Lecture 3: Online mirror descent and density approximation
CSE 599S: Entropy optimality, Winter 2016
Instructor: James R. Lee **Last updated:** January 23, 2016

1 Mirror descent

Let's attempt to rephrase what we did last time in a more general setting. The idea is to view our algorithm as a sort of "regularized" local improvement algorithm. One should consult [Ch. 4, Bubeck, 2014] and [Ch. 5, Hazan, 2015] (and the references therein) for further information about online mirror descent and related algorithms coming from the convex optimization and machine learning. Our treatment in this section follows [Bubeck, 2014].

First, we introduce the notion of a Bregman divergence (of which the relative entropy is one example).

Bregman divergences. Given a differentiable, strictly convex function $F : \mathcal{D} \to \mathbb{R}$ on a convex set $\mathcal{D} \subseteq \mathbb{R}^n$, we can define the associated Bregman divergence $D_F : \mathcal{D} \times \mathcal{D} \to \mathbb{R}_+$ by

$$D_F(x, y) = F(x) - \Big(F(y) + \langle \nabla F(y), x - y \rangle\Big) .$$

This is the "error" in using the first-order Taylor approximation of F at y to compute $F(x)$.

Exercise 1.1. Prove that $D_F(x, y) \geqslant 0$ for all $x, y \in \mathcal{D}$.

A basic example is $F(x) = \|x\|_2^2$ in which case $D_F(x, y) = \|x - y\|_2^2$. Another highly relevant example arises when $\mathcal{D} = \mathbb{R}^n_+$ and $F : \mathcal{D} \to \mathbb{R}$ is given by the negative entropy $F(x) = \sum_{i=1}^n x_i \log x_i$. In that case, one easily calculates

$$D_F(x, y) = \sum_{i=1}^n x_i \log \frac{x_i}{y_i} + \sum_{i=1}^n (y_i - x_i) .$$

Observe that if x and y are probability measures, i.e. $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i = 1$, then $D_F(x, y) = D(x \,\|\, y)$.

In general, Bregman divergences are not symmetric (as in the case of the relative entropy), but they share some nice properties of the squared Euclidean distance. One reason for this (especially in the treatment of Section 2.2) is that locally a Bregman divergence *is the square of a Euclidean norm.*

Local norms. For simplicity, let us consider a continuously differentiable and strictly convex $F : \mathcal{D} \to \mathbb{R}$, and assume that the $\mathcal{D}$ is an open convex set. Since F is convex, the Hessian $\nabla^2 F$ is positive semi-definite on $\mathcal{D}$ (see Lecture 1), and we have

$$F(y + h) = F(y) + \langle \nabla F(y), h \rangle + \frac{1}{2}\langle h, \nabla^2 F(y) h \rangle + O(\|h\|_2^3)$$

as $\|h\|_2 \to 0$.

Thus we can write

$$D_F(y + h, y) = \frac{1}{2}\langle h, \nabla^2 F(y) h \rangle + O(\|h\|_2^3) \approx \frac{1}{2}\|h\|^2_{\nabla^2 F(y)}$$

where the approximation is up to third order error, and we use the notation

$$\|h\|_A = \sqrt{\langle h, Ah \rangle}$$

when A is self-adjoint and positive semi-definite. When A is actually positive definite (as is the case for $\nabla^2 F(y)$ because F is strictly convex), this defines a norm on $\mathbb{R}^n$.

Now one can see the fundamental reason for the asymmetry of the divergence: $D_F(x, y)$ is computed using the local Euclidean geometry given by $\nabla^2 F$ at the point y. What's somewhat more interesting is that the divergence remains interesting for points x and y that are separated.

1.1 Bregman projection

Let $C \subseteq \mathbb{R}^n$ be a closed convex set. Given a Bregman divergence D_F, we can define the *Bregman projection* of a point $x \in \mathbb{R}^n$ on C by

$$\Pi_C^F(y) = \operatorname*{argmin}_{x \in C} D_F(x, y)\,.$$

By strong convexity of F, the projection is unique.

There is a corresponding "Pythagorean theorem" (it helps to think about the model case $D_F(x, y) = \|x - y\|^2$).

Lemma 1.2. *For all $x \in C$ and $y \in \mathbb{R}^n$, we have*

$$D_F(x, \Pi_C^F(y)) \leqslant D_F(x, y) - D_F(\Pi_C^F(y), y)\,.$$

A good way to think about this lemma: Think about x as the target, and our current point is y. Since $x \in C$, it makes sense that projecting to C will get us closer to x. The lemma gives us a quantitative version that says: The further away we were from C, the closer we get to the target by projecting.

Mirror maps. Let $\mathcal{D} \subseteq \mathbb{R}^n$ be an open convex set, and let $\Phi : \mathcal{D} \to \mathbb{R}$ be a strictly convex function. We call Φ a *mirror map* if it additionally satisfies:

1. Φ is differentiable on $\mathcal{D}$.
2. The range of $\nabla\Phi : \mathcal{D} \to \mathbb{R}$ is all of $\mathbb{R}^n$.
3. $\nabla\Phi(x) \to \infty$ as x approaches $\partial\mathcal{D}$ (i.e., $\nabla\Phi$ blows up on the boundary of $\mathcal{D}$).

Example 1.3. Two prominent scenarios are

1. $\mathcal{D} = \mathbb{R}^n$ and $\Phi(x) = \|x\|^2$ with $\nabla\Phi(x) = 2x$.
2. $\mathcal{D} = \mathbb{R}^n_{++} = \{x \in \mathbb{R}^n : x_1, \ldots, x_n > 0\}$ and $\Phi(x) = \sum_{i=1}^n x_i \log x_i$ with

$$\nabla\Phi(x) = (1 + \log x_1, \cdots, 1 + \log x_n)\,.$$

In general, it will be best to think about $\nabla\Phi$ as a "dual object." In finite-dimensional optimization settings, we have the space of points (potential solutions), and the space of directions (potential ways to improve). For us, both objects lie in $\mathbb{R}^n$, and thus they are often conflated, but it helps the mental picture sometimes to separate them. In this case, it's best to think about $\nabla\Phi$ as a *vector field* specifying a direction $\nabla\Phi(x)$ at every point $x \in \mathcal{D}$.

Optimization setup. Suppose now that C is a compact, convex set with $C \subseteq \overline{\mathcal{D}}$, and $f : C \to \mathbb{R}$ is a convex function. Our object of study will be the optimization

$$\min_{x \in C} f(x) \,. \tag{1.1}$$

But we may not know f (or in the next lecture, we may choose not to examine f entirely); instead at every discrete time step $t = 1, 2, \ldots$, we will have access to a *subgradient of* f, i.e. some direction along which we can improve a little.

Definition 1.4. If $U \subseteq \mathbb{R}^n$ is an open convex set and $f : U \to \mathbb{R}$ is convex on U, a vector $v \in \mathbb{R}^n$ is called a *subgradient of f at the point $x_0 \in U$* if

$$f(x) - f(x_0) \geqslant \langle v, x - x_0 \rangle$$

for all $x \in U$. The collection of subgradients of f at x_0 will be denoted $\partial f(x_0) \subseteq \mathbb{R}^n$.

Restating the definition: Any movement within U along the direction v increases the value of f. It helps to remember that in order to decrease f (we are minimizing it, after all) we should try to move in the direction $-v$.

Online mirror descent. We will generate a sequence of points $\{x_0, x_1, x_2, \ldots\}$. We assume we have a sequence $\{v_0, v_1, v_2, \ldots\}$ of directions satisfying $v_t \in \partial f(x_t)$. Initially, we choose $x_0 = \operatorname{argmin}_{x \in C} \Phi(x)$. Then given x_t, we put

$$x_{t+1} = \operatorname*{argmin}_{x \in C} \; D_\Phi(x, x_t) + \eta \langle v_t, x \rangle \,. \tag{1.2}$$

Here, $\eta > 0$ is a step size parameter we will specify carefully in a moment.

Exercise 1.5. Show that when Φ is the negative entropy and $C = \left\{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\right\}$ is the probability simplex, the solution to (1.2) is

$$x_{t+1}(i) = \frac{\exp\left(-\eta v_t(i)\right)}{\sum_{i=1}^n \exp\left(-\eta v_t(i)\right) x_t(i)} x_t(i) \,.$$

In other words, one recovers the "exponential weights" update algorithm we saw earlier (in greater generality now).

In general, we should think of (1.2) as a form of cautious (or, "regularized") "subgradient descent." We would like to move in the direction $-v_t$, but we also value remaining close to the previous point x_t in terms of the "distance" given by the Bregman divergence D_Φ.

In order to analyze this algorithm, we need a few quantitative definitions. Say that the function $\Phi : \mathcal{D} \to \mathbb{R}$ is *ρ-strongly convex with respect to the norm* $\|\cdot\|$ if for all $x, y \in \mathcal{D}$, we have

$$\Phi(y) \geqslant \Phi(x) + \langle \nabla \Phi(x), y - x \rangle + \frac{\rho}{2} \|x - y\|^2 \,.$$

Pinsker's inequality is precisely the assertion that Φ is 1-strongly convex on the probability simplex equipped with the ℓ_1 norm.

Say that the map $f : C \to \mathbb{R}$ is L-Lipschitz (with respect to the norm $\|\cdot\|$) if $\|f(x) - f(y)\| \leqslant L \cdot \|x - y\|$ for all $x, y \in C$. See [Th. 4.2, Bubeck, 2014] for a proof of the following statement. Note that the bound (1.3) is more general than what appears there, but follows from the last line of the proof.

Theorem 1.6. *Let $R = \sup_{x \in C} \Phi(x) - \Phi(x_0)$. If Φ is ρ-strongly convex mirror map and f is L-Lipschitz with respect to the norm $\|\cdot\|$, then for all $t \geqslant 1$, the algorithm specified by (1.2) with step size*

$$\eta = L^{-1}\sqrt{\frac{2\rho R}{t}}$$

yields a sequence of points $\{x_0, x_1, x_2, \ldots, x_{t-1}\}$ such that for any $x \in C$,

$$f\left(\frac{1}{t}\sum_{s=0}^{t-1} x_s\right) \leqslant f(x) + \sqrt{\frac{2D_\Phi(x, x_0)}{\rho t}}L\,. \tag{1.3}$$

2 Continuous dynamics

In the preceding lecture, we saw a continuous-time algorithm for matrix scaling. We now generalize that approach to the setting of the previous section. A few aspects become more intuitive, and the "targeting" principle in Lemma 2.1 will be a useful intuition.

2.1 Relative entropy and sparse approximation

Consider the setting of relative entropy, where: $\mathcal{D} = \mathbb{R}^n_{++}$, $\Phi(x) = \sum_{i=1}^n x(i)\log x(i)$, and $C = \left\{x \in \mathbb{R}^n : \sum_{i=1}^n x(i) = 1\right\}$.

We will now produce a continuous sequence $\{x_t : t \geqslant 0\} \subseteq \mathbb{R}^n$ of points and we assume we have a family $\{v_t : t \geqslant 0\} \subseteq \mathbb{R}^n$ of directions (we do not require them to be subgradients of f). We will choose our next point using an infinitesimal version of the update (1.2) where we send $\eta \to 0$. (For clarity of exposition, we have chosen to eliminate the negative sign, and thus our v_t here corresponds to $-v_t$ previously.)

Define

$$x_t(i) = \frac{\exp\left(\int_0^t v_s(i)\,ds\right)}{\sum_{j=1}^n \exp\left(\int_0^t v_s(i)\,ds\right)}x_0(i)\,. \tag{2.1}$$

The proof of the next lemma is a straightforward differentiation and is left to the reader.

Lemma 2.1. *For any $w \in C$, it holds that*

$$\frac{d}{dt}D(w \,\|\, x_t) = -\langle v_t, w - x_t\rangle\,.$$

Observe in (2.1) that x_t is moving in the direction of v_t exponentially. This lemma tells us that for any $w \in C$, as long as we are moving toward w along the direction v_t, the divergence from w to x_t is decreasing proportionally. If we care about the optimization (1.1), then we might think of $-v_t \in \partial f(x_t)$ and $w = x^* = \operatorname{argmin}_{x \in C} f(x)$. Since the relative entropy is always positive, Lemma 2.1 asserts that in this case we are always making progress toward x^*.

It's useful to now how much progress we make if we move in one direction for a period of time.

Lemma 2.2. *Suppose that for $t \in [t_0, t_1)$, it holds that $v_t = v$ for some fixed $v \in \mathbb{R}^n$. Then*

$$\frac{d}{dt}\langle v, x_t\rangle = \sum_{i=1}^{n} v(i)^2 x_t(i) - \langle v, x_t\rangle^2 .$$

In particular,

$$\frac{d}{dt}\langle v, x_t\rangle \leqslant \|v\|_\infty^2 .$$

Suppose that $\langle v, x_t\rangle < \langle v, w\rangle$ for some $v, w \in \mathbb{R}^n$. Let us set $v_t = v$ for $t \in [t_0, t_1]$, where

$$t_1 = \inf\{t \geqslant t_0 : \langle v, x_t\rangle \geqslant \langle v, w\rangle\} .$$

In other words, we move in the direction v until at time t_1, we have $\langle v, x_{t_1}\rangle = \langle v, w\rangle$. Our goal is to measure the change in the potential.

Lemma 2.3. *With the parameters, above we have*

$$D(w \,\|\, x_{t_0}) - D(w \,\|\, x_{t_1}) \geqslant \frac{\langle v, x_{t_0} - w\rangle^2}{2\|v\|_\infty^2} .$$

Proof. Let $f(t) = \langle v, w - x_t\rangle$. By Lemma 2.2, we know that $f'(t) \geqslant -\|v\|_\infty^2$. Combining this with Lemma 2.1 yields

$$D(w \,\|\, x_{t_0}) - D(w \,\|\, x_{t_1}) = \int_{t_0}^{t_1} f(t)\, dt \geqslant \int_{t_0}^{t_0 + f(t_0)/\|v\|_\infty^2} f(t_0) - t\|v\|_\infty^2 \, dt = \frac{f(t_0)^2}{2\|v\|_\infty^2} . \qquad \square$$

We can now use this to prove the existence of "dual sparse" solutions to systems of linear inequalities on probability distributions that are not too far from the uniform measure.

Theorem 2.4. *Let $C = \left\{x \in \mathbb{R}_+^n : \sum_{i=1}^n x(i) = 1\right\}$. Suppose that $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ are given. Let $A_1, \ldots, A_m \in \mathbb{R}^n$ denote the rows of A. If there exists an $x^* \in C$ such that $Ax^* \geqslant b$, then for every $\varepsilon > 0$, there is a vector $x \in C$ satisfying $Ax \geqslant b - \varepsilon(1, 1, \ldots, 1)^T$, and x is given by*

$$x(i) = \frac{\exp\left(\sum_{k=1}^m c_k A_k(i)\right)}{\sum_{j=1}^n \exp\left(\sum_{k=1}^m c_k A_k(j)\right)} .$$

for some constants $c_1, \ldots, c_m \geqslant 0$ with

$$\#\{i : c_i > 0\} \leqslant 2 \cdot D\left(x^* \,\|\, (\tfrac{1}{n}, \cdots, \tfrac{1}{n})\right) \frac{\max_i \|A_i\|_\infty^2}{\varepsilon^2} . \tag{2.2}$$

Proof. We use our continuous time algorithm to construct a family $\{x_t\}$. We choose v_t at every step as follows: If there exists an ε-violated constraint

$$\langle A_i, w\rangle > \langle A_i, x_t\rangle - \varepsilon ,$$

then we set $v_s = A_i$ for $s \in [t, t_1]$, where t_1 is the first time at which the constraint becomes satisfied. We repeat this until no more ε-violated constraints exist.

By Lemma 2.3, in each such iteration the relative entropy $D(x^* \,\|\, x_t)$ drops by at least

$$\frac{\varepsilon^2}{2L^2}$$

where $L = \max\{\|A_1\|_\infty, \ldots, \|A_m\|_\infty\}$. Since the relative entropy is always non-negative, the total number of iterations is bounded by $\frac{2L^2}{\varepsilon^2} \cdot D(x^* \,\|\, x_0)$, completing the proof. $\square$

Achieving a sparse solution with online mirror descent. Finally, we observe that Theorem 1.6 yields a similar statement. We set

$$f(x) = \max\left\{(\langle A_i, x\rangle - b_i)_+ : i = 1, 2, \ldots, m\right\}.$$

Here we have used the notation $x_+ = \max(x, 0)$. Note that f is L-Lipschitz, where $L = \max\{\|A_1\|_\infty, \ldots, \|A_m\|_\infty\}$. Using the fact that Φ is 1-strongly convex, we can apply Theorem 1.6 with $x_0 = (\frac{1}{n}, \ldots, \frac{1}{n})$ and parameter $t = \frac{2L^2}{\varepsilon^2} D(x^* \,\|\, x_0)$ to obtain a distribution

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_t}{t} \in C,$$

where each x_i is of the form (2.2), with the identical bound $\frac{2L^2}{\varepsilon^2} D(x^* \,\|\, x_0)$ on the dual sparsity of each x_i (by "sparsity" here we mean the number of constraints that are touched in obtaining the solution $\bar{x}$).

2.2 General version

Much of the preceding section generalizes to the setting of strictly convex, twice-differentiable $\Phi : \mathcal{D} \to \mathbb{R}$ and the associated divergence D_Φ. If we have a function $f : C \to \mathbb{R}$ on a convex set C, one can consider the following continuous-time mirror descent:

$$x'(t) = \mathbf{J}\Pi(x(t)) \left(\nabla^2 \Phi(x(t))\right)^{-1} \left(-\nabla f(x(t))\right).$$

where $\Pi = \Pi_C^\Phi$ is the corresponding Bregman projection. It turns out (I think) that if Φ is a mirror map on $\mathcal{D}$, then this can be thought of as gradient descent on C as an embedded submanifold of $\mathcal{D}$, where the latter domain is equipped with the Riemannian metric coming from $\nabla^2\Phi$.

[More notes to be added here. Dear reader, if you know of a reference for this perspective, please let me know.]

The primal-dual view. There is a primal-dual perspective on the algorithm described in (1.2) that may be helpful. Every iteration involves three steps: (i) Moving from the primal to the dual via the map $x \mapsto \nabla\Phi(x)$, (ii) improving in the dual, (iii) projecting back to the feasible region.

As before, let $x_0 = \operatorname{argmin}_{x \in C} \Phi(x)$. Now given $x_t \in C$, we choose $y_{t+1} \in \mathcal{D}$ so that

$$\nabla\Phi(y_{t+1}) = \nabla\Phi(x_t) - \eta v_t \,. \tag{2.3}$$

Such a y_{t+1} exists by property (ii) of a mirror map. Finally, we define $x_{t+1} = \Pi_C^\Phi(y_{t+1})$ as the Bregman projection of $y_{t+1} \in \mathbb{R}^n$ back to the feasible region.

To see that this gives the same sequence $\{x_0, x_1, x_2, \ldots\}$ we saw before, observe that

$$\begin{aligned}
x_{t+1} = \Pi_C^\Phi(y_{t+1}) &= \operatorname*{argmin}_{x \in C} \; D_\Phi(x, y_{t+1}) \\
&= \operatorname*{argmin}_{x \in C} \; \Phi(x) - \Phi(y_{t+1}) - \langle \nabla\Phi(y_{t+1}), x - y_{t+1}\rangle \\
&= \operatorname*{argmin}_{x \in C} \; \Phi(x) - \langle \nabla\Phi(y_{t+1}), x\rangle && \text{using (2.3)} \\
&= \operatorname*{argmin}_{x \in C} \; \Phi(x) - \langle \nabla\Phi(x_t) - \eta v_t, x\rangle \\
&= \operatorname*{argmin}_{x \in C} \; \eta\langle x, v_t\rangle + D_\Phi(x, x_t)\,,
\end{aligned}$$

just as in (1.2).

3 Density approximation

For applications in the next few weeks, I want to move to a more functional setting (i.e., we will replace vectors by functions on a finite set).

Let X be a finite set, equipped with a measure μ. For a function $f : X \to \mathbb{R}$, we write

$$\mathop{\mathbb{E}}_{\mu}[f] = \sum_{x \in X} \mu(x) f(x)$$

for the expected value of f with respect to μ. We say that f is a *density with respect to* μ if $f(x) \geqslant 0$ for all $x \in X$, and $\mathbb{E}_\mu[f] = 1$. Let Δ_X denote the set of all densities on X.

For $f \in \Delta_X$, we introduce the notation

$$\mathrm{Ent}_\mu(f) = D(f\mu \,\|\, \mu) = \mathop{\mathbb{E}}_{\mu}[f \log f] \,.$$

If $g \in \Delta_X$ as well, define a relative entropy between the respective densities:

$$D_\mu(f \,\|\, g) = \mathop{\mathbb{E}}_{\mu}\left[f \log \frac{f}{g} \right] .$$

Recall that (as we have seen in Lecture 1), $\mathrm{Ent}_\mu(f)$ is a convex function of f: For $f, g \in \Delta_X$ and $\lambda \in [0,1]$, we have

$$\mathrm{Ent}_\mu(\lambda f + (1-\lambda) g) \leqslant \lambda \mathrm{Ent}_\mu(f) + (1-\lambda)\mathrm{Ent}_\mu(g) \,.$$

We will work in the inner product space $L^2(X, \mu)$ whose elements are functions $f : X \to \mathbb{R}$. The inner product of $f, g \in L^2(X, \mu)$ is given by

$$\langle f, g \rangle = \mathop{\mathbb{E}}_{\mu}[fg] \,.$$

Let us now restate Theorem 2.4 in the functional setting. Although the proof of a very similar statement follows rather immediately from Theorem 1.6; the origin of the theorem lies in the paper [Lee-Raghenvedra-Steurer 2015].

Theorem 3.1 (Dual-sparse approximation)**.** *Consider some $\mathcal{F} \subseteq L^2(X, \mu)$. Let $f \in \Delta_X$ and $\varepsilon > 0$ be given. Then there exist non-negative constants $\{c_\varphi : \varphi \in \mathcal{F}\}$ such that*

$$\#\{c_\varphi > 0 : \varphi \in \mathcal{F}\} \leqslant 2 \frac{\max_{\varphi \in \mathcal{F}} \|\varphi\|_\infty^2}{\varepsilon^2} \mathrm{Ent}_\mu(f) \,,$$

and the density

$$\tilde{f} = \frac{\exp\left(\sum_{\varphi \in \mathcal{F}} c_\varphi \varphi\right)}{\mathbb{E}_\mu \exp\left(\sum_{\varphi \in \mathcal{F}} c_\varphi \varphi\right)}$$

satisfies $\langle \tilde{f}, \varphi \rangle \geqslant \langle f, \varphi \rangle - \varepsilon$ for all $\varphi \in \mathcal{F}$.