Lecture 7: Steepest Descent and Descent Lemmas

Lecture 6 used separation to update a containing convex set. That viewpoint is fundamentally global and fundamentally convex: one keeps shrinking a localization set while preserving the promise that it still contains the target set, so without convexity the cutting-plane picture no longer has the same meaning. Its complexity is therefore governed by how fast volume shrinks in ambient dimension n. Lecture 7 turns instead to point-based first-order methods for smooth functions. Once a norm is fixed, smoothness gives a local quadratic upper model around the current iterate, and steepest descent is obtained by choosing a point that makes this local model small. We will first write the update through this quadratic proxy, and then use the linear minimization oracle on the norm ball to decompose that proxy minimizer into a direction and a magnitude. The structural point of the lecture is that the descent lemma uses only smoothness, so it already gives local improvement even for nonconvex objectives. Convexity enters only afterwards, when we want to turn the one-step decrease into a bound on the global suboptimality $f(x) - f(x^\star)$. We will do this in two different ways: one through strong convexity, and one through a radius bound on the relevant sublevel set. In that sense the resulting rates are expressed through problem constants such as L, μ, and R, rather than directly through the ambient dimension.

Unless explicitly stated Euclidean, let E be a finite-dimensional real normed space and let E^* be its dual. Throughout the general-norm part of this lecture, the symbol $\nabla f(x)$ is to be read as the differential $Df(x) \in E^*$, and every expression $\langle \nabla f(x), h \rangle$ is a dual pairing.

7.1 Smoothness and Quadratic Upper Bounds

Definition 7.1 (Smoothness with respect to a norm)**.** Let $f : E \to \mathbb{R}$ be differentiable. We say that f is L-smooth with respect to $\|\cdot\|$ if

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

Lemma 7.1 (Gradient Lipschitzness implies smoothness)**.** *Let $f : E \to \mathbb{R}$ be differentiable. Assume that*

$$\forall x, y \in E, \qquad \|\nabla f(y) - \nabla f(x)\|_* \le L \|y - x\| .$$

Then f is L-smooth with respect to $\|\cdot\|$.

Proof of Lemma 7.1. Fix $x, y \in E$, and write $\Delta := y - x$. Consider the path

$$\gamma(t) := x + t\Delta, \qquad t \in [0, 1].$$

By the fundamental theorem of calculus,

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

Subtracting $\langle \nabla f(x), \Delta \rangle$ gives

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

Hence, by Hölder's inequality and the assumed gradient-Lipschitz bound,

$$\begin{aligned} |f(y) - f(x) - \langle \nabla f(x), \Delta \rangle| &\leq \int_0^1 \|\nabla f(\gamma(t)) - \nabla f(x)\|_* \, \|\Delta\| \, dt \\ &\leq \int_0^1 Lt \, \|\Delta\|^2 \, dt = \frac{L}{2} \|\Delta\|^2 . \end{aligned}$$

This is exactly the smoothness inequality. □

For every base point x and every parameter $\eta > 0$, consider the quadratic proxy on the original space

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

It is convenient to separate the affine first-order model

$$\ell_x(x') := f(x) + \langle \nabla f(x), x' - x \rangle ,$$

so that

$$Q_{x,\eta}(x') = \ell_x(x') + \frac{1}{2\eta} \|x' - x\|^2 .$$

We will call η the learning rate, or equivalently the stepsize. Thus a natural update is to choose x' minimizing this quadratic proxy. When $\eta \leq 1/L$, the upper half of smoothness gives

$$f(x') \leq Q_{x,\eta}(x'),$$

so this proxy is indeed a valid upper bound on the objective itself. In the Euclidean norm, the minimizer can be computed explicitly:

$$\arg\min_{x' \in E} \left\{ f(x) + \langle \nabla f(x), x' - x \rangle + \frac{1}{2\eta} \|x' - x\|_2^2 \right\} = \{ x - \eta \nabla f(x) \} .$$

So the update $x^+ = x - \eta \nabla f(x)$ is exactly the usual ℓ_2-gradient descent step. The rest of the lecture asks how to express the same proxy minimizer once the Euclidean norm is replaced by an arbitrary norm.

7.2 Norm-Ball Oracles and Steepest-Descent Updates

Definition 7.2 (Linear minimization oracle on a compact convex set)**.** Let $K \subseteq E$ be nonempty, compact, and convex. By a linear minimization oracle on K, we mean any map

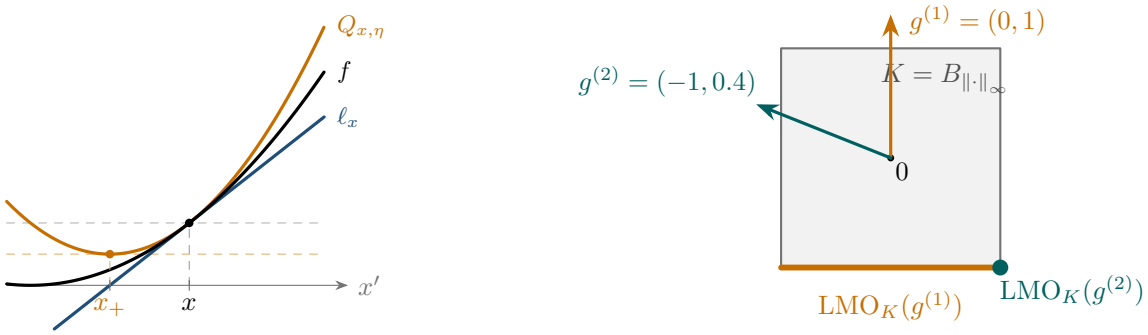
$$\mathrm{LMO}_K : E^* \to E$$

such that

$$\forall g \in E^*, \qquad \mathrm{LMO}_K(g) \in \arg\min_{s \in K} \langle g, s \rangle .$$

Since K is compact, such a map exists. The vector $\mathrm{LMO}_K(g) \in K$ is called the LMO output for the linear functional g over K.

In this lecture we use the LMO only on the norm ball $B_{\|\cdot\|}$, where it provides a convenient geometric description of steepest-descent steps.



(a) Proxy step in the Euclidean norm.

(b) LMO on the $\|\cdot\|_\infty$-ball.

Figure 1: (a) With $f(y) = \frac{1}{2}y^2$, $x = 1$, and $\eta = 1/L = 1/2$, the affine lower bound ℓ_x (blue) and the quadratic upper model $Q_{x,\eta}$ (orange) both pass through $(x, f(x))$; the minimizer of $Q_{x,\eta}$ is the proxy step $x_+ = x - \eta\nabla f(x)$. (b) For $K = B_{\|\cdot\|_\infty} \subset \mathbb{R}^2$, the LMO $\mathrm{LMO}_K(g) = \arg\min_{s \in K} \langle g, s \rangle$ may return an entire face (orange) or a unique corner (teal).

Fix a norm $\|\cdot\|$ on E. We write its closed unit ball as

$$B_{\|\cdot\|} := \{v \in E : \|v\| \le 1\},$$

and we write the associated dual norm on E^* as

$$\forall g \in E^*, \qquad \|g\|_* := \sup_{\|v\| \le 1} \langle g, v \rangle .$$

Proposition 7.2 (Characterizing all steepest-descent proxy minimizers via the norm-ball LMO)**.** *Let $f : E \to \mathbb{R}$ be differentiable with respect to $\|\cdot\|$. Fix $x \in E$ and let $\eta > 0$. Then*

$$\arg\min_{x' \in E} Q_{x,\eta}(x') = \left\{ x + \eta \|\nabla f(x)\|_* \; s : s \in \arg\min_{u \in B_{\|\cdot\|}} \langle \nabla f(x), u \rangle \right\}.$$

In particular, if $\mathrm{LMO}_{B_{\|\cdot\|}} : E^* \to E$ is any linear minimization oracle on the unit ball, then

$$x^+ := x + \eta \|\nabla f(x)\|_* \, \mathrm{LMO}_{B_{\|\cdot\|}}(\nabla f(x))$$

is one such minimizer. If f is L-smooth and $\eta \le 1/L$, then every such update minimizes a valid quadratic upper bound on f at x.

Proof of Proposition 7.2. Write $g := \nabla f(x)$, and write $x' = x + h$. Since

$$Q_{x,\eta}(x+h) = f(x) + \langle g, h\rangle + \frac{1}{2\eta}\|h\|^2,$$

minimizing $Q_{x,\eta}(x')$ over $x' \in E$ is equivalent to minimizing

$$\langle g, h\rangle + \frac{1}{2\eta}\|h\|^2$$

over $h \in E$. Every $h \in E$ can be written as $h = ru$ with $r := \|h\| \geq 0$ and $u \in B_{\|\cdot\|}$. Then

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

For fixed r, the direction u should minimize $\langle g, u\rangle$ over $B_{\|\cdot\|}$. Since $B_{\|\cdot\|}$ is centrally symmetric, $\min_{v\in B_{\|\cdot\|}} \langle g, v\rangle = -\max_{v\in B_{\|\cdot\|}} \langle g, v\rangle = -\|g\|_*$. So the problem reduces to

$$\min_{r\geq 0}\left\{-r\|g\|_* + \frac{r^2}{2\eta}\right\},$$

whose unique minimizer is $r = \eta\|g\|_*$. Hence

$$\arg\min_{x'\in E} Q_{x,\eta}(x') = \left\{x + \eta\|g\|_*\ s : s \in \arg\min_{u\in B_{\|\cdot\|}} \langle g, u\rangle\right\},$$

which is the claimed formula. The final sentence follows from $f(x') \leq Q_{x,\eta}(x')$ when f is L-smooth and $\eta \leq 1/L$. □

We already met the dual norm in Example 4.3; here it reappears in algorithmic form. The key point is that

$$\|g\|_* = \sup_{v\in B_{\|\cdot\|}} \langle g, v\rangle = \mathbf{1}^*_{B_{\|\cdot\|}}(g),$$

so $\|\cdot\|_*$ is the support function, equivalently the convex conjugate of the indicator, of the primal unit ball. Under the finite-dimensional identification $E^{**} \simeq E$, the maximizers of $\langle g, v\rangle$ over $B_{\|\cdot\|}$ are exactly the subgradients of $\|\cdot\|_*$ at g. Therefore

$$\arg\min_{v\in B_{\|\cdot\|}} \langle g, v\rangle = -\,\partial\,\|\cdot\|_*(g),$$

so normalized steepest-descent directions are precisely the negatives of dual-norm subgradients. Likewise, Proposition 7.2 is really computing

$$-\,\partial\left(\frac{1}{2\eta}\|\cdot\|^2\right)^*(g) = -\,\partial\left(\frac{\eta}{2}\|\cdot\|_*^2\right)(g) = -\,\eta\|g\|_*\,\partial\,\|\cdot\|_*(g).$$

This is the quadratic case of a more general primal-dual power-pair identity, which we leave as an exercise below.

Definition 7.3 (Normalized and unnormalized steepest-descent updates)**.** Let $f : E \to \mathbb{R}$ be differentiable, fix a linear minimization oracle $\mathrm{LMO}_{B_{\|\cdot\|}} : E^* \to E$ on the norm ball, and let

$\eta > 0$.

- Steepest-Descent (Unnormalized): $x_{t+1} = x_t + \eta \|\nabla f(x_t)\|_* \operatorname{LMO}_{B_{\|\cdot\|}}(\nabla f(x_t))$.
- Steepest-Descent (Normalized): $x_{t+1} = x_t + \eta \operatorname{LMO}_{B_{\|\cdot\|}}(\nabla f(x_t))$.

Thus the same norm-ball LMO output gives rise to two related algorithms. The normalized method uses only the direction information, while the unnormalized method rescales that direction by $\|\nabla f(x_t)\|_*$. Proposition 7.2 shows that the unnormalized update is exactly the steepest-descent proxy minimizer, so it is this update that we analyze in the rest of the lecture. The normalized variant will return later when we study Frank–Wolfe and related linear-oracle methods.

7.3 The Descent Lemma and Its Two Bridges

Lemma 7.3 (One-step descent lemma for unnormalized steepest descent)**.** *Let $f : E \to \mathbb{R}$ be differentiable and L-smooth with respect to $\|\cdot\|$. For every $x \in E$ and every $\eta \geq 0$, define*

$$x_+ := x + \eta \|\nabla f(x)\|_* \operatorname{LMO}_{B_{\|\cdot\|}}(\nabla f(x)).$$

Then

$$f(x_+) \leq f(x) - \eta \left(1 - \frac{L\eta}{2}\right) \|\nabla f(x)\|_*^2. \tag{1}$$

Up to this point, no convexity has been used: Lemma 7.3 is a pure smoothness statement, and it already gives a one-step decrease for every smooth function. To turn this into a convergence statement for the global suboptimality $f(x_t) - f(x^\star)$, we still need a bridge from the local quantity $\|\nabla f(x_t)\|_*^2$ to the global gap. The rest of the lecture develops two such bridges. Strong convexity gives a direct linear relation and hence a linear rate, while plain convexity together with a radius bound gives a quadratic recursion and hence an $O(1/T)$ rate.

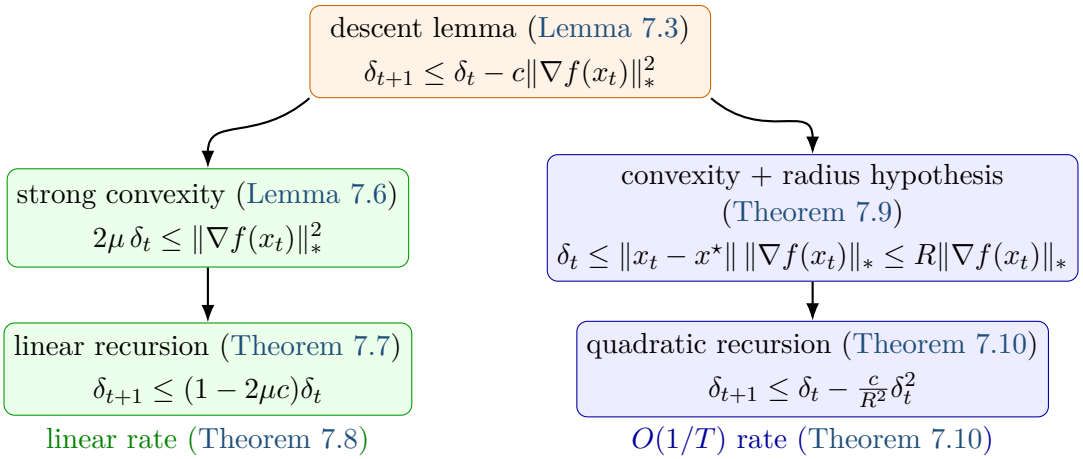


Figure 2: Two ways to turn the same descent lemma into a convergence theorem. The left branch is the strong-convexity route used in Theorems 7.7 and 7.8: strong convexity gives a direct lower bound on $\|\nabla f(x_t)\|_*^2$ in terms of the gap δ_t, which closes to a linear recursion. The right branch is the plain-convexity route used in Theorems 7.9 and 7.10: convexity gives $\delta_t \le \|x_t - x^\star\|\,\|\nabla f(x_t)\|_*$, and the extra radius hypothesis $\|x_t - x^\star\| \le R$ converts this into a quadratic recursion.

7.4 Strongly Convex Functions

Definition 7.4 (μ-strong convexity with respect to a norm)**.** Let $\mu > 0$. We say that $f : E \to (-\infty, +\infty]$ is μ-strongly convex with respect to $\|\cdot\|$ if

$$\forall x, y \in E,\ \forall \theta \in [0,1], \qquad f\big((1-\theta)x + \theta y\big) \le (1-\theta)f(x) + \theta f(y) - \frac{\mu}{2}\theta(1-\theta)\,\|x-y\|^2 .$$

Equivalently, $\operatorname{dom} f$ is convex and the restriction

$$f|_{\operatorname{dom} f} : \operatorname{dom} f \to \mathbb{R}$$

is μ-strongly convex with respect to $\|\cdot\|$ in the ordinary real-valued sense.

Lemma 7.4 (Differentiable characterization of strong convexity)**.** *Let $f : E \to \mathbb{R}$ be differentiable and let $\mu > 0$. Then the following are equivalent:*

(i) *f is μ-strongly convex with respect to $\|\cdot\|$ in the sense of Definition 7.4.*

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

Proof idea of Lemma 7.4. For differentiable real-valued f, the implication (i)⇒(ii) comes from applying strong convexity along the segment $x+t(y-x)$, rearranging, and letting $t \downarrow 0$. Conversely, (ii)⇒(i) follows by applying the first-order lower bound at the midpoint $z = (1-\theta)x+\theta y$ separately to x and y, and then averaging.

Lemma 7.5 (Proper closed strongly convex functions have unique minimizers)**.** *Let $f : E \to (-\infty, +\infty]$ be proper, closed, and μ-strongly convex with respect to $\|\cdot\|$, where $\mu > 0$. Then f attains its minimum at a unique point $x^\star \in E$.*

Proof idea of Lemma 7.5. If $\operatorname{dom} f = E$ and f is differentiable, then the earlier first-order strong-convexity inequality already gives a quadratic lower bound

$$f(y) \geq f(x_0) + \langle \nabla f(x_0), y - x_0 \rangle + \frac{\mu}{2} \|y - x_0\|^2,$$

so any minimizer must lie in a bounded region. In the general extended-value setting, one first proves the same bounded-sublevel conclusion using a subgradient at a point in $\operatorname{ri}(\operatorname{dom} f)$, and then uses a compact truncated slice of $\operatorname{epi} f$ to show that the infimum is attained.

Lemma 7.6. *Let $f : E \to \mathbb{R}$ be differentiable, L-smooth with respect to $\|\cdot\|$, and μ-strongly convex with respect to $\|\cdot\|$. Let $x^\star \in E$ be the unique minimizer given by Lemma 7.5. Then*

$$\forall x \in E, \qquad 2\mu(f(x) - f(x^\star)) \leq \|\nabla f(x)\|_*^2 \leq 2L(f(x) - f(x^\star)).$$

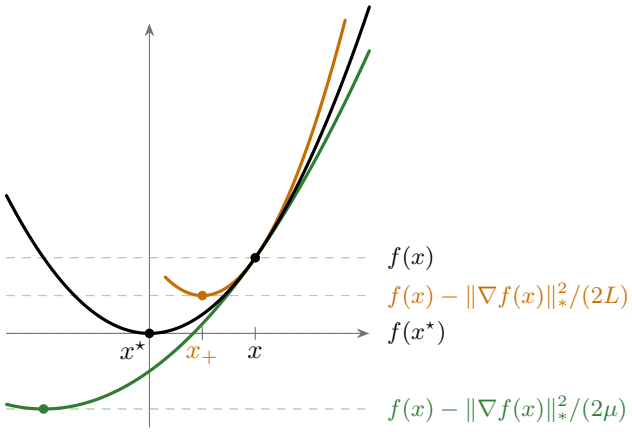


Figure 3: Picture behind Lemma 7.6. The L-smooth upper model (orange) and the μ-strong-convexity lower model (green) are both tangent to f at x. The orange parabola's minimum $f(x) - \|\nabla f(x)\|_*^2/(2L)$, attained at the proxy step x_+, sits above $f(x^\star)$; the green parabola's minimum $f(x) - \|\nabla f(x)\|_*^2/(2\mu)$ sits below. Together they sandwich the gap $f(x) - f(x^\star)$.

Theorem 7.7 (One-step suboptimality recursion under strong convexity)**.** *Let $f : E \to \mathbb{R}$ be differentiable, L-smooth with respect to $\|\cdot\|$, and μ-strongly convex with respect to $\|\cdot\|$. Let $x^\star \in E$ be the unique minimizer given by Lemma 7.5. Fix $\eta \in (0, 2/L]$, and define the unnormalized steepest-descent iterates by*

$$\forall t \geq 0, \qquad x_{t+1} = x_t + \eta d(x_t),$$

where $d(x_t)$ is any unnormalized steepest-descent update at x_t. For every $t \geq 0$, define

$$\delta_t := f(x_t) - f(x^\star).$$

Then

$$\forall t \geq 0, \qquad \delta_{t+1} \leq \left(1 - 2\mu\eta\left(1 - \frac{L\eta}{2}\right)\right)\delta_t.$$

Theorem 7.8 (Linear rate under strong convexity)**.** *Under the hypotheses of Theorem 7.7,*

$$\forall T \geq 0, \qquad \delta_T \leq \left(1 - 2\mu\eta\left(1 - \frac{L\eta}{2}\right)\right)^T \delta_0.$$

In particular, if $\eta = 1/L$, then

$$\forall T \geq 0, \qquad f(x_T) - f(x^\star) \leq \left(1 - \frac{\mu}{L}\right)^T (f(x_0) - f(x^\star)).$$

The strong-convexity branch closes because Lemma 7.6 controls the gradient norm directly by the gap. If strong convexity is unavailable, the descent lemma still holds, but we now need a different way to lower-bound $\|\nabla f(x_t)\|_*^2$. Convexity alone gives only

$$\delta_t \leq \|x_t - x^\star\| \, \|\nabla f(x_t)\|_* ,$$

so a radius bound becomes the missing ingredient.

7.5 Convex Functions without Strong Convexity

Theorem 7.9 (One-step suboptimality recursion under convexity)**.** *Let $f : E \to \mathbb{R}$ be convex, differentiable, and L-smooth with respect to $\|\cdot\|$. Assume that f attains its minimum at some $x^\star \in E$. Fix $\eta \in (0, 2/L]$, and define the unnormalized steepest-descent iterates by*

$$\forall t \geq 0, \qquad x_{t+1} = x_t + \eta d(x_t),$$

where $d(x_t)$ is any unnormalized steepest-descent update at x_t. For every $t \geq 0$, define

$$\delta_t := f(x_t) - f(x^\star).$$

Then

$$\forall t \geq 0, \qquad \text{if } x_t \neq x^\star, \text{ then} \qquad \delta_{t+1} \leq \delta_t - \frac{\eta\left(1 - \frac{L\eta}{2}\right)}{\|x_t - x^\star\|^2}\delta_t^2.$$

Theorem 7.10 (Fixed-step general-norm steepest-descent rate under a radius hypothesis)**.** *Under the hypotheses of Theorem 7.9, assume in addition that there exists $R > 0$ such that*

$$\forall t \geq 0, \qquad \|x_t - x^\star\| \leq R.$$

Then

$$\forall T \geq 0, \qquad \delta_T \leq \frac{1}{\delta_0^{-1} + \frac{\eta\left(1 - \frac{L\eta}{2}\right)}{R^2}T}.$$

In particular, if $\eta = 1/L$, then

$$\forall T \geq 1, \qquad f(x_T) - f(x^\star) \leq \frac{2LR^2(f(x_0) - f(x^\star))}{2LR^2 + T(f(x_0) - f(x^\star))} \leq \frac{2LR^2}{T}.$$

Corollary 7.11 (Bounded initial sublevel set implies the radius hypothesis)**.** *Under the hypotheses of Theorem 7.10, assume instead that the initial sublevel set*

$$S_0 := \{x \in E : f(x) \leq f(x_0)\}$$

is bounded in the norm $\|\cdot\|$, and define

$$R := \sup\{\|x - x^\star\| : x \in S_0\} < +\infty.$$

Then the conclusions of Theorem 7.10 hold with this value of R.

Proof of Corollary 7.11. By Lemma 7.3,

$$f(x_{t+1}) \leq f(x_t) - \eta\left(1 - \frac{L\eta}{2}\right)\|\nabla f(x_t)\|_*^2 \leq f(x_t),$$

since $\eta \in (0, 2/L]$. Hence every iterate belongs to S_0, so

$$\forall t \geq 0, \qquad \|x_t - x^\star\| \leq R.$$

Applying Theorem 7.10 yields the claimed bounds. □

The only remaining issue is whether this radius hypothesis is automatic in useful situations. For general norms it need not be, so Theorem 7.10 is stated conditionally. In the Euclidean norm, however, fixed-step gradient descent enjoys a stronger monotonicity property: the distance to any minimizer never increases.

Theorem 7.12 (Unconditional Euclidean gradient-descent rate)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex, differentiable, and L-smooth with respect to $\|\cdot\|_2$. Let $x^\star \in \arg\min f$. Fix $\eta \in (0, 1/L]$, and define*

$$\forall t \geq 0, \qquad x_{t+1} = x_t - \eta\nabla f(x_t).$$

Then

$$\forall t \geq 0, \qquad \|x_{t+1} - x^\star\|_2 \leq \|x_t - x^\star\|_2.$$

Thus, taking $R := \|x_0 - x^\star\|_2$ in Theorem 7.10, we obtain

$$\forall T \geq 1, \qquad f(x_T) - f(x^\star) \leq \frac{1}{\delta_0^{-1} + \frac{\eta\left(1 - \frac{L\eta}{2}\right)}{\|x_0 - x^\star\|_2^2} T}.$$

In particular, if $\eta = 1/L$, then

$$\forall T \geq 1, \qquad f(x_T) - f(x^\star) \leq \frac{2L\|x_0 - x^\star\|_2^2 (f(x_0) - f(x^\star))}{2L\|x_0 - x^\star\|_2^2 + T(f(x_0) - f(x^\star))} \leq \frac{2L\|x_0 - x^\star\|_2^2}{T}.$$

Proof of Theorem 7.12. First we show that the distance to the minimizer is nonincreasing. Write $g_t := \nabla f(x_t)$. Expanding the squared distance gives

$$\|x_{t+1} - x^\star\|_2^2 = \|x_t - x^\star\|_2^2 - 2\eta \langle g_t, x_t - x^\star \rangle + \eta^2 \|g_t\|_2^2 .$$

Rearranging,

$$\langle g_t, x_t - x^\star \rangle = \frac{1}{2\eta}\Big(\|x_t - x^\star\|_2^2 - \|x_{t+1} - x^\star\|_2^2\Big) + \frac{\eta}{2}\|g_t\|_2^2 .$$

By convexity,

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

By Lemma 7.3 in the Euclidean norm and because $\eta \le 1/L$,

$$f(x_{t+1}) \le f(x_t) - \eta\left(1 - \frac{L\eta}{2}\right)\|g_t\|_2^2 \le f(x_t) - \frac{\eta}{2}\|g_t\|_2^2 .$$

Combining the last two displays yields

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

Hence

$$\|x_{t+1} - x^\star\|_2 \le \|x_t - x^\star\|_2 .$$

Thus $\|x_t - x^\star\|_2 \le \|x_0 - x^\star\|_2$ for all $t \ge 0$, so Theorem 7.10 applies with $R := \|x_0 - x^\star\|_2$ and yields the displayed bounds. □

Proofs

Proof of Lemma 7.3. Write $g := \nabla f(x)$ and $s := \mathrm{LMO}_{B_{\|\cdot\|}}(g)$. Since $\langle g, s\rangle = \min_{\|v\|\le 1}\langle g, v\rangle = -\max_{\|v\|\le 1}\langle g, v\rangle = -\|g\|_*$ by central symmetry of $B_{\|\cdot\|}$, the vector $u := \|g\|_* \, s$ satisfies

$$\langle g, u\rangle = -\|g\|_*^2 , \qquad \|u\| = \|g\|_* .$$

Since $x_+ = x + \eta u$, applying smoothness at x with $h = \eta u$ gives

$$f(x_+) \le f(x) + \eta\langle g, u\rangle + \frac{L\eta^2}{2}\|u\|^2 = f(x) - \eta\left(1 - \frac{L\eta}{2}\right)\|g\|_*^2 ,$$

which is (1). □

Proof of Theorem 7.7. For every $t \ge 0$, Lemma 7.3 gives

$$\delta_{t+1} \le \delta_t - \eta\left(1 - \frac{L\eta}{2}\right)\|\nabla f(x_t)\|_*^2 .$$

By the lower bound in Lemma 7.6,

$$\|\nabla f(x_t)\|_*^2 \ge 2\mu\delta_t .$$

Substituting this inequality proves

$$\delta_{t+1} \le \left(1 - 2\mu\eta\left(1 - \frac{L\eta}{2}\right)\right)\delta_t .$$

□

Proof of Theorem 7.8. Set

$$q := 1 - 2\mu\eta\left(1 - \frac{L\eta}{2}\right).$$

By Theorem 7.7,

$$\forall t \geq 0, \qquad \delta_{t+1} \leq q\,\delta_t.$$

If $\delta_0 = 0$, then every iterate is already optimal and the claim is trivial. So assume $\delta_0 > 0$. Because $\eta \in (0, 2/L]$, we have

$$0 \leq \eta\left(1 - \frac{L\eta}{2}\right) \leq \frac{1}{2L}.$$

Applying Lemma 7.6 at x_0 gives

$$2\mu\delta_0 \leq \|\nabla f(x_0)\|_*^2 \leq 2L\delta_0,$$

so $\mu \leq L$. Hence

$$0 \leq 2\mu\eta\left(1 - \frac{L\eta}{2}\right) \leq \frac{\mu}{L} \leq 1,$$

so $q \in [0, 1]$. Iterating the one-step recurrence yields

$$\delta_T \leq q^T \delta_0.$$

If $\eta = 1/L$, then

$$q = 1 - \frac{\mu}{L},$$

which gives the displayed specialization. □

Proof of Theorem 7.9. Fix $t \geq 0$. If $x_t = x^\star$, then $\delta_t = 0$. Since $x^\star$ minimizes f, we have

$$0 \leq \delta_{t+1} \leq \delta_t = 0,$$

so x_{t+1} is optimal as well. Thus it remains to consider the case $x_t \neq x^\star$. By convexity,

$$\delta_t = f(x_t) - f(x^\star) \leq \langle \nabla f(x_t), x_t - x^\star \rangle \leq \|x_t - x^\star\|\,\|\nabla f(x_t)\|_*.$$

Therefore

$$\|\nabla f(x_t)\|_*^2 \geq \frac{\delta_t^2}{\|x_t - x^\star\|^2}.$$

By Lemma 7.3,

$$\delta_{t+1} \leq \delta_t - \eta\left(1 - \frac{L\eta}{2}\right)\|\nabla f(x_t)\|_*^2.$$

Substituting the previous lower bound gives

$$\delta_{t+1} \leq \delta_t - \frac{\eta\left(1 - \frac{L\eta}{2}\right)}{\|x_t - x^\star\|^2}\delta_t^2.$$

□

Proof of Theorem 7.10. Set

$$c := \frac{\eta\left(1 - \frac{L\eta}{2}\right)}{R^2}.$$

If $x_t = x^\star$, then $\delta_t = \delta_{t+1} = 0$, so $\delta_{t+1} \le \delta_t - c\delta_t^2$ is automatic. If $x_t \neq x^\star$, then $\|x_t - x^\star\| \le R$, so Theorem 7.9 gives

$$\delta_{t+1} \le \delta_t - \frac{\eta\left(1 - \frac{L\eta}{2}\right)}{\|x_t - x^\star\|^2}\delta_t^2 \le \delta_t - c\delta_t^2.$$

Hence

$$\forall t \ge 0, \qquad \delta_{t+1} \le \delta_t - c\delta_t^2.$$

If $\delta_t = 0$ for some t, then x_t is optimal and the recurrence implies $\delta_{t'} = 0$ for all $t' \ge t$. So it remains to consider the case $\delta_t > 0$. Then

$$\delta_{t+1} \le \delta_t(1 - c\delta_t).$$

Since $\delta_{t+1} \ge 0$, we have $0 \le c\delta_t \le 1$. Therefore

$$\frac{1}{\delta_{t+1}} \ge \frac{1}{\delta_t(1 - c\delta_t)} = \frac{1}{\delta_t} \cdot \frac{1}{1 - c\delta_t} \ge \frac{1}{\delta_t} + c,$$

where the last step uses $1/(1-z) \ge 1 + z$ for all $z \in [0, 1]$. Summing from $t = 0$ to $t = T - 1$ gives

$$\frac{1}{\delta_T} \ge \frac{1}{\delta_0} + cT.$$

Taking reciprocals yields

$$\delta_T \le \frac{1}{\delta_0^{-1} + cT} = \frac{1}{\delta_0^{-1} + \frac{\eta\left(1 - \frac{L\eta}{2}\right)}{R^2}T}.$$

If $\eta = 1/L$, then $c = 1/(2LR^2)$ and hence for every $T \ge 1$,

$$\delta_T \le \frac{1}{\delta_0^{-1} + \frac{T}{2LR^2}} = \frac{2LR^2\delta_0}{2LR^2 + T\delta_0} \le \frac{2LR^2}{T}.$$

□

Exercises

1. Compute $\mathrm{LMO}_K(g)$ explicitly when K is the ℓ_2-, ℓ_1-, or ℓ_∞-unit ball.
2. Let $1 < p < \infty$, and let q be defined by $\frac{1}{p} + \frac{1}{q} = 1$. Show that

   $$\left(\frac{1}{p}\|\cdot\|^p\right)^* = \frac{1}{q}\|\cdot\|_*^q.$$

   Also identify the endpoint cases $p = 1, q = \infty$ and $p = \infty, q = 1$, namely

   $$(\|\cdot\|)^* = \mathbf{1}_{\{g \in E^* : \|g\|_* \le 1\}} \qquad \text{and} \qquad \left(\mathbf{1}_{B_{\|\cdot\|}}\right)^* = \|\cdot\|_*.$$

   Deduce that

   $$\left(\frac{1}{2\eta}\|\cdot\|^2\right)^* = \frac{\eta}{2}\|\cdot\|_*^2.$$

3. Starting from Theorem 7.9, derive Theorem 7.10.

4. Let $f(x) = \frac{1}{2}x^\top Qx$ with $Q \succeq 0$. Work out the normalized steepest-descent directions and the unnormalized steepest-descent updates under the ℓ_∞ norm.

Deferred Proofs

Proof of Lemma 7.4. We first prove (i)⇒(ii). Fix $x, y \in E$, write $\Delta := y - x$, and for $t \in (0, 1]$ set

$$x_t := x + t\Delta = (1 - t)x + ty.$$

By μ-strong convexity,

$$f(x_t) \le (1 - t)f(x) + tf(y) - \frac{\mu}{2}t(1 - t) \|\Delta\|^2 .$$

Rearranging gives

$$f(y) \ge f(x) + \frac{f(x_t) - f(x)}{t} + \frac{\mu}{2}(1 - t) \|\Delta\|^2 .$$

Letting $t \downarrow 0$ and using differentiability at x yields

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

Next we prove (ii)⇒(i). Fix $x, y \in E$ and $\theta \in [0, 1]$, and set

$$z := (1 - \theta)x + \theta y.$$

Applying (ii) at the base point z once with x and once with y,

$$f(x) \ge f(z) + \langle \nabla f(z), x - z\rangle + \frac{\mu}{2} \|x - z\|^2 ,$$

$$f(y) \ge f(z) + \langle \nabla f(z), y - z\rangle + \frac{\mu}{2} \|y - z\|^2 .$$

Multiply the first inequality by $1 - \theta$ and the second by θ, then add them. Because

$$(1 - \theta)(x - z) + \theta(y - z) = 0,$$

the linear terms cancel, while

$$\|x - z\| = \theta \|x - y\| , \qquad \|y - z\| = (1 - \theta) \|x - y\| .$$

Therefore

$$(1 - \theta)f(x) + \theta f(y) \ge f(z) + \frac{\mu}{2}\theta(1 - \theta) \|x - y\|^2 ,$$

which is exactly (i). □

Proof of Lemma 7.5. By properness, $\operatorname{dom} f \neq \emptyset$. By Definition 7.4, $\operatorname{dom} f$ is convex, so $\operatorname{ri}(\operatorname{dom} f) \neq \emptyset$ in finite dimension. Choose $x_0 \in \operatorname{ri}(\operatorname{dom} f)$. Since f is proper, closed, and convex, Theorem 2.10 gives some $g_0 \in \partial f(x_0)$.
Fix $y \in E$. If $y \notin \operatorname{dom} f$, then $f(y) = +\infty$, so the lower bound below is automatic. If $y \in \operatorname{dom} f$,

then for each $t \in (0,1]$ the point

$$x_t := (1-t)x_0 + tyx_t := (1-t)x_0 + ty$$

lies in $\operatorname{dom} f$, and μ-strong convexity gives

$$f(x_t) \le (1-t)f(x_0) + tf(y) - \frac{\mu}{2}t(1-t)\left\|y - x_0\right\|^2.$$

Rearranging,

$$f(y) \ge f(x_0) + \frac{f(x_t) - f(x_0)}{t} + \frac{\mu}{2}(1-t)\left\|y - x_0\right\|^2.$$

Because $g_0 \in \partial f(x_0)$, convexity implies

$$f(x_t) \ge f(x_0) + \langle g_0, x_t - x_0\rangle = f(x_0) + t\left\langle g_0, y - x_0\right\rangle.$$

Substituting and letting $t \downarrow 0$ yields

$$f(y) \ge f(x_0) + \langle g_0, y - x_0\rangle + \frac{\mu}{2}\left\|y - x_0\right\|^2.$$

Hence

$$f(y) \ge f(x_0) - \left\|g_0\right\|_* \left\|y - x_0\right\| + \frac{\mu}{2}\left\|y - x_0\right\|^2.$$

The right-hand side tends to $+\infty$ as $\|y - x_0\| \to \infty$, so every sublevel set of f is bounded. Since f is closed, its sublevel sets are closed; in finite dimension they are therefore compact. Because f is proper, there exists $y_0 \in E$ with $f(y_0) < +\infty$, so the sublevel set

$$K := \{y \in E : f(y) \le f(y_0)\}K := \{y \in E : f(y) \le f(y_0)\}$$

is nonempty and compact. The continuous function

$$y \mapsto f(x_0) - \left\|g_0\right\|_* \left\|y - x_0\right\| + \frac{\mu}{2}\left\|y - x_0\right\|^2 y \mapsto f(x_0) - \left\|g_0\right\|_* \left\|y - x_0\right\| + \frac{\mu}{2}\left\|y - x_0\right\|^2$$

therefore attains its minimum on K; let $m \in \mathbb{R}$ denote that minimum. Then $f(y) \ge m$ for every $y \in K$. Consider the truncated epigraph slice

$$\mathcal{E}_K := \{(y,t) \in K \times [m, f(y_0)] : f(y) \le t\}.\mathcal{E}_K := \{(y,t) \in K \times [m, f(y_0)] : f(y) \le t\}.$$

Because f is closed, $\operatorname{epi} f \subseteq E \times \mathbb{R}$ is closed, and

$$\mathcal{E}_K = (\operatorname{epi} f) \cap \big(K \times [m, f(y_0)]\big)\mathcal{E}_K = (\operatorname{epi} f) \cap \big(K \times [m, f(y_0)]\big)$$

is compact. The second-coordinate projection

$$\pi_2(\mathcal{E}_K) = \{t \in \mathbb{R} : \exists\, y \in K,\ f(y) \le t\}\pi_2(\mathcal{E}_K) = \{t \in \mathbb{R} : \exists\, y \in K,\ f(y) \le t\}$$

is therefore compact, so it contains its minimum $t^\star$. Choose $x^\star \in K$ with $(x^\star, t^\star) \in \mathcal{E}_K$. Then $f(x^\star) \le t^\star$. Since $(x^\star, f(x^\star)) \in \operatorname{epi} f$ and $m \le f(x^\star) \le t^\star \le f(y_0)$, we also have $(x^\star, f(x^\star)) \in \mathcal{E}_K$. By minimality of $t^\star$, this forces $f(x^\star) = t^\star$. Hence $x^\star$ minimizes f on K, and therefore globally. For uniqueness, if $x^\star \ne y^\star$ were two minimizers, then μ-strong convexity with $\theta = \frac{1}{2}$ would give

$$f\Big(\frac{x^\star + y^\star}{2}\Big) \le \frac{f(x^\star) + f(y^\star)}{2} - \frac{\mu}{8}\left\|x^\star - y^\star\right\|^2 < \frac{f(x^\star) + f(y^\star)}{2},$$

contradicting minimality. □

Proof of Lemma 7.6. Fix $x \in E$. Taking the infimum over $y \in E$ in the smooth upper bound and the strong-convex lower bound gives

$$f(x)+\inf_{y\in E}\left\{\langle \nabla f(x), y-x\rangle + \frac{\mu}{2}\|y-x\|^2\right\} \le f(x^\star) \le f(x)+\inf_{y\in E}\left\{\langle \nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|^2\right\}.$$

By the same one-dimensional reduction as in Proposition 7.2,

$$\inf_{h\in E}\left\{\langle g, h\rangle + \frac{\alpha}{2}\|h\|^2\right\} = -\frac{1}{2\alpha}\|g\|_*^2 \qquad (\alpha > 0).$$

Applying this with $g = \nabla f(x)$ and $\alpha = \mu, L$ yields the three-term inequality

$$f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_*^2 \le f(x^\star) \le f(x) - \frac{1}{2L}\|\nabla f(x)\|_*^2.$$

Rearranging gives

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

□

References

[GLS88] Martin Grötschel, László Lovász, and Alexander Schrijver. The ellipsoid method. In *Geometric Algorithms and Combinatorial Optimization*, volume 2 of *Algorithms and Combinatorics*, pages 64–101. Springer, Berlin, Heidelberg, 1988.

[Grü60] Branko Grünbaum. Partitions of mass-distributions and of convex bodies by hyperplanes. *Pacific Journal of Mathematics*, 10(4):1257–1261, 1960.

[MSZ18] Sergiy Myroshnychenko, Matthew Stephen, and Nicole Zhang. Grünbaum's inequality for sections. *Journal of Functional Analysis*, 275(9):2516–2537, 2018.