CS2910 - Optimization Summer 2023

Lecture 4 — Mirror Descent

Lecturer: Yunwei Ren *Scribed by Zhidan Li*

Contents

1 Overview

In this lecture, we introduce a more general and efficient method of gradient descent. We firstly give an alternative view of the gradient descent, and generalize the mirror descent.

Additionally, we might assume all functions are C^1 for convenience.

2 An Alternative View of Gradient Descent

Now we turn back to the gradient descent. Consider the optimization problem

$$\min_{x\in\mathbb{R}^d} f(x).$$

Recall that the form of every update is

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

This can be viewed as the local linear approximation around x_k. Then we observe that the update rule is exactly a solution to some optimization problem.

Observation 1. *The update rule of gradient descent*

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

is equivalent to

$$x_{k+1} = \operatorname*{argmin}_{x\in\mathbb{R}^d}\left\{f(x_k) + \langle\nabla f(x_k), x - x_k\rangle + \frac{1}{2\eta}\|x - x_k\|^2\right\}.$$

Proof. Let $x_* = \operatorname{argmin}_{x\in\mathbb{R}^d}\left\{f(x_k) + \langle\nabla f(x_k), x - x_k\rangle + \frac{1}{2\eta}\|x - x_k\|^2\right\}$. Then by direct calculation,

$$\begin{aligned} x_* &= \operatorname*{argmin}_{x\in\mathbb{R}^d}\left\{f(x_k) + \langle\nabla f(x_k), x - x_k\rangle + \frac{1}{2\eta}\|x - x_k\|^2\right\} \\ &= \operatorname*{argmin}_{x\in\mathbb{R}^d}\left\{\langle\nabla f(x_k), x\rangle + \frac{1}{2\eta}\|x - x_k\|^2\right\}. \end{aligned}$$

Then, we know

$$\nabla f(x_k) + \frac{1}{\eta}(x_* - x_k) = 0.$$

Rearranging terms we prove the observation. □

Now we put our eyes on the term

$$\operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{\langle\nabla f(x_k), x - x_k\rangle + \frac{1}{2\eta}\|x - x_k\|^2\right\}$$

We can see, the term $\langle\nabla f(x_k), x - x_k\rangle$ is the global linear approximation, and we name the left term $\frac{1}{2\eta}\|x - x_k\|^2$ the *regularization*.

Remark 1. The introduction of the regularization is to make the hard constraint 'x is around x_k' a soft one.

2.1 Bregman divergence

Observe that, the regularization is not necessarily ℓ_2-norm. Now we replace the ℓ_2-norm with Bregman divergence, to establish the method of mirror descent.

Definition 2. *Let $g : \mathbb{R}^d \to \mathbb{R}$ be a C^1 convex function. We define its Bregman divergence $D_g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as*

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

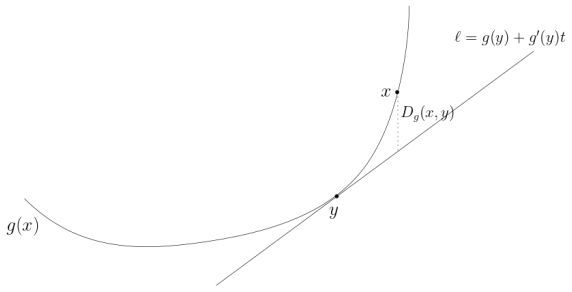

Figure 1: an example on $\mathbb{R}^2$

Remark 2. The Bregman divergence has the following properties.

- Since g is convex, $D_g(x, y) \geq 0$ for all $x, y \in \mathbb{R}^d$.

- For all $x \in \mathbb{R}^d$, $D_g(x, x) = 0$.
- If g is strictly convex, $D_g(x, y) > 0$ for all $x \neq y$.
- $D_g(x, y)$ is convex on x.
- In general Bregman divergence is not symmetric.

Examples:

- When $g(x) = \frac{1}{2}\|x\|^2$, its Bregman divergence is $D_g(x) = \frac{1}{2}\|x\|^2$
- Let $x \in \Delta_n \stackrel{\triangle}{=} \{x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1 \wedge \forall i \in [n], x_i \geq 0\}$. When $g(x) = \sum_{i=1}^n x_i \log x_i$, $x \in \Delta_n$, its Bregman divergence is

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

  This kind of divergence is also called **Kullback-Leibler divergence** (KL divergence for short) or **relative entropy**.

3 Mirror Descent

Similarly to gradient descent, we establish the art of mirror descent. We introduce the update rule of mirror descent as

$$x_{k+1} = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \langle \nabla f(x_k), x \rangle + \frac{1}{\eta} D_g(x, x_k) \right\}.$$

Since the function $x \mapsto \langle \nabla f(x_k), x \rangle + \frac{1}{\eta} D_g(x, x_k)$ is convex, consider its gradient:

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

Rearranging terms we obtain

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

Then

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

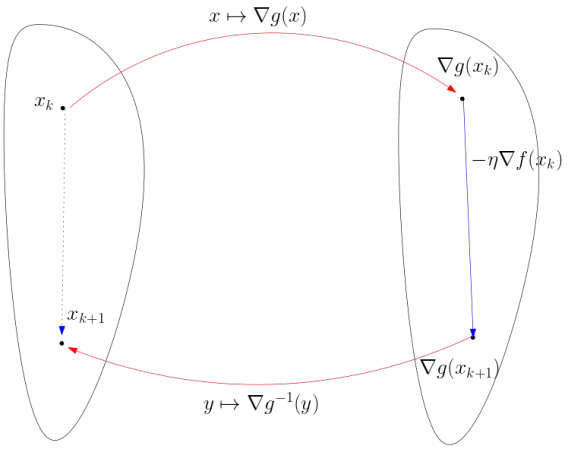

Figure 2: one step of update in mirror descent. The update can viewed as a step of update of gradient descent in its dual space generated by $x \mapsto \nabla g(x)$.

3.1 Convex conjugate

Note that, to implement mirror descent, we need to show the ∇g^{-1} can be computed efficiently. We introduce the convex conjugate to compute it.

Definition 3 (convex conjugate)**.** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a convex function. We define its convex conjugate $f^* : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ as*

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

Note that f^* is always convex.

Recall that, the convex function is equivalent to (or, can be expressed by) its supporting affine function. For $u \in \mathbb{R}^d$, we write $h_{u,b}(x) = \langle u, x \rangle + b$. Consider the 'highest' supporting affine function generated by u. Let

$$b^*(u) = \sup \{b \in \mathbb{R} : f(x) \geq \langle u, x \rangle + b\}.$$

It is not easy to show $b^*(u) = -f^*(u)$, and by definition the affine function $h_{u,b^*(u)}$ is the 'highest' supporting affine function of f.

Examples:

- When $f(x) = ax + b$, we compute that

$$f^*(u) = \sup_{x \in \mathbb{R}} \{ux - f(x)\} = \sup_{x \in \mathbb{R}} \{(u - a)x - b\} = \begin{cases} -b & u = a, \\ \infty & u \neq a. \end{cases}$$

- When $f(x) = \frac{1}{2}\|x\|^2$, we compute

$$\begin{aligned} f^*(u) &= \sup_{x \in \mathbb{R}} \left\{ \langle u, x \rangle - \frac{1}{2}\|x\|^2 \right\} \\ &= \frac{1}{2}\|x\|^2. \end{aligned}$$

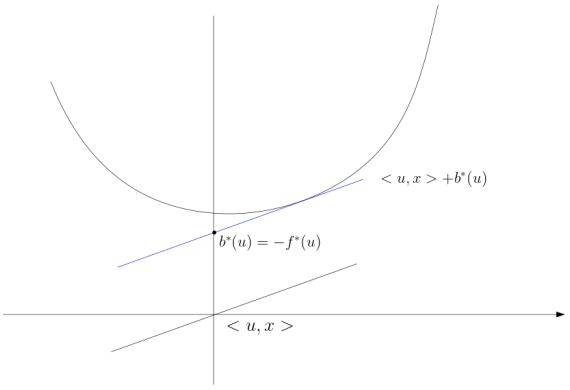

Figure 3: an example in $\mathbb{R}^2$. The blue line achieves the highest b.

Then we establish some properties of the convex conjugate.

Lemma 4 (Fenchel inequality)**.** *For all $x, u \in \mathbb{R}^d$, we have*

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

Additionally, the equality holds if and only if $u = \nabla f(x)$.

Proof. By definition of convex conjugate, for every $x, u \in \mathbb{R}^d$, it holds

$$f^*(u) \geq \langle u, x \rangle - f(x).$$

Rearranging terms we prove the inequality.

When the equality holds, we know

$$\begin{aligned} \langle u, x \rangle = f(x) + f^*(u) &\iff \langle u, x \rangle \geq f(x) + \langle u, z \rangle - f(z), \forall z \in \mathbb{R}^d \\ &\iff f(z) \geq f(x) + \langle u, z - x \rangle, \forall z \in \mathbb{R}^d. \end{aligned}$$

Then by lower linear bound, the last inequality is equivalent to $u = \nabla f(x)$. □

Then we introduce the Fenchel-Moreau theorem.

Theorem 5 (Fenchel-Moreau)**.** *If f is a convex function, then $f = f^{**}$.*

Proof. For all $u, x \in \mathbb{R}^d$, it holds that

$$f^*(u) = \sup_{x' \in \mathbb{R}^d} \left\{ \langle u, x' \rangle - f(x') \right\} \geq \langle u, x \rangle - f(x).$$

Thus we obtain

$$f(x) \geq \langle u, x \rangle - f^*(u), \forall x, u \in \mathbb{R}^d.$$

This means

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

Also, by Lemma 4, it holds that

$$f(x) + f^*(\nabla f(x)) = \langle \nabla f(x), x \rangle, \forall x \in \mathbb{R}^d.$$

This means

$$f(x) = \langle \nabla f(x), x \rangle - f^*(\nabla f(x)) \leq \sup_{y \in \mathbb{R}^d} \{\langle y, x \rangle - f^*(y)\} = f^{**}(x).$$

□

Corollary 6. *If f is strictly convex, then*

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

Proof. Let $u = \nabla f(x)$. Then, by Lemma 4,

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

By Theorem 5, it holds that

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

Then by Lemma 4 it holds that

$$\nabla f^*(u) = x.$$

□

Note that, when f is not a strictly convex function on $\mathbb{R}^d$, but still strictly convex on its domain, the result also holds.

Theorem 7 (Theorem 26.5 in [Roc70])**.** *Let $f : \mathrm{dom}(f) \subseteq \mathbb{R}^d \to \mathbb{R}$ be a closed strictly convex function. Suppose that $\|\nabla f(x)\| \to \infty$ as $x \to \partial \mathrm{dom}(f)$. Then $f^* : \mathrm{dom}(f^*) \subseteq \mathbb{R}^d \to \mathbb{R}$ is also a C^1 closed strictly convex function, satisfying the same properties. Moreover, $x \mapsto \nabla f(x)$ is a bijection from $\mathrm{int}(\mathrm{dom}(f))$ onto $\mathrm{int}(\mathrm{dom}(f^*))$ and $(\nabla f)^{-1} = \nabla f^*$.*

3.2 Convergence of mirror descent

Now we return to mirror descent. Combining the discussion above, we can write the update rule of mirror descent as

$$x_{k+1} = \nabla g^* \left(\nabla g(x_k) + \eta \nabla f(x_k) \right).$$

In this section we will show the convergence rate of mirror descent (under some regular conditions). We will extend our analysis of gradient descent to mirror descent. Firstly we extend the definition of smooth functions.

Definition 8. *Let $g \in C^1$ be a convex function. A function $f : \mathrm{dom}(g) \subseteq \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz with respect to D_g if for all $x, y \in \mathrm{dom}(g)$*

$$|\langle \nabla f(x), y - x \rangle| \leq L \sqrt{D_g(y, x)}.$$

Remark 3. Note that, let $g = \|x - y\|^2$, the definition contains the Lipschitz property (with a little difference in constant).

Theorem 9. *Suppose that f is convex and L-Lipschitz with respect to D_g. For any $T \geq 1$, if we choose*

$$\eta = \frac{2}{L}\sqrt{\frac{D_g(x_*, x_0)}{T}},$$

then

$$\frac{1}{T}\sum_{k=0}^{T-1} f(x_k) \leq f(x_*) + \frac{L}{2}\sqrt{\frac{D_g(x_*, x_0)}{T}}.$$

To prove Theorem 9, we need the following lemma which is an extension of law of cosines to deduce the mirror descent lemma.

Lemma 10 (law of cosines for Bregman divergence)**.** *Let $g \in C^1$ be a convex function. For all $x, y, z \in \mathrm{dom}(g)$, it holds that*

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

Proof. The result comes directly from the definition. □

Lemma 11 (mirror descent lemma)**.** *Suppose that f is convex. Let $\{x_k\}$ be the sequence generated by mirror descent of f. Then, for all y,*

$$f(x_k) \leq f(y) + \frac{1}{\eta}\left(D_g(y, x_k) - D_g(y, x_{k+1}) + D_g(x_k, x_{k+1})\right).$$

The proof of Lemma 11 is similar to the proof of basic mirror descent lemma we stated in Lecture 2. We leave it as a mental training.

Proof of Theorem 9. By Lemma 11, it holds that

$$f(x_k) \leq f(x_*) + \frac{1}{\eta}\left(D_g(x_*, x_k) - D_g(x_*, x_{k+1}) + D_g(x_k, x_{k+1})\right).$$

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

$$\sum_{k=0}^{T-1} f(x_k) \leq T f(x_*) + \frac{1}{\eta}\left(D_g(x_*, x_0) - D_g(x_*, x_T) + \sum_{k=0}^{T-1} D_g(x_k, x_{k+1})\right).$$

By definition,

$$\begin{aligned} D_g(x_k, x_{k+1}) &= g(x_k) - g(x_{k+1}) - \langle \nabla g(x_{k+1}), x_k - x_{k+1} \rangle, \\ D_g(x_{k+1}, x_k) &= g(x_{k+1}) - g(x_k) - \langle \nabla g(x_k), x_{k+1} - x_k \rangle. \end{aligned}$$

Then it holds that

$$\begin{aligned} D_g(x_k, x_{k+1}) + D_g(x_{k+1}, x_k) &= \langle \nabla g(x_k) - \nabla g(x_{k+1}), x_k - x_{k+1} \rangle \\ &= \langle \eta \nabla f(x_k), x_k - x_{k+1} \rangle \\ &\leq \eta L \sqrt{D_g(x_{k+1}, x_k)}. \end{aligned}$$

Then

$$D_g(x_k, x_{k+1}) \leq -D_g(x_{k+1}, x_k) + \eta L \sqrt{D_g(x_{k+1}, x_k)} \leq \frac{\eta^2 L^2}{4}.$$

Plugging it into above, we obtain

$$\begin{aligned} \frac{1}{T} \sum_{k=0}^{T-1} f(x_k) &\leq f(x_*) + \frac{1}{\eta T} \left(D_g(x_*, x_0) + T \frac{\eta^2 L^2}{4} \right) \\ &\leq f(x_*) + \frac{1}{\eta T} D_g(x_*, x_0) + \frac{\eta L^2}{4}. \end{aligned}$$

Choose $\eta = \frac{2}{L} \sqrt{\frac{D_g(x_*, x_0)}{T}}$, and we obtain the desired bound. □

References

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