DSCC/CSC 435 & ECE 412 Optimization for Machine Learning — Lecture 5

Mirror Descent

Lecturer: Jiaming Liang

September 19, 2023

1 Conjugate functions

Definition 1. *Let $f : \mathbb{R}^n \to [-\infty, \infty]$ be an extended real-valued function. The conjugate function of f is defined as*

$$f^*(x) = \max_y \{\langle x, y\rangle - f(y)\}.$$

Theorem 1. *Let f be a closed and convex function. Then, the biconjugate function $f^{**} = f$.*

Theorem 2. *Let f be a closed and convex function. Then, for any $x, y \in \mathbb{R}^n$, the following statements are equivalent:*

(i) $\langle x, y\rangle = f(x) + f^*(y)$;

(ii) $y \in \partial f(x)$;

(iii) $x \in \partial f^*(y)$.

Corollary 1. *Let f be a closed and convex function. Then, for any $x, y \in \mathbb{R}^n$,*

$$\partial f(x) = \mathrm{Argmax}_{\tilde{y}} \{\langle x, \tilde{y}\rangle - f^*(\tilde{y})\}$$

and

$$\partial f^*(y) = \mathrm{Argmax}_{\tilde{x}} \{\langle y, \tilde{x}\rangle - f(\tilde{x})\}.$$

Proposition 1. *Let f be a closed and strictly convex function. Then, f^* is differentiable, and for any $y \in \mathbb{R}^n$,*

$$\nabla f^*(y) = \mathrm{argmax}_x \{\langle y, x\rangle - f(x)\}.$$

The concept of strong convexity extends and parametrizes the notion of strict convexity. A strongly convex function is also strictly convex, but not vice versa.

An extremely useful connection between smoothness and strong convexity is given in the conjugate correspondence theorem.

Theorem 3. *If f is closed and μ-strongly convex, then f^* is $(1/\mu)$-smooth. On the other hand, if f is L-smooth, then f^* is $(1/L)$-strongly convex.*

It is worth noting that in this case, for every $y \in \mathbb{R}^n$,

$$\nabla f^*(y) = (\nabla f)^{-1}(y). \tag{1}$$

2 Mirror descent

We are interested in the same convex nonsmooth optimization problem as in Lecture 4

$$\min_{x \in Q} f(x)$$

where Q is a closed convex set. Recall that the convergence rate by the projected subgradient method is

$$\min_{0 \le i \le k-1} f(x_i) - f_* \le \frac{MR}{\sqrt{k}}.$$

One of the basic assumptions made in Lecture 4 is that the underlying space is Euclidean, meaning that $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$. In order to establish the above dimension-free convergence rate, we need to make another assumption that the objective function f and the constraint set Q are well-behaved in the Euclidean norm: that means for all points $x \in Q$ and all subgradients $f'(x) \in \partial f(x)$, we have $\|x\|$ and $\|f'(x)\|$ are independent of the ambient dimension n. If this assumption is not met then we lose the dimension-free convergence rate. For instance, Q is the unit simplex $\Delta_n = \left\{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x(i) = 1\right\}$ and f has subgradients bounded in ℓ_∞-norm, e.g., $\|f'(x)\|_\infty \le 1$. Then, $\|f'(x)\|_\infty \le \sqrt{n}$ and $R \le \sqrt{2}$, so the convergence rate becomes

$$\min_{0 \le i \le k-1} f(x_i) - f_* \le \frac{\sqrt{2n}}{\sqrt{k}}.$$

But if we use mirror descent in this lecture, the convergence rate will be improved to $\mathcal{O}(\sqrt{\log(n)/k})$. This improvement relies on changing the space to be non-Euclidean.

In non-Euclidean spaces, $x \in \mathbb{E}$ and $f'(x) \in \mathbb{E}^*$, hence the subgradient method

$$x_{k+1} = \mathrm{proj}_Q \left(x_k - h_k f'(x_k)\right)$$

does not make sense. This issue motivates us to generalize the projected subgradient method to better suite the non-Euclidean setting.

Let us take another look at the projected subgradient method. It can be equivalently written as

$$x_{k+1} = \mathrm{argmin}_{x \in Q} \left\{ f(x_k) + \langle f'(x_k), x - x_k \rangle + \frac{1}{2h_k} \|x - x_k\|_2^2 \right\}. \tag{2}$$

The idea in the non-Euclidean case is to replace the Euclidean distance function $\frac{1}{2}\|x - x_k\|_2^2$ by a different "distance". This non-Euclidean distance is the *Bregman divergence.*

Definition 2. *For an arbitrary norm $\| \cdot \|$ in $\mathbb{E}$, the dual norm equipped in $\mathbb{E}^*$ is defined as*

$$\|s\|_* = \max_{x \in \mathbb{E}} \left\{ \langle s, x \rangle : \|x\| \le 1 \right\}, \quad s \in \mathbb{E}^*.$$

By the Cauchy-Schwartz inequality, for $x \in \mathbb{E}$ and $s \in \mathbb{E}^*$, we have

$$\langle s, x\rangle \le \|s\|_* \|x\|.$$

E.g., let $\|\cdot\|$ be the ℓ_p-norm and $\|\cdot\|_*$ be the ℓ_q norm where $1 \le p, q \le \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$, then by Hölder's inequality

$$\langle s, x\rangle \le \|sx\|_1 \le \|x\|_p \|s\|_q, \quad \forall x \in \mathbb{E}, s \in \mathbb{E}^*,$$

i.e.,

$$\sum_{k=1}^n x_k s_k \le \sum_{k=1}^n |x_k s_k| \le \left(\sum_{k=1}^n |x_k|^p\right)^{1/p} \left(\sum_{k=1}^n |s_k|^q\right)^{1/q}.$$

Let $w : \mathbb{R}^n \to (-\infty, \infty]$ be a proper closed convex function satisfying

- w is differentiable on $\mathrm{int}(\mathrm{dom}\, w) = W^o$;
- $Q \subset \mathrm{dom}\,(w)$;
- w is ρ-strongly convex on Q w.r.t. $\|\cdot\|$ (here $\|\cdot\|$ is an arbitrary norm in $\mathbb{E}$).

Definition 3. *For a function w satisfying the above assumptions, the Bregman divergence associated with w is the fucntion $D_w : \mathrm{dom}\, w \times W^o \to \mathbb{R}$ given by*

$$D_w(x, y) := w(x) - w(y) - \langle \nabla w(y), x - y\rangle.$$

The function w is called the distance generating fucntion.

A few properties of D_w: let $x \in Q$ and $y \in Q \cap W^o$, then

- $D_w(x, y) \ge \frac{\rho}{2}\|x - y\|^2$ for every $x \in Q$ and $y \in Q \cap W^o$;
- $D_w(x, y) \ge 0$;
- $D_w(x, y) = 0$ if and only if $x = y$;
- $D_w(x, y) = D_{w^*}(x^*, y^*)$ where w^* is the Fenchel conjugate and $x^* = \nabla w(x)$ and $y^* = \nabla w(y)$.

Bregman divergence does not satisfy symmetry nor triangle inequality, and hence it is not a metric.

Now we replace the Euclidean distance in (2) by the Bregman divergence, then we obtain an iteration of the *mirror descent*

$$x_{k+1} = \mathrm{argmin}_{x \in Q} \left\{ f(x_k) + \langle f'(x_k), x - x_k\rangle + \frac{1}{h_k} D_w(x, x_k) \right\}. \tag{3}$$

(Note that Lemma 9.7 and Theorem 9.8 of Amir Beck's book guarantees that $x_{k+1} \in Q \cap W^o$, hence $\nabla w(x_{k+1})$ exists in the next iteration and mirror descent is well-defined.) Hence, $x_{k+1} = \text{proj}_Q(y_{k+1})$ and y_{k+1} satisfies

$$0 = f'(x_k) + \frac{1}{h_k}\left(\nabla w(y_{k+1}) - \nabla w(x_k)\right),$$

where we use the fact that $\nabla_x D_w(x, y) = \nabla w(x) - \nabla w(y)$. Thus,

$$y_{k+1} = (\nabla w)^{-1}\left(\nabla w(x_k) - h_k f'(x_k)\right) = \nabla w^*\left(\nabla w(x_k) - h_k f'(x_k)\right)$$

where the second equality is due to (1). Below is another way to derive the formula for y_{k+1}

$$\begin{aligned}
y_{k+1} &= \text{argmin}_{x\in\mathbb{R}^n}\left\{f(x_k) + \langle f'(x_k), x - x_k\rangle + \frac{1}{h_k}D_w(x, x_k)\right\} \\
&= \text{argmin}_{x\in\mathbb{R}^n}\left\{\langle h_k f'(x_k) - \nabla w(x_k), x\rangle + w(x)\right\} \\
&= \text{argmax}_{x\in\mathbb{R}^n}\left\{\langle -h_k f'(x_k) + \nabla w(x_k), x\rangle - w(x)\right\} \\
&= \nabla w^*\left(\nabla w(x_k) - h_k f'(x_k)\right).
\end{aligned}$$

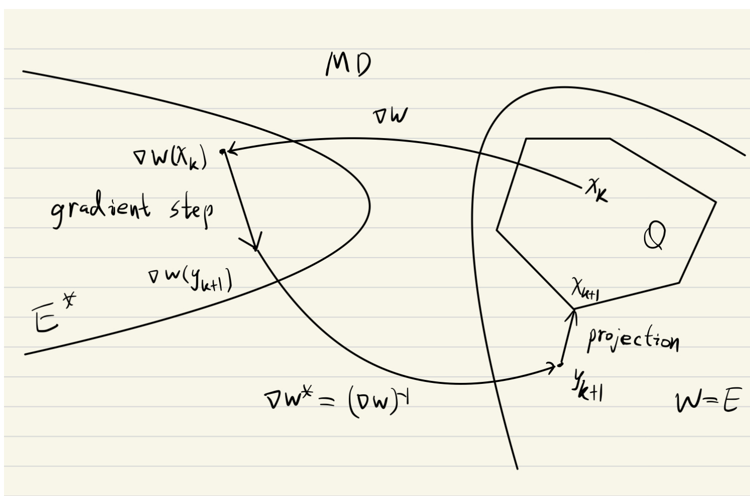

Figure 1: Mirror descent

The search point x_k is mapped from the primal space into the dual space using ∇w, the gradient step is then performed in the dual space $\nabla w(x_k) - h_k f'(x_k)$, and the point thus obtained is finally mapped back into the primal space using ∇w^*. The distance generating function w is also called the mirror map. See Figure 1 for an illustration.

Algorithm 1 Mirror descent

Input: Initial point $x_0 \in Q \cap W^o$

for $k \geq 0$ **do**

Step 1. Choose $h_k > 0$.

Step 2. Comput $y_{k+1} = \nabla w^* \left(\nabla w(x_k) - h_k f'(x_k)\right)$.

Step 3. Compute $x_{k+1} = \mathrm{proj}_Q(y_{k+1})$.

end for

Lemma 1. *For every $k \geq 0$,*

$$h_k f'(x_k) + \nabla w(x_{k+1}) - \nabla w(x_k) + N_Q(x_{k+1}) \ni 0$$

or

$$f'(x_k) + \frac{\nabla w(x_{k+1}) - \nabla w(x_k)}{h_k} + n_k = 0, \quad n_k \in N_Q(x_{k+1}),$$

where $N_Q(x_{k+1})$ is the normal cone of Q at x_{k+1}

$$N_Q(x_{k+1}) = \{g \in \mathbb{R}^n : 0 \geq \langle g, x - x_{k+1}\rangle, \quad \forall x \in Q\}.$$

Proof. The iteration (3) can be reformulated as

$$x_{k+1} = \mathrm{argmin}_{x \in \mathbb{R}^n} \left\{ f(x_k) + \langle f'(x_k), x - x_k\rangle + \frac{1}{h_k} D_w(x, x_k) + I_Q(x) \right\},$$

where $I_Q(\cdot)$ is the indicator functino of Q, i.e.,

$$I_Q(x) = \begin{cases} 0, & \text{if } x \in Q, \\ \infty, & \text{otherwise.} \end{cases}$$

The optimality condition reads as

$$\begin{aligned} 0 &\in \nabla f(x_k) + \frac{1}{h_k}\left(\nabla w(x_{k+1}) - \nabla w(x_k)\right) + \partial I_Q(x_{k+1}) \\ &= \nabla f(x_k) + \frac{1}{h_k}\left(\nabla w(x_{k+1}) - \nabla w(x_k)\right) + N_Q(x_{k+1}). \end{aligned}$$

□

Lemma 2. ***(Three points lemma)*** *Let w be a function satisfying the conditions above Definition 3. For every $z_0, z \in W^o$ and $x \in \mathrm{dom}\, w$, we have*

$$D_w(x, z_0) - D_w(z, z_0) - \langle \nabla D_w(z, z_0), x - z\rangle = D_w(x, z).$$

Lemma 3. *Assume that* $\|f'(x)\|_* \le M$ *for every* $x \in Q \cap domw$*. For every* $k \ge 0$ *and* $x \in \operatorname{dom} w$*, we have*

$$D_w(x, x_k) - D_w(x, x_{k+1}) \ge -\frac{h_k^2 M^2}{2\rho} + h_k[f(x_k) - f(x)].$$

Proof. Using Lemmas 1 and 2, the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
D_w(x, x_k) - D_w(x, x_{k+1}) &= D_w(x_{k+1}, x_k) - \langle \nabla D_w(x_{k+1}, x_k), x_{k+1} - x\rangle \\
&= D_w(x_{k+1}, x_k) + \langle \nabla w(x_k) - \nabla w(x_{k+1}), x_{k+1} - x\rangle \\
&= D_w(x_{k+1}, x_k) + h_k\langle f'(x_k) + n_k, x_{k+1} - x\rangle \\
&\ge D_w(x_{k+1}, x_k) + h_k\langle f'(x_k), x_{k+1} - x\rangle \\
&\ge \frac{\rho}{2}\|x_{k+1} - x_k\|^2 + h_k\langle f'(x_k), x_{k+1} - x_k\rangle + h_k\langle f'(x_k), x_k - x\rangle \\
&\ge \frac{\rho}{2}\|x_{k+1} - x_k\|^2 - h_k\|f'(x_k)\|_*\|x_{k+1} - x_k\| + \lambda_k\langle f'(x_k), x_k - x\rangle \\
&\ge -\frac{h_k^2\|s_k\|_*^2}{2\rho} + h_k\langle f'(x_k), x_k - x\rangle \\
&\ge -\frac{h_k^2\|s_k\|_*^2}{2\rho} + h_k[f(x_k) - f(x)],
\end{aligned}$$

where the last inequality is due to the subgradient inequality. □

Theorem 4.

$$f(\bar{x}_k) - f_* \le \frac{D_w(x_*, x_0) + \frac{M^2}{2\rho}\sum_{i=0}^{k-1} h_i^2}{\sum_{i=0}^{k-1} h_i}$$

where $\bar{x}_k$ *is any point satisfying*

$$f(\bar{x}_k) \le \frac{\sum_{i=0}^{k-1} h_i f(x_i)}{\sum_{i=0}^{k-1} h_i}.$$

Moreover, for a given $\varepsilon > 0$*, if* $h_k = h$*, then*

$$f(\bar{x}_k) - f_* \le \frac{D_w(x_*, x_0)}{kh} + \frac{M^2 h}{2\rho}.$$

3 Standard setups for mirror descent

Ball: The distance generating function is

$$w(x) = \frac{1}{2}\|x\|_2^2$$

is 1-strongly convex w.r.t. $\|\cdot\|_2$ and the associated Bregman divergence is given by

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

In this case, mirror descent is equivalent to projected subgradient method.
Simplex: The distance generating function is given by the negative entropy

$$w(x) = \sum_{i=1}^{n} x(i) \log x(i).$$

Note that $W^o = \mathbb{R}^n_{++}$ and w is 1-strongly convex w.r.t. $\|\cdot\|_1$ on Δ_n. The associated Bregman divergence is given by

$$D_w(x, y) = \sum_{i=1}^{n} x(i) \log \frac{x(i)}{y(i)} - \sum_{i=1}^{n} (x(i) - y(i)),$$

where the first summation is known as the relative entropy or Kullback-Leibler divergence

$$\mathrm{KL}(x, y) = \sum_{i=1}^{n} x(i) \log \frac{x(i)}{y(i)}.$$

The strong convexity property of w can be stated as for any $x, y \in \Delta_n$,

$$D_w(y, x) = \mathrm{KL}(x, y) \geq \frac{1}{2} |x - y|_1^2,$$

which is also known as the Pinsker's inequality. The projection onto simplex Δ_n w.r.t. the Bregman divergence is as simple as

$$\mathrm{proj}_{\Delta_n}(x_0) = \frac{x_0}{\|x_0\|_1}.$$

Corollary 2. *Assume $\|f'(x)\|_\infty \leq M$, $\forall x \in \Delta_n$. Let $x_0 = \mathrm{argmin}_{x \in \Delta_n} w(x)$ (in the simplex setup, $x_0 = (1/n, \ldots, 1/n)^\top$). Then, mirror descent with $h = \frac{1}{M}\sqrt{\frac{2 \log n}{k}}$ satisfies*

$$f(\bar{x}_k) - f_* \leq M \sqrt{\frac{2 \log n}{k}}.$$

Proof. We first note that since $x_0 = \mathrm{argmin}_{x \in \Delta_n} w(x)$, it holds

$$\langle \nabla w(x_0), x_* - x_0 \rangle \geq 0.$$

Then, we have

$$\begin{aligned} D_w(x_*, x_0) &= w(x_*) - w(x_0) - \langle \nabla w(x_0), x_* - x_0 \rangle \\ &\leq w(x_*) - w(x_0) \\ &\leq \max_{x \in \Delta_n} w(x) - \min_{x \in \Delta_n} w(x). \end{aligned}$$

Using the fact that

$$-\log n \leq w(x) \leq 0, \quad \forall x \in \Delta_n,$$

we have

$$D_w(x_*, x_0) \le \log n.$$

It follows from Theorem 4 that

$$f(\bar{x}_k) - f_* \le \frac{D_w(x_*, x_0)}{kh} + \frac{M^2 h}{2} \le \frac{\log n}{kh} + \frac{M^2 h}{2}.$$

Taking $h = \frac{1}{M}\sqrt{\frac{2\log n}{k}}$, we have

$$f(\bar{x}_k) - f_* \le M\sqrt{\frac{2\log n}{k}}.$$

□