Lecture 8: Mirror Descent and Bregman Geometry

TTIC 31070 / CMSC 35470 / BUSF 36903 / STAT 31015

Convex Optimization

Prof. Zhiyuan Li

Spring 2026

From Fixed-Norm to Bregman Geometry

Lecture 7: fixed norm $\|\cdot\|$ gives a fixed quadratic upper model

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

Problem. For barrier/entropy-like objectives on an open domain, $\nabla^2 f$ blows up near the boundary $\Rightarrow$ *no finite* L works with any fixed norm.

Fix. Allow the upper-model penalty to depend on the current point:

$$x_{t+1} \in \operatorname*{argmin}_x \{\eta_t \langle g_t, x - x_t \rangle + V_{x_t}(x)\}.$$

Structured case: $V_{x_t}(x) = D_\Phi(x, x_t)$, the **Bregman divergence**.

Mirror descent is obtained by the same recipe as gradient descent: *minimize the local upper model*.

Why We Need This: The Log-Barrier

Example 8.1 (Log-barrier objective). $\Phi(x) = -\log x - \log(1-x)$ on $(0,1)$ (extended by $+\infty$ outside), and for $c \in \mathbb{R}$,

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

Boundary blow-up: $\Phi''(x) = \frac{1}{x^2} + \frac{1}{(1-x)^2} \to +\infty$ as $x \to 0^+$ or $x \to 1^-$.

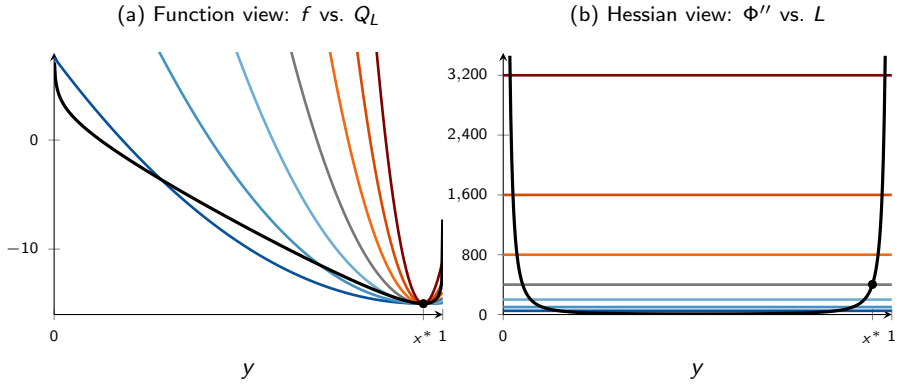
Fixed-norm impossibility. Every norm on $\mathbb{R}$ has the form $\|s\| = \gamma|s|$. If Lecture 7's smoothness inequality held with constant L, then

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

contradicting $\sup_x f''(x) = +\infty$.

This is exactly the pathology that forces us to move beyond fixed quadratic geometry. Barrier/entropy objectives are not rare — they appear whenever we have open-domain constraints.

No Finite L Works: Function vs. Hessian View



With $c = -19$, $x^* \approx 0.95$ and $f''(x^*) \approx 403$. Seven quadratic models Q_L at $L = 50 \cdot 2^k$, $k = 0, \ldots, 6$ (cool $\rightarrow$ warm). **(a)** no L dominates f everywhere; **(b)** Φ'' blows up at both boundaries — no constant $L \geq \Phi''$ works.

Bregman Divergence

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

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

How to read $D_\Phi(x, y)$. The *gap* between $\Phi(x)$ and the tangent affine approximation to Φ at y.

- Φ convex $\Rightarrow D_\Phi(x, y) \geq 0$
- Φ strictly convex $\Rightarrow D_\Phi(x, y) = 0$ iff $x = y$
- **Not a metric**: usually asymmetric, no triangle inequality

Role. A geometry *penalty*, not a distance in the metric-space sense.

Euclidean case $\Phi(x) = \frac{1}{2}\|x\|_2^2$: $D_\Phi(x, y) = \frac{1}{2}\|x - y\|_2^2$. (Symmetric only because Φ is quadratic.)

Relative Smoothness and Strong Convexity

Let Φ be proper convex, differentiable on $\operatorname{int}(\operatorname{dom}\Phi)$, and $f : \operatorname{dom}\Phi \to \mathbb{R}$ convex, differentiable on $\operatorname{int}(\operatorname{dom}\Phi)$.

Definition 8.2 (Relative smoothness / strong convexity). For all $x \in \operatorname{int}(\operatorname{dom}\Phi)$, $y \in \operatorname{dom}\Phi$:

L-smooth relative to Φ ($L > 0$):

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

μ-strongly convex relative to Φ ($\mu > 0$):

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

Direct analogues of Euclidean smoothness and strong convexity, with $\frac{1}{2}\|y - x\|^2$ replaced by $D_\Phi(y, x)$. Specializes to Lecture 7 when $\Phi(x) = \frac{1}{2}\|x\|_2^2$.

Bregman Reformulations

Lemma 8.1 (Bregman reformulations). For $x \in \operatorname{int}(\operatorname{dom}\Phi)$, $y \in \operatorname{dom}\Phi$:

(i) f is L-smooth rel. $\Phi \iff D_{L\Phi - f}(y, x) \geq 0$. In particular, $L\Phi - f$ is convex on $\operatorname{int}(\operatorname{dom}\Phi)$.

(ii) f is μ-strongly convex rel. $\Phi \iff D_{f-\mu\Phi}(y, x) \geq 0$. In particular, $f - \mu\Phi$ is convex on $\operatorname{int}(\operatorname{dom}\Phi)$.

Proof. The defining inequality

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq L\big[\Phi(y) - \Phi(x) - \langle \nabla\Phi(x), y - x\rangle\big]$$

rearranges directly to $D_{L\Phi - f}(y, x) \geq 0$. Restricting $y \in \operatorname{int}(\operatorname{dom}\Phi)$ gives the first-order convexity criterion for $L\Phi - f$. Similarly for (ii). □

Relative smoothness/SC reduce to *Bregman nonnegativity* of the modified potential $L\Phi - f$ (resp. $f - \mu\Phi$). Much easier to verify than pointwise inequalities.

Log-Barrier Revisited: An Exact Match

Back to Ex 8.1: $\Phi(x) = -\log x - \log(1-x)$, $f(x) = \Phi(x) + cx$ on $(0,1)$.

Verify relative smoothness. $f - \Phi = cx$ is affine (hence convex and concave), so

$$L\Phi - f = (L-1)\Phi - cx \text{ convex} \iff L \geq 1.$$

Thus f is 1-**smooth relative to** Φ.

Verify relative strong convexity. $f - \mu\Phi = (1-\mu)\Phi + cx$ convex $\iff \mu \leq 1$. So the best $\mu = 1$: f is 1-strongly convex relative to Φ.

Exact match. Since $f - \Phi$ is affine, it drops out of D_f: $D_f(y,x) = D_\Phi(y,x)$. Hence

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + D_\Phi(y,x) \qquad \text{(no slack!)}$$

The Bregman model is *equal* to f, not just an upper bound.

Any f differing from Φ by an affine term has $L_{\text{rel}} = \mu = 1$. Preview: mirror descent will reach x^* in **one step** (Thm 8.7).

Mirror Map

Definition 8.3 (Mirror map). $\Phi : E \to (-\infty, +\infty]$ is a *mirror map* if:

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

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

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

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

Role of each structural assumption.

- **Strict convexity (H1):** Bregman subproblem has *at most* one minimizer.
- **Range condition (H4):** after a dual-coordinate move $\nabla\Phi(x) - \eta g$, *some* primal point in $\operatorname{int}(\operatorname{dom}\Phi)$ realizes that mirror coordinate.

Together: strict convexity $\Rightarrow$ uniqueness; (H4) $\Rightarrow$ global existence for unconstrained updates.

Mirror Step: Primal and Dual Forms

Definition 8.4 (Unconstrained mirror step). For $x \in \operatorname{int}(\operatorname{dom}\Phi)$, $g \in E^*$, $\eta > 0$:

$$x^+ \in \operatorname*{argmin}_{z \in E} \{\Phi(z) - \langle \nabla\Phi(x) - \eta g, z\rangle\} = \operatorname*{argmin}_{z \in \operatorname{dom}\Phi} \{\eta\langle g, z - x\rangle + D_\Phi(z, x)\}.$$

Proposition 8.2 (Dual-coordinate implementation). $\nabla\Phi : \operatorname{int}(\operatorname{dom}\Phi) \to E^*$ is a bijection, and the unique solution is

$$x^+ = (\nabla\Phi)^{-1}\big(\nabla\Phi(x) - \eta g\big), \quad \text{i.e.} \quad \boxed{\nabla\Phi(x^+) = \nabla\Phi(x) - \eta g.}$$

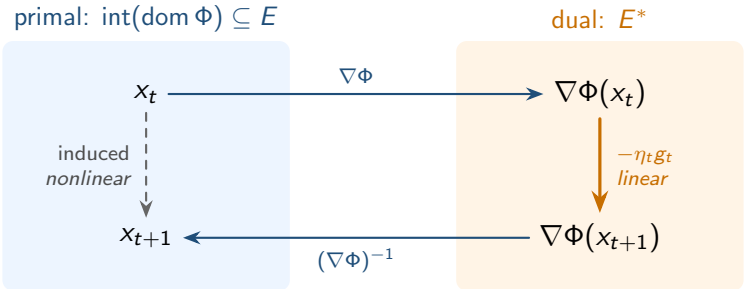
Proof sketch. Strict convexity of $\Phi \Rightarrow \nabla\Phi$ is *injective* on $\operatorname{int}(\operatorname{dom}\Phi)$. (H4) gives surjectivity onto E^*, so $\nabla\Phi$ is bijective.

Fenchel equality $\Phi(x^+) + \Phi^*(\nabla\Phi(x) - \eta g) = \langle \nabla\Phi(x) - \eta g, x^+\rangle$ identifies x^+ as the unique minimizer. $\square$

Mirror Descent: Change of Coordinates

Algorithm (Unconstrained mirror descent). Start $x_1 \in \operatorname{int}(\operatorname{dom}\Phi)$. For $t = 1, 2, \ldots$:

$$\nabla\Phi(x_{t+1}) = \nabla\Phi(x_t) - \eta_t g_t, \qquad g_t \in E^*,\ \eta_t > 0.$$



Dual trajectory: plain linear shift. **Primal** trajectory: curved by Φ. (H4) keeps the dual shift inside the coordinate system.

Example: Euclidean Geometry = Gradient Descent

Example 8.2. $E = \mathbb{R}^n$, $\Phi(x) = \frac{1}{2}\|x\|_2^2$, $\operatorname{dom}\Phi = \mathbb{R}^n$.

Ingredients:

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

Mirror step. Complete the square:

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

Minimizing: $x_{t+1} = x_t - \eta_t g_t$, the classical **gradient descent**.

Relative smoothness w.r.t. this Φ $\iff$ classical L-smoothness in ℓ_2. Thm 8.6 recovers Lecture 7's $O(1/T)$ rate.

Affine Slices via Centered Coordinates

Challenge. Many domains are affine slices (simplex, spectrahedron), not linear spaces. Mirror-map theory is stated on a linear E.

Fix. Translate: if $A = x_c + E^\circ$, identify $x \in A$ with $x' = x - x_c \in E^\circ$ and transport $\Phi'(x) = \Phi(x - x_c)$.

Invariance. Bregman divergence and linear pairing transfer unchanged: $D_{\Phi'}(x, y) = D_\Phi(x', y')$, $\langle g, x - u \rangle = \langle g, x' - u' \rangle$. All constants (L, μ) stay the same; only the ambient update formula is rewritten at the end.

Two standard centered models:

- **Simplex:** $\Delta_n - \frac{1}{n}\mathbf{1} \subseteq \{x' \in \mathbb{R}^n : \sum x'_i = 0\}$
- **Spectrahedron:** $\mathcal{S}_n - \frac{1}{n} I_n \subseteq \{X' = X'^\top : \mathrm{tr}(X') = 0\}$

Example: Simplex — Setup and Mirror Map

Example 8.3. $\Delta_n = \{x \in \mathbb{R}^n_+ : \sum_i x_i = 1\}$ with ambient entropy $\Phi'(x) = \sum_i x_i \log x_i$.

Centered linear model. $E^\circ = \{x' \in \mathbb{R}^n : \sum_i x'_i = 0\}$, $X^\circ = \Delta_n - \frac{1}{n}\mathbf{1} \subseteq E^\circ$, $\Phi(x') := \Phi'(\frac{1}{n}\mathbf{1} + x')$.

Verify H4. For $x = \frac{1}{n}\mathbf{1} + x'$ with $x_i > 0$ and $v' \in E^\circ$ (so $\sum_i v'_i = 0$):

$$\langle \nabla\Phi(x'), v' \rangle = \sum_i (\log x_i + 1) v'_i = \sum_i (\log x_i)\, v'_i.$$

Converse: given $g \in (E^\circ)^*$ with representative $\widetilde{g} \in \mathbb{R}^n$, set $x_i = e^{\widetilde{g}_i} / \sum_j e^{\widetilde{g}_j} \in (0,\infty)^n \cap \Delta_n$. Then $\langle \nabla\Phi(x'), v' \rangle = \sum_i \widetilde{g}_i v'_i = \langle g, v' \rangle$. ✓

So $\nabla\Phi$ is the restriction of $\log(\cdot)$ to E°, and $\nabla\Phi(\mathrm{int}(\mathrm{dom}\,\Phi)) = (E^\circ)^*$: **(H4) holds**.

Example: Simplex $\to$ Multiplicative Weights

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

$$D_\Phi(x', y') = D_{\Phi'}(x, y) = \sum_i x_i \log(x_i/y_i) = \mathrm{KL}(x\|y).$$

Mirror-step Lagrangian. Minimize $\eta_t \sum_i g_{t,i}(x_i - x_{t,i}) + \mathrm{KL}(x\|x_t)$ subject to $\sum_i x_i = 1$, $x_i \geq 0$. With multiplier λ for the equality:

$$\eta_t g_{t,i} + \log(x_i/x_{t,i}) + \lambda = 0.$$

Solving: $x_i \propto x_{t,i}\, e^{-\eta_t g_{t,i}}$. Normalize by $\sum_i x_i = 1$:

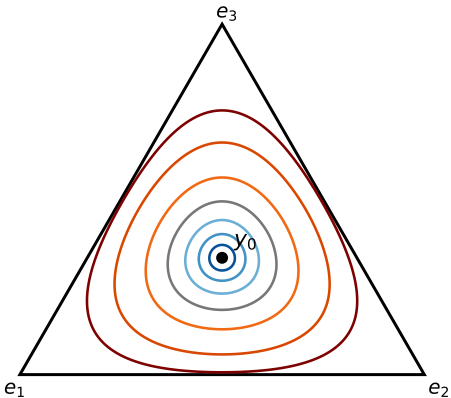
$$\boxed{x_{t+1,i} = \frac{x_{t,i}\, e^{-\eta_t g_{t,i}}}{\sum_j x_{t,j}\, e^{-\eta_t g_{t,j}}}}$$

the **multiplicative-weights** (exponential-weights) update.

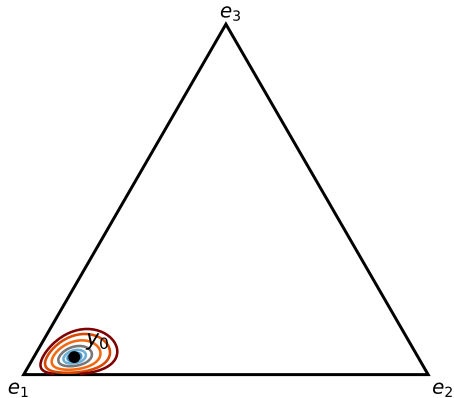
Dual form: $\nabla\Phi(x_{t+1}) = \nabla\Phi(x_t) - \eta_t g_t$ becomes $\log x_{t+1,i} = \log x_{t,i} - \eta_t g_{t,i} + \mathrm{const}$.

Geometry of KL on the Simplex

(a) $y_0 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ (center)



(b) $y_0 = (0.85, 0.10, 0.05)$ (near corner e_1)



Example: Spectrahedron — Setup and Mirror Map

Example 8.4. $\mathcal{S}_n = \{X \in \mathbb{R}^{n\times n} : X = X^\top,\ X \succeq 0,\ \operatorname{tr}(X) = 1\}$ with ambient matrix entropy $\Phi'(X) = \operatorname{tr}(X \log X)$.

Centered linear model. $E^\circ = \{X' = X'^\top : \operatorname{tr}(X') = 0\}$, $\mathcal{S}_n^\circ = \mathcal{S}_n - \frac{1}{n} I_n \subseteq E^\circ$, $\Phi(X') := \Phi'(\frac{1}{n} I_n + X')$.

Verify H4. For $X = \frac{1}{n} I_n + X' \succ 0$ and $V' \in E^\circ$ (so $\operatorname{tr}(V') = 0$):

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

Converse: given $G \in (E^\circ)^*$ with symmetric representative $\widetilde{G}$, set $X = \exp(\widetilde{G})/\operatorname{tr}\exp(\widetilde{G}) \succ 0$. Then $\log X = \widetilde{G} - (\log \operatorname{tr}\exp \widetilde{G})\, I_n$, so $\langle \nabla\Phi(X'), V' \rangle = \operatorname{tr}(\widetilde{G}\, V') = \langle G, V' \rangle$. ✓

$\nabla\Phi$ is the restriction of $\log(\cdot)$ to E°, and $\nabla\Phi(\operatorname{int}(\operatorname{dom}\Phi)) = (E^\circ)^*$: **(H4) holds**. (Matrix log via eigendecomposition.)

Example: Spectrahedron → Matrix MW

Bregman divergence (quantum relative entropy). For $X = \frac{1}{n} I_n + X' \succ 0$, $Y = \frac{1}{n} I_n + Y' \succ 0$:

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

Mirror-step Lagrangian. Minimize $\eta_t \operatorname{tr}(G_t(X - X_t)) + D_{\Phi'}(X, X_t)$ over $X \succeq 0$, $\operatorname{tr}(X) = 1$. With scalar multiplier λ:

$$\eta_t G_t + \log X - \log X_t + \lambda I_n = 0 \;\Rightarrow\; X \propto \exp\big(\log X_t - \eta_t G_t\big).$$

Normalize by $\operatorname{tr}(X) = 1$:

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

the **matrix multiplicative weights** update.

Each step: one matrix exponential (eigendecompose $\log X_t - \eta_t G_t$). Drives Arora–Kale SDP solvers, spectral sparsification, quantum state estimation.

The Three-Point Identity

Lemma 8.3 (Three-point identity). For $x, y \in \operatorname{int}(\operatorname{dom}\Phi)$, $z \in \operatorname{dom}\Phi$:

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

Proof. Expand each Bregman divergence:

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

Subtract: Φ-terms cancel; the $\langle \nabla\Phi(y), z - y \rangle - \langle \nabla\Phi(y), x - y \rangle = \langle \nabla\Phi(y), z - x \rangle$ combines with $\langle \nabla\Phi(x), z - x \rangle$ to give $\langle \nabla\Phi(x) - \nabla\Phi(y), z - x \rangle$. □

The Bregman analogue of the parallelogram law — an **algebraic identity** *(no convexity beyond differentiability). Engine of every mirror-descent bound.*

Unconstrained One-Step Equality

Theorem 8.4 (Unconstrained one-step equality). Let $\nabla\Phi(x_{t+1}) = \nabla\Phi(x_t) - \eta_t g_t$. Then for all $u \in \operatorname{dom}\Phi$,

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

Proof. From the update: $\eta_t g_t + \nabla\Phi(x_{t+1}) - \nabla\Phi(x_t) = 0$. Hence

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

Apply Lemma 8.3 with $(x, y, z) = (x_{t+1}, x_t, u)$:

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

Substitute. □

Master equality. Adding an objective upper model produces the convergence bound. Adding a lower model upgrades it to linear. Under constraints, "=" becomes "≤" but the algebra is the same.

Descent under Relative Smoothness

Proposition 8.5 (Descent). If f is L-smooth rel. Φ, $g_t = \nabla f(x_t)$, $0 < \eta_t \le 1/L$, then

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

Proof. By Prop 8.2, x_{t+1} minimizes the mirror subproblem, so comparing values at x_{t+1} vs x_t:

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

By rel. smoothness at (x_t, 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 $\eta_t \le 1/L$, $L \le 1/\eta_t$, so $L\, D_\Phi(x_{t+1}, x_t) \le \frac{1}{\eta_t} D_\Phi(x_{t+1}, x_t)$. Combining with $(*)$:

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

$O(1/T)$ Last-Iterate Rate

Theorem 8.6. Prop 8.5 setting, $x^* \in \operatorname{argmin}_{\operatorname{dom}\Phi} f$. One-step telescope:

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

Summing $t = 1, \dots, T$ and using monotonicity of $f(x_t)$:

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

Proof. Three ingredients with $g_t = \nabla f(x_t)$:

(A) Convexity: $\eta_t(f(x_t) - f(x^*)) \le \eta_t \langle g_t, x_t - x^* \rangle$.

(B) Rel. smoothness$\times \eta_t \le 1/L$: $\eta_t(f(x_{t+1}) - f(x_t)) \le \eta_t \langle g_t, x_{t+1} - x_t \rangle + D_\Phi(x_{t+1}, x_t)$.

Add (A)+(B), substitute Thm 8.4: $\eta_t(f(x_{t+1}) - f(x^*)) \le D_\Phi(x^*, x_t) - D_\Phi(x^*, x_{t+1})$. $\square$

Linear Rate under Relative Strong Convexity

Theorem 8.7. f both L-smooth and μ-strongly convex rel. Φ, $0 < \eta_t \le 1/L$:

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

With $\eta_t = 1/L$: $D_\Phi(x^*, x_{T+1}) \le (1 - \mu/L)^T D_\Phi(x^*, x_1)$ and $f(x_{T+1}) - f(x^*) \le L(1 - \mu/L)^T D_\Phi(x^*, x_1)$.

Why $\mu \le L$. Lemma 8.1 $\Rightarrow$ $L\Phi - f$ and $f - \mu\Phi$ both convex, so $(L - \mu)\Phi$ is convex. If $L < \mu$ then $-\Phi$ convex $\Rightarrow$ Φ both convex & concave, contradicting strict convexity (H1). So $\kappa := L/\mu \ge 1$.

$T = O(\kappa \log(1/\varepsilon))$ iterations. Bregman analogue of Lecture 7's linear rate.

Proof of Theorem 8.7

Apply Thm 8.4 with $u = x^*$, $g_t = \nabla f(x_t)$.

(A') Rel. *strong* convexity at (x_t, x^*):

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

(B) Rel. smoothness at (x_t, x_{t+1}), multiplied by $\eta_t \le 1/L$:

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

(A')+(B), then substitute Thm 8.4:

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

Rearrange:

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

Drop the nonnegative $\eta_t\big(f(x_{t+1}) - f(x^*)\big)$ term $\Rightarrow$
$D_\Phi(x^*, x_{t+1}) \le (1 - \mu\eta_t)\, D_\Phi(x^*, x_t)$.
Iterate T times. $\square$

Constrained Mirror Descent

Definition 8.5 (Constrained mirror step). $X \subseteq \operatorname{dom}\Phi$ closed convex, $X \cap \operatorname{int}(\operatorname{dom}\Phi) \neq \varnothing$, $\Psi := \Phi + \delta_X$. For $x \in X \cap \operatorname{int}(\operatorname{dom}\Phi)$, $g \in E^*$, $\eta > 0$:

$$x^+ \in \operatorname*{argmin}_{z \in X} \{\eta\langle g, z - x\rangle + D_\Phi(z, x)\}.$$

Lemma 8.8 (Well-posedness). x^+ exists, is unique, and lies in $X \cap \operatorname{int}(\operatorname{dom}\Phi)$. It is also the unique maximizer in $\Psi^*(\nabla\Phi(x) - \eta g)$ and $\partial\Psi^*(\nabla\Phi(x) - \eta g) = \{x^+\}$.

Warning. $\Psi = \Phi + \delta_X$ is **not** a mirror map: δ_X kills differentiability at ∂X. It's an *implementation device* via Ψ^*.

Constrained One-Step Inequality

Theorem 8.9 (Constrained one-step inequality). For $x_t \in X \cap \operatorname{int}(\operatorname{dom}\Phi)$, $u \in X$:

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

Proof idea. First-order optimality of x_{t+1} on X:
$\langle \eta_t g_t + \nabla\Phi(x_{t+1}) - \nabla\Phi(x_t), u - x_{t+1} \rangle \ge 0$ for all $u \in X$ (*variational inequality*).
Rearrange and apply Lemma 8.3. $\square$

Constrained relative geometry. Read Def 8.2 with $y \in X$: f is L-smooth rel. Φ on $X \iff L\Phi - f$ convex on X. Lemma 8.8 keeps iterates in $X \cap \operatorname{int}(\operatorname{dom}\Phi)$, so the proofs of Thms 8.6 & 8.7 transfer **verbatim** — only Thm 8.4 is replaced by Thm 8.9.

Constrained Rates under Relative Smoothness / Strong Convexity

Theorem 8.10 (Constrained rates). Run constrained mirror descent $x_{t+1} \in \operatorname{argmin}_{x \in X}\{\eta_t \langle g_t, x - x_t\rangle + D_\Phi(x, x_t)\}$ with $x_1 \in X \cap \operatorname{int}(\operatorname{dom}\Phi)$ and $\eta_t \equiv 1/L$.

- If f is convex on X and L-smooth rel. Φ on $X \cap \operatorname{int}(\operatorname{dom}\Phi)$, and $x^* \in \operatorname{argmin}_{x \in X} f(x)$, then

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

- If additionally f is μ-strongly convex rel. Φ on $X \cap \operatorname{int}(\operatorname{dom}\Phi)$, then

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

Proof. Same proofs as Thms 8.6 & 8.7, with *Thm 8.9* in place of Thm 8.4. The Bregman telescoping and relative-convexity steps are identical. □

Example: ℓ_2 Ball = Projected Gradient Descent

Example 8.5. $X = B_2(R) = \{x : \|x\|_2 \le R\}$, $\Phi(x) = \frac{1}{2}\|x\|_2^2$.

Constrained mirror step:

$$x_{t+1} = \operatorname*{argmin}_{x \in B_2(R)} \{\eta_t \langle g_t, x - x_t\rangle + \tfrac{1}{2}\|x - x_t\|_2^2\} = \Pi_{B_2(R)}(x_t - \eta_t g_t),$$

the **projected gradient** update (completing the square).

Rate (from Thm 8.10 with $\eta_t = 1/L$): $f(x_{T+1}) - f(x^*) \le \frac{L\|x^* - x_1\|_2^2}{T}$.

Takeaway.

- ℓ_2 ball with Euclidean $\Phi \to$ projected GD
- Simplex with entropy $\Phi \to$ MW (no Euclidean projection!)
- Spectrahedron with matrix entropy $\to$ matrix MW

The *geometry of* Φ dictates the algorithm, and matching Φ to the domain replaces expensive projections by cheap closed-form updates.

Summary & What's Next

Today: Bregman geometry replaces fixed-norm geometry.

- $D_\Phi(x, y) = \Phi(x) - \Phi(y) - \langle \nabla\Phi(y), x - y \rangle$ (Def 8.1)
- Relative L-smoothness / μ-strong convexity as *ordinary* convexity of $L\Phi - f$ and $f - \mu\Phi$ (Lemma 8.1)
- Mirror step: dual shift $\nabla\Phi(x^+) = \nabla\Phi(x) - \eta g$ (Def 8.4, Prop 8.2)
- Three-point identity + one-step equality (Lem 8.3, Thm 8.4)
- $O(1/T)$ rate (Thm 8.6) and linear rate $(1 - \mu/L)^T$ (Thm 8.7)
- Euclidean = GD, simplex = MW, spectrahedron = matrix MW
- Constraints: VI-based one-step (Thm 8.9) gives the same rates (Thm 8.10)

Next lecture:

- Online convex optimization: pathwise regret bounds from Thm 8.4
- Online-to-stochastic reduction: SGD and stochastic mirror descent