CS395T: Continuous Algorithms, Part III
Mirror descent

Kevin Tian

1 Convex duality

This lecture focuses on developing the mirror descent algorithm, which is motivated by generalizing the gradient descent methods of Part II to problem settings exhibiting non-Euclidean geometries. These geometries can arise due to the structure of a constraint set of interest, or non-Euclidean function regularity properties. Specifically, consider a constrained optimization problem

$$\min_{x \in \mathcal{X}} f(x).$$

If the natural way to capture regularity of $\mathcal{X}$ is by measuring it according to a norm $\|\cdot\|$ (e.g. $\mathcal{X}$ is a ball in $\|\cdot\|$), then from the perspective of Banach space theory, it is unnatural to treat "primal points" $x \in \mathcal{X}$ the same way as "dual points," which are linear operators acting on $\mathcal{X}$. In particular, when f is differentiable, the gradient $\nabla f(x)$ is naturally a dual object, and hence regularity of $\nabla f(x)$ should be measured in the dual norm $\|\cdot\|_*$. In the Euclidean setting (which was the setting for most of Part II), this problem is not encountered because ℓ_2 norms are self-dual. To gain some intuition for this treatment of the dual space, recall that Lemma 12, Part I shows that $\langle g, x\rangle \le \|g\|_* \|x\|$, so if we have control over $\|x\|$ then dual regularity is best measured in $\|\cdot\|_*$.

Mirror descent is an algorithm based on treating updates to primal and dual variables asymmetrically; the two spaces are linked through a primal-dual mapping. It is crucial to first understand properties of this mapping, described through the language of convex duality. Specifically, just as every norm $\|\cdot\|$ on $\mathbb{R}^d$ has a corresponding *dual norm* $\|\cdot\|_*$, every (closed, proper) convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ has a corresponding *convex conjugate*, denoted f^* and defined below.

Definition 1 (Convex conjugate)**.** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be convex, closed, and proper. We denote the* convex conjugate[1] *of f by f^*, defined as*

$$f^*(y) := \sup_{x \in \mathbb{R}^d} \langle y, x\rangle - f(x), \text{ for all } y \in \mathbb{R}^d.$$

Our development will use the following basic facts about suprema of convex functions.

Lemma 1. *For a set S, let $f_s : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a convex function for all $s \in S$. Then $f_S(x) := \sup_{s \in S} f_s(x)$ is convex, and if $s^\star \in S$ attains the supremum, we have $\partial f_{s^\star}(x) \subseteq \partial f_S(x)$.*

Proof. To see the first claim, for all $s \in S$, $x, x' \in \mathbb{R}^d$, and $\lambda \in [0, 1]$, we have

$$f_s((1-\lambda)x + \lambda x') \le (1-\lambda) f_s(x) + \lambda f_s(x') \le (1-\lambda) f_S(x) + \lambda f_S(x').$$

Supremizing over s on the left-hand side gives the claim. For the second, let $g \in \partial f_{s^\star}(x)$ for some $x \in \mathbb{R}^d$, where $s^\star \in \text{argmax}_{s \in S} f_s(y)$. Then for any $x' \in \mathbb{R}^d$, we have the desired

$$f_S(x) + \langle g, x' - x\rangle = f_{s^\star}(x) + \langle g, x' - x\rangle \le f_{s^\star}(x') \le f_S(x').$$

□

[1] Sometimes, f^* is also called the *Fenchel dual* or *Legendre transform* of f.

As a corollary of Lemma 1, we observe that f^* is convex, as a supremum over convex (indeed, linear) functions in its argument y.[2] Further, Lemma 1 also characterizes the maximizing argument of the convex conjugate definition. We also note the converse holds, i.e. if $x \in \partial f^*(y)$, we have[3]

$$\begin{aligned} f^*(w) \geq f^*(y) + \langle x, w - y \rangle &\implies \langle y, x \rangle - f^*(y) \geq \langle w, x \rangle - f^*(w) \text{ for all } w \in \mathbb{R}^d \\ &\implies \langle y, x \rangle - f^*(y) \geq f^{**}(x) = f(x) \\ &\implies x \in \text{argmax}_{x \in \mathbb{R}^d} \langle y, x \rangle - f(x), \text{ since } \langle y, x \rangle - f(x) \geq f^*(y). \end{aligned}$$

We state the following claim from [Roc70], slightly extending these arguments, without proof.

Fact 1 ([Roc70])**.** *If f^* is the convex conjugate of convex, closed, and proper $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, it is convex, closed, and proper. Moreover, for $y \in \mathbb{R}^d$, $x \in \text{argmax}_{x \in \mathbb{R}^d} \langle y, x \rangle - f(x)$ iff $x \in \partial f^*(y)$.*

The following fact is also immediate from the definition of the convex conjugate.

Fact 2. *If f^* is the convex conjugate of convex, closed, and proper $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$,*

$$f^*(y) + f(x) \geq \langle y, x \rangle \text{ for all } x, y \in \mathbb{R}^d.$$

We next prove several additional useful properties of convex conjugates. The first generalizes Lemma 12, Part II, and shows that the conjugation operation is its own inverse.

Lemma 2. *For closed, proper, and convex $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, $f^{**} = f$.*

Proof. We first note that $f^{**} \leq f$ pointwise, which follows from[4]

$$\begin{aligned} f^{**}(x) := \sup_{y \in \mathbb{R}^d} \langle y, x \rangle - \left(\sup_{z \in \mathbb{R}^d} \langle y, z \rangle - f(z) \right) &= \sup_{y \in \mathbb{R}^d} \inf_{z \in \mathbb{R}^d} \langle y, x - z \rangle + f(z) \\ \leq \inf_{z \in \mathbb{R}^d} \sup_{y \in \mathbb{R}^d} \langle y, x - z \rangle + f(z) = f(x). \end{aligned}$$

Next, suppose for the sake of contradiction that $f^{**}(x) < f(x)$ for $x \in \mathbb{R}^d$, which also means $f^{**}(x) + \epsilon < f(x)$ for some $\epsilon > 0$. By the separating hyperplane theorem (Corollary 1, Part I), there exists $(u, c) \in \mathbb{R}^d \times \mathbb{R}$ such that $(x, f^{**}(x) + \epsilon) \notin \text{epi}(f)$ is separated from all $(z, a) \in \text{epi}(f)$, i.e. for all $z \in \mathbb{R}^d$ and $a \geq f(z)$, $\langle u, z \rangle + ca > \langle u, x \rangle + c(f^{**}(x) + \epsilon)$. By considering $(x, f(x)) \in \text{epi}(f)$, it is clear that $c > 0$, so we may assume $c = 1$ by scaling u appropriately. Thus, for all $z \in \mathbb{R}^d$,

$$\langle u, z \rangle + f(z) > \langle u, x \rangle + f^{**}(x) + \epsilon \iff \langle -u, x \rangle - f^{**}(x) - \epsilon > \langle -u, z \rangle - f(z).$$

Supremizing the right-hand side over z shows $\langle -u, x \rangle - f^{**}(x) > f^*(-u)$, contradicting Fact 2. □

As an immediate corollary of Lemma 2 and the last part of Fact 1, we have the following remarkable duality characterization for conjugate pairs (f, f^*) which are both differentiable.

Corollary 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, closed, and proper. Then for all $x \in \mathbb{R}^d$ and $y \in \partial f(x)$, $\langle y, x \rangle = f(x) + f^*(y)$, and $x \in \partial f^*(y)$ if $\partial f^*(y) \neq \emptyset$.*

Proof. The first claim follows from Fact 1 and Lemma 2. To see the second, if $x \notin \partial f^*(y)$, letting $z \in \partial f^*(y)$, we would have $f^*(y) = \langle y, z \rangle - f(z) > \langle y, x \rangle - f(x)$, a contradiction. □

Thus, ∇f and ∇f^* are inverses when both always exist, and further, these operations act by mapping points to the maximizing arguments in the definitions of f and f^*. In the constrained case, we give an example of a statement one can show regarding bijectivity of ∇f and ∇f^*.

Lemma 3 ([Roc70])**.** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be convex and of Legendre type (f is differentiable everywhere in $\mathcal{X} := \text{int}(\{x \in \mathbb{R}^d \mid f(x) < \infty\}) \neq \emptyset$, and $\nabla f \to \infty$ as x approaches the boundary of $\mathcal{X}$). Then ∇f and ∇f^* are bijections between $\mathcal{X}$ and $\mathcal{X}^* := \text{int}(\{y \in \mathbb{R}^d \mid f^*(y) < \infty\})$.*

[2] In fact, this shows that the convex conjugate of a nonconvex f is also convex.

[3] Here we used that the conjugate of the conjugate is the original function, see Lemma 2.

[4] A mnemonic to remember the inequality $\sup\inf \leq \inf\sup$ is that it is preferable to go first in a game.

When f and f^* are twice-differentiable, differentiating the identity $\nabla f^*(\nabla f(x)) = x$ in x shows

$$\left(\nabla^2 f^*(\nabla f(x))\right) \nabla^2 f(x) = \mathbf{I}_d \iff \nabla^2 f^*(\nabla f(x)) = \left(\nabla^2 f(x)\right)^{-1}.$$

This suggests that conjugate pairs have inverse second-order regularity properties. We formalize this in the following, without requiring twice-differentiability of the functions. The proof is somewhat technical, as it requires handling edge cases due to non-differentiabilty, so we omit it.

Lemma 4 ([KST09])**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, closed, and proper. Then if f is L-smooth in $\|\cdot\|$, f^* is $\frac{1}{L}$-strongly convex in $\|\cdot\|_*$, and if f^* is $\frac{1}{L}$-strongly convex in $\|\cdot\|$, f^* is L-smooth in $\|\cdot\|_*$.*

One interesting consequence of Lemma 4 is the following fact.

Corollary 2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, closed, proper, and L-smooth in $\|\cdot\|$. Then*

$$f(x') - f(x) - \langle \nabla f(x), x' - x\rangle \ge \frac{1}{2L} \|\nabla f(x') - \nabla f(x)\|_*^2 \text{ for all } x, x' \in \mathbb{R}^d.$$

Proof. We first observe, by using Corollary 1, that

$$\begin{aligned} f(x') - f(x) - \langle \nabla f(x), x' - x\rangle &= \left(\langle \nabla f(x'), x'\rangle - f^*(\nabla f^*(x'))\right) \\ &\quad - \left(\langle \nabla f(x), x\rangle - f^*(\nabla f(x))\right) - \langle \nabla f(x), x' - x\rangle \\ &= f^*(\nabla f(x)) - f^*(\nabla f(x')) - \langle x', \nabla f(x) - \nabla f(x')\rangle . \end{aligned} \tag{1}$$

The claim follows from strong convexity of f^*, and $x' \in \partial f^*(\nabla f(x'))$ (see Remark 2, Part II). □

Adding the conclusions of Corollary 2 with x, x' interchanged, we see that

$$\langle \nabla f(x') - \nabla f(x), x' - x\rangle \ge \frac{1}{L} \|\nabla f(x') - \nabla f(x)\|_*^2, \tag{2}$$

which is sometimes known as co-coercivity of the gradient.

We conclude the section with a few examples of conjugate pairs often used in algorithm design.

1. Let $f(x) = \frac{1}{2}\|x\|^2$ for a norm $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$. Then $f^*(y) = \frac{1}{2}\|y\|_*^2$.
2. Let $\frac{1}{p} + \frac{1}{q} = 1$ for $p, q \ge 1$. Then if $f(x) = \frac{1}{p}\|x\|_p^p$, $f^*(y) = \frac{1}{q}\|y\|_q^q$.
3. Let $f(x) = \|x\|$ for a norm $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$. Then $f^*(y) = \chi_{\mathbb{B}_{\|\cdot\|_*}(1)}(y)$, the indicator function of the unit dual norm ball.
4. Let $f(x) = \sum_{i\in[d]} x_i(\log x_i - 1)$ for $x \in \mathbb{R}^d_{\ge 0}$. Then $f^*(y) = \sum_{i\in[d]} \exp(y_i)$.

Remark 1. *When f is not full-dimensional (i.e. it takes on finite values only on a low-dimensional subspace), appropriate generalizations of Corollary 1 and Lemma 4 may still hold true, but one must take more care. For example, when $f(x) := \sum_{i\in[d]} x_i \log x_i$ is the* entropy *function, defined only on the probability simplex $\mathcal{X} := \{x \in \mathbb{R}^d_{\ge 0} \mid \|x\|_1 = 1\}$ (i.e. $f(x) = \infty$ for $x \notin \mathcal{X}$), its conjugate is the* softmax *function $f^*(y) := \log(\sum_{i\in[n]} \exp(y_i))$. It is true that $\nabla f^*(\nabla f(x)) = x$ for any $x \in \mathcal{X}$, but $\nabla f(\nabla f^*(y))$ is only equal to y up to an additive shift by a multiple of $\mathbb{1}_d$.*

For the remainder of the lecture, to avoid repetitiveness, we will always assume that convex functions in question are closed and proper, which does not pose an issue in applications.

2 Proximal point methods

In this section, we introduce a conceptual framework which will motivate mirror descent. Throughout, let $f : \mathcal{X} \to \mathbb{R}$ and $\varphi : \mathcal{X} \to \mathbb{R}$ both be convex. We design an implicit method for minimizing f, subject to certain relative regularity conditions between f and φ, described below. We think of φ as a *regularizer* function, which guides our algorithm design by capturing the geometry of f.

Definition 2 (Relative conditioning)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ and $\varphi : \mathcal{X} \to \mathbb{R}$ both be convex. We say f is* L-relatively smooth *with respect to φ, or L-relatively smooth in φ, if $L\varphi - f$ is convex.*[5] *Similarly, we say f is* μ-relatively strongly convex *with respect to (or in) φ if $f - \mu\varphi$ is convex.*

[5] In this lecture, we only use relative strong convexity, but mention that much of Part II generalizes to the setting of relative smoothness, which can be a weaker restriction than smoothness. For more on this, see [BBT17, LFN18].

Our relative conditioning definitions recover our definitions in the Euclidean setting from Part II.

Lemma 5. *If $f : \mathcal{X} \to \mathbb{R}$ be convex, it is μ-strongly convex iff it is μ-relatively strongly convex in $\varphi := \frac{1}{2}\|\cdot\|_2^2$, and it is L-smooth iff it is L-relatively smooth in φ.*

Proof. The strong convexity equivalence is immediate from Definition 4, Part II, and the equality

$$\frac{1}{2}\|(1-\lambda)x+\lambda x'\|_2^2 = \frac{1-\lambda}{2}\|x\|_2^2 + \frac{\lambda}{2}\|x'\|_2^2 - \frac{\lambda(1-\lambda)}{2}\|x-x'\|_2^2. \tag{3}$$

Indeed, letting $f_\mu := f - \mu\varphi$, if f is μ-strongly convex then by definition and (3) f_μ is also convex, and if f_μ is convex then adding (3) to the definition of convexity shows that f is μ-strongly convex. Next, if $f_L := L\varphi - f$ is convex, then f is L-smooth because for all $x, x' \in \mathbb{R}^d$,

$$\begin{aligned} \frac{L}{2}\|x'\|_2^2 - f(x') = f_L(x') &\geq f_L(x) + \langle \nabla f_L(x), x'-x\rangle \\ &= \frac{L}{2}\|x\|_2^2 - f(x) + \langle Lx - \nabla f(x), x'-x\rangle \\ \iff f(x') &\leq f(x) + \langle \nabla f(x), x'-x\rangle + \frac{L}{2}\|x'-x\|_2^2, \end{aligned}$$

where we use the equivalent characterization in Lemma 6, Part II. Conversely, if f is L-smooth, reversing the sequence of claims in the above display shows f_L is convex. □

The relative conditioning notions in Definition 2 can also be captured in the language of Bregman divergences, which are a key concept in mirror descent, as a way to measure distances.

Definition 3 (Bregman divergence)**.** *Let $\varphi : \mathcal{X} \to \mathbb{R}$ be convex and differentiable. The* Bregman divergence *induced by φ is defined as*[6]

$$D_\varphi(x\|\bar{x}) := \varphi(x) - \varphi(\bar{x}) - \langle \nabla\varphi(\bar{x}), x - \bar{x}\rangle.$$

Note that $D_\varphi(x\|\bar{x})$ is the amount that the first-order extrapolation about $\bar{x}$ underestimates $\varphi(x)$, so it is always nonnegative. Additionally, for fixed $\bar{x}$, $D_\varphi(\cdot\|\bar{x})$ differs from φ by only a linear term. We record some basic observations about D_φ in the following.

Fact 3. *Let $\varphi : \mathcal{X} \to \mathbb{R}$ be convex and differentiable. Then $D_\varphi(x\|\bar{x}) \geq 0$ for all $x, \bar{x} \in \mathcal{X}$, and $D(\cdot\|\bar{x})$ is convex for any fixed $\bar{x} \in \mathcal{X}$. In general, $D_\varphi(x\|\cdot)$ is not necessarily convex for fixed $x \in \mathcal{X}$.*

In the language of Definition 3, relative smoothness is equivalent to $D_f(\cdot\|\cdot) \leq LD_\varphi(\cdot\|\cdot)$ pointwise, and relative strong convexity is equivalent to $\mu D_\varphi \leq D_f$ pointwise. Bregman divergences can be thought of as a generalized distance, since when $\varphi(x) = \frac{1}{2}\|x\|_2^2$, $D_\varphi(x\|\bar{x}) = \frac{1}{2}\|x-\bar{x}\|_2^2$. In general, however, Bregman divergences need not be symmetric in their arguments, as illustrated by Fact 3. We have already observed another fact about Bregman divergences in (1).

Fact 4. *If f and f^* are both differentiable, $D_f(x\|x') = D_{f^*}(\nabla f(x')\|\nabla f(x))$.*

We now analyze the proximal point method, which iterates updates of the form

$$x_{t+1} \leftarrow \text{argmin}_{x\in\mathcal{X}}\left\{\eta f(x) + D_\varphi(x\|x_t)\right\}, \tag{4}$$

for a step size parameter $\eta > 0$. Intuitively, each proximal point step trades off the goal of minimizing f with proximity to the previous iterate, as measured by D_φ. Of course, as $\eta \to \infty$ the proximal point method simply sets the first iterate to the minimizer of the function f, which is as hard as our original goal. In general, we will not view the proximal point method as an actual algorithm, because oracle access to (4) is frequently difficult to simulate. However, the principles behind its analysis will be very useful in Section 3 when designing algorithms.

[6] A mnemonic to remember the order of the arguments is that the function φ is convex in its first argument, so we usually think of D_φ as a function of it, parameterized by the second argument.

Theorem 1 (Proximal point method)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$, and let $f : \mathcal{X} \to \mathbb{R}$ and $\varphi : \mathcal{X} \to \mathbb{R}$ both be convex and of Legendre type.*[7] *Consider iterating the update (4) for $0 \le t < T$, from $x_0 \in \mathcal{X}$ with $\eta > 0$, and let $\bar{x} := \frac{1}{T}\sum_{t\in[T]} x_t$. Then letting $x^\star \in \text{argmin}_{x\in\mathcal{X}} f(x)$,*

$$f(\bar{x}) - f(x^\star) \le \frac{D_\varphi(x^\star \| x_0)}{\eta T}.$$

If f is further μ-relatively strongly convex with respect to φ for $\mu > 0$,

$$f(x_T) - f(x^\star) \le \frac{D_\varphi(x^\star \| x_{T-1})}{\eta}, \text{ and } D_\varphi(x^\star \| x_t) \le (1 + \eta\mu)^{-t} D_\varphi(x^\star \| x_0) \text{ for all } t \in [T].$$

Proof. Because f, φ are of Legendre type, it follows that all iterates x_t for $0 \le t \le T$, as well as $x^\star$, lie in relint$(\mathcal{X})$, since the existence of Legendre functions over the set implies $\mathcal{X}$ is open. We also observe that the first-order optimality condition on $x^\star$ shows that

$$\langle \nabla f(x^\star), x^\star - x \rangle \le 0 \text{ for all } x \in \mathcal{X}. \tag{5}$$

Next, we derive that for all $0 \le t < T$,

$$\begin{aligned} \eta(f(x_{t+1}) - f(x^\star)) &\le \langle \eta \nabla f(x_{t+1}), x_{t+1} - x^\star \rangle \\ &\le \langle -\nabla D_\varphi(x_{t+1} \| x_t), x_{t+1} - x^\star \rangle \\ &= \langle \nabla\varphi(x_t) - \nabla\varphi(x_{t+1}), x_{t+1} - x^\star \rangle \\ &= D_\varphi(x^\star \| x_t) - D_\varphi(x^\star \| x_{t+1}) - D_\varphi(x_{t+1} \| x_t) \\ &\le D_\varphi(x^\star \| x_t) - D_\varphi(x^\star \| x_{t+1}). \end{aligned} \tag{6}$$

The second-to-last line applied the following *three-point equality*, which holds for all $x, y, z \in \mathcal{X}$, and can be seen by rearranging the definition of the Bregman divergence:

$$\langle \nabla\varphi(x) - \nabla\varphi(y), y - z \rangle = D_\varphi(z \| x) - D_\varphi(z \| y) - D_\varphi(y \| x). \tag{7}$$

Summing (6) across all iterations and dividing both sides by ηT, we have the first claim, since

$$f(\bar{x}) - f(x^\star) \le \frac{1}{T}\sum_{t\in[T]} f(x_t) - f(x^\star) \le \frac{D_\varphi(x^\star \| x_0)}{\eta T}.$$

Next, in the relatively strongly convex case, we derive

$$\begin{aligned} \eta D_f(x^\star \| x_{t+1}) &\le \eta D_f(x_{t+1} \| x^\star) + \eta D_f(x^\star \| x_{t+1}) \\ &= \eta \langle \nabla f(x_{t+1}) - \nabla f(x^\star), x_{t+1} - x^\star \rangle \\ &\le \langle \eta \nabla f(x_{t+1}), x_{t+1} - x^\star \rangle \\ &\le \langle -\nabla D_\varphi(x_{t+1} \| x_t), x_{t+1} - x^\star \rangle. \end{aligned}$$

The first inequality used nonnegativity of the Bregman divergence, the second used (5), and the last used the first-order optimality condition on x_{t+1} in (4). We further have, by (7) and relative strong convexity combined with the above display, that

$$\eta\mu D_\varphi(x^\star \| x_{t+1}) \le D_\varphi(x^\star \| x_t) - D_\varphi(x^\star \| x_{t+1}).$$

Iteratively applying this equation yields the second claim.

□

Theorem 1 demonstrates that the average iterate of the proximal point method enjoys a $\frac{1}{T}$ rate of convergence in suboptimality error for f, where the initial bound is dictated by the size of the Bregman divergence from x_0 to $x^\star$. Similarly, it shows that, for strongly convex functions, the last iterate enjoys a geometric rate of convergence in the Bregman divergence. This motivates finding φ with small additive ranges over sets $\mathcal{X}$ of interest, with favorable regularity properties

[7] It is straightforward to check that another situation where this proof goes through is when f, φ are differentiable on all of $\mathbb{R}^d$, and we consider their restrictions to $\mathcal{X}$, i.e. we are solving a constrained optimization problem.

for algorithms. Interestingly, this last iterate vs. average iterate discrepancy frequently appears in analyses within the mirror descent family. We also observe that letting x_{t+1} be defined by the update rule in (4), x_{t+1} also satisfies

$$x_{t+1} = \mathrm{argmax}_{x \in \mathcal{X}} \left\{ \langle -\eta \nabla f(x_{t+1}), x \rangle + \langle \nabla \varphi(x_t), x \rangle - \varphi(x) \right\},$$

so if φ^* is differentiable, Fact 1 shows that

$$x_{t+1} = \nabla \varphi^* \left(\nabla \varphi(x_t) - \eta \nabla f(x_{t+1}) \right). \tag{8}$$

We can view $\nabla\varphi$ and $\nabla\varphi^*$ as *mirror maps* linking a primal space and a dual space. Note in particular that gradient updates are only made to $\nabla\varphi(x_t)$, rather than to x_t itself. We will adopt this perspective in Section 3, where gradients of f are used to guide a dual variable.

Remark 2. *When $\mathcal{X} = \mathbb{R}^d$ and $\varphi = \frac{1}{2}\|\cdot\|_2^2$, we can rearrange the optimality condition on x_{t+1} to derive that (4) results in the update rule $x_{t+1} \leftarrow x_t - \eta \nabla f(x_{t+1})$. We can view this as a discretization of the gradient flow (see Part II), where instead of using $\nabla f(x) \approx \nabla f(x_t)$ for a short time interval of length η, we use $\nabla f(x) \approx \nabla f(x_{t+1})$, i.e. the approximation is made at the ending point rather than the starting point. This is known as a "backward Euler discretization" (as opposed to the more conventional forward Euler), which of course cannot be implemented in closed form in general as we do not know x_{t+1} in advance. Conventional numerical analysis wisdom often states that (implicit implementations of) backward Euler schemes are more stable than forward Euler schemes. Theorem 1 gives quantitative evidence of this, as it yields a T^{-1} rate of convergence, as opposed to the slower $T^{-1/2}$ rate of the forward Euler discretization in Theorem 2, Part II.*

3 Mirror descent

In this section, we develop a discretization of the proximal point method in Section 2, which is implementable under access to a subgradient oracle for f, and as long as we can solve certain regularized subproblems in $\varphi : \mathcal{X} \to \mathbb{R}$. Concretely, we assume that for any $g \in \mathbb{R}^d$,

$$\mathrm{argmin}_{x \in \mathcal{X}} \left\{ \langle g, x \rangle + \varphi(x) \right\} \tag{9}$$

can be computed in closed form. We remark that this oracle access to φ is a special limiting case of the proximal oracle access in Definition 7, Part II, where $v \leftarrow -\frac{g}{\lambda}$ and $\lambda \to 0$ (i.e. we take the quadratic part of the proximal oracle to 0, but keep the linear portion). Moreover, by Fact 3, to implement (9) we can simply query $\nabla\varphi^*(-g)$ when we have a closed form formula for $\nabla\varphi^*$.

The standard mirror descent analysis (Theorem 2) simulates the proximal point update (4) with a subgradient oracle, in what is best viewed as a forward Euler discretization scheme. If we assume the optimization objective f is L-Lipschitz, the fact that $\|g\|_* \le L$ for any subgradient g of f (Lemma 13, Part II) gives a way to pay for the discretization error of mirror descent by using a Bregman divergence term we discarded in (6) while proving Theorem 1. Specifically, we use the following basic fact about strongly convex regularizers, immediate from Lemma 14, Part II.

Fact 5. *Let $\mathcal{X} \subseteq \mathbb{R}^d$, and let $\varphi : \mathcal{X} \to \mathbb{R}$ be 1-strongly convex in $\|\cdot\|$. For any $x, \bar{x} \in \mathcal{X}$,*

$$D_\varphi(x\|\bar{x}) \ge \frac{1}{2}\|x - \bar{x}\|^2.$$

Theorem 2 (Mirror descent)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$, let $f : \mathcal{X} \to \mathbb{R}$ be convex and L-Lipschitz in $\|\cdot\|$, and let $\varphi : \mathcal{X} \to \mathbb{R}$ be 1-strongly convex in $\|\cdot\|$ and of Legendre type.*[8] *Consider iterating the update*[9]

$$x_{t+1} \leftarrow \mathrm{argmin}_{x \in \mathcal{X}} \left\{ \langle \eta g_t, x \rangle + D_\varphi(x\|x_t) \right\}, \text{ for } g_t \in \partial f(x_t), \text{ for } 0 \le t < T, \tag{10}$$

from $x_0 \in \mathcal{X}$ with $\eta > 0$, and let $\bar{x} := \frac{1}{T}\sum_{0 \le t < T} x_t$. Then letting $x^\star \in \mathrm{argmin}_{x\in\mathcal{X}} f(x)$,

$$f(\bar{x}) - f(x^\star) \le \frac{D_\varphi(x^\star\|x_0)}{\eta T} + \frac{\eta L^2}{2}.$$

Letting $\Theta \ge D_\varphi(x^\star\|x_0)$ and $\eta \leftarrow \frac{1}{L} \cdot \sqrt{2\Theta}T^{-1/2}$, the right-hand side above is at most $\sqrt{2\Theta}LT^{-1/2}$.

[8] As with Theorem 1, this proof generalizes to the case where φ is differentiable everywhere on $\mathbb{R}^d$.

[9] We ignore the subtlety of subgradients of f at the boundary of $\mathcal{X}$, because Lipschitzness of f means that the proof is not meaningfully affected by moving points into $\mathrm{relint}(\mathcal{X})$.

Proof. We pattern our derivation from the proof of Theorem 1 in (6), up to dropping a nonnegative Bregman divergence term. Specifically, by rearranging the first-order optimality condition on x_{t+1},

$$\begin{aligned}\eta(f(x_t) - f(x^\star)) &\le \langle \eta g_t, x_t - x^\star\rangle = \langle \eta g_t, x_{t+1} - x^\star\rangle + \langle \eta g_t, x_t - x_{t+1}\rangle \\ &\le \langle -\nabla D_\varphi(x_{t+1}\|x_t), x_{t+1} - x^\star\rangle + \langle \eta g_t, x_t - x_{t+1}\rangle \\ &= D_\varphi(x^\star\|x_t) - D_\varphi(x^\star\|x_{t+1}) - D_\varphi(x_{t+1}\|x_t) + \langle \eta g_t, x_t - x_{t+1}\rangle\,.\end{aligned} \tag{11}$$

The first inequality used the definition of the subgradient, and the second inequality used first-order optimality of x_{t+1} (recall the last line follows from (7)). Next, we have

$$\begin{aligned}-D_\varphi(x_{t+1}\|x_t) + \langle \eta g_t, x_t - x_{t+1}\rangle &\le -\frac{1}{2}\left\|x_{t+1} - x_t\right\|^2 + \langle \eta g_t, x_t - x_{t+1}\rangle \\ &\le -\frac{1}{2}\left\|x_{t+1} - x_t\right\|^2 + \left\|x_{t+1} - x_t\right\|\left\|\eta g_t\right\|_* \\ &\le \frac{1}{2}\left\|\eta g_t\right\|_*^2 \le \frac{\eta^2 L^2}{2}.\end{aligned}$$

The first inequality used Fact 5, the second used the generalized Hölder's inequality for dual norms (Lemma 12, Part II), the third used Young's inequality, and the last used that Lipschitz functions have bounded subgradients (Lemma 13, Part II). Finally, combining the above displays shows that

$$\eta(f(x_t) - f(x^\star)) \le D_\varphi(x^\star\|x_t) - D_\varphi(x^\star\|x_{t+1}) + \frac{\eta^2 L^2}{2} \text{ for all } 0 \le t < T,$$

and the rest of the proof follows as in Theorem 1, by summing, dividing by ηT, and using convexity. □

As discussed in Remark 1, Theorem 2 results in a slower convergence rate than its implicit counterpart Theorem 1, due to the discretization error term $\frac{\eta L^2}{2}$. When φ is a quadratic, Theorem 2 exactly reduces to the analysis of projected gradient descent in Theorem 2, Part II.

Remark 3. *As suggested by the discussion after Theorem 1, mirror descent implements the update*

$$x_{t+1} \leftarrow \nabla\varphi^*(\nabla\varphi(x_t) - \eta g_t), \text{ for } g_t \in \partial f(x_t),$$

the forward Euler variant of the update (8). Another variant of mirror descent, sometimes called "lazy mirror descent" or "dual mirror descent," explicitly maintains a dual variable s_t, and iterates

$$s_t \leftarrow s_{t-1} - \eta g_t,\ x_{t+1} \leftarrow \nabla\varphi^*(s_t), \text{ for } g_t \in \partial f(x_t).$$

When $\nabla\varphi^*$ and $\nabla\varphi$ are inverses, these updates are the same, but in certain settings (such as in projected gradient descent) they can differ, which can make a difference in applications [DAW12].

In general, Theorem 2 can be modified to obtain faster convergence rates when f is strongly convex (as suggested by the second part of Theorem 1), but necessarily falls short of a linear convergence rate in this regime due to the $\approx \frac{1}{T}$-type lower bound in Remark 1, Part II. In the following lecture, we give an example of a setting, beyond Lipschitzness of the objective, where linear convergence rates can nonetheless be attained through a mirror descent-type algorithm.

Remark 4 (Online regret minimization)**.** *In some cases, mirror descent is motivated in a more general setting as an* online regret minimization algorithm*. In particular, suppose we receive a sequence $\{g_t\}_{0\le t<T} \in \mathbb{R}^d$, and we wish to play the following repeated game for turns $0 \le t < T$.*

1. *We choose a point $x_t \in \mathcal{X}$ (we think of $\mathcal{X}$ as describing a set of actions a player can take).*
2. *We then observe $g_t \in \mathbb{R}^d$, and incur loss $\langle g_t, x_t\rangle$.*

This game is called "online," a term used to mean that we do not know the loss vector g_t in advance, and must choose our action $x_t \in \mathcal{X}$ before observing the loss vector. The regret of a player in this game is defined as $(\frac{1}{T}\sum_{0\le t<T}\langle g_t, x_t\rangle) - (\inf_{x^\star\in\mathcal{X}}\frac{1}{T}\sum_{0\le t<T}\langle g_t, x^\star\rangle)$. In other words, the regret compares the average incurred loss by the player to the loss incurred by the best action in hindsight, had we known all the $\{g_t\}_{0\le t<T}$ (but were forced to repeatedly play the same $x^\star$).

By repeating the proof of Theorem 2 in (11), short of using the step $f(x_t) - f(x^\star) \le \langle g_t, x_t - x^\star \rangle$, we see that mirror descent gives a game strategy which incurs regret

$$\frac{1}{T}\sum_{0\le t<T} \langle g_t, x_t \rangle - \left(\inf_{x^\star \in \mathcal{X}} \frac{1}{T} \sum_{0\le t<T} \langle g_t, x^\star \rangle \right) = O\left(\frac{L\sqrt{\Theta}}{\sqrt{T}} \right),$$

provided that for some norm $\|\cdot\|$, all $\|g_t\|_* \le L$, and there is a regularizer φ which is 1-strongly convex in $\|\cdot\|$ with Bregman diameter Θ over $\mathcal{X}$ from a starting point x_0. In the special case where the $\{g_t\}_{0\le t<T}$ are taken to be subgradients of a convex function, we recover Theorem 2, so online regret minimization generalizes convex optimization. More generally, online regret minimization is often used in game theory to model repeated game dynamics with linear losses. Correspondingly, mirror descent is termed a no-regret algorithm*, because when $T \to \infty$ mirror descent gives vanishing regret, i.e. a strategy with performance comparable to the best fixed action in hindsight. We refer the reader to [Sha12] for an expanded discussion of this perspective.*

4 Multiplicative weights

We now give a concrete application of mirror descent to a prominent setting which is often useful in algorithm design; the resulting framework is termed the multiplicative weights update method. This setting is so widespread that an entire monograph about it can be found in [AHK12]. Let

$$\mathcal{X} := \left\{ x \in \mathbb{R}^d_{>0} \mid \|x\|_1 = 1 \right\} \tag{12}$$

be the interior of the d-dimensional probability simplex, and define the entropy regularizer

$$\varphi(x) := \sum_{i\in[d]} x_i \log x_i \text{ for } x \in \mathcal{X}. \tag{13}$$

We specialize our discussion in this section to these particular choices of $\mathcal{X}, \varphi$, and note that as $\nabla\varphi(x) = \log x + \mathbb{1}_d$ and φ is differentiable in $\mathcal{X}$, it is indeed of Legendre type since $\log x \to \infty$ as $x \to 0$. We mention that $\mathcal{X}$ has a natural interpretation as the set of probability distributions on $[d]$ placing positive probability on all elements. In light of Theorem 2, the reason for choosing entropy as our mirror descent regularizer over $\mathcal{X}$ is because of the following two facts.

Lemma 6. *Defining $\mathcal{X}, \varphi$ as in (12), (13), φ is 1-strongly convex with respect to $\|\cdot\|_1$ over $\mathcal{X}$.*

Proof. By Lemma 14, Part II, it suffices to prove that $\nabla^2\varphi(x)[v,v] \ge \|v\|_1^2$ for all $v \in \mathbb{R}^d$. To see this, note that $\nabla^2\varphi(x) = \mathbf{diag}\left(x^{-1}\right)$ where inversion is entrywise, so

$$\nabla^2\varphi(x)[v,v] = \sum_{i\in[d]} \frac{v_i^2}{x_i} = \left(\sum_{i\in[d]} \frac{v_i^2}{x_i} \right) \left(\sum_{i\in[d]} x_i \right) \ge \left(\sum_{i\in[d]} |v_i| \right)^2 = \|v\|_1^2.$$

The second equality used $x \in \mathcal{X}$, and the inequality follows due to Cauchy-Schwarz. □

Lemma 7. *If $x_0 := \frac{1}{d}\mathbb{1}_d$ is the uniform distribution on $[d]$, $D_\varphi(x\|x_0) < \log d$ for all $x \in \mathcal{X}$.*

Proof. We first derive a formula for $D_\varphi(x\|\bar{x})$ for any $x, \bar{x} \in \mathcal{X}$, where log and division are entrywise:

$$D_\varphi(x\|\bar{x}) = \langle x, \log x \rangle - \langle \bar{x}, \log \bar{x} \rangle - \langle x - \bar{x}, \log \bar{x} + \mathbb{1}_d \rangle = \left\langle x, \log \frac{x}{\bar{x}} \right\rangle = \sum_{i\in[d]} x_i \log \frac{x_i}{\bar{x}_i}. \tag{14}$$

Here we used that since $x, \bar{x} \in \mathbb{1}_d$, we have $\langle x - \bar{x}, \mathbb{1}_d \rangle = 0$. In the special case $\bar{x} = x_0$,

$$D_\varphi(x\|\bar{x}) = \langle x, \log x \rangle + \log(d) \langle x, \mathbb{1}_d \rangle < \log(d), \text{ since } \log z < 0 \text{ for all } z \in (0,1).$$

□

Remark 5. *The Bregman divergence in the entropic regularizer φ, defined in (14), is also known as the* Kullback-Leibler *(KL) divergence, and is a fundamental object in information theory. It is not symmetric, but happens to be jointly convex in its arguments, which is often useful.*

By combining Lemmas 6 and 7 with Theorem 2, we have an algorithm for optimizing Lipschitz functions in the ℓ_1 norm over the probability simplex. All that is left is implementing the updates required by the algorithm, which by Fact 3 and the definition in (10), amounts to computing $\nabla\varphi^*$. In the entropic setting (12), (13), the conjugate φ^* is straightforward to compute in closed form.

Lemma 8. *Defining $\mathcal{X}, \varphi$ as in (12), (13), we have*

$$\varphi^*(y) := \log\left(\sum_{i\in[d]} \exp(y_i)\right) \text{ for all } y \in \mathbb{R}^d,\ \nabla_i\varphi^*(y) = \frac{\exp(y_i)}{\sum_{j\in[d]}\exp(y_j)} \text{ for all } i \in [d].$$

Proof. By considering the Lagrangian formulation of the problem defining φ^*, we have

$$\varphi^*(y) = \max_{x\in\mathcal{X}} \langle y, x\rangle - \sum_{i\in[d]} x_i \log x_i = \min_{\lambda\in\mathbb{R}} \max_{x\in\mathbb{R}_{>0}} \langle y, x\rangle - \sum_{i\in[d]} x_i \log x_i + \lambda\left(1 - \langle \mathbb{1}_d, x\rangle\right),$$

where strong duality is by Slater's condition. The KKT conditions show that the optimal x has

$$\log x = y - \alpha\mathbb{1}_d \iff x = \exp(y - \alpha\mathbb{1}_d),$$

for some $\alpha = 1 + \lambda \in \mathbb{R}$, where log and exp are applied entrywise. Since $x \in \mathcal{X}$, we derive that $\alpha = \log(\sum_{i\in[d]}\exp(y_i))$, and the first conclusion follows by plugging in the optimal choice of x:

$$\varphi^*(y) = \langle y, x\rangle - \sum_{i\in[d]} x_i\,(y_i - \alpha) = \alpha = \log\left(\sum_{i\in[d]}\exp(y_i)\right).$$

The second conclusion follows by a direct calculation. □

In other words, $\nabla\varphi^*$ induces a probability distribution in $\mathcal{X}$ proportional to the exponential of its input. This has a particularly nice interpretation in the setting of "learning from experts," which is often how multiplicative weights is motivated. Namely, suppose that there is a panel of d purported experts, identified with elements of $[d]$, and on each day $0 \le t < T$ we need to choose a distribution of experts to trust.[10] After we choose our action $x_t \in \mathcal{X}$, the performance of each expert is revealed through a penalty vector $g_t \in [-L, L]^d$, and we incur a cost for the day given by $\sum_{i\in[d]}[x_t]_i[g_t]_i = \langle x_t, g_t\rangle$, i.e. the average performance of the experts according to our chosen distribution. Combining Lemmas 6 and 7 with Remark 4, mirror descent gives a strategy yielding

$$\begin{aligned}\frac{1}{T}\sum_{0\le t<T}\langle g_t, x_t\rangle &\le \left(\min_{x^\star\in\mathcal{X}}\frac{1}{T}\sum_{0\le t<T}\langle g_t, x^\star\rangle\right) + \frac{L\sqrt{2\log(d)}}{\sqrt{T}}\\ &= \left(\min_{i\in[d]}\frac{1}{T}\sum_{0\le t<T}[g_t]_i\right) + \frac{L\sqrt{2\log(d)}}{\sqrt{T}},\end{aligned}$$

so as long as we follow the prescription of mirror descent, we asymptotically perform as well as the best expert. In fact, we only need about $T \approx \log d$ observations before mirror descent starts giving strong guarantees. Interestingly, all penalties can be chosen completely arbitrarily (even with knowledge of our strategy) and this guarantee still holds. This is a very powerful and perhaps surprising observation, and we will see an application of it to game theory in the following lecture.

Remark 6. *More generally, choosing x_0 to minimize φ (just as $\frac{1}{d}\mathbb{1}_d$ minimizes entropy) in Theorem 2 makes it simple to bound Θ by the additive range of φ, since first-order optimality shows*

$$D_\varphi(x^\star\|x_0) = \varphi(x^\star) - \varphi(x_0) - \langle\nabla\varphi(x_0), x^\star - x_0\rangle \le \varphi(x^\star) - \varphi(x_0).$$

Now, let us derive the strategy that mirror descent with the entropy regularizer prescribes in this setting. By writing in closed form the updates (10), in light of Lemma 8, we have that

$$[x_t]_i = \frac{\exp([s_t]_i)}{\sum_{j\in[d]}\exp([s_t]_j)}, \text{ where } s_t = -\sum_{0\le\tau<t}\eta g_\tau.$$

[10] A simple example of this setting is betting markets, or predicting the weather.

In other words, entropic mirror descent exponentiates a rescaling of the negated sum of all penalties seen so far, and tells us to choose an expert $i \in [d]$ proportionally to this exponential. We can think of the update to this prescribed distribution as multiplying a set of weights pointwise by $\exp(-\eta g_t)$ and renormalizing onto the simplex $\mathcal{X}$, explaining the origin of the name "multiplicative weights." In the machine learning literature, this strategy is also known as "Hedge" or "boosting."

Remark 7. *In light of the learning from experts example, we cannot hope for a variant of regret minimization (Remark 4) where the benchmark $x^\star$ is allowed to change each day. For example, suppose there are just two experts, and each day one incurs loss L and one incurs loss $-L$. If the benchmark can change daily, it will always choose to take a penalty of $-L$, and it is straightforward to see that it is impossible for any player to achieve expected negative loss in the online setting.*

We conclude with some brief discussion on the utility of φ^*, the conjugate of entropy, in algorithm design. As mentioned in Remark 1, φ^* in Lemma 8 is often called the softmax function, and can be interpreted as a smooth approximation of the max function, which takes an input $y \in \mathbb{R}^d$ to its largest coordinate value. Indeed, without regularization, $\max_{x \in \mathcal{X}} \langle y, x\rangle$ exactly implements max when $\mathcal{X}$ is the probability simplex. More generally, we can define the family of softmax functions

$$\text{smax}_\eta(y) := \max_{x \in \mathcal{X}} \langle y, x\rangle - \eta\varphi(x) = \eta\left(\max_{x \in \mathcal{X}} \left\langle \frac{y}{\eta}, x\right\rangle - \varphi(x)\right) = \eta \log\left(\sum_{i \in [d]} \exp\left(\frac{1}{\eta} y_i\right)\right), \quad (15)$$

for any $\eta > 0$. By using Lemmas 6 and 7 with Lemma 4, we derive some useful consequences.

Corollary 3. *Let $\eta > 0$. We have for any $y \in \mathbb{R}^d$ that $\max(y) \le \text{smax}_\eta(y) \le \max(y) + \eta \log(d)$ where $\max(y) := \max_{i \in [d]} y_i$, and smax_η is $\frac{1}{\eta}$-smooth in $\|\cdot\|_\infty$.*

Oftentimes in robust machine learning (see e.g. [SW16]), minimizing a maximum of loss functions is a useful primitive,[11] but the maximum is not a smooth function. By using smax_η as an approximation to max, and applying algorithms suited for the ℓ_∞ geometry (e.g. our generalized gradient descent method in Theorem 6, Part II), we can nonetheless obtain efficient algorithms for these max-type objectives. This strategy was famously pioneered by the influential work [Nes05].

Remark 8. *It is sometimes useful to consider variants of the multiplicative weights update strategy over ℓ_p norm balls, where $p \in (1, 2)$ (as opposed to $p = 1$, as in this section). To this end, we remark that $\frac{1}{p(p-1)} \|\cdot\|_p^p$ is 1-strongly convex over $\mathbb{B}_{\|\cdot\|_p}(1)$, and $\frac{1}{2(p-1)} \|\cdot\|_p^2$ is 1-strongly convex over $\mathbb{R}^d$, both in the ℓ_p norm [BCL94]. The former is convenient because it is coordinatewise-separable, whereas the latter can be preferable due to its global strong convexity property.*

5 Stochastic mirror descent

We next explore an advantage of the mirror descent framework, compared to the gradient descent methods in Part II: its tolerance of randomness. In particular, consider the setting where we wish to optimize convex $f : \mathcal{X} \to \mathbb{R}$ with a stochastic subgradient estimator, defined as follows.

Definition 4 (Stochastic subgradient estimator)**.** *We say $\tilde{g} : \mathcal{X} \to \mathbb{R}$ is a* stochastic subgradient estimator *for convex $f : \mathcal{X} \to \mathbb{R}$ if for all $x \in \mathcal{X}$, $\mathbb{E}\tilde{g}(x) \in \partial f(x)$.*

In other words, $\tilde{g}$ is a stochastic subgradient estimator if it is an unbiased estimate of a subgradient of f. To motivate this definition, consider the problem of *stochastic convex optimization*, a fundamental paradigm in machine learning. In this problem, there is a distribution $\mathcal{D}$, each $s \in \text{supp}(\mathcal{D})$ induces a convex loss function $f(x; s)$ over $\mathcal{X}$, and we wish to (approximately) compute

$$x^\star := \text{argmin}_{x \in \mathcal{X}} f(x), \text{ where } f(x) := \mathbb{E}_{s \sim \mathcal{D}}\left[f(x; s)\right].$$

For example, we can model mean estimation, i.e. estimation of $\mathbb{E}_{s \sim \mathcal{D}}[s]$, as an instance of stochastic convex optimization by defining $f(x; s) := \frac{1}{2}\|x - s\|_2^2$. It is straightforward to see by first-order optimality that the minimizer of $f(x) := \mathbb{E}_{s \sim \mathcal{D}} \frac{1}{2}\|x - s\|_2^2$ is indeed the mean of $\mathcal{D}$. Similarly, in linear regression we view samples from a distributions as feature-label pairs $(a, b) \in \mathbb{R}^d \times \mathbb{R}$, and the goal is to minimize $f(x) = \mathbb{E}_{s = (a,b) \sim \mathcal{D}} f(x; s)$ where $f(x; s) := \frac{1}{2}(\langle a, x\rangle - b)^2$. In stochastic

[11] For example, we may want to ensure our classifier achieves a uniformly-bounded loss over groups of examples.

convex optimization, we can design a stochastic subgradient estimator $\tilde{g}(x)$ by simply drawing a fresh sample $s \sim \mathcal{D}$, and returning $\partial f(x; s)$. Linearity of expectation shows $\mathbb{E}\partial f(x; s) = \partial f(x)$, but exact access to subgradients of the population average f is impossible from finite samples.

Another example of a setting where the stochastic subgradient model yields benefits is the related problem of *empirical risk minimization*, where we draw a dataset of samples $\{s_i\}_{i\in[n]} \sim \mathcal{D}$, and directly wish to minimize the empirical risk (as opposed to the population risk), defined by

$$f(x) := \frac{1}{n}\sum_{i\in[n]} f_i(x), \text{ where } f_i(x) := f(x; s_i) \text{ for all } x \in \mathcal{X}.$$

It is common in machine learning practice to take multiple passes over the dataset, at which point there are dependencies between iterates and the dataset, and sample subgradients are no longer unbiased for the population loss. However, if the minimizer of the empirical risk has sufficiently good generalization properties, directly aiming to solve for it may still be a meaningful target. In this setting, we can design a stochastic subgradient estimator for f by returning

$$\tilde{g}(x) := \partial f_i(x), \text{ for } i \sim_{\text{unif.}} [n]. \tag{16}$$

Here, the benefit is computational rather than statistical; returning the subgradient estimator defined above only requires one subgradient query of the form $\partial f(\cdot; s)$, whereas the standard way of returning a subgradient of the empirical risk f is to average n subgradients $\partial f(\cdot; s)$. The latter oracle can be significantly more expensive to implement, requiring a full dataset pass.

To give our algorithm using the access in Definition 4, we make the simple observation that the proof of Theorem 2 extends readily to the setting where we only have access to a stochastic subgradient estimator (rather than an exact subgradient oracle), assuming a second moment bound.

Corollary 4 (Stochastic mirror descent)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$, let $f : \mathcal{X} \to \mathbb{R}$ be convex and admit a stochastic subgradient estimator $\tilde{g} : \mathcal{X} \to \mathbb{R}^d$, and let $\varphi : \mathcal{X} \to \mathbb{R}$ be 1-strongly convex in $\|\cdot\|$ and of Legendre type. Consider iterating the update*

$$x_{t+1} \leftarrow \text{argmin}_{x\in\mathcal{X}} \left\{\langle \eta\tilde{g}_t, x\rangle + D_\varphi(x\|x_t)\right\}, \text{ for } \tilde{g}_t \leftarrow \tilde{g}(x_t), \text{ for } 0 \le t < T,$$

from $x_0 \in \mathcal{X}$ with $\eta > 0$, and let $\bar{x} := \frac{1}{T}\sum_{0\le t<T} x_t$. Further, suppose that $\mathbb{E}\|\tilde{g}(x)\|_*^2 \le L^2$ for all $x \in \mathcal{X}$. Then letting $x^\star \in \text{argmin}_{x\in\mathcal{X}} f(x)$,

$$\mathbb{E}\left[f(\bar{x}) - f(x^\star)\right] \le \frac{D_\varphi(x^\star\|x_0)}{\eta T} + \frac{\eta L^2}{2}.$$

Letting $\Theta \ge D_\varphi(x^\star\|x_0)$ and $\eta \leftarrow \frac{1}{L}\cdot\sqrt{2\Theta}T^{-1/2}$, the right-hand side above is at most $\sqrt{2\Theta}LT^{-1/2}$.

Proof. Similarly to the proof of Theorem 2, we have that for all $0 \le t < T$,

$$\langle \eta\tilde{g}_t, x_t - x^\star\rangle \le D_\varphi(x^\star\|x_t) - D_\varphi(x^\star\|x_{t+1}) + \frac{\eta^2\|\tilde{g}_t\|_*^2}{2}.$$

Taking expectations on both sides, and using independence of $\tilde{g}$ from x_t and $x^\star$ (as $x^\star$ is a deterministic point) yields, conditioned on x_t, that

$$\begin{aligned}\langle \eta g_t, x_t - x^\star\rangle &= \mathbb{E}\left[\langle \eta\tilde{g}_t, x_t - x^\star\rangle \mid x_t\right] \\ &\le D_\varphi(x^\star\|x_t) - \mathbb{E}\left[D_\varphi(x^\star\|x_{t+1}) \mid x_t\right] + \frac{\eta^2 L^2}{2}, \text{ for } g_t \in \partial f(x_t).\end{aligned}$$

Here, we used the second moment bound assumption on $\tilde{g}$. At this point, the remainder of the proof is identical to Theorem 2, where we apply the law of iterated expectations:

$$\mathbb{E}\left[f(\bar{x}) - f(x^\star)\right] \le \frac{1}{T}\mathbb{E}\left[\sum_{0\le t<T} f(x_t) - f(x^\star)\right] \le \frac{1}{T}\sum_{0\le t<T}\mathbb{E}\left[\langle \tilde{g}_t, x_t - x^\star\rangle \mid x_t\right].$$

□

As we can see from the proof of Corollary 4, one benefit of the mirror descent framework is that it decouples the effect of the subgradient estimator from the randomness of iterates in a way that preserves independence. This is in contrast to gradient descent proofs, where convexity is typically applied between iterates (e.g. Corollary 2, Part II), which are less robust to stochastic estimation. Indeed, it turns out that the rate of Corollary 4 is optimal even in Euclidean geometries under the stochastic oracle access stated in the problem; see e.g. [Duc18] for a proof.

6 Variance reduction

In this section, we give an application where empirical risk minimization can be solved much more efficiently than stochastic convex optimization (i.e. optimization of the population loss), by going beyond the black-box guarantees of Corollary 4. Consider an empirical risk minimization problem

$$\min_{x\in\mathbb{R}^d} f(x), \text{ where } f(x) := \frac{1}{n}\sum_{i\in[n]} f_i(x).$$

Further, suppose that f is μ-strongly convex, and each f_i is individually L-smooth. If we let $\mathcal{T}_{\text{grad}}$ denote the time it takes to query a gradient oracle of f_i for any $i\in[n]$, it is clear that we can implement a gradient oracle for f in time $n\cdot\mathcal{T}_{\text{grad}}$. Therefore, Theorem 4, Part II (i.e. well-conditioned gradient descent) shows that f can be minimized to additive error ϵ in time

$$O\left(\left(\frac{L}{\mu}\log\left(\frac{f(x_0)-f(x^\star)}{\epsilon}\right)\right)\cdot\left(n\cdot\mathcal{T}_{\text{grad}}\right)\right), \tag{17}$$

where the former term is the number of iterations and the latter is the cost per iteration. On the other hand, Corollary 4 with the estimator in (16) yields an incomparable complexity of

$$O\left(\frac{\Lambda^2R^2}{\epsilon^2}\cdot\mathcal{T}_{\text{grad}}\right),$$

where Λ^2 is a variance bound on (16), and we assume the minimizer $x^\star := \text{argmin}_{x\in\mathbb{R}^d} f(x)$ is contained in a known ball of radius R. This begs the question, can we give an algorithm which inherits both the linear convergence rate of well-conditioned gradient descent, and the improved dependence of stochastic mirror descent on the dataset size n?

This question was addressed by a sequence of works [SRB17, SZ13, JZ13], which designed *variance-reduced* gradient estimators going beyond (16), and gave corresponding custom convergence analyses, specifically for empirical risk minimization problems of the form described above. We give an overview of this variance reduction technique, primarily borrowing from arguments in [JZ13], because of the generality of this problem setting and the strong convergence guarantees this strategy can achieve. The key helper tool we use in this endeavor is the following bound.

Lemma 9. *Let $x,\bar{x}\in\mathbb{R}^d$, let $f(x)=\frac{1}{n}\sum_{i\in[n]}f_i(x)$ where each $f_i:\mathbb{R}^d\to\mathbb{R}$ is L-smooth and convex, and consider the stochastic gradient estimator*

$$\tilde{g}(x) := \nabla f_i(x) - \nabla f_i(\bar{x}) + \nabla f(\bar{x}), \text{ for } i \sim_{\text{unif.}} [n]. \tag{18}$$

Then $\mathbb{E}\left\lVert\tilde{g}\right\rVert_2^2 \le 4L\left(f(x)-f(x^\star)\right)+4L\left(f(\bar{x})-f(x^\star)\right)$.

Proof. Let $x^\star := \text{argmin}_{x\in\mathbb{R}^d} f(x)$, so that by first-order optimality, $\nabla f(x^\star)=\mathbb{0}_d$. Moreover, note that for any random vector $v\in\mathbb{R}^d$, we have

$$\mathbb{E}\left\lVert v-\mathbb{E}v\right\rVert_2^2 = \mathbb{E}\left\lVert v\right\rVert_2^2 - \left\lVert\mathbb{E}v\right\rVert_2^2 \le \mathbb{E}\left\lVert v\right\rVert_2^2. \tag{19}$$

We hence derive our desired bound:

$$\begin{aligned}\mathbb{E}\left\lVert\tilde{g}(x)\right\rVert_2^2 &\le 2\mathbb{E}\left\lVert\nabla f_i(x)-\nabla f_i(x^\star)\right\rVert_2^2 + 2\mathbb{E}\left\lVert\nabla f_i(\bar{x})-\nabla f(\bar{x})+\nabla f_i(x^\star)-\nabla f(x^\star)\right\rVert_2^2\\ &\le 2\mathbb{E}\left\lVert\nabla f_i(x)-\nabla f_i(x^\star)\right\rVert_2^2 + 2\mathbb{E}\left\lVert\nabla f_i(\bar{x})-\nabla f_i(x^\star)\right\rVert_2^2\\ &\le 4L\mathbb{E}\left[f_i(x)-f_i(x^\star)-\langle\nabla f_i(x^\star),x-x^\star\rangle\right]+4L\mathbb{E}\left[f_i(\bar{x})-f_i(x^\star)-\langle\nabla f_i(x^\star),\bar{x}-x^\star\rangle\right]\\ &= 4L\left(f(x)-f(x^\star)\right)+4L\left(f(\bar{x})-f(x^\star)\right).\end{aligned}$$

The first inequality used $\left\lVert a+b\right\rVert_2^2\le 2\left\lVert a\right\rVert_2^2+2\left\lVert b\right\rVert_2^2$, the second used (19), the third used Corollary 2 twice, and the fourth used linearity of expectation and $\nabla f(x^\star)=\mathbb{0}_d$. □

Lemma 9 shows that we can relate the variance of the estimator in (18) to the function error at the current point, plus the function error at an "anchor point" $\bar{x}$. The stochastic variance reduction scheme of [JZ13] repeatedly computes $\nabla f(\bar{x})$ for a current anchor point using n gradient queries, runs stochastic mirror descent using the estimator (18), and then resets $\bar{x}$. As we will see, the resulting gradient query complexity dramatically improves, because we only infrequently query gradients of the full empirical risk, and most iterations only require 2 new sample gradient queries.

Lemma 10. *Let $\bar{x} \in \mathbb{R}^d$, and let $f(x) = \frac{1}{n}\sum_{i\in[n]} f_i(x)$ where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and convex, and f is μ-strongly convex. Consider iterating the update, for $\tilde{g}$ in (18) and $\eta \leftarrow \frac{1}{6L}$,*

$$x_{t+1} \leftarrow x_t - \eta \tilde{g}(x_t),$$

from $x_0 \leftarrow \bar{x}$, and let $\bar{x}' := \frac{1}{T}\sum_{0\le t<T} x_t$. Then letting $x^\star := \mathrm{argmin}_{x\in\mathbb{R}^d} f(x)$, if $T \ge \frac{36L}{\mu}$,

$$\mathbb{E}\left[f(\bar{x}') - f(x^\star)\right] \le \frac{3}{4}\left(f(\bar{x}) - f(x^\star)\right).$$

Proof. We first note that the update is equivalent to the update in Corollary 4 using the stochastic gradient oracle in (18), where $\mathcal{X} = \mathbb{R}^d$ and $\varphi = \frac{1}{2}\|\cdot\|_2^2$. Therefore, repeating the proof of Corollary 4, and using the variance bound in Lemma 9, we have for all $0 \le t < T$ that conditioned on x_t,

$$\begin{aligned}\eta\left(f(x_t) - f(x^\star)\right) &\le \langle \eta\mathbb{E}\tilde{g}(x_t), x_t - x^\star\rangle \\ &\le \frac{1}{2}\|x_t - x^\star\|_2^2 - \mathbb{E}\left[\frac{1}{2}\|x_{t+1} - x^\star\|_2^2\right] + \frac{\eta^2}{2}\mathbb{E}\|\tilde{g}(x_t)\|_2^2 \\ &\le \frac{1}{2}\|x_t - x^\star\|_2^2 - \mathbb{E}\left[\frac{1}{2}\|x_{t+1} - x^\star\|_2^2\right] \\ &+ 2\eta^2 L\left(\left(f(x_t) - f(x^\star)\right) + \left(f(\bar{x}) - f(x^\star)\right)\right).\end{aligned}$$

Rearranging and using our choice of η thus shows that

$$\frac{2\eta}{3}\left(f(x_t) - f(x^\star)\right) \le \frac{1}{2}\|x_t - x^\star\|_2^2 - \mathbb{E}\left[\frac{1}{2}\|x_{t+1} - x^\star\|_2^2\right] + \frac{\eta}{3}\left(f(\bar{x}) - f(x^\star)\right).$$

Summing the above inequality across all iterations, dividing by $\frac{2}{3}\eta T$, and applying convexity,

$$\begin{aligned}\mathbb{E}\left[f(\bar{x}') - f(x^\star)\right] &\le \frac{3}{4\eta T}\|x_0 - x^\star\|_2^2 + \frac{1}{2}\left(f(\bar{x}) - f(x^\star)\right) \\ &\le \left(\frac{3}{2\eta\mu T} + \frac{1}{2}\right)\left(f(\bar{x}) - f(x^\star)\right),\end{aligned}$$

where we used strong convexity of f (Lemma 9, Part II). Our choice of T then yields the claim. □

In other words, Lemma 10 shows that stochastic variance reduction run for $O(\frac{L}{\mu})$ iterations decreases function error by a constant factor. Recursively applying this result then yields a linearly-convergent algorithm for well-conditioned empirical risk minimization, which dramatically improves upon the naïve query complexity of gradient descent (17), summarized as follows.

Theorem 3 (Stochastic variance reduction)**.** *Let $x_0 \in \mathbb{R}^d$, let $f(x) = \frac{1}{n}\sum_{i\in[n]} f_i(x)$ where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and convex, and f is μ-strongly convex. Let $\kappa := \frac{L}{\mu}$ and $\epsilon > 0$. Consider the following algorithm, initialized from $\bar{x}_0 \leftarrow x_0$ and for $0 \le k < K$.*

1. *Call the algorithm in Lemma 10 for $\lceil 36\kappa\rceil$ iterations, with $\bar{x} \leftarrow \bar{x}_k$.*
2. *Set $\bar{x}_{k+1}$ to the output, and update $k \leftarrow k + 1$.*

Then letting $x^\star := \mathrm{argmin}_{x\in\mathbb{R}^d} f(x)$, if $K \ge \log_4(\frac{f(x_0) - f(x^\star)}{\epsilon})$, we have

$$\mathbb{E}\left[f(\bar{x}_K) - f(x^\star)\right] \le \epsilon.$$

In particular, in the notation of (17), note that the cost of implementing a single call to Lemma 10 (for $O(\kappa)$ iterations) is only $O((n+\kappa)\cdot\mathcal{T}_{\mathrm{grad}})$. This should be compared to the cost of naïvely implementing gradient descent for $O(\kappa)$ iterations, which is $O(n\kappa\cdot\mathcal{T}_{\mathrm{grad}})$.

We mention that one benefit of having a linearly-convergent randomized algorithm is that in expectation guarantees can be boosted to high probability ones with small overhead. For example, suppose we wanted a variant of Theorem 3 which succeeded with probability $1 - \delta$, for some $\delta \in (0, 1)$. Because $f(x) - f(x^\star)$ is a nonnegative random variable, if x is a random point satisfying

$$\mathbb{E}\left[f(x) - f(x^\star)\right] \le \delta\epsilon,$$

Markov's inequality shows that in fact $f(x) - f(x^\star) \le \epsilon$ except with probability δ. The cost of producing such a random point x using Theorem 3 only loses an additive logarithmic factor (in the number of calls to Lemma 10) over the cost of achieving expected error ϵ.

Source material

Portions of this lecture are based on reference material in [Roc70, Sha12, Bub15, Sid23], as well as the author's own experience working in the field.

References

[AHK12] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory Comput.*, 8(1):121–164, 2012.

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

[BCL94] Keith Ball, Eric A. Carlen, and Elliott H. Lieb. Sharp uniform convexity and smoothness estimates for trace norms. *Inventiones mathematicae*, 115(1):463–482, 1994.

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[DAW12] John C. Duchi, Alekh Agarwal, and Martin J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Trans. Autom. Control.*, 57(3):592–606, 2012.

[Duc18] John C Duchi. Introductory lectures on stochastic optimization. *The Mathematics of Data*, pages 99–186, 2018.

[JZ13] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, pages 315–323, 2013.

[KST09] Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Applications of strong convexity–strong smoothness duality to learning with matrices. *CoRR*, abs/0910.0610, 2009.

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

[Nes05] Yurii Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.

[Roc70] R. Tyrell Rockafellar. *Convex Analysis.* Princeton University Press, 1970.

[Sha12] Shai Shalev-Shwartz. Online learning and online convex optimization. *Found. Trends Mach. Learn.*, 4(2):107–194, 2012.

[Sid23] Aaron Sidford. *Optimization Algorithms.* 2023.

[SRB17] Mark W. Schmidt, Nicolas Le Roux, and Francis R. Bach. Minimizing finite sums with the stochastic average gradient. *Math. Program.*, 162(1-2):83–112, 2017.

[SW16] Shai Shalev-Shwartz and Yonatan Wexler. Minimizing the maximal loss: How and why. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 793–801. JMLR.org, 2016.

[SZ13] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research*, 14(1):567–599, 2013.