Machine Learning Theory
Lecture 20: Mirror Descent

Nicholas Harvey

November 21, 2018

In this lecture we will present the Mirror Descent algorithm, which is a common generalization of Gradient Descent and Randomized Weighted Majority. This will require some preliminary results in convex analysis.

1 Conjugate Duality

A good reference for the material in this section is [5, Part E].

Definition 1.1. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a function. Define $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ by

$$f^*(y) \;=\; \sup_{x \in \mathbb{R}^n} \big(y^\mathsf{T} x - f(x) \big).$$

This is the ***convex conjugate*** or ***Legendre-Fenchel transform*** or ***Fenchel dual*** of f.

For each linear functional y, the convex conjugate $f^*(y)$ gives the the greatest amount by which y exceeds the function f. Alternatively, we can think of $f^*(y)$ as the downward shift needed for the linear function y to just touch or "support" epi f.

1.1 Examples

Let us consider some simple one-dimensional examples.

Example 1.2. Let $f(x) = cx$ for some $c \in \mathbb{R}$. We claim that $f^* = \delta_{\{c\}}$, i.e.,

$$f^*(x) \;=\; \begin{cases} 0 & (\text{if } x = c) \\ +\infty & (\text{otherwise}) \end{cases}.$$

This is called the *indicator function* of $\{c\}$.

Note that f is itself a linear functional that obviously supports epi f; so $f^*(c) = 0$. Any other linear functional $x \mapsto yx - r$ cannot support epi f for any r (we have $\sup_x (yx - cx) = \infty$ if $y \neq c$), so $f^*(y) = \infty$ if $y \neq c$. Note here that a line (f) is getting mapped to a single point (f^*). ■

Example 1.3. Let $f(x) = |x|$. We claim that $f^* = \delta_{[-1,1]}$ (the indicator function of $[-1,1]$).

For any $y \in [-1,1]$, the linear functional $x \mapsto yx$ supports epi f at the point $(0,0)$; so $f^*(y) = 0$. On the other hand, if $y > 1$ then the linear functional $x \mapsto yx - r$ cannot support epi f for any r (we have $\sup_x(yx - |x|) = \infty$ for $y > 1$), so $f^*(y) = \infty$. Similarly for $y < -1$. ■

Example 1.4. Let $f(x) = \frac{1}{2}x^\mathsf{T}x$. We claim that $f^* = f$.

We have

$$f^*(y) \;=\; \sup_{x\in\mathbb{R}^n} \left(y^\mathsf{T}x - \tfrac{1}{2}x^\mathsf{T}x\right) \;\le\; \sup_{x\in\mathbb{R}^n} \left(\|y\|_2 \|x\|_2 - \tfrac{1}{2}\|x\|_2^2 \right).$$

This upper bound is maximized when $\|x\|_2 = \|y\|_2$, and the inequality is achieved when $x = y$. Thus $f^*(y) = \frac{1}{2}y^\mathsf{T}y = f(y)$, so $f = f^*$. ■

Example 1.5 (Negative entropy). Define $f : \mathbb{R}^n_{>0} \to \mathbb{R}$ by $f(x) = \sum_{i=1}^n x_i \ln x_i$. We saw in our earlier lectures on convexity that f is convex. We claim that $f^*(y) = \sum_{i=1}^n e^{y_i - 1}$. By Claim 1.9, proving the result for $n = 1$ also establishes the general result.

By definition $f^*(y) = \sup_{z>0}(yz - z\ln z)$. The derivative of $yz - z\ln z$ is $y - \ln z - 1$. The unique critical point satisfies $z = e^{y-1}$ and it is a maximizer. Thus $f^*(y) = ye^{y-1} - e^{y-1}(y-1) = e^{y-1}$. ■

Example 1.6. Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ and let Let $f(x) = \frac{1}{2}\|x\|^2$. Then $f^* = \frac{1}{2}\|x\|_*^2$. ■

References. [3, Example 3.27].

1.2 Properties

Claim 1.7 (Young-Fenchel Inequality). For any $x, y \in \mathbb{R}^n$,

$$y^\mathsf{T}x \;\le\; f(x) + f^*(y).$$

Proof.

$$f^*(y) + f(x) \;=\; \sup_{x'\in\mathbb{R}^n} \left(y^\mathsf{T}x' - f(x')\right) + f(x) \;\ge\; \left(y^\mathsf{T}x - f(x)\right) + f(x) \;=\; y^\mathsf{T}x. \quad \square$$

Claim 1.8. f^* is closed and convex (regardless of whether f is).

Proof. For each x, define $g_x(y) = y^\mathsf{T}x - f(x)$. Note that g_x is an affine function of y, so g_x is closed and convex. As $f^* = \sup_{x\in\mathbb{R}^n} g_x$, Lemma 5.8 implies that f^* is closed and convex. $\square$

Claim 1.9 (Conjugate of Separable Function). Let $f : \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R}$ be defined by $f(x_1, x_2) = f_1(x_1) + f_2(x_2)$. Then $f^*(x_1, x_2) = f_1^*(x_1) + f_2^*(x_2)$.

Proof. Straight from the definitions, we have

$$\begin{aligned} f^*(y_1, y_2) &= \sup_{(z_1,z_2)\in\mathbb{R}^a\times\mathbb{R}^b} \big((y_1, y_2)^\mathsf{T}(z_1, z_2) - f(z_1, z_2)\big) \\ &= \sup_{(z_1,z_2)\in\mathbb{R}^a\times\mathbb{R}^b} \big(y_1^\mathsf{T} z_1 + y_2^\mathsf{T} z_2 - f_1(z_1) - f_2(z_2)\big) \\ &= \sup_{z_1\in\mathbb{R}^a} \big(y_1^\mathsf{T} z_1 - f_1(z_1)\big) \;+\; \sup_{z_2\in\mathbb{R}^b} \big(y_2^\mathsf{T} z_2 - f_2(z_2)\big) \\ &= f_1^*(y_1) \;+\; f_2^*(y_2). \end{aligned}$$

□

Claim 1.10. Suppose f is a closed, convex function. Then $f^{**} = f$.

References. [2, Proposition 7.1.1], [3, Exercise 3.39].

The following claim shows that vectors x and y achieving inequality in Claim 1.7 are rather special.

Claim 1.11. Suppose that f is closed and convex. The following are equivalent:

$$y \in \partial f(x) \tag{1.1a}$$
$$x \in \partial f^*(y) \tag{1.1b}$$
$$\langle\, y,\, x \,\rangle = f(x) + f^*(y) \tag{1.1c}$$

References. See [7, Slide 7-15], [5, Part E, Corollary 1.4.4]. In the differentiable setting, (1.1a) $\Longleftrightarrow$ (1.1c) appears in [3, pp. 95].

Proof.

(1.1a)⇒(1.1c): Suppose $y \in \partial f(x)$. Then $f^*(y) = \sup_u \big(\langle\, y,\, u \,\rangle - f(u)\big) = \langle\, y,\, x \,\rangle - f(x)$, by the subgradient inequality.

(1.1c)⇒(1.1b): For any $v \in \mathbb{R}^n$, we have

$$\begin{aligned} f^*(v) &= \sup_u \big(\langle\, v,\, u \,\rangle - f(u)\big) \\ &\geq \langle\, v,\, x \,\rangle - f(x) \\ &= \langle\, v - y,\, x \,\rangle - f(x) + \langle\, x,\, y \,\rangle \\ &= \langle\, v - y,\, x \,\rangle + f^*(y), \end{aligned}$$

by (1.1c). This shows that $x \in \partial f^*(y)$.

(1.1b)⇒(1.1a): Let $g = f^*$. Then g is closed and convex by Claim 1.8. If $x \in \partial g(y)$ then $y \in \partial g^*(x)$, by the implication (1.1a)⇒(1.1c). But $g^* = f$ by Claim 1.10, so this establishes the desired result. □

2 Bregman Divergence

A good reference for the material in this section is [8].

Let $\mathcal{X}$ be a closed convex set. Let $f : \mathcal{X} \to \mathbb{R}$ be a continuously-differentiable and convex function. The first-order approximation of f at x is

$$f(x) \;\approx\; f(y) + \langle\, \nabla f(y),\, x - y \,\rangle.$$

Since f is convex, the subgradient inequality implies that the left-hand side is at least the right-hand side. The amount by which the left-hand side exceeds the right-hand side is the Bregman divergence.

Definition 2.1. The ***Bregman divergence*** is defined to be

$$D_f(x, y) \;=\; f(x) - f(y) - \langle\, \nabla f(y),\, x - y \,\rangle.$$

2.1 Examples

Example 2.2. Define $f : \mathbb{R}^n \to \mathbb{R}$ by $f(x) = \|x\|_2^2$. Then

$$\begin{aligned} D_f(x, y) &= f(x) - f(y) - \langle\, \nabla f(y),\, x - y \,\rangle \\ &= \|x\|_2^2 - \|y\|_2^2 - 2\langle\, y,\, x - y \,\rangle \\ &= \|x\|_2^2 + \|y\|_2^2 - 2\langle\, y,\, x \,\rangle \\ &= \|x - y\|_2^2 . \end{aligned}$$

■

Example 2.3 (Negative entropy). Recall that the *negative entropy function* is $f : \mathbb{R}^n_{>0} \to \mathbb{R}$ defined by $f(x) = \sum_{i=1}^n x_i \ln x_i$. Then

$$\begin{aligned} D_f(x, y) &= f(x) - f(y) - \nabla f(y)^{\mathsf{T}}(x - y) \\ &= \sum_{i=1}^n x_i \ln x_i - \sum_{i=1}^n y_i \ln y_i - \sum_{i=1}^n (\ln y_i + 1)(x_i - y_i) \\ &= \sum_{i=1}^n x_i \ln x_i - \sum_{i=1}^n x_i \ln y_i - \sum_{i=1}^n x_i + \sum_{i=1}^n y_i \\ &= \sum_{i=1}^n x_i \ln(x_i / y_i) - \sum_{i=1}^n x_i + \sum_{i=1}^n y_i \\ &= D_{\mathrm{KL}}(x \,\|\, y), \end{aligned} \tag{2.1}$$

the *generalized KL-divergence* between x and y, which we introduced in Lecture 16. In the case that $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i = 1$, this is the ordinary KL divergence (or "relative entropy") between x and y. ■

Negative entropy will be particularly important to us, so we prove one property of it now.

Claim 2.4. Negative entropy is 1-strongly convex with respect to the ℓ_1 norm.

To prove this, we require the following theorem.

Theorem 2.5 (Pinsker's Inequality). For any distributions p, q, we have $D_{\mathrm{KL}}(p \parallel q) \geq \frac{1}{2} \|p - q\|_1^2$.

References. Wikipedia, Lecture notes of Sanjeev Khudanpur.

Proof (of Claim 2.4). As in Example 2.3, let $f(x) = \sum_{i=1}^n x_i \ln x_i$. Then,

$$\begin{aligned} f(y) &\geq f(x) + \langle\, \nabla f(x),\, y - x \,\rangle + D_{\mathrm{KL}}(y \parallel x) \qquad \text{(by (2.1))} \\ &= f(x) + \langle\, \nabla f(x),\, y - x \,\rangle + \tfrac{1}{2} \|x - y\|_1^2 \qquad \text{(by Theorem 2.5).} \end{aligned}$$ □

There are also some interesting examples involving matrices.

Example 2.6. Let $f(X) = \mathrm{tr}(X \log X)$. Then $D_f(X, y) = \mathrm{tr}(X \log X - X \log Y - X + Y)$. This is called the von Neumann divergence, or quantum relative entropy. ■

Example 2.7. Let $f(X) = -\log\det X$. Then $D_f(X, Y) = \mathrm{tr}(XY^{-1} - I) - \log\det(XY^{-1})$. This is called the log-det divergence. ■

2.2 Properties

Claim 2.8. $D_f(x, y)$ is convex in x.

Proof. This is immediate from the definition since $f(x)$ is convex in x and $-\langle\, \nabla f(y),\, x - y \,\rangle$ is linear in x. □

Note. $D_f(x, y)$ is not generally convex in y. Consider the case $f(x) = \exp(x)$ and $x = 4$. Then

$$\begin{aligned} D_f(4, 0) &= e^4 - 5 < 50 \\ D_f(4, 1) &= e^4 - 4e > 43 \\ D_f(4, 2) &= e^4 - 3e^2 < 33 \end{aligned}$$

As $D_f(4, 1) > \big(D_f(4, 0) + D_f(4, 2)\big)/2$, $D_f(x, y)$ is not convex in y.

Lemma 2.9. Let f be closed, convex and differentiable. Fix any $x, y \in \mathcal{X}$. Define $\hat{x} = \nabla f(x)$ and $\hat{y} = \nabla f(y)$. Then

$$\nabla f^*(\hat{x}) = x \tag{2.2}$$
$$D_f(x, y) = D_{f^*}(\hat{y}, \hat{x}) \tag{2.3}$$

References. [8, Eq. (6)].

Proof. By Claim 1.11,

$$f^*(\hat{x}) = \langle\, \hat{x},\, x \,\rangle - f(x), \qquad f^*(\hat{y}) = \langle\, \hat{y},\, y \,\rangle - f(y), \qquad \text{and} \qquad \nabla f^*(\hat{x}) = x.$$

This proves (2.2). Thus

$$\begin{aligned} D_{f^*}(\hat{y}, \hat{x}) &= f^*(\hat{y}) - f^*(\hat{x}) - \langle\, \nabla f^*(\hat{x}),\, \hat{y} - \hat{x}\,\rangle \\ &= \big(\langle\, \nabla f(y),\, y\,\rangle - f(y)\big) - \big(\cancel{\langle\, \nabla f(x),\, x\,\rangle} - f(x)\big) - \langle\, x,\, \nabla f(y) - \cancel{\nabla f(x)}\,\rangle \\ &= f(x) - f(y) - \langle\, \nabla f(y),\, x - y\,\rangle \\ &= D_f(x, y) \end{aligned}$$

This proves (2.3). □

Lemma 2.10 (Generalized Pythagoras Identity).

$$D_f(x, y) + D_f(z, x) - D_f(z, y) \;=\; \big(\nabla f(x) - \nabla f(y)\big)^{\mathsf{T}}(x - z).$$

Example 2.11. Consider the case of $f(x) = \|x\|_2^2$. Applying (5.1) with $a = x - z$ and $b = x - y$ (so that $a - b = y - z$), we obtain

$$\|x - y\|_2^2 + \|z - x\|_2^2 - \|z - y\|_2^2 \;=\; 2(x - y)^{\mathsf{T}}(x - z) \;=\; \big(\nabla f(x) - \nabla f(y)\big)^{\mathsf{T}}(x - z).$$

■

Proof (of Lemma 2.10).

$$\begin{aligned} & D_f(x, y) + D_f(z, x) - D_f(z, y) \\ = \; & \Big(\cancel{f(x)} - \cancel{f(y)} - \langle\, \nabla f(y),\, x - \cancel{y}\,\rangle\Big) \;+\; \Big(\cancel{f(z)} - \cancel{f(x)} - \langle\, \nabla f(x),\, z - x\,\rangle\Big) \\ & - \Big(\cancel{f(z)} - \cancel{f(y)} - \langle\, \nabla f(y),\, z - \cancel{y}\,\rangle\Big) \\ = \; & \big(\nabla f(x) - \nabla f(y)\big)^{\mathsf{T}}(x - z) \end{aligned}$$

□

Claim 2.12. $\nabla_x D_f(x, y) = \nabla f(x) - \nabla f(y)$.

Proof.

$$\nabla_x\Big(f(x) - f(y) - \langle\, \nabla f(y),\, x - y\,\rangle\Big) \;=\; \nabla_x f(x) - \nabla_x\langle\, \nabla f(y),\, x\,\rangle \;=\; \nabla f(x) - \nabla f(y)$$

□

2.2.1 Projections

Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed, convex set. Assume that $f : \mathcal{X} \to \mathbb{R}$ is a *strictly* convex function, which implies that $D_f(x, y)$ is strictly convex in x.

Definition 2.13. The projection of y onto $\mathcal{X}$ under the Bregman divergence is

$$\Pi_{\mathcal{X}}^{f}(y) \;=\; \operatorname*{argmin}_{x \in \mathcal{X}} D_f(x, y).$$

The minimizer is uniquely determined since $D_f(x, y)$ is strictly convex in x.

The next claim gives optimality conditions for the Bregman projection, which is analogous to results that we discussed for Euclidean projections.

Claim 2.14. Suppose that f is differentiable. Fix any $y \in \mathbb{R}^n$ and let $\pi = \Pi_{\mathcal{X}}^f(y)$. Then

$$\big(\nabla f(y) - \nabla f(\pi)\big)^{\mathsf{T}}(w - \pi) \;\leq\; 0 \qquad \forall w \in \mathcal{X}.$$

Proof. Since $\pi = \operatorname{argmin}_{x \in \mathcal{X}} D_f(x, y)$, Recall the optimality conditions for minimizing a convex function over a convex set:

$$\nabla_x D_f(\pi, y)^{\mathsf{T}}(\pi - w) \;\leq\; 0 \quad \forall w \in \mathcal{X}.$$

By Claim 2.12, $\nabla_x D_f(\pi, y) = \nabla f(\pi) - \nabla f(y)$, which proves the result. □

The following corollary is the main optimality condition we will use for Bregman divergences. (Compare with [6, Lemma 14.9].)

Corollary 2.15. Fix any $y \in \mathbb{R}^n$ and let $\pi = \Pi_{\mathcal{X}}^f(y)$. Then

$$D_f(w, y) \;\geq\; D_f(w, \pi) \qquad \forall w \in \mathcal{X}.$$

Proof. By Lemma 2.10 and Claim 2.14,

$$D_f(\pi, y) + D_f(w, \pi) - D_f(w, y) \;=\; \big(\nabla f(\pi) - \nabla f(y)\big)^{\mathsf{T}}(\pi - w) \;\leq\; 0$$

for all $w \in \mathcal{X}$. Since $D_f(\pi, y) \geq 0$, the result follows by rearranging. □

3 The Mirror Descent Algorithm

A good reference for the material in this section is [4, Chapter 4].

Mirror descent has several motivations.

- Our preceding analyses of gradient descent assume that the function f is Lipschitz with respect to the Euclidean norm. What if f is L-Lipschitz with respect to another norm, e.g., $\|\cdot\|_\infty$? It follows that f is $(\sqrt{n}L)$-Lipschitz with respect to $\|\cdot\|_2$, but this factor $\sqrt{n}$ may be undesirably large.
- Gradient descent is troubling in that it does not explicitly distinguish between the "primal" and "dual" vector spaces. By this we mean that the iterates x_i are vectors in $\mathbb{R}^n$, whereas the gradients $\nabla f(x_i)$ are technically *linear functionals* on $\mathbb{R}^n$. So, technically, $\nabla f(x_i)$ lives in the *dual space*. We are somewhat justified in conflating the $\mathbb{R}^n$ and its dual space, since they are isomorphic (by the transpose operation). Nevertheless, gradient descent computes a linear combination of the iterate x_i and the gradient $\nabla f(x_i)$, without calling attention to the fact that these objects technically lie in different vector spaces.

The main idea of Mirror Descent is to explicitly distinguish between the primal and dual spaces, and to specify a useful bijection between the two of them. The bijection is determined by a ***mirror function*** $\Phi : \mathbb{R}^n \to \mathbb{R}$. A canonical example of Φ is the negative entropy function, which we have discussed before in our lecture on convex functions.

The bijection between the primal space and the dual space is as follows: a primal point x maps to the dual point $\nabla\Phi(x)$. To map back from the dual space to the primal space, we will use the inverse of the mapping $x \mapsto \nabla\Phi(x)$. (In fact, this inverse mapping is simply $y \mapsto \nabla\Phi^*(y)$, the gradient of the Fenchel dual of Φ. This is proven in (2.2).)

Mirror descent really only has two main ideas:

- **The main idea.** Instead of taking gradient steps in the primal space, mirror descent takes gradient steps in the dual space. The bijection $\nabla\Phi$ and its inverse $\nabla\Phi^*$ are used to map back and forth between primal points and dual points.
- **Ensuring feasibility.** The goal is *constrained* optimization over a convex set $\mathcal{X}$ in the primal space, so a concern is that our gradient step may have produced a point outside of $\mathcal{X}$. The next main idea is to project that point onto $\mathcal{X}$ under the Bregman divergence D_Φ, as discussed in Section 2.2.1.

3.1 The mirror map Φ

In order for this scheme to work out, we need to make some formal assumptions. There is an open set $\mathcal{D} \subseteq \mathbb{R}^n$ and the mirror map $\Phi : \mathcal{D} \to \mathbb{R}$.

P1: Φ is strictly convex and differentiable on all of $\mathcal{D}$.

P2: The dual space is all of $\mathbb{R}^n$. That is, $\{ \nabla\Phi(x) : x \in \mathcal{D} \} = \mathbb{R}^n$.

P3: The gradient of Φ diverges on the boundary of $\mathcal{D}$. That is, $\lim_{x \to \partial\mathcal{D}} \|\nabla(x)\| = +\infty$.

Example 3.1. Returning to our negative entropy example, we will take $\mathcal{D} = \mathbb{R}^n_{>0}$, the positive orthant. Recall that $\nabla\Phi(x)_j = 1 + \ln x_j$ and $\nabla^2\Phi(x) = \mathrm{diag}(x)^{-1}$, so P1 is satisfied. P2 is satisfied because, for any point $y \in \mathbb{R}^n$, we have $y = \nabla f(x)$ where $x_j = e^{y_j - 1}$. (This can also be seen by (2.2): the inverse of $\nabla\Phi$ is $\nabla\Phi^*$, and Example 1.5 showed that $\nabla\Phi^*(y)_j = e^{y_j - 1}$.) P3 is satisfied because $|\nabla\Phi(x)_i| = |1 + \ln x_i| \to +\infty$ as $x_i \searrow 0$. ■

Claim 3.2. Suppose that assumption- P1 and P2 hold. Then $\Phi : \mathcal{D} \to \mathbb{R}^n$ is a bijection.

Proof. By P1, $x \mapsto \nabla\Phi(x)$ is a well-defined function on $\mathcal{D}$.

By P2, Φ is a surjection.

By P2, for every $y \in \mathbb{R}^n$, there exists $x \in \mathcal{D}$ such that $y = \nabla\Phi(x)$. By Claim 1.11, $x \in \partial\Phi^*(y)$. If there existed any other $x' \in \partial\Phi^*(y)$, then Claim 1.11 would imply that $y = \nabla\Phi(x')$, which contradicts Claim 5.7. Thus, Φ^* is the inverse map of Φ, so Φ is a bijection. □

3.1.1 The constraint set $\mathcal{X}$

Typically we will optimize over a constraint set $\mathcal{X}$. It must satisfy the following formal assumptions.

P4: $\mathcal{X}$ is closed and convex.

P5: $\mathcal{X} \subseteq \overline{\mathcal{D}}$, the closure of $\mathcal{D}$.

P6: $\mathcal{X} \cap \mathcal{D} \neq \emptyset$.

Example. Returning to our negative entropy example, we will take $\mathcal{X}$ to be the simplex

$$\Delta^n \;=\; \left\{ x \in \mathbb{R}^n_{\geq 0} \,:\, \sum_i x_i = 1 \right\}.$$

Note that $\mathcal{X} \subseteq \overline{\mathcal{D}}$, the non-negative orthant. Obviously $\mathcal{X}$ is closed, convex and intersects with $\mathcal{D}$.

3.1.2 The projection

A primal point $y \in \mathcal{D}$ will be projected back onto the constraint set by computing the Bregman projection

$$\Pi^{\Phi}_{\mathcal{X}\cap\mathcal{D}}(y) \in \operatorname*{argmin}_{x\in\mathcal{X}\cap\mathcal{D}} D_\Phi(x,y).$$

There is a small annoyance: $\mathcal{X} \cap \mathcal{D}$ is not closed, so the argmin is not necessarily defined. However, I believe that since $y \in \mathcal{D}$, this should not matter... For now, let's sweep this under the rug.

Let us determine the Bregman projection in our negative entropy example.

Claim 3.3. Let $\Phi(x) = \sum_i x_i \ln x_i$. Suppose that $y \in \mathcal{D} \setminus \mathcal{X}$, i.e., $y \in \mathbb{R}^n_{>0}$, but $\sum_i y_i \neq 1$. Then $\Pi^{\Phi}_{\mathcal{X}\cap\mathcal{D}}(y) = y/\left\|y\right\|_1$.

Proof. As shown in Example 2.3, $D_\Phi(x,y)$ is the generalized KL divergence $D_{\mathrm{KL}}(x \parallel y) = \sum_i \big(x_i \ln(x_i/y_i) - x_i + y_i\big)$. Thus,

$$\operatorname*{argmin}_{x\in\mathcal{X}\cap\mathcal{D}} D_\Phi(x,y) \;=\; \operatorname*{argmin}_{x\in\mathcal{X}\cap\mathcal{D}} \sum_{i=1}^n \Big(x_i \ln\frac{x_i}{y_i} - x_i + y_i\Big) \;=\; \operatorname*{argmin}_{x\in\mathcal{X}\cap\mathcal{D}} \sum_{i=1}^n x_i \ln\frac{x_i}{y_i}$$

since $\sum_{i=1}^n x_i = 1$ for all $x \in \mathcal{X}$. Letting $f(x) = x\ln x$, which is convex, Jensen's inequality yields

$$\sum_{i=1}^n x_i \ln\frac{x_i}{y_i} \;=\; \sum_{i=1}^n y_i f(x_i/y_i) \;\geq\; f\Big(\frac{\sum_{i=1}^n x_i}{\left\|y\right\|_1}\Big)\left\|y\right\|_1 \;=\; \ln(1/\left\|y\right\|_1),$$

and equality holds iff all x_i/y_i are equal, i.e., $x_i = y_i/\left\|y\right\|_1$. Thus, $y/\left\|y\right\|_1 = \operatorname{argmin}_{x\in\mathcal{X}\cap\mathcal{D}} D_\Phi(x,y)$. □

3.2 The algorithm

Algorithm 1 The online mirror descent algorithm. The initial point is any $x_1 \in \mathcal{X} \cap \mathcal{D}$.

procedure MIRRORDESCENT($x_1 \in \mathcal{X} \cap \mathcal{D}$, η)
  for $i \leftarrow 1, 2, \ldots$ **do**
    ▷ Incur cost $f_i(x_i)$, receive a subgradient $g_i \in \partial f_i(x_i)$
    $\hat{x}_i \leftarrow \nabla\Phi(x_i)$ (map primal point to dual)
    $\hat{y}_{i+1} \leftarrow \hat{x}_i - \eta g_i$ (take gradient step in the dual)
    $y_{i+1} \leftarrow \nabla\Phi^*(\hat{y}_{i+1})$ (map new dual point back to a primal point in $\mathcal{D}$)
    $x_{i+1} \leftarrow \Pi^{\Phi}_{\mathcal{X}\cap\mathcal{D}}(y_{i+1})$ (project new point onto feasible region)

Example 3.4. In the negative entropy setting with $x_1 = (1/n, \ldots, 1/n)$, this algorithm is identical to Algorithm 2 of Lecture 16. To see this, recall from our preceding discussion that $\nabla\Phi(x_i)_j = 1 + \ln x_{i,j}$ and $\nabla\Phi^*(\hat{y}_{i+1})_j = \exp(\hat{y}_{i+1,j} - 1)$, so

$$y_{i+1,j} \;=\; \exp(\hat{y}_{i+1,j} - 1) \;=\; \exp(\hat{x}_{i,j} - \eta g_{i,j} - 1) \;=\; \exp(\ln x_{i,j} - \eta g_{i,j}) \;=\; x_{i,j}\exp(-\eta g_{i,j}).$$

Also, as mentioned above, $\Pi_{\mathcal{X}\cap\mathcal{D}}(y_{i+1}) = y_{i+1}/\left\|y_{i+1}\right\|_1$. Thus, the algorithms are identical.

We may think of the dual space as the "logarithmic weight space", in which we make additive updates (using a subgradient of f_i), and the primal space as the "normalized weight space" in which we make multiplicative updates to the weights (and then immediately renormalize them). ■

Theorem 3.5 (Online Mirror Descent). Given are any $\eta > 0$ and:

- A mirror map $\Phi : \mathcal{D} \to \mathbb{R}$ satisfying P1, P2, P3, where $\mathcal{D}$ is an open subset of $\mathbb{R}^n$.
- A feasible region $\mathcal{X} \subseteq \mathbb{R}^n$ satisfying P4, P5, P6.
- Convex functions $f_1, f_2, \ldots : \mathcal{X} \to \mathbb{R}$.

We assume that Φ is ρ-strongly convex with respect to a norm $\|\cdot\|$. Then Algorithm 1 satisfies

$$\sum_{i=1}^{t}\big(f_i(x_i) - f_i(x^*)\big) \;\le\; \frac{D_\Phi(x^*, x_1)}{\eta} + \sum_{i=1}^{t}\left(\langle\, g_i,\, x_i - y_{i+1}\,\rangle - \frac{\rho}{2\eta}\left\|x_i - y_{i+1}\right\|^2\right) \tag{3.1}$$

$$\le\; \frac{D_\Phi(x^*, x_1)}{\eta} + \frac{\eta}{2\rho}\sum_{i=1}^{t}\left\|g_i\right\|_*^2. \tag{3.2}$$

References. [4, Theorem 4.2].

Remark. The reason that Theorem 3.5 has two inequalities is that (3.1) allows us to prove *multiplicative* error for RWM, as in Corollary 4.2 below. In contrast, the simpler inequality (3.2) is convenient for analyzing *additive* error, as in Corollary 3.6

Corollary 3.6. Consider the offline setting where each $f_i = f$, and f is L-Lipschitz with respect to $\|\cdot\|$. Define $R^2 = \sup_{x\in\mathcal{X}\cap\mathcal{D}} D_\Phi(x, x_1)$. Then

$$f\left(\frac{1}{t}\sum_{i=1}^{t} x_i\right) - f(x^*) \;\le\; \frac{D_\Phi(x^*, x_1)}{\eta t} + \frac{\eta L^2}{2\rho} \;\le\; RL\sqrt{\frac{2}{\rho t}},$$

by choosing $\eta = \frac{R}{L}\sqrt{\frac{2\rho}{t}}$.

Proof. Recall from Theorem 5.6 that f is L-Lipschitz with respect to $\|\cdot\|$ if and only if $\|g\|_* \le L$ for all subgradients L. Thus, from Theorem 3.5, Jensen's inequality we have

$$f\left(\frac{1}{t}\sum_{i=1}^{t} x_i\right) - f(x^*) \;\le\; \frac{D_\Phi(x^*, x_1)}{\eta t} + \frac{\eta L^2}{2\rho}$$

The hypothesis implies $D_\Phi(x^*, x_1) \le R^2$. Substituting for η yields the result. □

Theorem 3.5 may be viewed as a generalization of Theorem 2.1 of Lecture 16. The differences are highlighted below. The first half of the proof is very similar, using the Bregman divergence $D_\Phi(\cdot,\cdot)$ instead of the KL divergence $D_{\mathrm{KL}}(\cdot \,\|\, \cdot)$. The second half of the proof is different, as strong convexity is used.

Proof (of Theorem 3.5). For any $z \in \mathcal{X}$,

$$f_i(x_i) - f_i(z) \;\le\; g_i^{\mathsf{T}}(x_i - z) \qquad (\text{since } g_i \in \partial f_i(x_i))$$

By the gradient step, $g_i = (\hat{x}_i - \hat{y}_{i+1})/\eta$, so

$$\begin{aligned}
&= \frac{1}{\eta}\big(\nabla\Phi(x_i) - \nabla\Phi(y_{i+1})\big)^{\mathsf{T}}(x_i - z) \\
&= \frac{1}{\eta}\Big(D_\Phi(x_i, y_{i+1}) + D_\Phi(z, x_i) - D_\Phi(z, y_{i+1})\Big) \qquad \text{(by Lemma 2.10)} \\
&\le \frac{1}{\eta}\Big(D_\Phi(x_i, y_{i+1}) + D_\Phi(z, x_i) - D_\Phi(z, x_{i+1})\Big) \qquad \text{(by Corollary 2.15)}.
\end{aligned}$$

Summing over i, the last two terms telescope:

$$\sum_{i=1}^{t}\big(f_i(x_i) - f_i(z)\big) \;\le\; \frac{1}{\eta}\Big(D_\Phi(z, x_1) + \sum_{i=1}^{t} D_\Phi(x_i, y_{i+1})\Big). \tag{3.3}$$

The summands on the right-hand side are bounded by strong convexity:

$$\begin{aligned}
D_\Phi(x_i, y_{i+1}) &= \Phi(x_i) - \Phi(y_{i+1}) - \langle\, \nabla\Phi(y_{i+1}),\, x_i - y_{i+1}\,\rangle \qquad (\text{by definition of } D_\Phi) \\
&= \underbrace{\Phi(x_i) - \Phi(y_{i+1}) + \langle\, \nabla\Phi(x_i),\, y_{i+1} - x_i\,\rangle}_{\text{use strong convexity}} + \langle\, \nabla\Phi(x_i) - \nabla\Phi(y_{i+1}),\, x_i - y_{i+1}\,\rangle \\
&\le -\frac{\rho}{2}\|x_i - y_{i+1}\|^2 + \eta\langle\, g_i,\, x_i - y_{i+1}\,\rangle.
\end{aligned}$$

by strong convexity of Φ and the gradient step. Combining this with (3.3) and setting $z = x^*$ proves (3.1). Next, use Claim 5.2 to obtain

$$D_\Phi(x_i, y_{i+1}) \;\le\; \eta\,\|g_i\|_*\,\|x_i - y_{i+1}\| - \frac{\rho}{2}\|x_i - y_{i+1}\|^2 \;\le\; \frac{\eta^2\,\|g_i\|_*^2}{2\rho},$$

since $\max_z(az - bz^2) = a^2/4b$. Combining this with (3.3) proves (3.2). □

4 Mirror Descent Examples

4.1 Example: Online convex optimization over the simplex

Let us again discuss the scenario where Φ is the negative entropy function and $\mathcal{X}$ is the simplex Δ^n. Our first corollary concerns online convex optimization over the simplex for functions that are Lipschitz with respect to the ℓ_1-norm. In this setting, I believe that the online mirror descent algorithm is called the *Exponentiated Gradient Algorithm* of Kivinen and Warmuth.

Corollary 4.1. Suppose that $f_1, f_2, ... : \mathcal{X} \to \mathbb{R}$ are convex functions, each of which is 1-Lipschitz with respect to $\|\cdot\|_1$. Run the mirror descent algorithm with Φ being the negative entropy function, $x_1 = (1/n, ..., 1/n)$ and $\eta = \sqrt{2\ln(n)/t}$. Then

$$\sum_{i=1}^{t} \big(f_i(x_i) - f_i(x^*)\big) \;\le\; \sqrt{2t \ln n}.$$

References. [1, Theorem 7].

Proof. Recall from Claim 2.4 that Φ is 1-strongly convex with respect to $\|\cdot\|_1$. Thus we may apply Theorem 3.5 with $\rho = 1$. Then (3.2) gives that

$$\sum_{i=1}^{t} \big(f_i(x_i) - f_i(x^*)\big) \;\le\; \frac{D_\Phi(x^*, x_1)}{\eta} + \frac{\eta}{2} \sum_{i=1}^{t} \|g_i\|_\infty^2 .$$

As shown in Example 2.3, $D_\Phi(x^*, x_1) = D_{\mathrm{KL}}(x^* \,\|\, \mathbf{1}/n) \le \ln n$, by Claim 5.5. Since f_i is 1-Lipschitz with respect to $\|\cdot\|_1$, and $\|\cdot\|_\infty$ is the dual norm of $\|\cdot\|_1$, Theorem 5.6 implies that $\|g_i\|_\infty \le 1$ for all $i \ge 1$. Thus,

$$\sum_{i=1}^{t} \big(f_i(x_i) - f_i(x^*)\big) \;\le\; \frac{\ln n}{\eta} + \frac{\eta t}{2}.$$

Substituting the value of η completes the proof. □

4.2 Example: Randomized weighted majority

Now let us consider the setting of learning with experts discussed in Lectures 15 & 16. As before, the cost of expert j at time i is $c_{i,j}$. Let f_i be the linear function $f_i(x) = \langle\, c_i,\, x \,\rangle$. We now derive a form[1] of our RWM analysis as a corollary of Theorem 3.5.

Corollary 4.2. Let j^* be the expert with minimum total cost. For any costs $c_{i,j} \in [0, 1]$, the mirror descent algorithm achieves

$$\sum_{i=1}^{t} \langle\, c_i,\, x_i \,\rangle \;\le\; (1 + 2\eta) \sum_{i=1}^{t} c_{i,j^*} + \frac{(1 + 2\eta) \ln n}{\eta}. \tag{4.1}$$

[1] Unfortunately the parameters are slightly worse, since (3.1) is stated abstractly in terms of the norm $\|\cdot\|$ (which requires the use of Pinsker's inequality), and this loses some sharpness.

Proof. As before, we will use the fact and that $D_\Phi(x^*, x_1) \le \ln n$ (see Claim 5.5). Note that

$$\min_{1\le j\le n} \sum_{i=1}^{t} c_{i,j} \;=\; \min_{1\le j\le n} \sum_{i=1}^{t} f_i(e_j) \;=\; \min_{x\in\mathcal{X}} \sum_{i=1}^{t} f_i(x)$$

since $\sum_{i=1}^{t} f_i$ is a linear function, so its minimum must occur at one of its vertices $\{e_1, ..., e_n\}$. We apply Theorem 3.5 with $\rho = 0$, so (3.1) gives

$$\sum_{i=1}^{t} \langle\, c_i,\, x_i \,\rangle \;=\; \sum_{i=1}^{t} f_i(x_i) \;\le\; \sum_{i=1}^{t} c_{i,j^*} + \frac{\ln n}{\eta} + \sum_{i=1}^{t} \langle\, g_i,\, x_i - y_{i+1} \,\rangle$$

Note that

$$x_{i,j} - y_{i+1,j} \;=\; x_{i,j}\big(1 - \exp(-\eta c_{i,j})\big) \;=\; x_{i,j}\big(1 - \exp(-\eta)\big) \;\le\; \eta x_{i,j}$$

by Claim 5.3 and using $c_{i,j} \in [0, 1]$. Since f_i is linear, $g_i = c_i$ for all i. Thus

$$\sum_{i=1}^{t} c_i^{\mathsf{T}} x_i \;\le\; \sum_{i=1}^{t} c_{i,j^*} + \frac{\ln n}{\eta} + \eta \sum_{i=1}^{t} c_i^{\mathsf{T}} x_i.$$

Rearranging (as in Corollary 2.7 of Lecture 16), and using Claim 5.4 completes the proof. □

5 Standard facts

Claim 5.1 (Law of cosines).

$$\|a - b\|_2^2 = \|a\|_2^2 - 2a^{\mathsf{T}} b + \|b\|_2^2 \qquad \forall a, b \in \mathbb{R}^n. \tag{5.1}$$

Claim 5.2. For $w \in V$ and $z \in V^*$, we have $\langle\, z,\, w \,\rangle \le \|w\| \cdot \|z\|_*$.

Claim 5.3. $1 + x \le e^x$ for all $x \in \mathbb{R}$.

Claim 5.4.

$$1 + x \;\le\; \frac{1}{1 - x} \;\le\; 1 + 2x.$$

The first inequality holds for $x < 1$; the second holds for $x \in [0, 1/2]$.

Claim 5.5. Suppose that $p \in \mathbb{R}^n_{\ge 0}$ is a distribution. Let $q = (1/n, ..., 1/n) \in \mathbb{R}^n_{\ge 0}$ be the uniform distribution. Then $D_{\mathrm{KL}}(p \,\|\, q) \le \ln n$.

Theorem 5.6. Let $\mathcal{X}$ be convex and open. Let $f : \mathcal{X} \to \mathbb{R}$ be convex. Let $\|\cdot\|$ be an arbitrary norm. The following conditions are equivalent.

- $f : \mathcal{X} \to \mathbb{R}$ is L-Lipschitz with respect to $\|\cdot\|$:

$$|f(x) - f(y)| \;\le\; L\,\|x - y\| \qquad \forall x, y \in \mathcal{X}. \tag{5.2}$$

- f has bounded subgradients:

$$\|g\|_* \le L \qquad \forall w \in X,\; g \in \partial f(w). \tag{5.3}$$

Claim 5.7. Let $f : \mathcal{X} \to \mathbb{R}$ be strictly convex function. Let $x, x' \in \mathcal{X}$ be distinct. Then $\partial f(x) \cap \partial f(x') = \emptyset$.

Lemma 5.8. Let $\mathcal{X}$ be a convex set. For each $a \in \mathcal{A}$, let $f_a : \mathcal{X} \to \mathbb{R}$ be a convex function. Define $\hat{f}(x) = \sup_{a \in \mathcal{A}} f_a(x)$. Then $\hat{f}$ is convex. Furthermore, if each f_a is closed then $\hat{f}$ is closed.

References

[1] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: A meta algorithm and its applications. *Theory of Computing*, 8(6):121–164, 2012.

[2] Dmitris Bertsekas. *Convex Analysis and Optimization*. Athena Scientific, 2003.

[3] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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

[5] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer-Verlag, 2004.

[6] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[7] Lieven Vandenberghe. Conjugate functions. `http://www.seas.ucla.edu/~vandenbe/236C/lectures/conj.pdf`.

[8] Xinhua Zhang. Bregman divergence and mirror descent. `http://www.cs.uic.edu/~zhangx/teaching/bregman.pdf`.