DSC 211 Introduction to Optimization Winter 2024 Instructor: Jun-Kun Wang
Scribe: Parsa Mirtaheri February 15, 2024
Editors/TAs: Merlin Chang, Marialena Sfyraki

Lecture 12: Mirror Descent

1 Mirror Descent

1.1 Algorithm

Many results in optimization are relatively new. Mirror Descent method is one of them which was introduced by Arkadi Nemirovsky in 1983.

We are already familiar with the Projected Gradient Descent (PGD) method for solving a constrained optimization problem $\min_{x \in C} f(x)$:

Algorithm 1 Projected Gradient Descent

```
for $k = 1, 2, \ldots$ do
    $x_{k+1} = \mathrm{Proj}_C \left[ x_k - \eta \nabla_k f(x_k) \right]$.
end for
```

which was shown to be equivalent to the following in the first homework:

Algorithm 2 Projected Gradient Descent

```
for $k = 1, 2, \ldots$ do
    $x_{k+1} = \arg\min_{x \in C} \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\eta} \|x - x_k\|_2^2$.
end for
```

This formulation motivates considering other notions of distance instead of the L^2.

Definition 1. *Let $\phi(\cdot) : C \to R$ be a convex and differentiable function. The Bregman divergence induced by $\phi(\cdot)$ is defiend as:*

$$D_x^{\phi}(y) = \phi(y) - \phi(x) - \langle \nabla \phi(x), y - x \rangle$$

The function ϕ is called the distance generating function. If instead of L^2 distance, we use Bregman Divergence, which in fact is not a distance because it is asymmetric, we get another variant of the optimization method called Mirror Descent:

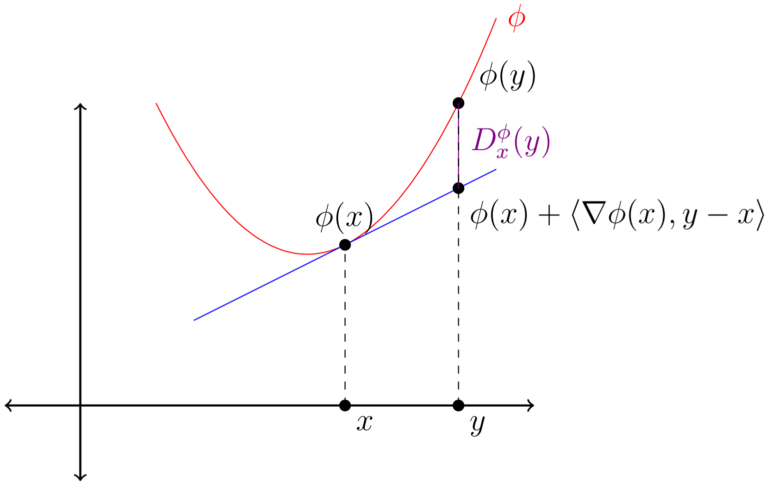

Figure 1: Bregman Divergence between two points x and y with respect to a convex function ϕ.

Algorithm 3 Mirror Descent

```
for $k = 1, 2, \ldots$ do
    $x_{k+1} = \arg\min_{x \in C} \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\eta} D^\phi_{x_k}(x)$.
end for
```

Remark: You can recover Projected Gradient Descent by letting $\phi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$ to get:

$$D_x^\phi(y) = \phi(y) - \phi(x) - \langle \nabla\phi(x), y - x\rangle = \frac{\|y\|_2^2}{2} - \frac{\|x\|_2^2}{2} - \langle x, y - x\rangle = \frac{\|y - x\|^2}{2}$$

which means that PGD is an instance of Mirror Descent.

Example: Suppose $C := \{x \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1, x_i \geq 0\}$ is the probability simplex. If we define $\phi(x) = \sum_{i=1}^d x_i \log x_i$ to be negative entropy, then we get KullbackLeibler (KL) divergence as $D_y^\phi(x)$:

$$\begin{aligned} D_x^\phi(y) &= \phi(y) - \phi(x) - \langle \nabla\phi(x), y - x\rangle \\ &= \sum_{i=1}^d y_i \log y_i - \sum_{i=1}^d x_i \log x_i - \sum_{i=1}^d (1 + \log x_i)(y_i - x_i) \\ &= \sum_{i=1}^d y_i \log y_i - \sum_{i=1}^d y_i \log x_i - \sum_{i=1}^d (y_i - x_i) = \sum_{i=1}^d y_i \log \frac{y_i}{x_i} = D_{\mathrm{KL}}(y \parallel x) \end{aligned}$$

Then the optimization problem for the update at each step becomes:

$$x_{k+1} = \arg\min_{x \in C} \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\eta} \sum_{i=1}^{d} x_i \log \frac{x_i}{x_{k,i}}$$

which has a closed-form solution and we can find it using what we have learned about duality theory and the optimality conditions. We first construct the Lagrangian function in terms of the primal variable x and dual variables λ and μ:

$$L(x, \lambda, \mu) = \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\eta} \sum_{i=1}^{d} x_i \log \frac{x_i}{x_{k,i}} - \sum_{i=1}^{d} \lambda_i x_i + \mu \left(\sum_{i=1}^{d} x_i - 1 \right)$$

The stationarity condition then implies:

$$\nabla_x L(x, \lambda, \mu) = 0 \Rightarrow \forall i : \left[\nabla_x L(x, \lambda, \mu) \right]_i = \left[\nabla f(x_k) \right]_i + \frac{1}{\eta} \log \frac{x_i}{x_{k,i}} + \frac{1}{\eta} - \lambda_i + \mu = 0$$
$$\Rightarrow x_i = x_{k,i} \exp\left(-\eta \left[\nabla f(x_k) \right]_i + \eta(\lambda_i - \mu) - 1 \right)$$

Assuming $x_i \neq 0$, complementary slackness ($\lambda_i x_i = 0$) implies $\lambda_i = 0$ and therefore:

$$x_i = x_{k,i} \exp\left(-\eta \left[\nabla f(x_k) \right]_i - \eta\mu - 1 \right) \tag{1}$$

By the primal feasibility ($\sum_{i=1}^{d} x_i = 1$) we have:

$$\sum_{i=1}^{d} x_i = \frac{\sum_{i=1}^{d} x_{k,i} \exp\left(-\eta \left[\nabla f(x_k) \right]_i \right)}{\exp\left(\eta\mu + 1 \right)} = 1 \tag{2}$$

$$\Leftrightarrow \mu = \frac{1}{\eta} \log \left(\sum_{i=1}^{d} x_{k,i} exp \left(-\eta \left[\nabla f(x_k) \right]_i \right) \right) - 1 \tag{3}$$

By (1) and (3) we get that the solution to the following optimization problem for the update:

$$x_{k+1} = \arg\min_{x \in C} \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\eta} \sum_{i=1}^{d} x_i \log \frac{x_i}{x_{k,i}}$$

is given by:

$$x_i = x_{k+1,i} = \frac{x_{k,i} \exp\left(-\eta \left[\nabla f(x_k) \right]_i \right)}{\sum_{i=j}^{d} x_{k,j} \exp\left(-\eta \left[\nabla f(x_k) \right]_j \right)} \ , \ \forall i \in [d]$$

1.2 An alternative view of Mirror Descent

We have seen that the update of Mirror Descent is:

$$x_{k+1} = \arg\min_{x\in C} \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{2\eta}\|x - x_k\|_2^2$$

Equivalently, one can express the update step of Mirror Descent as:

$$\nabla\phi(y_{k+1}) = \nabla\phi(x_k) - \eta\nabla f(x_k) \tag{4}$$

$$x_{k+1} = \min_{x\in C} D^{\phi}_{y_{k+1}}(x). \tag{5}$$

Recall the following theorem from the Fenchel Conjugate:

Theorem 1. *If $\phi(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is closed and convex, then*

$$y = \nabla\phi(x) \quad \Longleftrightarrow \quad x = \nabla\phi^{\,*}(y).$$

Observe that we can rewrite:

$$y = \nabla\phi(x) = \nabla\phi\left(\nabla\phi^{\,*}(y)\right),$$
$$x = \nabla\phi^{\,*}(y) = \nabla\phi^{\,*}\left(\nabla\phi(x)\right).$$

Hence, we may write the step:

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

as:

$$y_{k+1} = \nabla\phi^{*}\left(\nabla\phi(x_k) - \eta\nabla f(x_k)\right).$$

Remark: When $\phi(x) = \sum_{i=1}^{d} x_i \log x_i$ is negative entropy,

- $\phi(x)$ is defined only the probabiltiy simplex $C := \{x \in \mathbb{R}^d : \sum_{i=1}^{d} x_i = 1, x_i \geq 0\}$.
- $\phi^*(x) = \log\left(\sum_{i=1}^{d} \exp(x_i)\right)$.
- $\nabla\phi^*(\nabla\phi(x)) = x$ for any $x \in C$.
- $\nabla\phi(\nabla\phi^*(y)) = y$ up to an additive translation of $\mathbf{1}_d$.

A proof of the equivalency of the Mirror Descent update steps is stated below:
Proof.

$$\begin{aligned}
x_{k+1} &= \arg\min_{x\in C} \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\eta} D^{\phi}_{x_k}(x) \\
&= \arg\min_{x\in C} \eta \nabla f(x_k)^{\top} x + \phi(x) - \phi(x_k) - \langle \nabla\phi(x_k), x - x_k \rangle \\
&= \arg\min_{x\in C} \phi(x) - (\nabla\phi(x_k) - \eta\nabla f(x_k))^{\top} x \\
&= \arg\min_{x\in C} \phi(x) - (\nabla\phi(y_{k+1}))^{\top} x \\
&= \arg\min_{x\in C} \phi(x) - \phi(y_{k+1}) - \langle \nabla\phi(y_{k+1}), x - y_{k+1} \rangle \\
&= \arg\min_{x\in C} D^{\phi}_{y_{k+1}}(x)
\end{aligned}$$

1.3 Geometric Picture

The update steps of Mirror Descent:
Mirror Descent

$$\begin{aligned}
y_{k+1} &= \nabla\phi^* \big(\nabla\phi(x_k) - \eta\nabla f(x_k)\big) \\
x_{k+1} &= \min_{x\in C} D^{\phi}_{y_{k+1}}(x).
\end{aligned}$$

can be visualized geometrically as in Figure 2.

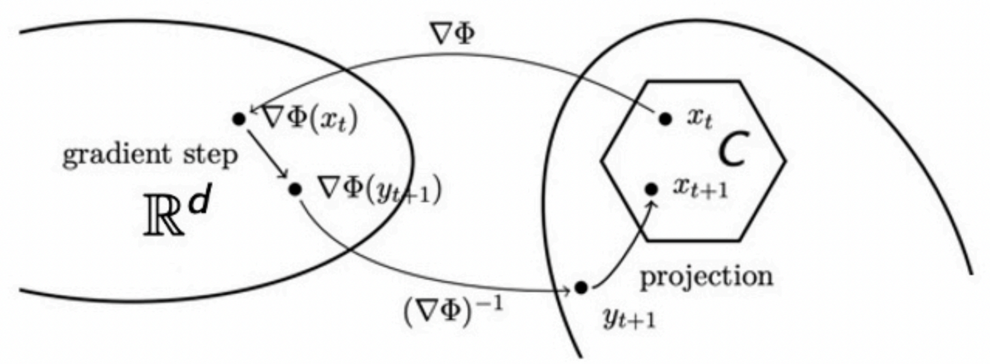

Figure 2: Geometric picture of Mirror Descent.

1.4 Non-differentiable case

Now, assume that $f(x)$ is convex but not necessarily differentiable. Let $g_k \in \partial f(x_k)$ be the subgradient of $f(\cdot)$ at x_k. Then, the algorithm can be rewritten as:

```
for $k = 1, 2, \ldots$ do
    $x_{k+1} = \arg\min_{x \in C} \langle g_k, x - x_k \rangle + \frac{1}{\eta} D^{\phi}_{x_k}(x)$.
end for
Output: $\bar{x} := \frac{\sum_{k=1}^{K} x_K}{K}$.
```

Recall the definition of the dual norm. Given a norm $\|\cdot\|$, the dual norm $\|\cdot\|_*$ is defined as

$$\|y\|_* := \sup_{x:\|x\|=1} x^\top y.$$

Additionally, recall the definition of the l_p-norm, for any $p \geq 1$:

$$\|x\|_p := \left(\sum_{i=1}^{d} (x_i)^p \right)^{1/p}.$$

The dual norm is related to the l_p-norm by the following theorem:

Theorem 2. *If $p, q \in [1, \infty]$ and $\frac{1}{p} + \frac{1}{q} = 1$, then $\|\cdot\|_p$ and $\|\cdot\|_q$ are dual with each other.*

Example. $\|\cdot\|_\infty$ and $\|\cdot\|_1$ are dual to each other.

Now, we can introduce the following theorem, which we will use later:

Theorem 3. *Choose a generating function $\phi(x)$ that is 1-strongly convex w.r.t. $\|\cdot\|$. Then, Mirror Descent has*

$$\sum_{k=1}^{K} f(x_k) - f(x_*) \leq \frac{1}{\eta} D^{\phi}_{x_1}(x_*) + \sum_{k=1}^{K} \frac{\eta}{2} \|g_k\|_*^2.$$

1.5 Mirror Descent v.s. Projected Gradient Descent

Consider the following convex constrained optimization problem:

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

where C is a simplex, with $\Delta_d := \{x \in \mathbb{R}^d : \sum_{i=1}^{d} x[i] = 1, x[i] \geq 0, \forall i\}$. If we let K be the number of iterations, then:

Projected Gradient Descent	$\epsilon = O\left(\sqrt{\frac{d}{K}}\right)$
Mirror Descent	$\epsilon = O\left(\sqrt{\frac{\log d}{K}}\right)$

Observe that the negative entropy $\phi(x) = \sum_{i=1}^{d} x_i \log x_i$ is 1-strongly convex with respect to $\|\cdot\|_1$. Hence, from Theorem 3 we have that

$$\sum_{k=1}^{K} f(x_k) - f(x_*) \leq \frac{1}{\eta} D_{x_1}^{\phi}(x_*) + \sum_{k=1}^{K} \frac{\eta}{2} \|g_k\|_*^2.$$

Let $x_1 = \frac{1}{d}\mathbf{1}_d$. Then

$$\begin{aligned} D_{x_1}^{\phi}(x_*) &= KL(x_*||x_1) \\ &= \sum_{i=1}^{d} x_{*,i} \log \frac{x_{*,i}}{x_{1,i}} \\ &= \underbrace{\sum_{i=1}^{d} x_{*,i} \log x_{*,i}}_{\leq 0} + \underbrace{\sum_{i=1}^{d} x_{*,i} \log \frac{1}{x_{1,i}}}_{=\sum_{i=1}^{d} x_{*,i} \log d} \\ &\leq \log d. \end{aligned}$$

Now, suppose that $\|g_k\|_\infty^2 \leq 1$. Then,

$$\begin{aligned} \sum_{k=1}^{K} f(x_k) - f(x_*) &\leq \frac{1}{\eta} D_{x_1}^{\phi}(x_*) + \sum_{k=1}^{K} \frac{\eta}{2} \|g_k\|_\infty^2 \\ &\leq \frac{1}{\eta} \log d + \frac{\eta}{2} K \\ &= O\left(\sqrt{K \times \log d}\right) \text{ , for } \eta = \sqrt{\frac{\log d}{K}}. \end{aligned}$$

Additionally, for $\bar{x} := \frac{\sum_{k=1}^{K} x_K}{K}$, when $f(\cdot)$ is convex, using Jensen's inequality we can write:

$$f(\bar{x}) - f(x_*) = O\left(\sqrt{\frac{\log d}{K}}\right).$$

Now let us look at the Projected Gradient Descent:

$$\begin{aligned}\sum_{k=1}^{K} f(x_k) - f(x_*) &\leq \frac{1}{\eta} D^{\phi}_{x_1}(x_*) + \sum_{k=1}^{K} \frac{\eta}{2} \|g_k\|_*^2 \\ &= \frac{1}{\eta}\left(\frac{1}{2}\|x_1 - x_*\|_2^2\right) + \sum_{k=1}^{K} \frac{\eta}{2}\|g_k\|_2^2 \\ &\leq \frac{1}{\eta} D + \frac{\eta}{2} K\sqrt{d} \\ &= O\left(\sqrt{dKD}\right) \text{ , for } \eta = \sqrt{\frac{D}{K}},\end{aligned}$$

where D is the bound of the inital distance. The second inequality follows from the fact that for any vector $z \in \mathbb{R}^d$,

$$\|z\|_\infty \leq \|z\|_2 \leq \sqrt{d}\|z\|_\infty.$$

Additionally, for $\bar{x} := \frac{\sum_{k=1}^{K} x_K}{K}$, when $f(\cdot)$ is convex, using Jensen's inequality we can write:

$$f(\bar{x}) - f(x_*) = O\left(\sqrt{\frac{d}{K}}\right).$$

1.6 Proof of Theorem 3

We will now present the proof of Theorem 3.

Proof. We have that

$$\begin{aligned} f(x_k) - f(x_*) &\leq \langle g_k, x_k - x_* \rangle \text{ , by convexity of } f(\cdot) \\ &= \langle g_k, x_{k+1} - x_* \rangle + \langle g_k, x_k - x_{k+1} \rangle.\end{aligned}$$

Since

$$x_{k+1} = \arg\min_{x \in C} \langle g_k, x \rangle + \frac{1}{\eta} D^{\phi}_{x_k}(x),$$

by the optimality condition,

$$\left\langle g_k + \frac{1}{\eta}(\nabla\phi(x_{k+1}) - \nabla\phi(x_k)), x - x_{k+1} \right\rangle \geq 0, \ \forall x \in C.$$

In particular, this inequality holds for $x = x_* \in C$. Thus, we have

$$\begin{aligned} f(x_k) - f(x_*) &\leq \langle g_k, x_{k+1} - x_* \rangle + \langle g_k, x_k - x_{k+1} \rangle \\ &\leq \frac{1}{\eta} \langle \nabla\phi(x_{k+1}) - \nabla\phi(x_k), x_* - x_{k+1} \rangle + \langle g_k, x_k - x_{k+1} \rangle. \end{aligned}$$

We will use the three point inequality, which states that:

For any $x_{k+1}, x_k, x_* \in C$,

$$\langle \nabla\phi(x_{k+1}) - \nabla\phi(x_k), x_* - x_{k+1} \rangle = D^\phi_{x_k}(x_*) - D^\phi_{x_{k+1}}(x_*) - D^\phi_{x_k}(x_{k+1}).$$

Therefore,

$$f(x_k) - f(x_*) \leq \frac{1}{\eta} \left(D^\phi_{x_k}(x_*) - D^\phi_{x_{k+1}}(x_*) - D^\phi_{x_k}(x_{k+1}) \right) + \langle g_k, x_k - x_{k+1} \rangle.$$

We will now use the following fact:

Fact: $\langle u, v \rangle \leq \frac{\eta}{2}\|u\|^2 + \frac{1}{2\eta}\|v\|_*^2$, *by Fenchel-Young inequality*

Thus,

$$\langle g_k, x_k - x_{k+1} \rangle \leq \frac{\eta}{2}\|g_k\|_*^2 + \frac{1}{2\eta}\|x_k - x_{k+1}\|^2.$$

Hence, we have

$$\begin{aligned} f(x_k) - f(x_*) \leq \frac{1}{\eta} &\left(D^\phi_{x_k}(x_*) - D^\phi_{x_{k+1}}(x_*) - D^\phi_{x_k}(x_{k+1}) \right) \\ &+ \frac{\eta}{2}\|g_k\|_*^2 + \frac{1}{2\eta}\|x_k - x_{k+1}\|^2. \end{aligned}$$

From 1-strong convexity of $\phi(\cdot)$ w.r.t. a norm $\|\cdot\|$, we have that

$$\begin{aligned} &D^\phi_{x_{k+1}}(x_k) \geq \frac{1}{2}\|x_k - x_{k+1}\|^2 \\ &\Leftrightarrow \phi(x_{k+1}) - \phi(x_k) - \langle \nabla\phi(x_k), x_{k+1} - x_k \rangle \geq \frac{1}{2} \left\|x_k - x_{k+1}\right\|_2^2. \end{aligned}$$

Thus,

$$f(x_k) - f(x_*) \leq \frac{1}{\eta} \left(D^\phi_{x_k}(x_*) - D^\phi_{x_{k+1}}(x_*) \right) + \frac{\eta}{2}\|g_k\|_*^2.$$

Summing over $k = 1, 2, \ldots, K$, we get

$$\sum_{k=1}^{K} f(x_k) - f(x_*) \leq \frac{1}{\eta} D^\phi_{x_1}(x_*) + \sum_{k=1}^{K} \frac{\eta}{2}\|g_k\|_*^2. \tag{6}$$

□

Bibliographic notes

For more details about mirror descent, see e.g., [1] and [2].

References

[1] Arkadi Nemirovsky and David Yudin. Problem Complexity and Method Efficiency in Optimization. Wiley-Interscience series in discrete mathematics, 1983

[2] Sebastien Bubeck Convex Optimization: Algorithms and Complexity. Foundations and Trends in Machine Learning, 2015