Lecture 8: Mirror Descent and Bregman Geometry

Lecture 7 used a fixed norm, and in the smooth case this meant a fixed quadratic upper model. The first point of Lecture 8 is to go beyond fixed-norm geometry: for some important open-domain objectives, the smoothness constant with respect to any fixed norm is unbounded, so the local model should be allowed to depend on the current point x_t. The replacement upper model is relative smoothness, where the quadratic penalty is replaced by a Bregman divergence D_Φ. Mirror descent is then obtained in exactly the same way as gradient descent in Lecture 7: minimize the local upper model. The later mirror-map assumptions are algorithmic assumptions; they make this Bregman model minimization well posed and give the clean dual-coordinate update.

8.1 Going Beyond Fixed-Norm Geometry

Lecture 7 measured local curvature using a fixed norm. In its smooth form, this means a global quadratic upper model of the form

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

with one finite constant L. This is the right model when the curvature of f is uniformly bounded in the chosen norm. Its limitation is exposed by open-domain objectives whose curvature becomes unbounded near the boundary. These objectives are not pathological; they are exactly the kind of barrier-like or entropy-like functions one wants to optimize over constrained domains.

Example 8.1 (A log-barrier objective that breaks fixed-norm smoothness)**.** Let

$$\Phi(x) = -\log x - \log(1 - x),$$

the standard log-barrier on the open interval $(0, 1)$, extended to $\mathbb{R}$ by setting $\Phi(x) = +\infty$ outside $(0, 1)$. Define, for $c \in \mathbb{R}$,

$$f(x) = \Phi(x) + cx \qquad \text{for } x \in (0, 1).$$

Its curvature is

$$\Phi''(x) = \frac{1}{x^2} + \frac{1}{(1-x)^2}, \qquad f''(x) = \Phi''(x).$$

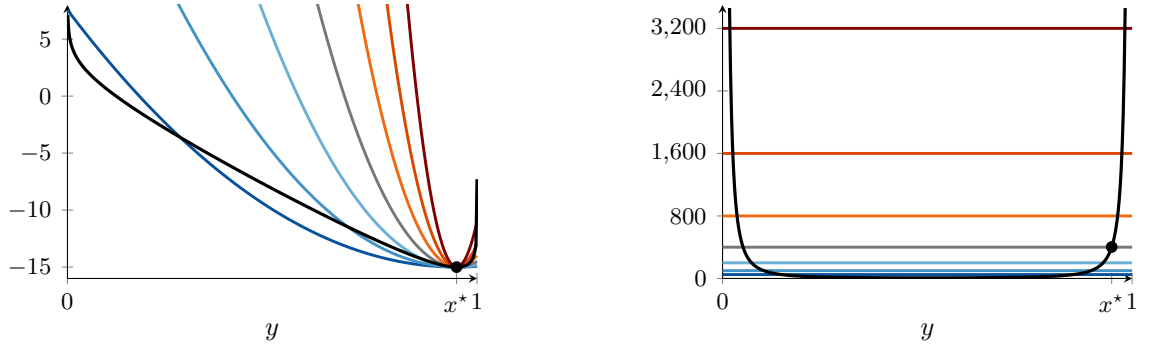
Thus $f''(x) \to +\infty$ as $x \downarrow 0$ or $x \uparrow 1$, and $f(x)$ itself diverges to $+\infty$ at both boundaries. Consequently, f is not globally smooth with respect to any fixed norm on $\mathbb{R}$. Indeed, every norm on $\mathbb{R}$ has the form

$$\|s\| = \gamma |s| \qquad \text{for some } \gamma > 0.$$

If the Lecture 7 smoothness inequality held with a finite constant L, then the one-dimensional second-order expansion would imply

$$f''(x) \le L\gamma^2 \qquad \forall x \in (0, 1),$$

contradicting the boundary blow-up. So this example exhibits exactly the pathology that forces us to move beyond fixed quadratic geometry. We now introduce the replacement geometry and then return to this example.



(a) $f(y)$ (black) vs. Q_L at $L = 50 \cdot 2^k$, $k = 0, \dots, 6$. (b) $\Phi''(y)$ (black) vs. constants $L = 50 \cdot 2^k$, $k = 0, \dots, 6$.

Figure 1: With $c = -19$ in Example 8.1, $x^\star \approx 0.95$ sits near the right boundary and $f''(x^\star) \approx 403$. Seven quadratic candidates $Q_L(y) = f(x^\star) + \frac{L}{2}(y - x^\star)^2$ at $L = 50 \cdot 2^k$, $k = 0, \dots, 6$, cool to warm: (a) no L dominates f everywhere; (b) Φ'' blows up at both ends, so no constant $L \geq \Phi''$ works.

The general move is to replace the fixed quadratic penalty by a current-point dependent penalty

$$x_{t+1} \in \arg\min_x \{\eta_t \langle g_t, x - x_t \rangle + V_{x_t}(x)\}.$$

The penalty V_{x_t} can become steeper near the boundary or flatter in directions where the objective has low curvature. In this lecture we study the most important structured special case:

$$V_{x_t}(x) = D_\Phi(x, x_t).$$

The symbol D_Φ denotes the following Bregman divergence.

Definition 8.1 (Bregman divergence)**.** Let $\Phi : E \to (-\infty, +\infty]$. For $x, y \in \operatorname{dom} \Phi$, if Φ is differentiable at y, define

$$D_\Phi(x, y) := \Phi(x) - \Phi(y) - \langle \nabla\Phi(y), x - y \rangle.$$

Remark 8.1 (How to read D_Φ). The quantity $D_\Phi(x, y)$ is not a metric: it is usually asymmetric and does not satisfy the triangle inequality. Its role is instead to measure the gap between $\Phi(x)$ and the first-order affine approximation to Φ at y. Convexity gives $D_\Phi(x, y) \geq 0$, and strict convexity makes this gap positive unless $x = y$ in the interior. Thus $D_\Phi(\cdot, x_t)$ is a geometry penalty, not a distance in the metric-space sense.

Relative smoothness is the upper-model condition that makes this Bregman penalty useful for a single smooth objective. It is the direct analogue of Euclidean smoothness

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2,$$

with the quadratic penalty replaced by $L D_\Phi(y, x)$.

Definition 8.2 (Relative smoothness and relative strong convexity)**.** Let $\Phi : E \to (-\infty, +\infty]$ be a proper convex function, and assume that Φ is differentiable on $\operatorname{int}(\operatorname{dom}\Phi)$. Let $f : \operatorname{dom}\Phi \to \mathbb{R}$ be differentiable on $\operatorname{int}(\operatorname{dom}\Phi)$. We say:

- f is L-smooth relative to Φ, where $L > 0$, if, for all $x \in \operatorname{int}(\operatorname{dom}\Phi)$ and $y \in \operatorname{dom}\Phi$,
$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + LD_\Phi(y, x).$$
- f is μ-strongly convex relative to Φ, where $\mu > 0$, if, for all $x \in \operatorname{int}(\operatorname{dom}\Phi)$ and $y \in \operatorname{dom}\Phi$,
$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \mu D_\Phi(y, x).$$

Remark 8.2 (Historical note). The modern relative-smoothness viewpoint was developed in Bauschke–Bolte–Teboulle [BBT17] and Lu–Freund–Nesterov [LFN18]. The former emphasizes the convexity comparison behind a Bregman descent lemma, while the latter packages the same geometry as relative smoothness and relative strong convexity for first-order convex optimization.

Lemma 8.1 (Bregman reformulations)**.** *In the setup of Definition 8.2:*

- *f is L-smooth relative to Φ if and only if $D_{L\Phi - f}(y, x) \ge 0$ for all $x \in \operatorname{int}(\operatorname{dom}\Phi)$ and $y \in \operatorname{dom}\Phi$. In particular, $L\Phi - f$ is convex on $\operatorname{int}(\operatorname{dom}\Phi)$.*
- *f is μ-strongly convex relative to Φ if and only if $D_{f-\mu\Phi}(y, x) \ge 0$ for all $x \in \operatorname{int}(\operatorname{dom}\Phi)$ and $y \in \operatorname{dom}\Phi$. In particular, $f - \mu\Phi$ is convex on $\operatorname{int}(\operatorname{dom}\Phi)$.*

Proof of Lemma 8.1. For the first item, for $x \in \operatorname{int}(\operatorname{dom}\Phi)$ and $y \in \operatorname{dom}\Phi$,
$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + LD_\Phi(y, x) \iff D_{L\Phi - f}(y, x) \ge 0.$$
Restricting to $y \in \operatorname{int}(\operatorname{dom}\Phi)$ yields
$$(L\Phi - f)(y) \ge (L\Phi - f)(x) + \langle \nabla(L\Phi - f)(x), y - x\rangle,$$
so $L\Phi - f$ is convex on $\operatorname{int}(\operatorname{dom}\Phi)$.
For the second item, again for $x \in \operatorname{int}(\operatorname{dom}\Phi)$ and $y \in \operatorname{dom}\Phi$,
$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \mu D_\Phi(y, x) \iff D_{f-\mu\Phi}(y, x) \ge 0.$$
Restricting to $y \in \operatorname{int}(\operatorname{dom}\Phi)$ yields
$$(f - \mu\Phi)(y) \ge (f - \mu\Phi)(x) + \langle \nabla(f - \mu\Phi)(x), y - x\rangle,$$
so $f - \mu\Phi$ is convex on $\operatorname{int}(\operatorname{dom}\Phi)$. □

These two equivalences are often the fastest way to verify the relative-geometry constants. The first item is the upper-model test used directly in the algorithmic proof; the second is the matching lower-model test used later for linear convergence.

8.2 Mirror Maps and the Mirror Descent Algorithm

Once Definition 8.2 is in place, the algorithm is motivated exactly as in Lecture 7. At x_t, relative smoothness gives the upper model

$$f(z) \le f(x_t) + \langle \nabla f(x_t), z - x_t \rangle + L D_\Phi(z, x_t).$$

Ignoring the constant $f(x_t)$, the natural step is to minimize a linearized objective plus a Bregman penalty:

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

When $g_t = \nabla f(x_t)$ and $\eta_t = 1/L$, this is exactly minimization of the relative-smooth upper model; other positive η_t's are the usual stepsize variants. The remaining question is not why this is the right model; it is when this Bregman subproblem is well posed and easy to implement. That is the role of the mirror-map assumptions.

Definition 8.3 (Mirror map)**.** Let E be a finite-dimensional normed vector space. We say that a function

$$\Phi : E \to (-\infty, +\infty]$$

is a mirror map if the following four properties hold:

(H1) Φ is proper, closed, and strictly convex.

(H2) $\operatorname{int}(\operatorname{dom}\Phi) \neq \emptyset$.

(H3) Φ is differentiable on $\operatorname{int}(\operatorname{dom}\Phi)$.

(H4) $\nabla\Phi\big(\operatorname{int}(\operatorname{dom}\Phi)\big) = E^*$.

The two structural assumptions to keep in mind are strict convexity and the range condition (H4). Strict convexity makes the Bregman subproblem have at most one minimizer. The range condition $\nabla\Phi(\operatorname{int}(\operatorname{dom}\Phi)) = E^*$ says that after the dual-coordinate move $\nabla\Phi(x) - \eta g$, there is still a primal point in $\operatorname{int}(\operatorname{dom}\Phi)$ with that mirror coordinate. Thus strict convexity gives uniqueness, while (H4) gives global existence for unconstrained dual updates.

The point of starting here is that the mirror step is just a change of coordinates: move linearly in the dual coordinate $\nabla\Phi(x)$, then map back to the primal domain via the inverse of $\nabla\Phi$.

Definition 8.4 (Unconstrained mirror step)**.** Let $\Phi : E \to (-\infty, +\infty]$ be a mirror map, let $x \in \operatorname{int}(\operatorname{dom}\Phi)$, let $g \in E^*$, and let $\eta > 0$. Any solution of the global minimization problem

$$x^+ \in \arg\min_{z \in E} \Big\{ \Phi(z) - \langle \nabla\Phi(x) - \eta g, z \rangle \Big\}$$

is called an unconstrained mirror step. Equivalently, because $\Phi(z) = +\infty$ off $\operatorname{dom}\Phi$, this is the same as

$$x^+ \in \arg\min_{z \in \operatorname{dom}\Phi} \Big\{ \eta \langle g, z - x \rangle + D_\Phi(z, x) \Big\}.$$

In words, the next point minimizes the linearized loss plus the current Bregman geometry penalty.

Proposition 8.2 (Unconstrained dual-coordinate implementation)**.** *Let $\Phi : E \to (-\infty, +\infty]$ be a mirror map, let $x_t \in \operatorname{int}(\operatorname{dom}\Phi)$, let $g_t \in E^*$, and let $\eta_t > 0$. Then $\nabla\Phi : \operatorname{int}(\operatorname{dom}\Phi) \to E^*$ is bijective. Define*

$$x_{t+1} := (\nabla\Phi)^{-1}(\nabla\Phi(x_t) - \eta_t g_t).$$

Then $x_{t+1} \in \operatorname{int}(\operatorname{dom}\Phi)$, and x_{t+1} is the unique solution of the global minimization problem

$$\arg\min_{x\in E}\Big\{\Phi(x) - \langle \nabla\Phi(x_t) - \eta_t g_t, x\rangle\Big\}.$$

Proof of Proposition 8.2. Suppose $y_1, y_2 \in \operatorname{int}(\operatorname{dom}\Phi)$ satisfy $\nabla\Phi(y_1) = \nabla\Phi(y_2)$. First-order convexity of Φ at y_1 and at y_2 gives

$$\Phi(y_2) \geq \Phi(y_1) + \langle \nabla\Phi(y_1), y_2 - y_1\rangle\,,$$

$$\Phi(y_1) \geq \Phi(y_2) + \langle \nabla\Phi(y_2), y_1 - y_2\rangle\,.$$

Adding the two inequalities and using $\nabla\Phi(y_1) = \nabla\Phi(y_2)$ shows that both hold with equality. Strict convexity therefore forces $y_1 = y_2$. Thus $\nabla\Phi$ is injective on $\operatorname{int}(\operatorname{dom}\Phi)$. Together with (H4), this shows that $\nabla\Phi : \operatorname{int}(\operatorname{dom}\Phi) \to E^*$ is bijective. Hence

$$x_{t+1} := (\nabla\Phi)^{-1}\big(\nabla\Phi(x_t) - \eta_t g_t\big)$$

is well-defined and belongs to $\operatorname{int}(\operatorname{dom}\Phi)$. Since Φ is differentiable at x_{t+1}, Lemma 4.1 gives

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

Equivalently, x_{t+1} maximizes $\langle \nabla\Phi(x_t) - \eta_t g_t, z\rangle - \Phi(z)$, or minimizes

$$\Phi(z) - \langle \nabla\Phi(x_t) - \eta_t g_t, z\rangle$$

over $z \in E$. Because Φ is strictly convex by (H1), this minimizer is unique. Thus x_{t+1} is exactly the unique solution of the displayed global problem. □

Thus unconstrained mirror descent is the following recursion.

Algorithm 1 Unconstrained mirror descent with mirror map Φ

Require: A mirror map $\Phi : E \to (-\infty, +\infty]$, an initial point $x_1 \in \operatorname{int}(\operatorname{dom}\Phi)$, covectors $g_t \in E^*$, and stepsizes $\eta_t > 0$.

Ensure: A sequence of mirror-descent iterates $(x_t)_{t\geq 1}$.

1: Choose an initial point $x_1 \in \operatorname{int}(\operatorname{dom}\Phi)$.
2: **for** $t = 1, 2, \ldots$ **do**
3: Select a covector $g_t \in E^*$ and a stepsize $\eta_t > 0$.
4: Set

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

5: **end for**

Remark 8.3 (Returning to Example 8.1). For Example 8.1, this is mirror descent with respect to Φ. Because $f - \Phi = cx$ is affine, f is both 1-smooth and 1-strongly convex relative to Φ. Equivalently,

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + D_\Phi(y, x) \qquad \forall\, x, y \in (0, 1).$$

So the relative-smoothness upper model is exact, not merely an upper bound. With $\eta = 1$, one iteration reaches the minimizer.

By Proposition 8.2, line 4 of the algorithm can be implemented equivalently as

$$x_{t+1} = (\nabla\Phi)^{-1}(\nabla\Phi(x_t) - \eta_t g_t),$$

or, equivalently,

$$\nabla\Phi(x_{t+1}) = \nabla\Phi(x_t) - \eta_t g_t.$$

This is why mirror descent is often described as a linear update in mirror coordinates.

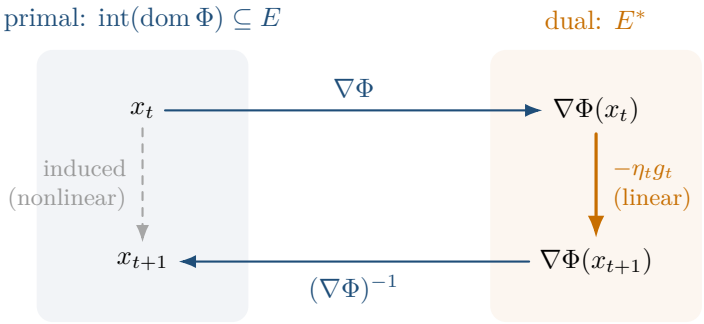


Figure 2: Mirror descent as a commutative square between primal and dual coordinates. The map $\nabla\Phi$ sends the primal iterate $x_t \in \operatorname{int}(\operatorname{dom}\Phi)$ to its dual coordinate $\nabla\Phi(x_t) \in E^*$; the update itself is a *linear* shift $\nabla\Phi(x_{t+1}) = \nabla\Phi(x_t) - \eta_t g_t$ in the dual (thick right arrow); the inverse $(\nabla\Phi)^{-1}$ returns to the primal, giving x_{t+1}. The induced primal trajectory (dashed) is generally nonlinear. Condition (H4) ($\nabla\Phi$ surjective onto E^*) guarantees that the dual shift stays inside the coordinate system.

The next examples should be read as a dictionary for the mirror update. For each geometry, the important data are the mirror map Φ, the resulting Bregman divergence D_Φ, and the closed-form update. The regret and convergence constants come later from the general theorems; here the point is simply to see what the algorithm looks like in familiar geometries.

Example 8.2 (Unconstrained Euclidean geometry)**.** Take $E = \mathbb{R}^n$, let $X = E$, and let

$$\Phi(x) = \frac{1}{2}\|x\|_2^2.$$

Then

$$D_\Phi(x, y) = \frac{1}{2}\|x - y\|_2^2,$$

and the mirror update can be written either as

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

or, equivalently, as

$$x_{t+1} = (\nabla\Phi)^{-1}(\nabla\Phi(x_t) - \eta_t g_t) = x_t - \eta_t g_t.$$

Indeed,

$$\eta_t \langle g_t, x - x_t\rangle + D_\Phi(x, x_t) = \eta_t \langle g_t, x - x_t\rangle + \frac{1}{2}\|x - x_t\|_2^2 = \frac{1}{2}\|x - (x_t - \eta_t g_t)\|_2^2 - \frac{1}{2}\eta_t^2 \|g_t\|_2^2,$$

so minimizing over $\mathbb{R}^n$ gives $x_t - \eta_t g_t$. Since $\nabla\Phi(x) = x$, this is exactly the inverse-gradient formula above.

Remark 8.4 (Affine-slice examples in centered linear coordinates). The main mirror-descent results are stated on a linear space E. For an affine-slice example, the clean way to apply them is to identify the slice with a centered linear model. Concretely, if

$$A = x_c + E^\circ,$$

then every

$$x \in A \qquad \text{corresponds to} \qquad x' := x - x_c \in E^\circ.$$

A function Φ' on A is transported to a function

$$\Phi(x') := \Phi'(x_c + x') \qquad \text{for } x' \in E^\circ.$$

Under this translation, the mirror-map hypotheses are read intrinsically on E°, the Bregman divergence is unchanged, and the linear pairing is unchanged as well: if $x = x_c + x'$ and $y = x_c + y'$, then

$$D_\Phi(x', y') = D_{\Phi'}(x, y),$$

and if an ambient covector g_t is restricted to the slice E°, then for ambient points $x_t = x_c + x_t'$ and $u = x_c + u'$,

$$\langle g_t, x_t' - u'\rangle = \langle g_t, x_t - u\rangle.$$

So the assumptions and conclusions transfer with the same constants; only the ambient formula for the update needs to be rewritten at the end. For the simplex, the natural centered model is $\Delta_n - \frac{1}{n}\mathbf{1} \subseteq \{x' \in \mathbb{R}^n : \sum_i x_i' = 0\}$. For the spectrahedron, the natural centered model is $\mathcal{S}_n - \frac{1}{n}I_n \subseteq \left\{x' \in \mathbb{R}^{n\times n} : x' = x'^\top,\ \operatorname{tr}(x') = 0\right\}$.

Example 8.3 (Negative entropy on the simplex)**.** Let

$$\Delta_n := \left\{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\right\}.$$

The centered linear model used to apply the main results is

$$E^\circ := \left\{x' \in \mathbb{R}^n : \sum_{i=1}^n x_i' = 0\right\}, \qquad X^\circ := \Delta_n - \frac{1}{n}\mathbf{1} \subseteq E^\circ.$$

Write the ambient entropy as

$$\Phi'(x) = \sum_{i=1}^{n} x_i \log x_i \qquad \text{for } x \in \Delta_n,$$

extended by $+\infty$ off Δ_n. Define the mirror map on the centered model by

$$\Phi : E^\circ \to (-\infty, +\infty], \qquad \Phi(x') := \Phi'\left(\frac{1}{n}\mathbf{1} + x'\right).$$

Then Φ is indeed a mirror map on E°: (H1)–(H3) are inherited from the ambient negative entropy on $\Delta_n \cap (0,\infty)^n$, and (H4) can be checked explicitly. Indeed, if $x = \frac{1}{n}\mathbf{1} + x' \in \Delta_n \cap (0,\infty)^n$, then for every $v' \in E^\circ$ we have $\sum_i v'_i = 0$, so

$$D\Phi(x')[v'] = \sum_{i=1}^{n} (\log x_i + 1) v'_i = \sum_{i=1}^{n} (\log x_i) v'_i.$$

Thus $\nabla\Phi(x')$ is the restriction of the ambient covector $(\log x_1, \dots, \log x_n)$ to E°. Conversely, let $g \in (E^\circ)^*$, choose any ambient representative $\widetilde{g} \in \mathbb{R}^n$, and set

$$x_i := \frac{e^{\widetilde{g}_i}}{\sum_{j=1}^{n} e^{\widetilde{g}_j}}, \qquad x' := x - \frac{1}{n}\mathbf{1}.$$

Then $x \in \Delta_n \cap (0,\infty)^n$, and for every $v' \in E^\circ$,

$$\langle \nabla\Phi(x'), v' \rangle = \sum_{i=1}^{n} (\log x_i) v'_i = \sum_{i=1}^{n} \widetilde{g}_i v'_i = \langle g, v' \rangle .$$

Hence $\nabla\Phi(\operatorname{int}(\operatorname{dom}\Phi)) = (E^\circ)^*$. Therefore $\operatorname{dom}\Phi = X^\circ$, while

$$\operatorname{int}(\operatorname{dom}\Phi) = \left\{ x' \in E^\circ : \frac{1}{n}\mathbf{1} + x' \in \Delta_n \cap (0,\infty)^n \right\}.$$

For $x', y' \in \operatorname{int}(\operatorname{dom}\Phi)$ and the corresponding ambient points

$$x := \frac{1}{n}\mathbf{1} + x', \qquad y := \frac{1}{n}\mathbf{1} + y' \in \Delta_n \cap (0,\infty)^n,$$

$$D_\Phi(x', y') = D_{\Phi'}(x, y) = \sum_{i=1}^{n} x_i \log\left(\frac{x_i}{y_i}\right) =: \mathrm{KL}(x \| y),$$

and, if $x_t \in \Delta_n \cap (0,\infty)^n$ and $x'_t := x_t - \frac{1}{n}\mathbf{1}$, the centered mirror update can be written either as

$$x'_{t+1} \in \operatorname*{arg\,min}_{x' \in X^\circ} \left\{ \eta_t \langle g_t, x' - x'_t \rangle + D_\Phi(x', x'_t) \right\}$$

or, equivalently, as

$$x'_{t+1} = (\nabla\Phi)^{-1}(\nabla\Phi(x'_t) - \eta_t g_t).$$

Written back in ambient simplex coordinates, this is multiplicative weights:

$$\forall i \in \{1, \dots, n\}, \qquad x_{t+1,i} = \frac{x_{t,i} e^{-\eta_t g_{t,i}}}{\sum_{j=1}^{n} x_{t,j} e^{-\eta_t g_{t,j}}}.$$

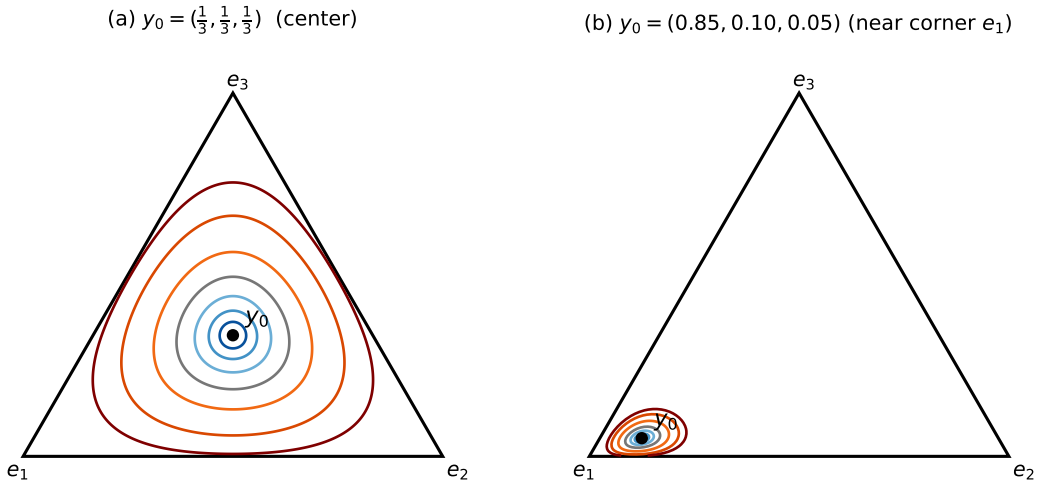


Figure 3: Level sets of $D_\Phi(x, y_0) = \mathrm{KL}(x\|y_0)$ on Δ_3, seven radii per panel (cool = small, warm = large). (a) y_0 at the center: 3-fold symmetric. (b) y_0 near e_1: Fisher diag$(1.18, 10, 20)$ elongates level sets along e_1–e_2 and pinches along e_3. D_Φ is *bounded* on Δ_n (maximum $-\log y_{0,i}$ at vertex e_i); interior preservation of the mirror step comes from $\nabla\Phi = \log$ sending $\partial\Delta_n \to -\infty$ in the dual, not from any boundary blow-up of D_Φ.

Example 8.4 (Matrix entropy on the spectrahedron)**.** Let

$$\mathcal{S}_n := \left\{ X \in \mathbb{R}^{n\times n} : X = X^\top,\ X \succeq 0,\ \mathrm{tr}(X) = 1 \right\}.$$

The centered linear model used to apply the main results is

$$E^\circ := \left\{ X' \in \mathbb{R}^{n\times n} : X' = X'^\top,\ \mathrm{tr}(X') = 0 \right\}, \qquad \mathcal{S}_n^\circ := \mathcal{S}_n - \frac{1}{n} I_n \subseteq E^\circ.$$

Write the ambient entropy as

$$\Phi'(X) = \mathrm{tr}(X \log X) \qquad \text{for } X \in \mathcal{S}_n,$$

extended by $+\infty$ off $\mathcal{S}_n$. Define the mirror map on the centered model by

$$\Phi : E^\circ \to (-\infty, +\infty], \qquad \Phi(X') := \Phi'\left(\frac{1}{n} I_n + X'\right).$$

Then Φ is indeed a mirror map on E°: (H1)–(H3) are inherited from the ambient matrix entropy on the positive-definite part of $\mathcal{S}_n$, and (H4) can again be checked explicitly. Indeed, if $X = \frac{1}{n} I_n + X' \in \mathcal{S}_n$ satisfies $X \succ 0$, then for every $V' \in E^\circ$ we have $\mathrm{tr}(V') = 0$, so

$$D\Phi(X')[V'] = \mathrm{tr}\big((\log X + I_n)V'\big) = \mathrm{tr}\big((\log X)V'\big).$$

Thus $\nabla\Phi(X')$ is the restriction of the ambient covector $\log X$ to E°. Conversely, let $G \in (E^\circ)^*$, choose any symmetric ambient representative $\widetilde{G} \in \mathbb{R}^{n\times n}$, and set

$$X := \frac{\exp(\widetilde{G})}{\mathrm{tr}(\exp(\widetilde{G}))}, \qquad X' := X - \frac{1}{n} I_n.$$

Then $X \in \mathcal{S}_n$ and $X \succ 0$, while

$$\log X = \widetilde{G} - \log\big(\operatorname{tr}(\exp(\widetilde{G}))\big) I_n.$$

Hence for every $V' \in E^\circ$,

$$\langle \nabla\Phi(X'), V' \rangle = \operatorname{tr}\big((\log X)V'\big) = \operatorname{tr}(\widetilde{G}V') = \langle G, V' \rangle .$$

Hence $\nabla\Phi(\operatorname{int}(\operatorname{dom}\Phi)) = (E^\circ)^*$. Therefore $\operatorname{dom}\Phi = \mathcal{S}_n^\circ$, while

$$\operatorname{int}(\operatorname{dom}\Phi) = \left\{ X' \in E^\circ : \frac{1}{n} I_n + X' \in \mathcal{S}_n,\ \frac{1}{n} I_n + X' \succ 0 \right\}.$$

For $X', Y' \in \operatorname{int}(\operatorname{dom}\Phi)$ and the corresponding ambient points

$$X := \frac{1}{n} I_n + X', \qquad Y := \frac{1}{n} I_n + Y' \in \mathcal{S}_n \quad \text{with } Y \succ 0,$$

$$D_\Phi(X', Y') = D_{\Phi'}(X, Y) = \operatorname{tr}\big(X(\log X - \log Y)\big),$$

and, if $X_t \in \mathcal{S}_n$ satisfies $X_t \succ 0$ and $X_t' := X_t - \frac{1}{n} I_n$, the centered mirror update can be written either as

$$X_{t+1}' \in \operatorname*{arg\,min}_{X' \in \mathcal{S}_n^\circ} \Big\{ \eta_t \operatorname{tr}\big(G_t(X' - X_t')\big) + D_\Phi(X', X_t') \Big\}$$

or, equivalently, as

$$X_{t+1}' = (\nabla\Phi)^{-1}(\nabla\Phi(X_t') - \eta_t G_t).$$

Written back in ambient matrix coordinates, this is matrix multiplicative weights:

$$X_{t+1} = \frac{\exp\big(\log X_t - \eta_t G_t\big)}{\operatorname{tr}\big(\exp\big(\log X_t - \eta_t G_t\big)\big)}.$$

8.3 The Three-Point Identity and the Unconstrained One-Step Equality

Lemma 8.3 (Three-point identity)**.** *Let $x, y \in \operatorname{int}(\operatorname{dom}\Phi)$ and let $z \in \operatorname{dom}\Phi$. Then*

$$\langle \nabla\Phi(x) - \nabla\Phi(y), z - x \rangle = D_\Phi(z, y) - D_\Phi(z, x) - D_\Phi(x, y).$$

Proof of Lemma 8.3. By Definition 8.1,

$$\begin{aligned} &D_\Phi(z, y) - D_\Phi(z, x) - D_\Phi(x, y) \\ &= \Phi(z) - \Phi(y) - \langle \nabla\Phi(y), z - y \rangle - \Phi(z) + \Phi(x) + \langle \nabla\Phi(x), z - x \rangle - \Phi(x) + \Phi(y) + \langle \nabla\Phi(y), x - y \rangle \\ &= \langle \nabla\Phi(x), z - x \rangle - \langle \nabla\Phi(y), z - x \rangle = \langle \nabla\Phi(x) - \nabla\Phi(y), z - x \rangle . \end{aligned}$$

□

Theorem 8.4 (Unconstrained one-step mirror descent equality)**.** *Let $\Phi : E \to (-\infty, +\infty]$ be a mirror map. Fix $t \in \mathbb{N}$, let $x_t \in \operatorname{int}(\operatorname{dom}\Phi)$, let $g_t \in E^*$, let $\eta_t > 0$, and let*

$$\nabla\Phi(x_{t+1}) = \nabla\Phi(x_t) - \eta_t g_t.$$

Then

$$\forall u \in \operatorname{dom}\Phi, \qquad \eta_t \langle g_t, x_t - u\rangle = D_\Phi(u, x_t) - D_\Phi(u, x_{t+1}) + \eta_t \langle g_t, x_t - x_{t+1}\rangle - D_\Phi(x_{t+1}, x_t).$$

Proof of Theorem 8.4. The update gives

$$\eta_t g_t + \nabla\Phi(x_{t+1}) - \nabla\Phi(x_t) = 0.$$

Therefore

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

Now apply Lemma 8.3 with $(x, y, z) = (x_{t+1}, x_t, u)$ to obtain

$$\langle \nabla\Phi(x_{t+1}) - \nabla\Phi(x_t), u - x_{t+1}\rangle = D_\Phi(u, x_t) - D_\Phi(u, x_{t+1}) - D_\Phi(x_{t+1}, x_t).$$

Substituting proves the claim. □

This is the master equality of the lecture. Relative smoothness will turn it into a last-iterate descent theorem. When constraints are added later, the equality becomes an inequality because first-order optimality becomes a variational inequality.

8.4 Convergence of Mirror Descent with Relative Smoothness

We now combine the one-step Bregman identity with the relative-smoothness upper model from Definition 8.2. The proof is the same pattern as Lecture 7: the algorithm minimizes the local model, and the upper model turns that local decrease into a decrease for the actual objective.

Remark 8.5 (Rate summary). Before proving them, here are the two basic rates for the unconstrained mirror-descent update with $\eta_t \equiv 1/L$:

- if f is convex and L-smooth relative to Φ, then

$$f(x_{T+1}) - f(x^\star) \le \frac{L D_\Phi(x^\star, x_1)}{T};$$

- if f is also μ-strongly convex relative to Φ, then

$$D_\Phi(x^\star, x_{T+1}) \le \left(1 - \frac{\mu}{L}\right)^T D_\Phi(x^\star, x_1),$$

  and hence

$$f(x_{T+1}) - f(x^\star) \le L\left(1 - \frac{\mu}{L}\right)^T D_\Phi(x^\star, x_1).$$

The rest of this subsection proves these two statements.

Proposition 8.5 (Relative smoothness gives descent). *Let $\Phi : E \to (-\infty, +\infty]$ be a mirror map, let $f : \operatorname{dom}\Phi \to \mathbb{R}$ be L-smooth relative to Φ. Let $x_t \in \operatorname{int}(\operatorname{dom}\Phi)$, let $g_t = \nabla f(x_t)$, and*

let

$$\nabla\Phi(x_{t+1}) = \nabla\Phi(x_t) - \eta_t \nabla f(x_t).$$

If $0 < \eta_t \le 1/L$, *then*

$$f(x_{t+1}) \le f(x_t).$$

Proof of Proposition 8.5. By Proposition 8.2, x_{t+1} minimizes the unconstrained mirror subproblem. Comparing its value with the value at x_t gives

$$\eta_t \langle \nabla f(x_t), x_{t+1} - x_t \rangle + D_\Phi(x_{t+1}, x_t) \le 0.$$

By Definition 8.2, applied with $x = x_t$ and $y = x_{t+1}$,

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

Since $0 < \eta_t \le 1/L$, we have $L \le 1/\eta_t$. Therefore

$$\begin{aligned} f(x_{t+1}) &\le f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{1}{\eta_t} D_\Phi(x_{t+1}, x_t) \\ &= f(x_t) + \frac{1}{\eta_t}\Big(\eta_t \langle \nabla f(x_t), x_{t+1} - x_t \rangle + D_\Phi(x_{t+1}, x_t)\Big) \le f(x_t). \end{aligned}$$

□

Theorem 8.6 (Last-iterate rate under relative smoothness)**.** *Let* $\Phi : E \to (-\infty, +\infty]$ *be a mirror map, let* $f : \operatorname{dom}\Phi \to \mathbb{R}$ *be convex and* L*-smooth relative to* Φ*. Assume that*

$$x^\star \in \operatorname*{arg\,min}_{x \in \operatorname{dom}\Phi} f(x)$$

exists. Run unconstrained mirror descent with $g_t = \nabla f(x_t)$ *and stepsizes* $0 < \eta_t \le 1/L$*. Then*

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

Consequently, for every $T \ge 1$,

$$f(x_{T+1}) - f(x^\star) \le \frac{D_\Phi(x^\star, x_1)}{\sum_{t=1}^T \eta_t}.$$

In particular, if $\eta_t \equiv 1/L$, *then*

$$f(x_{T+1}) - f(x^\star) \le \frac{L D_\Phi(x^\star, x_1)}{T}.$$

Proof of Theorem 8.6. Apply Theorem 8.4 with $u = x^\star$ and $g_t = \nabla f(x_t)$. Since f is convex on $\operatorname{dom}\Phi$ and differentiable at $x_t \in \operatorname{int}(\operatorname{dom}\Phi)$, Lemma 1.4 gives

$$\eta_t\big(f(x_t) - f(x^\star)\big) \le \eta_t \langle \nabla f(x_t), x_t - x^\star \rangle.$$

By Definition 8.2, applied with $x = x_t$ and $y = x_{t+1}$,

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

Since $0 < \eta_t \le 1/L$, multiplying by η_t gives

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

Adding these two inequalities and substituting the identity from Theorem 8.4 yields

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

This is the displayed one-step telescope. Summing over $t = 1, \dots, T$ gives

$$\sum_{t=1}^{T} \eta_t(f(x_{t+1}) - f(x^\star)) \le D_\Phi(x^\star, x_1).$$

By Proposition 8.5, the sequence $f(x_t)$ is nonincreasing. Since all stepsizes are positive,

$$\sum_{t=1}^{T} \eta_t(f(x_{t+1}) - f(x^\star)) \ge \left(\sum_{t=1}^{T} \eta_t\right)(f(x_{T+1}) - f(x^\star)),$$

which proves the last-iterate bound. □

The second item of Definition 8.2 is the lower-model analogue of relative smoothness. Combined with the one-step telescope from Theorem 8.6, it upgrades the sublinear Bregman bound to a linear contraction.

Theorem 8.7 (Linear convergence under relative smoothness and relative strong convexity)**.** *Let $\Phi : E \to (-\infty, +\infty]$ be a mirror map, let $f : \operatorname{dom}\Phi \to \mathbb{R}$ be L-smooth relative to Φ and μ-strongly convex relative to Φ, and assume that*

$$x^\star \in \operatorname*{arg\,min}_{x \in \operatorname{dom}\Phi} f(x)$$

exists. Run unconstrained mirror descent with $g_t = \nabla f(x_t)$ and stepsizes satisfying

$$0 < \eta_t \le \frac{1}{L}.$$

Then

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

Consequently,

$$D_\Phi(x^\star, x_{T+1}) \le D_\Phi(x^\star, x_1) \prod_{t=1}^{T} (1 - \mu\eta_t).$$

In particular, if $\eta_t \equiv 1/L$, then

$$D_\Phi(x^\star, x_{T+1}) \le \left(1 - \frac{\mu}{L}\right)^T D_\Phi(x^\star, x_1),$$

and

$$f(x_{T+1}) - f(x^\star) \le L\left(1 - \frac{\mu}{L}\right)^T D_\Phi(x^\star, x_1).$$

Proof of Theorem 8.7. Because $\operatorname{int}(\operatorname{dom}\Phi)$ is nonempty and open, choose distinct points $x, y \in \operatorname{int}(\operatorname{dom}\Phi)$. Strict convexity of Φ gives $D_\Phi(y, x) > 0$. Applying relative smoothness and relative strong convexity to this pair yields

$$\mu D_\Phi(y, x) \le f(y) - f(x) - \langle \nabla f(x), y - x \rangle \le L D_\Phi(y, x),$$

so $L \ge \mu$. Therefore $0 < \eta_t \le 1/L$ implies $\mu\eta_t \le \mu/L \le 1$.
Apply Theorem 8.4 with $u = x^\star$ and $g_t = \nabla f(x_t)$. By relative strong convexity, applied with $x = x_t$ and $y = x^\star$,

$$\eta_t(f(x_t) - f(x^\star)) + \mu\eta_t D_\Phi(x^\star, x_t) \le \eta_t \langle \nabla f(x_t), x_t - x^\star \rangle .$$

By Definition 8.2, applied with $x = x_t$ and $y = x_{t+1}$,

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

Since $0 < \eta_t \le 1/L$, multiplying by η_t gives

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

Adding the last two inequalities and substituting the identity from Theorem 8.4 gives

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

which is exactly the displayed one-step inequality. Dropping the nonnegative objective-gap term yields the product bound for $D_\Phi(x^\star, x_{T+1})$.
If $\eta_t \equiv 1/L$, this becomes

$$D_\Phi(x^\star, x_{T+1}) \le \left(1 - \frac{\mu}{L}\right)^T D_\Phi(x^\star, x_1).$$

The one-step inequality also gives

$$\frac{1}{L}(f(x_{T+1}) - f(x^\star)) \le \left(1 - \frac{\mu}{L}\right)^T D_\Phi(x^\star, x_1),$$

which rearranges to the displayed objective-gap bound. □

8.5 Mirror Descent on a Constrained Set

We now add an explicit feasible set X. This is the point where the notation becomes more delicate. The function $\Phi + \delta_X$ should not be viewed as another mirror map: the indicator usually destroys differentiability at the boundary of X. The primal Bregman subproblem remains the right definition; the dual formula through $(\Phi + \delta_X)^*$ is an implementation statement derived from the unique exposed maximizer of the constrained conjugate.

Definition 8.5 (Constrained mirror step and potential)**.** Let $\Phi : E \to (-\infty, +\infty]$ be a mirror map, let $X \subseteq \operatorname{dom}\Phi$ be closed and convex, and assume $X \cap \operatorname{int}(\operatorname{dom}\Phi) \neq \emptyset$. Define

$$\Psi := \Phi + \delta_X.$$

For $x \in X \cap \operatorname{int}(\operatorname{dom}\Phi)$, $g \in E^*$, and $\eta > 0$, any solution of the global minimization problem

$$x^+ \in \arg\min_{z \in E} \Big\{ \Psi(z) - \langle \nabla\Phi(x) - \eta g, z \rangle \Big\}.$$

is called a constrained mirror step. Equivalently, because $\Psi(z) = +\infty$ off X, this is the same as

$$x^+ \in \arg\min_{z \in X} \Big\{ \eta \langle g, z - x \rangle + D_\Phi(z, x) \Big\}.$$

Lemma 8.8 (Well-posedness of the constrained mirror step)**.** *In the setup of Definition 8.5, for every*

$$x \in X \cap \operatorname{int}(\operatorname{dom}\Phi), \qquad g \in E^*, \qquad \eta > 0,$$

the global minimization problem in Definition 8.5 has a unique solution

$$x^+ \in X \cap \operatorname{int}(\operatorname{dom}\Phi).$$

Moreover, this point x^+ is the unique maximizer in the definition of

$$\Psi^*(\nabla\Phi(x) - \eta g) = \sup_{z \in X} \{ \langle \nabla\Phi(x) - \eta g, z \rangle - \Phi(z) \},$$

and

$$\partial\Psi^*(\nabla\Phi(x) - \eta g) = \{x^+\}.$$

Theorem 8.9 (Constrained one-step mirror descent inequality)**.** *Let $\Phi : E \to (-\infty, +\infty]$ be a mirror map, let $X \subseteq \operatorname{dom}\Phi$ be closed and convex. Fix $t \in \mathbb{N}$, let $x_t \in X \cap \operatorname{int}(\operatorname{dom}\Phi)$, let $g_t \in E^*$, let $\eta_t > 0$, and let x_{t+1} be the constrained mirror step, so by Lemma 8.8,*

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

and

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

Then

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

Proof of Theorem 8.9. Let

$$Q_t(x) := \eta_t \langle g_t, x - x_t \rangle + D_\Phi(x, x_t).$$

By Lemma 8.8, $x_{t+1} \in \operatorname{int}(\operatorname{dom}\Phi)$, so Q_t is differentiable at x_{t+1}. Fix $u \in X$. Since X is convex, the segment

$$x_{t+1} + s(u - x_{t+1}) \in X \qquad \text{for } s \in [0, 1].$$

Define

$$\gamma_u(s) := Q_t(x_{t+1} + s(u - x_{t+1})), \qquad s \in [0, 1].$$

Because x_{t+1} minimizes Q_t over X, the scalar function γ_u has a minimum at $s = 0$. Hence its right derivative at 0 is nonnegative. Using Definition 8.1, this gives

$$\langle \eta_t g_t + \nabla\Phi(x_{t+1}) - \nabla\Phi(x_t), u - x_{t+1}\rangle \geq 0.$$

Equivalently,

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

Adding $\eta_t \langle g_t, x_t - x_{t+1}\rangle$ to both sides and applying Lemma 8.3 with $(x, y, z) = (x_{t+1}, x_t, u)$ gives the result. □

Thus constraints do not change the Bregman algebra; they only replace the unconstrained first-order equality by a variational inequality. In particular, the proofs of Theorems 8.6 and 8.7 use only the one-step estimate, so the same relative-smoothness and relative strong-convexity consequences carry over once the constrained one-step bound is used in place of Theorem 8.4.

Remark 8.6 (Constrained relative geometry). If $X \subseteq \operatorname{dom}\Phi$ is closed and convex, we say that f is L-smooth relative to Φ on $X \cap \operatorname{int}(\operatorname{dom}\Phi)$ when Definition 8.2 holds for all $x \in X \cap \operatorname{int}(\operatorname{dom}\Phi)$ and $y \in X$. Equivalently,

$$D_{L\Phi - f}(y, x) \geq 0$$

for all $x \in X \cap \operatorname{int}(\operatorname{dom}\Phi)$ and $y \in X$. Likewise, f is μ-strongly convex relative to Φ on $X \cap \operatorname{int}(\operatorname{dom}\Phi)$ when the second item of Definition 8.2 holds for all $x \in X \cap \operatorname{int}(\operatorname{dom}\Phi)$ and $y \in X$. Equivalently,

$$D_{f - \mu\Phi}(y, x) \geq 0$$

for all $x \in X \cap \operatorname{int}(\operatorname{dom}\Phi)$ and $y \in X$. This is the right constrained notion because Lemma 8.8 keeps the iterates in $X \cap \operatorname{int}(\operatorname{dom}\Phi)$, so gradients are only evaluated at interior iterates.

Theorem 8.10 (Constrained rates under relative smoothness and relative strong convexity)**.** *Consider constrained mirror descent*

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

with $x_1 \in X \cap \operatorname{int}(\operatorname{dom}\Phi)$ and $\eta_t \equiv 1/L$. Then:

- *if f is convex on X and L-smooth relative to Φ on $X \cap \operatorname{int}(\operatorname{dom}\Phi)$, and if $x^\star \in \arg\min_{x \in X} f(x)$, then*

$$f(x_{T+1}) - f(x^\star) \leq \frac{L D_\Phi(x^\star, x_1)}{T};$$

- *if f is also μ-strongly convex relative to Φ on $X \cap \operatorname{int}(\operatorname{dom}\Phi)$, then*

$$D_\Phi(x^\star, x_{T+1}) \leq \left(1 - \frac{\mu}{L}\right)^T D_\Phi(x^\star, x_1),$$

and hence

$$f(x_{T+1}) - f(x^\star) \leq L \left(1 - \frac{\mu}{L}\right)^T D_\Phi(x^\star, x_1).$$

This follows from the same proofs as Theorems 8.6 and 8.7, with Theorem 8.9 in place of Theorem 8.4, so we do not reprove it separately here.

The genuinely constrained Euclidean specialization is projected gradient descent.

Example 8.5 (Euclidean geometry on the ℓ_2 ball)**.** Take $E = \mathbb{R}^n$, let

$$B_2(R) := \{x \in \mathbb{R}^n : \|x\|_2 \le R\}, \qquad X := B_2(R), \qquad \Phi(x) = \frac{1}{2}\|x\|_2^2.$$

Then

$$D_\Phi(x, y) = \frac{1}{2}\|x - y\|_2^2,$$

and the constrained mirror step can be written as

$$x_{t+1} \in \arg\min_{x \in B_2(R)} \Big\{ \eta_t \langle g_t, x - x_t \rangle + D_\Phi(x, x_t) \Big\}$$

or, equivalently, as projected gradient descent

$$x_{t+1} = \Pi_{B_2(R)}(x_t - \eta_t g_t).$$

Indeed, the same completion of the square as above rewrites the proxy objective as

$$\frac{1}{2}\|x - (x_t - \eta_t g_t)\|_2^2 + \text{constant},$$

so minimizing over the ball $B_2(R)$ gives the Euclidean projection of $x_t - \eta_t g_t$ onto that ball.

The main role of this lecture is to introduce Bregman geometry as the smooth-descent analogue of Lecture 7's quadratic geometry. The same one-step inequality also has an online reading: if the covectors g_t are revealed losses or subgradients, summing the inequality gives a pathwise regret bound with no probabilistic structure. The next lecture starts from this online reading and then turns it into stochastic optimization guarantees by an online-to-stochastic reduction.

Exercises

1. Prove Lemma 8.3 without expanding all terms at once: first derive a formula for $D_\Phi(z, y) - D_\Phi(z, x)$, and then subtract $D_\Phi(x, y)$.

2. Check that Theorem 8.6 specializes to the Euclidean smooth gradient-descent geometry when $\Phi(x) = \frac{1}{2}\|x\|_2^2$ and $X = \mathbb{R}^n$. Compare the Bregman-telescope term with Theorem 7.12.

3. Let $X = B_2(R)$. Show that the Euclidean mirror step of Example 8.5 is exactly projected gradient descent onto the ℓ_2 ball.

4. Consider constrained mirror descent on the simplex

$$\Delta_n := \left\{ x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1 \right\}$$

with the full-dimensional mirror map

$$\Phi(x) := \sum_{i=1}^n x_i \log x_i \qquad \text{on } \mathbb{R}^n_+.$$

Write the KKT system of the constrained proxy subproblem and re-derive the multiplicative-weights update.

5. Consider constrained mirror descent on the spectrahedron

$$\mathcal{S}_d := \left\{ X \in \mathbb{S}_+^d : \ \operatorname{tr}(X) = 1 \right\}$$

with the full-dimensional mirror map

$$\Phi(X) := \operatorname{tr}(X \log X) \qquad \text{on } \mathbb{S}_+^d.$$

Write the KKT system of the constrained proxy subproblem and re-derive the matrix multiplicative-weights update in Example 8.4.

Deferred Proofs

Lemma 8.11 (Indicator sum rule under interior qualification)**.** *Let $f : E \to \mathbb{R} \cup \{+\infty\}$ be proper, closed, and convex, let $X \subseteq E$ be nonempty, closed, and convex, and assume*

$$0 \in \operatorname{int}(\operatorname{dom} f - X).$$

Then, for every $x \in X \cap \operatorname{dom} f$,

$$\partial(f + \delta_X)(x) = \partial f(x) + \partial \delta_X(x).$$

Proof of Lemma 8.11. The inclusion

$$\partial f(x) + \partial \delta_X(x) \subseteq \partial(f + \delta_X)(x)$$

is immediate by adding the two subgradient inequalities.
For the reverse inclusion, fix $\xi \in \partial(f + \delta_X)(x)$ and define

$$\tilde{f}(z) := f(z) - \langle \xi, z \rangle .$$

Then $\operatorname{dom} \tilde{f} = \operatorname{dom} f$, so

$$0 \in \operatorname{int}(\operatorname{dom} \tilde{f} - X).$$

Moreover, the subgradient inequality for ξ is exactly

$$\tilde{f}(z) + \delta_X(z) \geq \tilde{f}(x) + \delta_X(x) \qquad \forall z \in E,$$

so x minimizes $\tilde{f} + \delta_X$.
Now consider the perturbation

$$\Phi(z, u) := \tilde{f}(z) + \delta_X(z - u), \qquad (z, u) \in E \times E.$$

Its marginal value function is

$$p(u) := \inf_{z \in E} \Phi(z, u) = \inf_{z \in E} \left\{ \tilde{f}(z) + \delta_X(z - u) \right\}.$$

By definition,

$$p(0) = \inf_{z \in E} \left\{ \tilde{f}(z) + \delta_X(z) \right\} = \tilde{f}(x) + \delta_X(x) \in \mathbb{R}.$$

Also,

$$\begin{aligned} u \in \operatorname{dom} p &\iff \exists z \in E \text{ such that } \tilde{f}(z) + \delta_X(z-u) < +\infty \\ &\iff \exists z \in \operatorname{dom} \tilde{f} \text{ such that } z - u \in X \\ &\iff u \in \operatorname{dom} \tilde{f} - X. \end{aligned}$$

Hence $0 \in \operatorname{int}(\operatorname{dom} p)$. By part (4) of Theorem 4.4,

$$p(0) = \max_{y \in E^*} \{-\Phi^*(0,y)\}.$$

We compute $\Phi^*(0,y)$. By the definition of the conjugate,

$$\Phi^*(0,y) = \sup_{z,u \in E} \left\{ \langle y, u \rangle - \tilde{f}(z) - \delta_X(z-u) \right\}.$$

Writing $w := z - u \in X$, so $u = z - w$, gives

$$\begin{aligned} \Phi^*(0,y) &= \sup_{z \in E,\ w \in X} \left\{ \langle y, z-w \rangle - \tilde{f}(z) \right\} \\ &= \sup_{z \in E} \left\{ \langle y, z \rangle - \tilde{f}(z) \right\} + \sup_{w \in X} \{\langle -y, w \rangle\} \\ &= \tilde{f}^*(y) + \delta_X^*(-y). \end{aligned}$$

Therefore there exists $y^\star \in E^*$ such that

$$\tilde{f}(x) + \delta_X(x) + \tilde{f}^*(y^\star) + \delta_X^*(-y^\star) = 0.$$

By Lemma 4.1 for $\tilde{f}$ and for δ_X,

$$\tilde{f}(x) + \tilde{f}^*(y^\star) \geq \langle y^\star, x \rangle, \qquad \delta_X(x) + \delta_X^*(-y^\star) \geq \langle -y^\star, x \rangle.$$

The sum of these two inequalities is at least 0, and the displayed equality shows that the sum is exactly 0. Hence both inequalities hold with equality. By the equality case in Lemma 4.1,

$$y^\star \in \partial \tilde{f}(x) \qquad \text{and} \qquad -y^\star \in \partial \delta_X(x).$$

Since $\tilde{f} = f - \langle \xi, \cdot \rangle$, the first condition is equivalent to

$$y^\star + \xi \in \partial f(x).$$

Thus

$$\xi = (y^\star + \xi) + (-y^\star) \in \partial f(x) + \partial \delta_X(x),$$

which proves the reverse inclusion. □

Proof of Lemma 8.8. Write

$$\theta := \nabla \Phi(x) - \eta g,$$

and define

$$\varphi_\theta(z) := \Phi(z) - \langle \theta, z \rangle, \qquad F_\theta(z) := \Psi(z) - \langle \theta, z \rangle.$$

Then $F_\theta = \varphi_\theta + \delta_X$, so Definition 8.5 defines the constrained mirror step as the global minimization of F_θ over E.

We first record a consequence of Proposition 8.2. Fix any $\xi \in E^*$. Choose any $y_0 \in \operatorname{int}(\operatorname{dom}\Phi)$, and apply Proposition 8.2 with $x = y_0$, $\eta = 1$, and $g = \nabla\Phi(y_0) - \xi$. This yields a unique point $y_\xi \in \operatorname{int}(\operatorname{dom}\Phi)$ satisfying $\nabla\Phi(y_\xi) = \xi$. Since Φ is differentiable at y_ξ, Lemma 4.1 gives

$$y_\xi \in \partial\Phi^*(\xi) \qquad \text{and} \qquad \Phi^*(\xi) = \langle \xi, y_\xi \rangle - \Phi(y_\xi) < \infty.$$

Strict convexity of Φ makes this maximizer unique, so

$$\partial\Phi^*(\xi) = \{y_\xi\} \qquad \text{with} \qquad y_\xi \in \operatorname{int}(\operatorname{dom}\Phi).$$

Now let

$$B_* := \{u \in E^* : \|u\|_* \le 1\}.$$

Choose a basis $e_1, \dots, e_n$ of E, and let $e_1^*, \dots, e_n^*$ be the dual basis of E^*. If

$$u = \sum_{i=1}^n \alpha_i e_i^* \qquad \text{and} \qquad u \in B_*,$$

then

$$|\alpha_i| = |\langle u, e_i \rangle| \le \|u\|_* \, \|e_i\| \le \|e_i\| \qquad \forall i \in \{1, \dots, n\}.$$

Therefore

$$B_* \subseteq P := \left\{ \sum_{i=1}^n \alpha_i e_i^* : \ |\alpha_i| \le \|e_i\| \ \text{ for all } i \right\},$$

where P is the convex hull of its finitely many vertices $v_1, \dots, v_N$. Since $\Phi^*(\theta + v_j) \in \mathbb{R}$ for each j, the quantity

$$M_\theta := \max_{1 \le j \le N} \Phi^*(\theta + v_j)$$

is finite. Every $\theta + u$ with $u \in B_*$ lies in $\theta + P = \operatorname{conv}\{\theta + v_1, \dots, \theta + v_N\}$, so convexity of Φ^* gives

$$\Phi^*(\theta + u) \le M_\theta \qquad \forall u \in B_*.$$

For any $z \in E$, choose $u_z \in B_*$ with $\langle u_z, z \rangle = \|z\|$. Fenchel's inequality gives

$$\Phi(z) + \Phi^*(\theta + u_z) \ge \langle \theta + u_z, z \rangle,$$

hence

$$\varphi_\theta(z) = \Phi(z) - \langle \theta, z \rangle \ge \langle u_z, z \rangle - \Phi^*(\theta + u_z) \ge \|z\| - M_\theta.$$

Thus every sublevel set of φ_θ is bounded. Since Φ is closed and strictly convex by (H1), the same is true of φ_θ. Therefore F_θ is proper, closed, and strictly convex. Choose $x_0 \in X \cap \operatorname{int}(\operatorname{dom}\Phi)$. Then $F_\theta(x_0) = \varphi_\theta(x_0) < +\infty$, so the sublevel set

$$K := \{z \in E : F_\theta(z) \le F_\theta(x_0)\}$$

is nonempty. It is bounded by the estimate above, and it is closed because F_θ is closed. In finite dimension, K is therefore compact. By Lemma 7.13, F_θ attains its minimum at a unique point $\bar{x} \in E$. Since $F_\theta < +\infty$ exactly on X, this $\bar{x}$ is exactly the unique minimizer of φ_θ over X.
It remains to show that the minimizer lies in $\operatorname{int}(\operatorname{dom}\Phi)$. Because $x_0 \in X \cap \operatorname{int}(\operatorname{dom}\Phi)$, there exists $r > 0$ such that

$$x_0 + \{u \in E : \|u\| < r\} \subseteq \operatorname{dom}\Phi.$$

Since $x_0 \in X$, this implies

$$\{u \in E : \|u\| < r\} \subseteq \operatorname{dom} \Phi - X.$$

Hence

$$0 \in \operatorname{int}(\operatorname{dom} \Phi - X) = \operatorname{int}(\operatorname{dom} \varphi_\theta - X).$$

Now $\bar{x}$ minimizes $F_\theta = \varphi_\theta + \delta_X$, so

$$0 \in \partial F_\theta(\bar{x}).$$

Applying Lemma 8.11 to $f = \varphi_\theta$, we obtain

$$0 \in \partial \varphi_\theta(\bar{x}) + \partial \delta_X(\bar{x}).$$

Thus there exists $v \in \partial \delta_X(\bar{x})$ such that

$$\theta - v \in \partial \Phi(\bar{x}),$$

because $\partial \varphi_\theta(\bar{x}) = \partial \Phi(\bar{x}) - \theta$. Applying Fenchel reciprocity gives

$$\bar{x} \in \partial \Phi^*(\theta - v).$$

But the first part of the proof showed that

$$\partial \Phi^*(\theta - v) = \{y_{\theta - v}\} \subseteq \operatorname{int}(\operatorname{dom} \Phi).$$

Therefore $\bar{x} \in X \cap \operatorname{int}(\operatorname{dom} \Phi)$. Set

$$w := \nabla \Phi(x) - \eta g \qquad \text{and} \qquad x^+ := \bar{x}.$$

For $z \in X$,

$$\eta \langle g, z - x \rangle + D_\Phi(z, x) = \Phi(z) - \langle w, z \rangle - \Phi(x) + \langle w, x \rangle .$$

The last two terms are constant in z. Hence minimizing the constrained mirror objective is equivalent to maximizing

$$\langle w, z \rangle - \Phi(z) = \langle w, z \rangle - \Psi(z)$$

over $z \in X$. Since x^+ is the unique minimizer, it is also the unique maximizer, and

$$\Psi^*(w) = \left\langle w, x^+ \right\rangle - \Psi(x^+) < +\infty.$$

Applying Lemma 4.1 to the proper closed convex function Ψ gives

$$x^+ \in \partial \Psi^*(w).$$

Now let $y \in \partial \Psi^*(w)$. Another application of Lemma 4.1 yields

$$\Psi(y) + \Psi^*(w) = \langle w, y \rangle ,$$

so $y \in X$ and

$$\Psi^*(w) = \langle w, y \rangle - \Phi(y).$$

Thus y is also a maximizer of

$$\sup_{z \in X} \{\langle w, z \rangle - \Phi(z)\},$$

hence $y = x^+$. Therefore

$$\partial \Psi^*(\nabla \Phi(x) - \eta g) = \{x^+\}.$$

□

References

[BBT17] Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

[LFN18] Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.