moreover,

$$\frac{\lambda_t^2}{\Lambda_t} = \frac{t^2}{t(t+1)/2} = \frac{2t}{t+1} \le 2.$$

Therefore

$$\sum_{t=1}^{T} \frac{\lambda_t^2}{\Lambda_t} \le 2T.$$

Applying Theorem 9.6 yields

$$f(\widetilde{x}_T^{(\lambda)}) - f(x^\star) \le \frac{G^2}{2\alpha\mu\Lambda_T} \cdot 2T = \frac{G^2}{\alpha\mu} \cdot \frac{2T}{T(T+1)} = \frac{2G^2}{\alpha\mu(T+1)}.$$

□

Remark 9.3 (Why weighted averaging appears in the strongly convex case). For a general convex objective, Theorem 9.2 and Corollary 9.5 naturally lead to the equal-weight average when the stepsizes are constant. In the strongly convex setting, the proof instead controls a weighted sum

$$\sum_{t=1}^{T} \lambda_t (f(x_t) - f(x^\star)),$$

so Jensen naturally returns the weighted average $\widetilde{x}_T^{(\lambda)}$ rather than the uniform average $\bar{x}_T$. The linear choice $\lambda_t = t$ in Corollary 9.7 gives the $1/T$ rate. By contrast, equal weights $\lambda_t \equiv 1$ yield only

$$f(\widetilde{x}_T^{(\lambda)}) - f(x^\star) \le \frac{G^2}{2\alpha\mu T} \sum_{t=1}^{T} \frac{1}{t} = O\left(\frac{\log T}{T}\right).$$

So the faster rate comes from increasing weights, not from the standard equal-weight regret viewpoint.

9.5 Stochastic Mirror Descent

The pathwise inequality of Theorem 9.1 can be turned into a stochastic optimization theorem by taking expectations. The reduction itself is purely probabilistic: it does not depend on any special property of mirror descent. We state this reduction first, and then apply it to the pathwise mirror-descent estimates from Theorem 9.1 and Corollary 9.4.

Theorem 9.8 (General online-to-stochastic reduction)**.** *Let E be a finite-dimensional normed vector space. Let $X \subseteq E$ be nonempty, closed, and convex. Let $(\Xi, \mathcal{A}, \mathbb{P})$ be a probability space, let $f : X \times \Xi \to \mathbb{R}$ be measurable in ξ and convex in x, let $F : X \to \mathbb{R}$ be convex, and let $(\xi_t)_{t\geq 1}$ be i.i.d. samples from $\mathbb{P}$. For each $t \in \mathbb{N}$, let $x_t \in X$ be measurable with respect to $\sigma(\xi_1, \dots, \xi_{t-1})$, let $\eta_t > 0$, and let $\hat{g}_t \in \partial_x f(x_t, \xi_t)$. For $T \in \mathbb{N}$, define*

$$A_T := \sum_{t=1}^{T} \eta_t, \qquad \bar{x}_T^{(\eta)} := \frac{1}{A_T} \sum_{t=1}^{T} \eta_t x_t.$$

Assume moreover that, for every $u \in X$ and every $t \in \mathbb{N}$, the random variable $f(x_t, \xi_t) - f(u, \xi_t)$ is integrable and

$$\mathbb{E}\big[f(x_t, \xi_t) - f(u, \xi_t) \mid \xi_1, \dots, \xi_{t-1}\big] = F(x_t) - F(u).$$

Assume also that for every $u \in X$ one has the pathwise bound

$$\sum_{t=1}^{T} \eta_t \langle \hat{g}_t, x_t - u \rangle \leq \mathcal{R}_T(u)$$

for some integrable random variable $\mathcal{R}_T(u)$. Then, for every $u \in X$,

$$\mathbb{E}\big[F(\bar{x}_T^{(\eta)}) - F(u)\big] \leq \frac{1}{A_T} \mathbb{E}[\mathcal{R}_T(u)].$$

Proof of Theorem 9.8. By convexity of F and the definition of the weighted average,

$$F(\bar{x}_T^{(\eta)}) \leq \frac{1}{A_T} \sum_{t=1}^{T} \eta_t F(x_t).$$

Hence

$$A_T\big(F(\bar{x}_T^{(\eta)}) - F(u)\big) \leq \sum_{t=1}^{T} \eta_t \big(F(x_t) - F(u)\big).$$

Taking expectations in the previous display and using the assumed conditional expectation identity yields

$$A_T \mathbb{E}\big[F(\bar{x}_T^{(\eta)}) - F(u)\big] \leq \mathbb{E} \sum_{t=1}^{T} \eta_t \big(f(x_t, \xi_t) - f(u, \xi_t)\big).$$

Because $\hat{g}_t \in \partial_x f(x_t, \xi_t)$ and $f(\cdot, \xi_t)$ is convex,

$$f(x_t, \xi_t) - f(u, \xi_t) \leq \langle \hat{g}_t, x_t - u \rangle .$$

Therefore

$$A_T \mathbb{E}\big[F(\bar{x}_T^{(\eta)}) - F(u)\big] \leq \mathbb{E} \sum_{t=1}^{T} \eta_t \langle \hat{g}_t, x_t - u \rangle .$$

Now apply the assumed pathwise bound and divide by A_T. □

Combining Theorem 9.8 with the pathwise mirror-descent inequalities from Theorem 9.1 and Corollary 9.4 yields stochastic mirror-descent bounds. The corollary below records the standard nonsmooth bound under a bounded stochastic subgradient oracle.

Definition 9.3 (Unbiased stochastic subgradient oracle with bounded population subgradient and bounded noise)**.** Let E be a finite-dimensional normed vector space. Let $X \subseteq E$ be nonempty, closed, and convex. Let $F : X \to \mathbb{R}$ be a population risk. Let $(x_t)_{t\geq 1}$ be a sequence in X. Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration. A stochastic subgradient sequence $(\hat{g}_t)_{t\geq 1}$ is called unbiased with bounded population subgradient and bounded noise if there exist a predictable sequence of covectors $(g_t)_{t\geq 1}$ and constants $G \geq 0$ and $\sigma \geq 0$ such that, for every $t \in \mathbb{N}$,

$$\mathbb{E}[\hat{g}_t \mid \mathcal{F}_{t-1}] = g_t,$$

$$g_t \in \partial F(x_t), \qquad \|g_t\|_* \leq G,$$

and

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

Corollary 9.9 (A nonsmooth stochastic mirror bound under bounded population subgradients)**.** *Let E be a finite-dimensional normed vector space. Let $\Phi : E \to (-\infty, +\infty]$ be a mirror map, and let $X \subseteq \operatorname{dom}\Phi$ be nonempty, closed, and convex with $X \cap \operatorname{int}(\operatorname{dom}\Phi) \neq \emptyset$. Assume that Φ is α-strongly convex on $X \cap \operatorname{int}(\operatorname{dom}\Phi)$ with respect to a norm $\|\cdot\|$ for some $\alpha > 0$. Let $(\xi_t)_{t\geq 1}$ be i.i.d., let*

$$F(x) := \mathbb{E}[f(x, \xi_1)],$$

and let

$$x_1 \in X \cap \operatorname{int}(\operatorname{dom}\Phi).$$

For each $t \in \mathbb{N}$, let x_t be measurable with respect to $\sigma(\xi_1, \dots, \xi_{t-1})$, let $\eta_t \equiv \eta > 0$, let $\hat{g}_t \in \partial_x f(x_t, \xi_t)$, and let

$$x_{t+1} \in \operatorname*{arg\,min}_{x\in X}\Big\{\eta \langle \hat{g}_t, x - x_t\rangle + D_\Phi(x, x_t)\Big\}.$$

Assume moreover that, for every $u \in X$ and every $t \in \mathbb{N}$, the random variable $f(x_t, \xi_t) - f(u, \xi_t)$ is integrable and

$$\mathbb{E}[f(x_t, \xi_t) - f(u, \xi_t) \mid \xi_1, \dots, \xi_{t-1}] = F(x_t) - F(u).$$

Assume that $(\hat{g}_t)_{t\geq 1}$ is an unbiased stochastic subgradient oracle with bounded population subgradient and bounded noise in the sense of Definition 9.3. Then, for every $u \in X$ and every $T \in \mathbb{N}$,

$$\mathbb{E}[F(\bar{x}_T) - F(u)] \leq \frac{D_\Phi(u, x_1)}{\eta T} + \frac{\eta}{\alpha}(G^2 + \sigma^2), \qquad \bar{x}_T := \frac{1}{T}\sum_{t=1}^{T} x_t.$$

Consequently, if $x^\star \in \arg\min_X F$ and

$$\eta = \sqrt{\frac{\alpha D_\Phi(x^\star, x_1)}{T(G^2 + \sigma^2)}},$$

then

$$\mathbb{E}[F(\bar{x}_T) - F(x^\star)] \leq 2\sqrt{\frac{D_\Phi(x^\star, x_1)(G^2 + \sigma^2)}{\alpha T}}.$$

Proof of Corollary 9.9. By repeated application of Lemma 8.8, we have

$$x_t \in X \cap \operatorname{int}(\operatorname{dom}\Phi) \qquad \forall t \in \mathbb{N}.$$

By Corollary 9.4, for every $u \in X$,

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

Because each sample loss $f(\cdot, \xi_1)$ is convex, the population risk $F(x) = \mathbb{E}[f(x, \xi_1)]$ is also convex. Applying Theorem 9.8 with

$$\mathcal{R}_T(u) := D_\Phi(u, x_1) + \frac{\eta^2}{2\alpha} \sum_{t=1}^{T} \|\hat{g}_t\|_*^2$$

yields

$$\mathbb{E}[F(\bar{x}_T) - F(u)] \le \frac{D_\Phi(u, x_1)}{\eta T} + \frac{\eta}{2\alpha T} \sum_{t=1}^{T} \mathbb{E}\,\|\hat{g}_t\|_*^2 .$$

Fix $t \in \{1, \dots, T\}$, and let

$$g_t := \mathbb{E}[\hat{g}_t \mid \mathcal{F}_{t-1}] \in \partial F(x_t).$$

By the triangle inequality,

$$\|\hat{g}_t\|_* \le \|g_t\|_* + \|\hat{g}_t - g_t\|_* .$$

Squaring and using $(a+b)^2 \le 2a^2 + 2b^2$ gives

$$\|\hat{g}_t\|_*^2 \le 2\,\|g_t\|_*^2 + 2\,\|\hat{g}_t - g_t\|_*^2 .$$

Taking conditional expectations and using the assumptions,

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

Taking full expectations yields

$$\mathbb{E}\,\|\hat{g}_t\|_*^2 \le 2G^2 + 2\sigma^2.$$

Substituting into the previous display proves

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

The optimized choice of η balances the two terms. □

The nonsmooth theory above works with an arbitrary mirror map and arbitrary subgradients. The smooth stochastic case keeps the same constrained mirror step, but smoothness sharpens the local term: instead of paying for the full second moment of $\hat{g}_t$, one pays for the smooth descent part plus the pure noise. The result below is genuinely constrained: the comparator is only an optimizer over X, not a global minimizer over the ambient space.

Theorem 9.10 (Smooth stochastic mirror descent under relative smoothness)**.** *Let E be a finite-dimensional normed vector space. Let $\Phi : E \to (-\infty, +\infty]$ be a mirror map, and let*

$X \subseteq \operatorname{dom}\Phi$ be nonempty, closed, and convex with $X \cap \operatorname{int}(\operatorname{dom}\Phi) \neq \emptyset$. Assume that Φ is α-strongly convex on $X \cap \operatorname{int}(\operatorname{dom}\Phi)$ with respect to a norm $\|\cdot\|$ for some $\alpha > 0$. Let $F : X \to \mathbb{R}$ be convex, differentiable on $X \cap \operatorname{int}(\operatorname{dom}\Phi)$, and let $x^\star \in \arg\min_X F$. Assume that F is L-smooth relative to Φ on $X \cap \operatorname{int}(\operatorname{dom}\Phi)$, meaning that

$$\forall x \in X \cap \operatorname{int}(\operatorname{dom}\Phi),\ \forall y \in X, \qquad F(y) \le F(x) + \langle \nabla F(x), y - x\rangle + L D_\Phi(y, x). \tag{2}$$

Let $(\mathcal{F}_t)_{t\ge 0}$ be a filtration, let $x_1 \in X \cap \operatorname{int}(\operatorname{dom}\Phi)$, and for each $t \in \mathbb{N}$, assume that x_t is $\mathcal{F}_{t-1}$-measurable and that $\hat{g}_t \in E^*$ satisfies

$$\mathbb{E}[\hat{g}_t \mid \mathcal{F}_{t-1}] = \nabla F(x_t), \qquad \mathbb{E}\big[\|\hat{g}_t - \nabla F(x_t)\|_*^2 \mid \mathcal{F}_{t-1}\big] \le \sigma^2$$

for some $\sigma \ge 0$. Fix $\eta \in (0, 1/L)$, and for each $t \in \mathbb{N}$, define

$$x_{t+1} \in \arg\min_{x \in X} \Big\{ \eta \langle \hat{g}_t, x - x_t\rangle + D_\Phi(x, x_t) \Big\},$$

and for $T \in \mathbb{N}$, define the average of the post-update iterates

$$\bar{x}_T^+ := \frac{1}{T} \sum_{t=1}^{T} x_{t+1}.$$

Then

$$\mathbb{E}\big[F(\bar{x}_T^+) - F(x^\star)\big] \le \frac{D_\Phi(x^\star, x_1)}{\eta T} + \frac{\eta}{2\alpha(1 - \eta L)} \sigma^2. \tag{3}$$

In particular, if $D_\Phi(x^\star, x_1) \le R^2$ and $\sigma > 0$, then choosing

$$\eta = \min\left\{ \frac{1}{2L}, \frac{R\sqrt{\alpha}}{\sigma\sqrt{T}} \right\}$$

gives

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

Proof of Theorem 9.10. By repeated application of Lemma 8.8, each mirror step satisfies

$$x_{t+1} \in X \cap \operatorname{int}(\operatorname{dom}\Phi).$$

Hence

$$x_t \in X \cap \operatorname{int}(\operatorname{dom}\Phi) \qquad \forall t \in \{1, \dots, T+1\},$$

so Theorem 9.1 applies to the first T steps of the run. For each $t \in \{1, \dots, T\}$, write

$$\xi_t := \hat{g}_t - \nabla F(x_t), \qquad D_t := D_\Phi(x^\star, x_t), \qquad s_t := x_t - x_{t+1}.$$

Split the local term into its smooth part and its noise part:

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

By relative smoothness Equation (2),

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

so

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

Since $\eta < 1/L$, Young's inequality with parameter $\alpha(1 - \eta L) > 0$ gives

$$\eta \langle \xi_t, s_t \rangle \le \frac{\eta^2}{2\alpha(1 - \eta L)} \|\xi_t\|_*^2 + \frac{\alpha(1 - \eta L)}{2} \|s_t\|^2 .$$

Because Φ is α-strongly convex on $X \cap \operatorname{int}(\operatorname{dom} \Phi)$,

$$D_\Phi(x_{t+1}, x_t) \ge \frac{\alpha}{2} \|s_t\|^2 .$$

Hence

$$\frac{\alpha(1 - \eta L)}{2} \|s_t\|^2 \le (1 - \eta L) D_\Phi(x_{t+1}, x_t).$$

Combining the previous three displays yields

$$\eta \langle \hat{g}_t, s_t \rangle - D_\Phi(x_{t+1}, x_t) \le \eta(F(x_t) - F(x_{t+1})) + \frac{\eta^2}{2\alpha(1 - \eta L)} \|\xi_t\|_*^2 . \tag{5}$$

Applying Theorem 9.1 with $u = x^\star$, we first obtain

$$\sum_{t=1}^{T} \eta \langle \hat{g}_t, x_t - x^\star \rangle \le D_\Phi(x^\star, x_1) - D_\Phi(x^\star, x_{T+1}) + \sum_{t=1}^{T} \Big(\eta \langle \hat{g}_t, s_t \rangle - D_\Phi(x_{t+1}, x_t) \Big). \tag{6}$$

Using Equation (5) in Equation (6) and dropping the nonnegative term $D_\Phi(x^\star, x_{T+1})$, we obtain

$$\sum_{t=1}^{T} \eta \langle \hat{g}_t, x_t - x^\star \rangle \le D_\Phi(x^\star, x_1) + \eta \sum_{t=1}^{T} (F(x_t) - F(x_{t+1})) + \frac{\eta^2}{2\alpha(1 - \eta L)} \sum_{t=1}^{T} \|\xi_t\|_*^2 . \tag{7}$$

By convexity of F,

$$F(x_t) - F(x^\star) \le \langle \nabla F(x_t), x_t - x^\star \rangle = \langle \hat{g}_t, x_t - x^\star \rangle - \langle \xi_t, x_t - x^\star \rangle .$$

Multiply by η, sum over $t \in \{1, \dots, T\}$, and combine with Equation (7). The objective-difference terms telescope, so

$$\eta \sum_{t=1}^{T} (F(x_{t+1}) - F(x^\star)) \le D_\Phi(x^\star, x_1) + \frac{\eta^2}{2\alpha(1 - \eta L)} \sum_{t=1}^{T} \|\xi_t\|_*^2 - \eta \sum_{t=1}^{T} \langle \xi_t, x_t - x^\star \rangle . \tag{8}$$

Taking expectations, the last sum vanishes because x_t is $\mathcal{F}_{t-1}$-measurable and

$$\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0.$$

Using the conditional second-moment bound, we obtain

$$\eta \sum_{t=1}^{T} \mathbb{E}\big[F(x_{t+1}) - F(x^\star)\big] \le D_\Phi(x^\star, x_1) + \frac{\eta^2 T}{2\alpha(1 - \eta L)} \sigma^2. \tag{9}$$

Finally, convexity of F gives

$$F(\bar{x}_T^+) \le \frac{1}{T} \sum_{t=1}^{T} F(x_{t+1}).$$

Taking expectations and dividing by ηT proves the first claim.
For the second claim, if $\eta \le 1/(2L)$, then

$$1 - \eta L \ge \frac{1}{2},$$

so the first display implies

$$\mathbb{E}[F(\bar{x}_T^+) - F(x^\star)] \le \frac{R^2}{\eta T} + \frac{\eta}{\alpha}\sigma^2. \tag{10}$$

If $\eta = R\sqrt{\alpha}/(\sigma\sqrt{T})$, the right-hand side equals $2\sigma R/\sqrt{\alpha T}$. If instead $\eta = 1/(2L)$, then the condition

$$\frac{1}{2L} \le \frac{R\sqrt{\alpha}}{\sigma\sqrt{T}}$$

implies

$$\frac{\sigma^2}{\alpha} \le \frac{4L^2R^2}{T},$$

so

$$\frac{R^2}{\eta T} + \frac{\eta}{\alpha}\sigma^2 \le \frac{4LR^2}{T}.$$

Thus, in either case,

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

□

Dependency and proof sketch

1. Theorem 9.1 is the telescope of the constrained one-step mirror inequality from Lecture 8. This is the basic pathwise statement of the lecture.

2. An optional dual-Bregman reformulation of the local term is deferred to the exercises; it is not used in the main theorem spine.

3. Theorem 9.2 is the repeated-loss reading of the same pathwise inequality. Convexity plus Jensen turn weighted regret into an optimization theorem for the weighted average of the iterates.

4. Theorem 9.6 is the general weighted strongly convex template, and Corollary 9.7 is the linear-weight $1/T$ specialization. The faster rate comes from increasing weights rather than from the equal-weight regret regime.

5. Theorem 9.8 is the online-to-stochastic reduction itself: conditional expectation plus Jensen plus an arbitrary pathwise linearized-regret bound.

6. Corollary 9.9 is the standard nonsmooth stochastic mirror bound obtained by combining Theorem 9.8 with the norm-based pathwise estimate from Corollary 9.4.

7. Theorem 9.10 is the smooth stochastic closing theorem. Its proof reuses the constrained mirror-step inequality, but smoothness turns the deterministic part of the local term into descent and leaves only the pure noise term.

Exercises

1. Starting from Theorem 9.1, show that

$$\eta_t \langle g_t, x_t - x_{t+1} \rangle - D_\Phi(x_{t+1}, x_t) \leq D_{\Phi^*}(\nabla\Phi(x_t) - \eta_t g_t, \nabla\Phi(x_t))$$

   for every t, and deduce the corresponding dual-Bregman reformulation of the weighted linearized regret bound.

2. In Corollary 9.9, find the stepsize that minimizes the upper bound when the horizon T is unknown in advance and explain what goes wrong.

3. Specialize Theorem 9.10 to the Euclidean geometry $H = I_d$. Compare the result to the deterministic Euclidean theorem of Lecture 7.