Lecture 20. Bregman Divergence and Mirror Descent

20.1 Mirror descent: the proximal point view

Recall the gradient descent method, where we optimize

$$\tilde{f}(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\eta}\|x - x_k\|^2$$

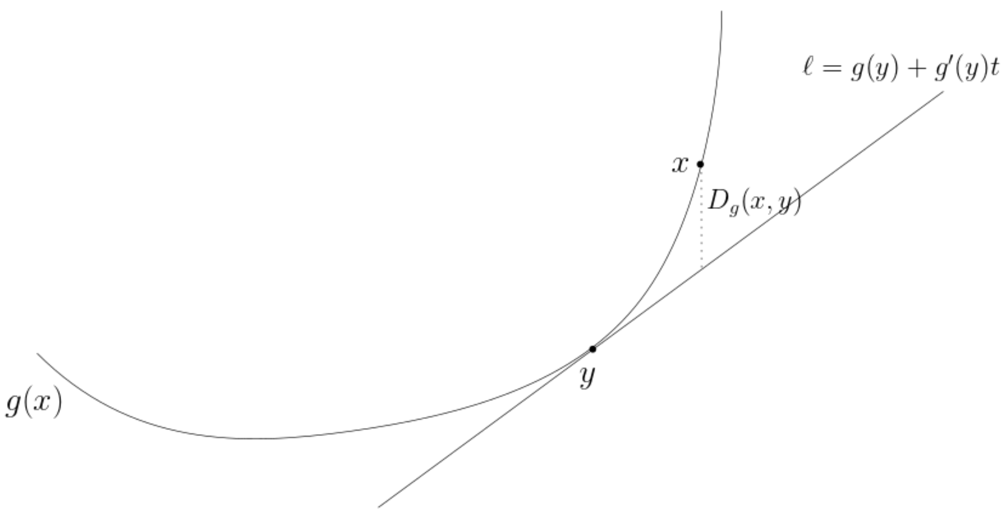
and let $x_{k+1} \leftarrow \arg\min\ \tilde{f}(x)$. A natural question is, can we use other functions instead of quadratic functions to approximate $f(x)$? Clearly, we hope the approximate function is easy to optimize, and somehow adapt the "geometry" of the problem.

The *mirror descent* framework allows us to do precisely this. Specifically, given an objective function f, we assume that there exists a convex function g to approximate f. Then we use the *Bregman divergence* with respect to g to replace the squared Euclidean norm in $\tilde{f}$ and still let $x_{k+1} \leftarrow \arg\min\ \tilde{f}(x)$, where the *Bregman divergence* is defined by

$$D_g(x, y) = g(x) - g(y) - \langle \nabla g(y),\ x - y \rangle$$

and thus $\tilde{f}$ can be expressed by

$$\tilde{f}(x) = f(x_k) + \langle \nabla f(x_k),\ x - x_k \rangle + \frac{1}{\eta} D_g(x, x_k)\,.$$



Dropping the constant terms (that only depends on x_k but not on x), the update step of the mirror descent is given by

$$x_{k+1} = \arg\min_{x} \left\{ \langle \nabla f(x_k),\, x \rangle + \frac{1}{\eta} D_g(x, x_k) \right\},$$

or equivalently,

$$x_{k+1} = \arg\min_{x} \left\{ \eta \langle \nabla f(x_k),\, x \rangle + D_g(x, x_k) \right\}.$$

Remark

What is the "right" choice of g to minimize the function f? A little thought shows that the "best" g should equal f , because adding $D_f(x, x_k)$ to the linear approximation of f at x_k gives us back exactly f. Of course, the update now requires us to minimize $f(x)$, which is the original problem. So we should choose a function g that is somehow "similar" to f, and make the update step tractable.

Bregman divergence

We now introduce more on Bregman divergence.

Definition (*Bregman divergence*)

Let $g : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable and strictly convex function. Then the *Bregman divergence* from y to x with respect to function g is defined by

$$D_g(x, y) = g(x) - g(y) - \langle \nabla g(y),\, x - y \rangle.$$

Here are some examples.

Example

1. *Euclidean distance*. Let $g(x) = \frac{1}{2}\|x\|_2^2$. Then the Bregman divergence with respect to g is

$$D_g(x, y) = \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - \langle y,\, x - y \rangle = \frac{1}{2}\|x - y\|_2^2.$$

2. *Negative entropy*. Let $\Delta_{n-1} \triangleq \left\{ \boldsymbol{x} \in \mathbb{R}^n \mid \sum_{i=1}^n x_i = 1 \text{ and } x_i > 0 \text{ for all } i = 1, \ldots, n \right\}$ be the (open) standard $(n-1)$-simplex, and $g(\boldsymbol{x}) : \Delta_{n-1} \to \mathbb{R} = \sum_{i=1}^n x_i \log x_i$ be the negative entropy function over Δ_{n-1}. Then the Bregman divergence with respect to g is

$$\begin{aligned} D_g(\boldsymbol{x}, \boldsymbol{y}) &= g(\boldsymbol{x}) - g(\boldsymbol{y}) - \langle \nabla g(\boldsymbol{y}), \, \boldsymbol{x} - \boldsymbol{y} \rangle \\ &= \sum_{i=1}^n x_i \log x_i - \sum_{i=1}^n y_i \log y_i - \sum_{i=1}^n (1 + \log y_i)(x_i - y_i) \\ &= \sum_{i=1}^n x_i (\log x_i - \log y_i) - \sum_{i=1}^n (x_i - y_i) \\ &= \sum_{i=1}^n x_i \log \frac{x_i}{y_i} \, . \end{aligned}$$

This is called the *relative entropy*, or *Kullback-Leibler divergence* (*KL divergence*) between probability distribution $\boldsymbol{x}$ and $\boldsymbol{y}$, measuring the expected number of extra bits required to code samples from distribution $\boldsymbol{x}$ using a code optimized for $\boldsymbol{y}$ rather than the code optimized for $\boldsymbol{x}$.

Since g is a strictly convex function, for any fixed y, we know that $D_g(x, y)$ is also a (strictly) convex function in the first argument x. But it is not convex in the second argument y in general.

Remark

It is clear that $D_g(x, x) = 0$ for all $x \in \mathbb{R}^n$. Since g is strictly convex, by the first order condition for convexity, we know that $D_g(x, y) > 0$ if $x \neq y$. Furthermore, if g is μ-strongly convex, then $D_g(x, y) \geq \frac{\mu}{2} \|x - y\|_2^2$ by definition. So the Bregman divergence somehow measures the (squared) distance from y to x. But we should note that in general the Bregman divergence is **NOT** symmetric. For example, see KL divergence.

Consider a well-known puzzle: given k points $x_1, \ldots, x_k$, the goal is to find a point y to minimize the total (squared) *distances* from y to $x_1, \ldots, x_k$. A natural idea is to choose the mean of $x_1, \ldots, x_k$. For example, in a triangle, the *centroid* is the point that minimizes the sum of the squared distances of a point from the three vertices. The Bregman divergence encodes a kind of (squared) distances that the mean of distribution works.

Lemma

Suppose $\boldsymbol{x}$ is a random variable over an open set with distribution μ. Then

$$\min_{\boldsymbol{y}\in S} \mathbb{E}_{\boldsymbol{x}\sim\mu}[D_g(\boldsymbol{x}, \boldsymbol{y})]$$

is optimized at $\boldsymbol{y}^* = \bar{\boldsymbol{x}} \triangleq \mathbb{E}_{\boldsymbol{x}\sim\mu}[\boldsymbol{x}] = \int_{\boldsymbol{x}\in S} \boldsymbol{x}\,\mathrm{d}\mu$.

Proof

For any $\boldsymbol{y} \in S$, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\mu}[D_g(\boldsymbol{x}, \boldsymbol{y})] - \mathbb{E}_{\boldsymbol{x}\sim\mu}[D_g(\boldsymbol{x}, \bar{\boldsymbol{x}})] \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mu}\Big[\big(g(\boldsymbol{x}) - g(\boldsymbol{y}) - \langle \nabla g(\boldsymbol{y}),\, \boldsymbol{x} - \boldsymbol{y}\rangle\big) - \big(g(\boldsymbol{x}) - g(\bar{\boldsymbol{x}}) - \langle \nabla g(\bar{\boldsymbol{x}}),\, \boldsymbol{x} - \bar{\boldsymbol{x}}\rangle\big)\Big] \\
&= g(\bar{\boldsymbol{x}}) - g(\boldsymbol{y}) + \langle \nabla g(\boldsymbol{y}),\, \boldsymbol{y}\rangle - \langle \nabla g(\bar{\boldsymbol{x}}),\, \bar{\boldsymbol{x}}\rangle + \mathbb{E}_{\boldsymbol{x}\sim\mu}\big[-\langle \nabla g(\boldsymbol{y}),\, \boldsymbol{x}\rangle + \langle \nabla g(\bar{\boldsymbol{x}}),\, \boldsymbol{x}\rangle\big] \\
&= g(\bar{\boldsymbol{x}}) - g(\boldsymbol{y}) + \langle \nabla g(\boldsymbol{y}),\, \boldsymbol{y}\rangle - \langle \nabla g(\bar{\boldsymbol{x}}),\, \bar{\boldsymbol{x}}\rangle + \Big\langle \nabla g(\bar{\boldsymbol{x}}) - \nabla g(\boldsymbol{y}),\, \mathbb{E}[\boldsymbol{x}]\Big\rangle \\
&= g(\bar{\boldsymbol{x}}) - g(\boldsymbol{y}) + \langle \nabla g(\boldsymbol{y}),\, \boldsymbol{y} - \bar{\boldsymbol{x}}\rangle \\
&= D_g(\bar{\boldsymbol{x}},\, \boldsymbol{y})\,.
\end{aligned}$$

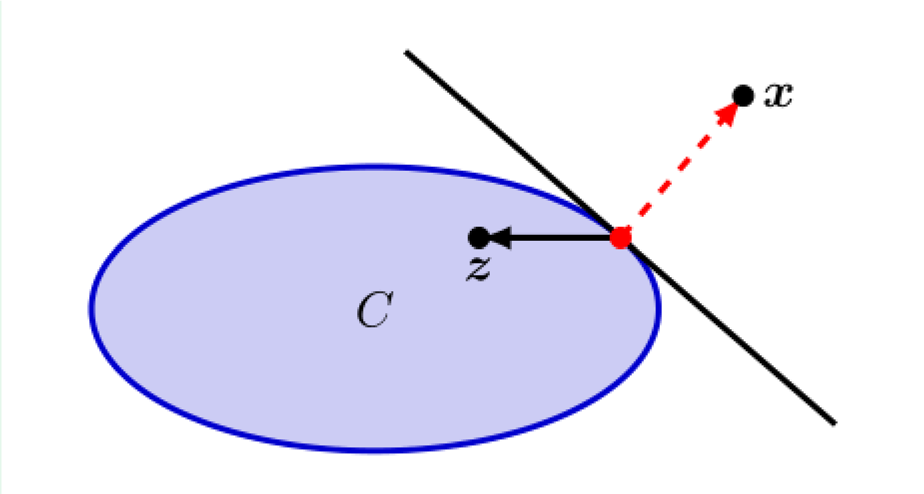
This must be nonnegative, and equal 0 if and only if $\boldsymbol{y} = \bar{\boldsymbol{x}}$.

Perhaps a surprising result is that Bregman divergence is an *exhaustive* notion for such (squared) distances. In other words, if a kind of distance satisfies the above lemma, then it must be a Bregman divergence. See, e.g., 1 or 2 for proof details.

The Bregman divergence is also a right way to describe the (squared) distance from a point to a convex set. Recall that, in Lecture 4, we show the following lemma, which means $\angle \boldsymbol{xyz}$ is obtuse.

Lemma

Let C be a nonempty, closed and convex set. Given $\boldsymbol{x}$ and $\boldsymbol{y} = \mathcal{P}_C(\boldsymbol{x})$, for any $\boldsymbol{z} \in C$, it holds that $\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{z} - \boldsymbol{y}\rangle \le 0$.



We now establish a similar result using Bregman divergence. If $\boldsymbol{x}^*$ is the projection of $\boldsymbol{x}_0$ onto a convex set C, namely,

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in C} D_g(\boldsymbol{x}, \boldsymbol{x}_0)\,.$$

Then for all $\boldsymbol{y} \in C$, it holds that

$$D_g(\boldsymbol{y}, \boldsymbol{x}_0) \geq D_g(\boldsymbol{y}, \boldsymbol{x}^*) + D_g(\boldsymbol{x}^*, \boldsymbol{x}_0)\,. \tag{♠}$$

In Euclidean case, it also means that the angle $\angle \boldsymbol{y}\boldsymbol{x}^*\boldsymbol{x}_0$ is obtuse, by the *generalized Pythagorean theorem* (*law of cosines*) $c^2 = a^2 + b^2 - 2ab\cos\gamma$. The proof is a simple application of the *law of cosines for Bregman divergence*. Since

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in C} D_g(\boldsymbol{x}, \boldsymbol{x}_0)\,,$$

we have

$$\left\langle \nabla D_g(\boldsymbol{x}, \boldsymbol{x}_0)\big|_{\boldsymbol{x}=\boldsymbol{x}^*},\, \boldsymbol{y} - \boldsymbol{x}^* \right\rangle \geq 0$$

for all $\boldsymbol{y} \in C$. Note that $\nabla D_g(\boldsymbol{x}, \boldsymbol{x}_0) = \nabla g(\boldsymbol{x}) - \nabla g(\boldsymbol{x}_0)$. So the above inequality is equivalent to

$$\langle \nabla g(\boldsymbol{x}^*) - \nabla g(\boldsymbol{x}_0),\, \boldsymbol{y} - \boldsymbol{x}^* \rangle \geq 0\,.$$

Then the proof concludes with the following lemma (by setting $x = \boldsymbol{y}$, $y = \boldsymbol{x}^*$, and $z = \boldsymbol{x}_0$).

Lemma (*Law of cosines for Bregman divergence*)

$$\begin{aligned} D_g(x,y) + D_g(y,z) &= g(x) - g(y) - \langle \nabla g(y),\, x - y\rangle + g(y) - g(z) - \langle \nabla g(z),\, y - z \\ &= g(x) - g(z) - \langle \nabla g(z),\, x - z\rangle - \langle \nabla g(z),\, y - x\rangle - \langle \nabla g(y),\, x \\ &= D_g(x,z) + \langle \nabla g(z) - \nabla g(y),\, x - y\rangle \end{aligned}$$

20.2 Mirror descent: the mirror map view

A different view of the mirror descent framework is the one originally presented by Arkadi Nemirovski and David Yudin. Recall that in the gradient descent, we update the iterate by $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \nabla f(\boldsymbol{x}_k)$. However, the gradient was actually defined as a *linear functional* on $\mathbb{R}^n$ (a linear map from the vector space $\mathbb{R}^n$ into its underlying field $\mathbb{R}$). Hence, $\nabla f(\boldsymbol{x})$ naturally belongs to the dual space of $\mathbb{R}^n$. The fact that we represent this functional as a vector is a matter of convenience, and highly depends on the choice of coordinates. In fact, that's why the gradient descent is not *affinely invariant*.

In the vanilla gradient descent method, we only consider $\mathbb{R}^n$ with L^2-norm, and this normed space is self-dual, so it is perhaps reasonable to combine points in the primal space (the iterates $\boldsymbol{x}_k$) with objects in the dual space (the gradients $f(\boldsymbol{x}_k)$). But when working with other normed spaces, adding a linear map $\nabla f(\boldsymbol{x}_k)$ to a vector $\boldsymbol{x}_k$ might not be the right thing to do.

Instead, Nemirovski and Yudin propose the following:

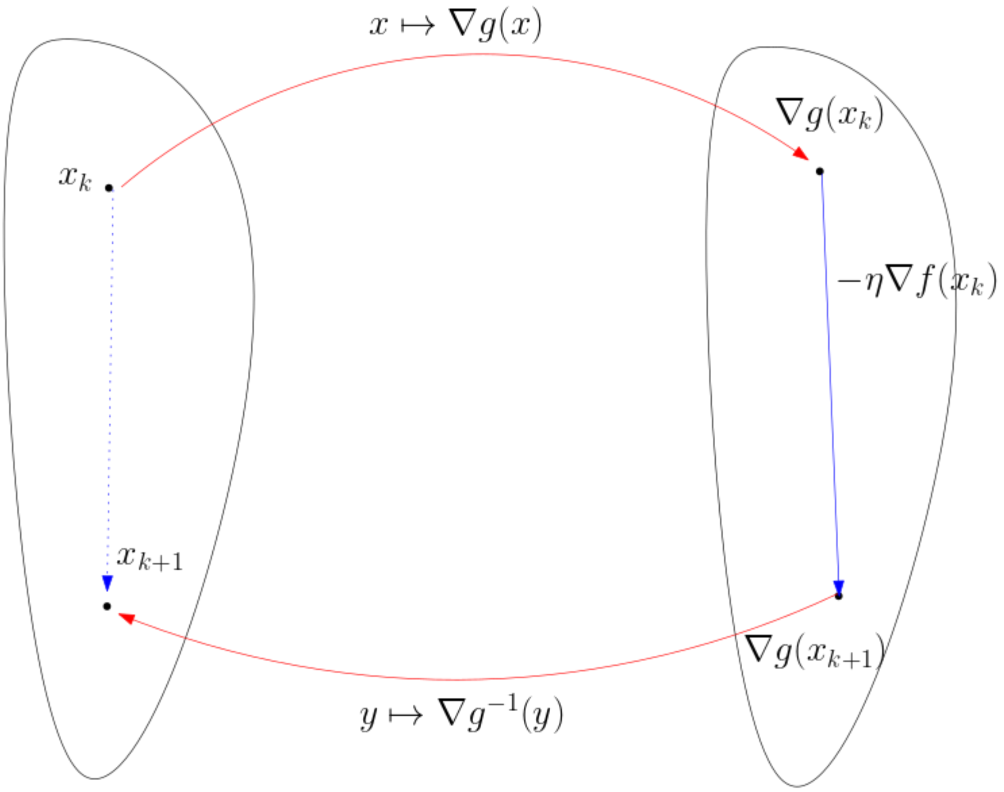
1. we map our current point $\boldsymbol{x}_k$ to a point $\boldsymbol{y}_k$ in the dual space using a *mirror map*.
2. Next, we take the gradient step $\boldsymbol{y}_{k+1} \leftarrow \boldsymbol{y}_k - \eta \nabla f(\boldsymbol{x}_k)$.
3. We map $\boldsymbol{y}_{k+1}$ back to a point in the primal space $\boldsymbol{x}'_{k+1}$ using the inverse of the mirror map from Step 1.
4. If we are in the constrained case, this point $\boldsymbol{x}'_{k+1}$ might not be in the convex feasible region Ω, so we still need to project $\boldsymbol{x}'_{k+1}$ back to a close point $\boldsymbol{x}_{k+1}$ in Ω.

How do we choose these mirror maps? Again, this comes down to understanding the geometry of the problem, the kinds of functions and feasible sets Ω we care about. We usually choose a proper differentiable and strongly convex function $g(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$, and define the *mirror map* by $\nabla g : \mathbb{R}^n \to \mathbb{R}^n$, that is,

$$\boldsymbol{x} \mapsto \nabla g(\boldsymbol{x}) \,.$$

Since g is differentiable and strongly convex, its gradient is "monotone", and thus the inverse mirror map exists. We can use these maps in the Nemirovski-Yudin process, namely, we set

$$\boldsymbol{y}_k = \nabla g(\boldsymbol{x}_k) \quad \text{and} \quad \boldsymbol{x}_{k+1} = \nabla g^{-1}(\boldsymbol{y}_{k+1})\,.$$



The name of the process comes from thinking of the *dual space as being a mirror image of the primal space*.

But why this view and the proximal point view give the same algorithm? We consider the update rule in the proximal point view

$$x_{k+1} = \arg\min_x \left\{ \langle \nabla f(x_k),\, x\rangle + \frac{1}{\eta} D_g(x, x_k) \right\},$$

and consider the gradient of Bregman divergence

$$\nabla D_g(x, x_k) = \nabla g(x) - \nabla g(x_k)\,.$$

Since $x \mapsto \langle \nabla f(x_k),\, x\rangle + \frac{1}{\eta} D_g(x, x_k)$ is a convex function, we obtain that

$$\nabla \Big(\langle \nabla f(x_k),\, x\rangle + \frac{1}{\eta} D_g(x, x_k) \Big)\Big|_{x = x_{k+1}} = \mathbf{0}\,,$$

which is

$$\nabla f(x_k) + \frac{1}{\eta}\Big(\nabla g(x_{k+1}) - \nabla g(x_k) \Big) = \mathbf{0}\,.$$

Rearranging terms it gives a step of update in the dual space

$$\nabla g(x_{k+1}) = \nabla g(x_k) - \eta \nabla f(x_k)\,.$$

Dual space and dual norm

Given any vector space V over a field $\mathbb{F}$, the (algebraic) *dual space* V^* is defined as the set of all linear map $\varphi : V \to \mathbb{F}$ (*linear functional*). Since linear maps are vector space homomorphisms, the dual space may be denoted $\hom(V, \mathbb{F})$. The dual space V^* itself becomes a vector space over $\mathbb{F}$ when equipped with an addition and scalar multiplication satisfying:

$$(\varphi + \psi)(x) = \varphi(x) + \psi(x)$$
$$(a\varphi)(x) = a(\varphi(x))$$

for all $\phi, \psi \in V^*$, $x \in V$, and $a \in \mathbb{F}$.

If V is finite-dimensional, then V^* has the same dimension as V. In particular, $\mathbb{R}^n$ can be interpreted as the space of columns of n real numbers, its dual space is typically written as the space of rows of n real numbers. Such a row acts on $\mathbb{R}^n$ as a linear functional by ordinary matrix multiplication. This is because a functional maps every n-vector $\boldsymbol{x}$ into a real number y. Then, seeing this functional as a matrix $\boldsymbol{M}$, and $\boldsymbol{x}$ as an $n \times 1$ matrix, and y a 1×1 matrix (trivially, a real number) respectively, if $\boldsymbol{M}\boldsymbol{x} = y$, then by dimension reasons, $\boldsymbol{M}$ must be a $1 \times n$ matrix, that is, a row vector. So there is an *isomorphism* between $\mathbb{R}^n$ (and any finite-dimensional vector space V) and its dual space.

However, it is not a *canonical isomorphism*. Informally, an isomorphism is a map that preserves sets and relations among elements. When this map or this correspondence is established with no choices involved, it is called *canonical isomorphism*. When we defined V^* from V we did so by picking a special basis (the dual basis), therefore the isomorphism from V to V^* is not canonical. But for the *double dual* V^{**} of a finite-dimensional vector space V (the dual of the normed vector space V^*), there is a canonical isomorphism. Indeed, the following map $\pi : V \to V^{**}$ defined as follows is a canonical isomorphism. For any $v \in V$, $\pi(v) \in V^{**}$ is a map from V^* to $\mathbb{F}$ given by

$$\forall\, \varphi \in V^* : V \to \mathbb{F}\,, \qquad \pi(v)(\varphi) \triangleq \varphi(v)\,.$$

Given a norm $\|\cdot\|$ on a vector space V, its *dual norm*, denoted by $\|\cdot\|_*$, is a function (a norm) of a linear functional φ belonging to V^* defined by

$$\|\varphi\|_* \triangleq \sup\{|\varphi(v)| : v \in V,\ \|v\| \leq 1\}\,.$$

In particular, for $\mathbb{R}^n$, a linear functional can be represented by a vector with inner product. Thus, the dual norm is given by

$$\|\boldsymbol{u}\|_* = \sup\{\langle \boldsymbol{u}, \boldsymbol{v}\rangle : \|\boldsymbol{v}\| \le 1\}\,.$$

By Cauchy-Schwarz inequality, the dual norm of the L^2-norm is again the L^2-norm. In general, the dual for the L^p-norm is the L^q-norm, where $1/p + 1/q = 1$ and we assume $1/\infty = 0$ for convenience.

Similar to the double dual space, for a finite-dimensional space with norm $\|\cdot\|$, we have $(\|\cdot\|_*)_* = \|\cdot\|$.

20.3 Convex conjugate

We now focus on how to implement mirror descent. We need to show that the inverse gradient $(\nabla g)^{-1}$ can be computed efficiently.

For any convex function f with domain $D \subseteq \mathbb{R}^n$, the gradient of f at some point x is a vector (actually a covector) v satisfying

$$f(y) \ge f(x) + \langle v,\, y - x\rangle$$

for all $y \in D$. More generally, the subgradients of f is the set of all such vectors, namely,

$$\partial f(x) = \{v \in \mathbb{R}^n \mid f(y) \ge f(x) + \langle v,\, y - x\rangle \text{ for all } y\}\,.$$

Rearranging terms we obtain

$$\langle v,\, y\rangle - f(y) \le \langle v,\, x\rangle - f(x)$$

for all $y \in D$. Note that $x \in D$. It gives that

$$\max_{y\in D} \langle v,\, y\rangle - f(y) = \langle v,\, x\rangle - f(x)\,.$$

Thus we can rewrite the subgradients as

$$\partial f(x) = \{v \in \mathbb{R}^n \mid \max_{y\in D}\{\langle v,\, y\rangle - f(y)\} = \langle v,\, x\rangle - f(x)\}\,.$$

We can now introduce the *convex conjugate* of a function.

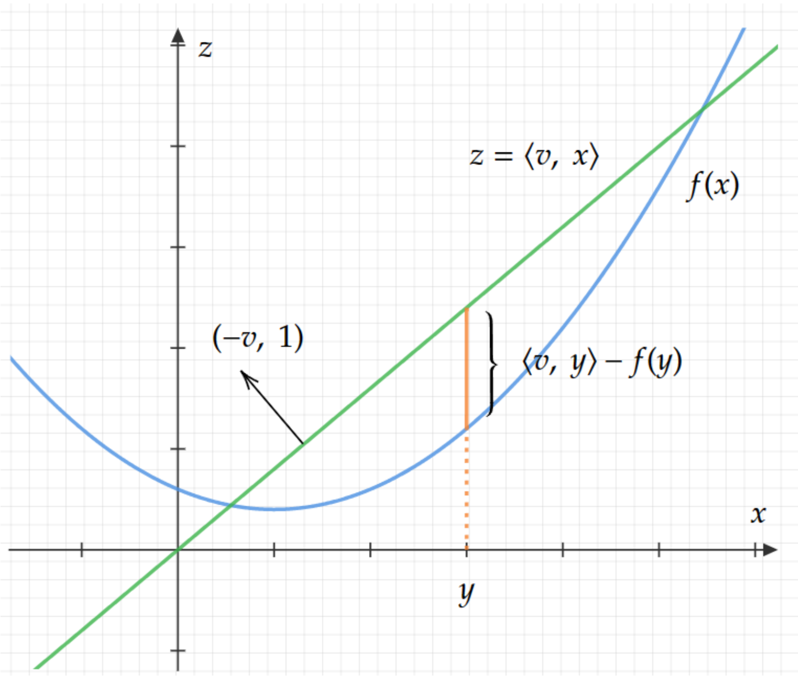
Definition (*Convex conjugate*)

Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a convex function. Its *convex conjugate* is the function $f^*(v) : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ given by

$$f^*(v) \triangleq \sup_{y \in D} \langle v,\, y \rangle - f(y)\,.$$

Note that for any fixed y, $\langle v,\, y \rangle - f(y)$ is an affine function of v. Thus $f^*(v)$ is a convex function of v (by the convexity of pointwise supremum).

In fact, f^* is defined on the dual space of $\mathbb{R}^n$. Roughly speaking, for each $\boldsymbol{v} \in \mathbb{R}^n$, one can think of it as the hyperplane $\{(\boldsymbol{x}, z)^\mathsf{T} \in \mathbb{R}^{n+1} \mid z = \langle \boldsymbol{v}, \boldsymbol{x} \rangle\}$ with the normal vector $(-\boldsymbol{v}, 1)^\mathsf{T}$. Then $f^*(\boldsymbol{v})$ gives the longest (directed) vertical distance between the hyperplane and the graph of $f(\boldsymbol{x})$. In other words, $f^*(\boldsymbol{v})$ is how far down you can translate the hyperplane so that the entire hyperplane is just below the graph of $f(\boldsymbol{x})$, namely, becomes the supporting hyperplane of the epigraph. So this definition can be interpreted as an encoding of the convex hull of the function's epigraph in terms of its supporting hyperplanes.



Example

We now see some examples.

1. Let $f(\boldsymbol{x}) = \langle \boldsymbol{a},\, \boldsymbol{x} \rangle - b$ be an affine function. Its convex conjugate is

$$f^*(\boldsymbol{v}) = \begin{cases} b\,, & \boldsymbol{v} = \boldsymbol{a} \\ +\infty\,, & \boldsymbol{v} \neq \boldsymbol{a}\,. \end{cases}$$

2. Let $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$ be a quadratic function. Its convex conjugate is

$$f^*(\boldsymbol{v}) = \sup_{\boldsymbol{x} \in \mathbb{R}^n} \left\{ \langle \boldsymbol{v}, \boldsymbol{x} \rangle - \frac{1}{2}\|\boldsymbol{x}\|^2 \right\} = \frac{1}{2}\|\boldsymbol{x}\|^2\,.$$

3. Let $f(x) = x \log x$. Its convex conjugate is

$$f^*(v) = \sup_{x \in \mathbb{R}} \; vx - x \log x = ve^{v-1} - f(e^{v-1}) = e^{v-1}\,.$$

4. Let $f(x) = e^x$. It convex conjugate is

$$f^*(v) = \sup_{x \in \mathbb{R}} vx - e^x = \begin{cases} v \log v - v\,, & v > 0 \\ 0\,, & v = 0 \\ +\infty\,, & v < 0\,. \end{cases}$$

It is easy to see that $v \in \partial f(x)$ (in particular, $v = \nabla f(x)$ if f is differentiable at x) if and only if $f^*(v) = \langle v,\, x\rangle - f(x)$. Otherwise ($v \neq \nabla f(x)$) we have $f^*(v) > \langle v,\, x\rangle - f(x)$, which gives the following Fenchel's inequality.

Theorem (*Fenchel's inequality*)

For all $x \in D$ and $v \in \mathbb{R}^n$, we have

$$f(x) + f^*(v) \geq \langle v,\, x \rangle\,.$$

The equality holds if and only if $v = \nabla f(x)$ (or $v \in \partial f(x)$ in general).

It is still not easy to compute $(\nabla f)^{-1}$ by Fenchel's inequality. We need the following direct corollary.

Theorem (*Fenchel-Moreau theorem*)

For any convex function $f : \mathbb{R}^n \to \mathbb{R}$, we have $f = f^{**}$.

Proving the theorem in full generality (the domain of f is given by $D \subseteq \mathbb{R}^n$) requires a bit of care. But it is relatively straightforward to show that the result holds on the interior of D. For simplicity, we only consider the case where $D = \mathbb{R}^n$. The proof consists of two parts: (1). proving that $f(x) \geq f^{**}(x)$ for all x; (2). proving that $f(x) \leq f^{**}(x)$.

Proof

By definition we have

$$f^{**}(x) = \sup_{v \in \mathbb{R}^n} \langle v, x \rangle - f^*(v) \,.$$

Note that $-f^*(v) = \inf_{y \in \mathbb{R}^n} f(y) - \langle v, y \rangle$. In particular,

$$-f^*(v) \leq f(x) - \langle v, x \rangle \,.$$

Thus, $f^{**}(x) \leq \sup_{v \in \mathbb{R}^n} \{\langle v, x \rangle + f(x) - \langle v, x \rangle\} = f(x)$.
For any $x \in \mathbb{R}^n$, let $u = \nabla f(x)$ (or $u \in \partial f(x)$ for general non-differentiable f). Then by Fenchel's inequality we have

$$\langle u, x \rangle = f(x) + f^*(u) \,.$$

So

$$f^{**}(x) = \sup_{v \in \mathbb{R}^n} \langle v, x \rangle - f^*(v) \geq \langle u, x \rangle - f^*(u) = f(x) \,.$$

Now we can show that

Corollary

If $f : \mathbb{R}^n \to \mathbb{R}$ is strictly convex, then

$$(\nabla f)^{-1} \equiv \nabla f^* \,.$$

More generally, if the domain of f is $D \subseteq \mathbb{R}^n$, then

$$\big(x \in D, v \in \partial f(x)\big) \quad \Longleftrightarrow \quad x \in \partial f^*(v) \cap D \,.$$

Proof

Let $v = \nabla f(x)$. By Fenchel's inequality we have

$$f(x) + f^*(v) = \langle v, x \rangle \,.$$

By Fenchel-Moreau theorem, it is equivalent to

$$f^*(v) + f^{**}(x) = \langle v, x \rangle \,,$$

which gives $x = \nabla f^*(v)$ if we apply Fenchel's inequality again.

Example

Consider $f(x) = x \log x$ on $(0, 1)$. We know that $f^*(v) = e^{v-1}$. If we take $v = \nabla f(x) = \log x + 1$, then

$$x = e^{v-1} = \nabla f^*(v) \,.$$

Now we put convex conjugate together with Bregman divergence. Let $g : \mathbb{R}^n \to \mathbb{R}$ be a differentiable and strictly convex function. Then $g^* : \mathbb{R}^n \to \mathbb{R}$ is also differentiable and strictly convex. The Bregman divergence with respect to g and g^* are

$$D_g(x, y) = g(x) - g(y) - \langle \nabla g(y),\, x - y \rangle \,,$$

and

$$D_{g^*}(u, v) = g^*(u) - g^*(v) - \langle \nabla g^*(v),\, u - v \rangle \,.$$

Let $u = \nabla g(x)$ and $v = \nabla g(y)$ in the Bregman divergence with respect to g^*. Then we have

$$g(x) + g^*(u) = \langle u,\, x \rangle \quad \text{and} \quad g(y) + g^*(v) = \langle v,\, y \rangle \,.$$

Thus $B_{g^*}(u, v)$ simplifies to

$$D_{g^*}(u, v) = \langle u,\, x \rangle - g(x) - \langle v,\, y \rangle + g(y) - \langle y,\, u - v \rangle = g(y) - g(x) + \langle u,\, x - y \rangle \,,$$

which gives the following result.

Theorem

Let $g : \mathbb{R}^n \to \mathbb{R}$ be a differentiable and strictly convex function. Then for any $x, y \in \mathbb{R}^n$ it holds that

$$D_{g^*}\big(\nabla g(x), \nabla g(y)\big) = D_g(y, x) \,.$$

20.4 Convergence of mirror descent

We now consider the convergence analysis of mirror descent. Similar to the analysis for gradient descent, we hope to establish the connection between $f(\boldsymbol{x}_k)$ and $f(\boldsymbol{x}^*)$ in terms of Bregman divergence. The basic ingredient is equation (♠). In general, given any convex function $L(\boldsymbol{x})$, let $\boldsymbol{x}^*$ be the following minimizer

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in C} \left\{ L(\boldsymbol{x}) + D_g(\boldsymbol{x}, \boldsymbol{x}_0) \right\}.$$

Then for all $\boldsymbol{y} \in C$, it holds that

$$L(\boldsymbol{y}) + D_g(\boldsymbol{y}, \boldsymbol{x}_0) \geq L(\boldsymbol{x}^*) + D_g(\boldsymbol{y}, \boldsymbol{x}^*) + D_g(\boldsymbol{x}^*, \boldsymbol{x}_0).$$

Recall the mirror descent update

$$\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x}} \left\{ \eta \langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x} \rangle + D_g(\boldsymbol{x}, \boldsymbol{x}_k) \right\}.$$

It gives that for all $\boldsymbol{y}$,

$$\eta \langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{y} \rangle + D_g(\boldsymbol{y}, \boldsymbol{x}_k) \geq \eta \langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x}_{k+1} \rangle + D_g(\boldsymbol{y}, \boldsymbol{x}_{k+1}) + D_g(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k).$$

Rearranging terms we obtain that

$$\eta \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{y} - \boldsymbol{x}_k \rangle \geq \eta \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle + D_g(\boldsymbol{y}, \boldsymbol{x}_{k+1}) + D_g(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) - D_g(\boldsymbol{y}, \boldsymbol{x}_k).$$

Note that $f(\boldsymbol{y}) \geq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{y} - \boldsymbol{x}_k \rangle$, and

$$\begin{aligned} D_g(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}) + D_g(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) &= -\langle \nabla g(\boldsymbol{x}_{k+1}),\, \boldsymbol{x}_k - \boldsymbol{x}_{k+1} \rangle - \langle \nabla g(\boldsymbol{x}_k),\, \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle \\ &= \langle \nabla g(\boldsymbol{x}_{k+1}) - \nabla g(\boldsymbol{x}_k),\, \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle \\ &= -\eta \langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle. \end{aligned}$$

Hence we have

$$f(\boldsymbol{y}) - f(\boldsymbol{x}_k) \geq \frac{1}{\eta} \Big(D_g(\boldsymbol{y}, \boldsymbol{x}_{k+1}) - D_g(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}) - D_g(\boldsymbol{y}, \boldsymbol{x}_k) \Big)$$

for all $\boldsymbol{y}$. Now we can give the following lemma.

Theorem

Let f be a convex and L-Lipschitz function with respect to some norm $\|\cdot\|$, and g be a σ-strongly convex function with respect to the same norm. Suppose $D_g(\boldsymbol{x}^*, \boldsymbol{x}_0)$ can be bounded by R. Then by selecting

$$\eta = \frac{\sigma}{L} \sqrt{\frac{2R}{\sigma T}},$$

it holds that

$$\min_{k=0,\dots,T-1} f(\boldsymbol{x}_k) \leq f(\boldsymbol{x}^*) + L \sqrt{\frac{2R}{\sigma T}}.$$

In other words, if we would like to obtain an (approximate) answer that is less than $f(\boldsymbol{x}^*) + \varepsilon$, it is sufficient to run the mirror descent $O(L^2R/\varepsilon^2)$ steps.

Proof

By previous analysis we have

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}_k) \geq \frac{1}{\eta}\Big(D_g(\boldsymbol{x}^*, \boldsymbol{x}_{k+1}) - D_g(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}) - D_g(\boldsymbol{x}^*, \boldsymbol{x}_k)\Big)\,.$$

Summing over both sides from 0 to $T-1$, it implies that

$$\sum_{k=0}^{T-1} f(\boldsymbol{x}) \leq Tf(\boldsymbol{x}^*) + \frac{1}{\eta}\Big(D_g(\boldsymbol{x}^*, \boldsymbol{x}_0) - D_g(\boldsymbol{x}^*, \boldsymbol{x}_T) + \sum_{k=0}^{T-1} D_g(\boldsymbol{x}_k, \boldsymbol{x}_{k+1})\Big)\,.$$

The remaining part is to bound $\sum D_g(\boldsymbol{x}_k, \boldsymbol{x}_{k+1})$.
We assume f is differentiable for convenience. Then f is L-Lipschitz with respect to some norm $\|\cdot\|$ if and only if its gradients are bounded by L with respect to the dual norm $\|\cdot\|_*$. Otherwise there exists $\boldsymbol{x}, \boldsymbol{v} \in \mathbb{R}^n$ such that $\langle \nabla f(\boldsymbol{x}), v\rangle > L$. So f is not L-Lipschitz for $\boldsymbol{x}$ and $\boldsymbol{x} + \delta\boldsymbol{v}$ with sufficiently small $\delta > 0$.
Since g is σ-strongly convex with respect to the same norm $\|\cdot\|$, we have

$$\begin{aligned} D_g(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}) &= \eta\,\langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x}_k - \boldsymbol{x}_{k+1}\rangle - D_g(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) \\ &\leq \eta L\|\boldsymbol{x}_k - \boldsymbol{x}_{k+1}\| - \frac{\sigma}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 \\ &\leq \frac{\eta^2 L^2}{2\sigma}\,. \end{aligned}$$

Thus we obtain that

$$\begin{aligned} \frac{1}{T}\sum_{k=0}^{T-1} f(\boldsymbol{x}_k) &\leq f(\boldsymbol{x}^*) + \frac{D_g(\boldsymbol{x}^*, \boldsymbol{x}_0)}{\eta T} + \frac{\eta L^2}{2\sigma} \\ &\leq f(\boldsymbol{x}^*) + \frac{R}{\eta T} + \frac{\eta L^2}{2\sigma} \\ &= f(\boldsymbol{x}^*) + L\sqrt{\frac{2R}{\sigma T}}\,. \end{aligned}$$

Note that this results holds even for non-differentiable f. We only need to replace $\nabla f(\boldsymbol{x}_k)$ by some subgradient $\boldsymbol{v} \in \partial f(\boldsymbol{x})$ in the previous analysis.

To see the advantage of mirror descent, suppose f is L-Lipschitz with respect to some norm (which means the gradient of f can be bounded by L with respect to its dual norm), and g is σ-strongly convex with respect to the same norm. Then f is $L\sqrt{2/\sigma}$-Lipschitz with respect to the Bregman divergence. We can choose a particular norm and a particular Bregman divergence to capture the geometry of the problem.

We now give an example. Suppose Δ_{n-1} is the (open) n-dimensional probability simplex, and we use KL-divergence for which g is 1-strongly convex with respect to the L^1 norm. The dual norm of the L^1-norm is the L^∞-norm. Then we can bound $D_g(\boldsymbol{x}^*, \boldsymbol{x}_0)$ by using KL divergence, and it is at most $\log n$ if we set $\boldsymbol{x}_0 = \frac{1}{n}\mathbf{1}$ and $\boldsymbol{x}^*$ lies in the probability simplex. Suppose the objective function f is L-Lipschitz with respect to L^1-norm (and thus is $L\sqrt{n}$-Lipschitz with respect to L^2-norm). So the mirror descent requires $O(L^2 \log n)$ time to approximate $\boldsymbol{x}^*$, which is smaller than that of subgradient descent by an order of $O(L^2 n)$. Note the saving of n term is from the norm of gradient by replacing the L^2-norm by the L^∞-norm (decreasing by an order of $\sqrt{n}$), at a slight cost of increasing $D_g(\boldsymbol{x}^*, \boldsymbol{x}_0)$ by $\log n$.

Furthermore, if f is L-smooth with respect to some norm $\|\cdot\|$ (the gradient of f is L-Lipschitz continuous), namely,

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \,,$$

then the convergence rate can be better.

Theorem

Let f be a convex and L-smooth function with respect to some norm $\|\cdot\|$, and g be a σ-strongly convex function with respect to the same norm. Suppose $D_g(\boldsymbol{x}^*, \boldsymbol{x}_0)$ can be bounded by R. Then by selecting

$$\eta = \frac{\sigma}{L} \,,$$

it holds that

$$\min_{k=0,\ldots,T-1} f(\boldsymbol{x}_k) \leq f(\boldsymbol{x}^*) + \frac{LR}{\sigma T} \,.$$

This result gives $O(1/T)$ convergence rate to obtain an (approximate) optimal value.

Proof

We start again from

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}_k) \geq \frac{1}{\eta}\Big(D_g(\boldsymbol{x}^*, \boldsymbol{x}_{k+1}) - D_g(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}) - D_g(\boldsymbol{x}^*, \boldsymbol{x}_k)\Big)\,. \qquad (\star)$$

Now we bound $D_g(\boldsymbol{x}_k, \boldsymbol{x}_{k+1})$ by $|f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1})|$. Since f is L-smooth and g is σ-strongly convex, we have

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) \leq \langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle + \frac{L}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2$$

and

$$D_g(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) \geq \frac{\sigma}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2\,.$$

Thus it follows that

$$\begin{aligned} D_g(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}) &= \eta\,\langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x}_k - \boldsymbol{x}_{k+1} \rangle - D_g(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) \\ &\leq \eta\left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) + \frac{L}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2\right) - \frac{\sigma}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 \\ &= \eta\left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1})\right) \end{aligned}$$

by selecting $\eta = \sigma/L$. Plugging in inequality $(\star)$ it gives that

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}_{k+1}) \geq \frac{1}{\eta}\Big(D_g(\boldsymbol{x}^*, \boldsymbol{x}_{k+1}) - D_g(\boldsymbol{x}^*, \boldsymbol{x}_k)\Big)\,.$$

The remaining part is the same as the previous proof. Summing over both sides from 0 to $T-1$, we obtain that

$$\frac{1}{T}\sum_{k=1}^{T} f(\boldsymbol{x}_k) \leq f(\boldsymbol{x}^*) + \frac{D_g(\boldsymbol{x}, \boldsymbol{x}_0)}{\eta T} \leq f(\boldsymbol{x}^*) + \frac{LR}{\sigma T}\,.$$

Reference

[1] A. Banerjee, Xin Guo and Hui Wang, "On the optimality of conditional expectation as a Bregman predictor," in *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2664-2669, July 2005.

[2] A. Banerjee, Srujana Merugu, Inderjit S. Dhillon, Joydeep Ghosh, and John Lafferty. "Clustering with Bregman divergences." *Journal of machine learning research* 6, no. 10 (2005).