ECE 273 Convex Optimization and Applications Instructor: Jun-Kun Wang
Scribe: Girish Krishnan, Mohamed Hamdy, and Shifeng Cheng May 9, 2024
Editor/TA: Marialena Sfyraki

Lecture 12: Mirror Descent

1 Projected Gradient Descent and Bregman Divergence

Recall from previous lectures that the Projected Gradient Descent (PGD) algorithm can be used to solve constrained optimization problems. Consider the following optimization problem:

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

PGD involves applying the following steps when initialized at $\mathbf{x}_1$ and using a step size of η:

Algorithm 1 Projected Gradient Descent

1: **for** $k = 1, 2, \cdots$ **do**
2: $\quad x_{k+1} = \text{Proj}_C[x_k - \eta\nabla_k f(x_k)]$
3: **end for**

As proved in the first homework assignment, the above expression is equivalent to the following:

Algorithm 2 Projected Gradient Descent

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

Notice that the second expression contains a term involving the squared Euclidean norm. A key idea of Mirror Descent is to generalize this algorithm to consider metrics other than the Euclidean norm, and we will do so using the notion of the Bregman Divergence.

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

$$D_y^\phi(x) = \phi(x) - \phi(y) - \langle\nabla\phi(y), x - y\rangle.$$

$\phi(x)$ is called the distance-generating function. The Bregman divergence is thus the difference between $\phi(x)$ and its linear approximation, $\phi(y) + \langle\nabla\phi(y), x - y\rangle$, at y.

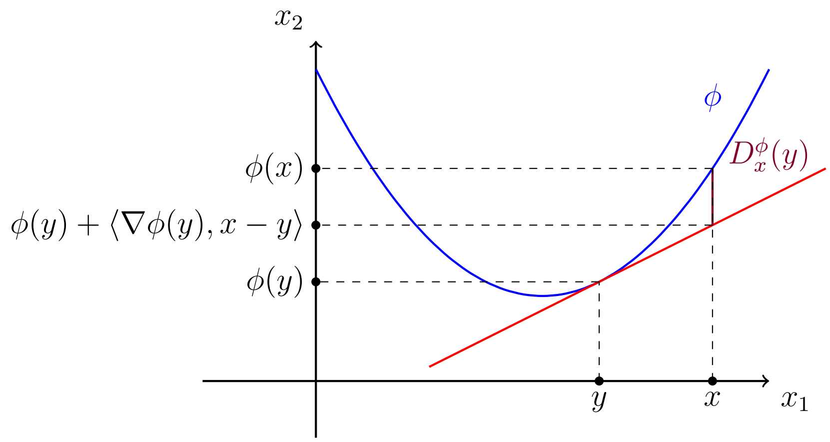

Figure 1: Bregman Divergence (shown in purple) between two points x and y with respect to a convex function ϕ (shown in blue). The tangent to the curve is shown in red.

2 Mirror Descent Algorithm

The Mirror Descent (MD) algorithm was proposed by Arkadi Nemirovsky and David Yudin in 1983, and is described below.

Algorithm 3 Mirror Descent

1: **for** $k = 1, 2, \cdots$ **do**
2: $\quad x_{k+1} = \arg\min_{x \in C} \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{\eta} D_{x_k}^{\phi}(x)$
3: **end for**

Remark: PGD is an instance of MD where $\phi(x) = \frac{1}{2}||x||_2^2$. In this instance, $\nabla\phi(x) = x$. Furthermore, we have:

$$\begin{aligned} D_{x_k}^{\phi}(x) &= \frac{1}{2}(\|x\|_2^2 - \|x_k\|_2^2) - \langle x_k, x - x_k\rangle = \frac{1}{2}(\|x\|_2^2 - \|x_k\|_2^2 - 2\langle x_k, x - x_k\rangle) \\ &= \frac{1}{2}(\|x\|_2^2 - \|x_k\|_2^2 - 2\langle x_k, x\rangle + 2\langle x_k, x_k\rangle) = \frac{1}{2}(\|x\|_2^2 + \|x_k\|_2^2 - 2\langle x_k, x\rangle) \\ &= \frac{1}{2}\|x - x_k\|_2^2. \end{aligned}$$

2.1 Example: Mirror Descent with Negative Entropy

Suppose we use the negative entropy function $\phi(x) = \sum_{i=1}^d x_i \log x_i$ for MD on a probability simplex set $C := \{x \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1, x_i \geq 0\}$. The Bregman

Divergence in this case is the Kullback-Leibler (KL) divergence:

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

The resulting MD update function on set $C = \{x \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1, x_i \geq 0\}$ is given by:

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

Based on this, we can formulate the following optimization problem:

$$\begin{gathered} \min_{x\in\mathbb{R}^d} \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{\eta}\sum_{i=1}^d x_i \log \frac{x_i}{x_{k,i}} \\ \text{s.t.} - x_i \leq 0,\ \forall i \in [d] \text{ and } \quad \sum_{i=1}^d x_i - 1 = 0. \end{gathered}$$

The first term in the objective function is linear in x and thus convex. The second term (involving negative entropy) is convex over the simplex, as proved in the first homework assignment. Therefore, the objective function is convex as it is a non-negative sum of two convex functions. The feasible set, the probability simplex, is convex as well. This renders the whole problem a convex problem. Therefore, strong duality holds and the KKT conditions can be used to determine the solution.

We will take the following steps to solve the optimization problem:

1. Find the **Lagrangian**:

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

2. **Stationary condition**: we require $\frac{\partial L}{\partial x}[i] = 0, \forall i \in [d]$, which means

$$\begin{aligned} &[\nabla f(x_k)]_i + \frac{1}{\eta}\left(\log \frac{x_i}{x_{k,i}} + 1\right) - \lambda_i + \mu = 0 \\ \Leftrightarrow \quad &\log\left(\frac{x_i}{x_{k,i}}\right) = -\eta([\nabla f(x_k)]_i - \lambda_i + \mu) - 1 \\ \Leftrightarrow \quad &x_i = x_{k,i} \exp\left(-\eta(\nabla[f(x_k)]_i - \lambda_i + \mu) - 1\right). \end{aligned} \tag{1}$$

3. **Complementary slackness**: $\lambda_i x_i = 0$ and $\forall i \in [d]$. In the case where $\lambda_i = 0, x_i \neq 0$, Equation 1 becomes

$$\begin{aligned} x_i &= x_{k,i} \exp\left(-\eta([\nabla f(x_k)]_i + \mu) - 1\right) \\ &= x_{k,i} \frac{\exp\left(-\eta[\nabla f(x_k)]_i\right)}{\exp\left(\eta\mu + 1\right)}. \end{aligned}$$

Primal feasibility requires the following:

$$\begin{aligned} \sum_{i=1}^{d} x_i = 1 \quad &\Leftrightarrow \quad \sum_{i=1}^{d} x_{k,i} \frac{\exp\left(-\eta[\nabla f(x_k)]_i\right)}{\exp(\eta\mu + 1)} = 1 \\ &\Rightarrow \quad \exp(\eta\mu + 1) = \sum_{j=1}^{d} x_{k,j} \exp\left(-\eta[\nabla f(x_k)]_j\right) \\ &\Rightarrow \quad x_i = x_{k,i} \frac{\exp\left(-\eta[\nabla f(x_k)]_i\right)}{\sum_{j=1}^{d} x_{k,j} \exp(-\eta[\nabla f(x_k)]_j)}. \end{aligned}$$

Therefore, the update at each coordinate $i \in [d]$ using negative entropy is given by:

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

This is known as the **exponentiated gradient**.
Note: When applying the complementary slackness condition, we ruled out the possibility that $x_i = 0$. This is because if $x_{1,i} \neq 0 \, \forall \, i \in [d]$, then according to the update step shown above, $x_{2,i}$ will not be zero unless $[\nabla f(x_k)]_i = \infty$.

3 Mirror Descent on Non-Differentiable Functions

Let $f(x)$ be a convex but not necessarily differentiable function, and let $g_k \in \partial f(x_k)$ be the subgradient of $f(\cdot)$ at x_k. The MD in this case is summarized in algorithm 4.

Algorithm 4 Mirror Descent, non-differentiable f

```
1: for k = 1, 2, ··· do
2:     x_{k+1} = arg min_{x∈C} ⟨g_k, x − x_k⟩ + (1/η) D^φ_{x_k}(x)
3: end for
4: Output: x̄ := (Σ_{k=1}^K x_k) / K.
```

Recall the definition of the dual norm:

Definition 2. ***(Dual norm)*** *Given a norm $\|\cdot\|$, the dual norm $\|\cdot\|_*$ is defined as*

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

For any $p \geq 1$, the l_p-norm is defined as:

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

Theorem 1. *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.*

For example, the l_1-norm, $||\cdot||_1$ is dual with the l_∞ norm, $||\cdot||_\infty$.

Theorem 2. *Consider a generating function $\phi(x)$ that 1-strongly convex w.r.t $\|\cdot\|$. Then, mirror descent has*

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

where $\|\cdot\|_*$ denotes the dual norm.

Remark: The above inequality is similar to the inequality we proved in lecture 7 for the expected optimality gap in Stochastic Gradient Descent (SGD) for convex functions.

4 Mirror Descent vs. Projected Gradient Descent

Consider the convex constrained optimization problem $\min_{x \in C} f(x)$, where C is the probability simplex defined by $C := \{x \in \mathbb{R}^d : \sum_{i=1}^{d} x_i = 1, x_i \geq 0\}$. In this problem, $\phi(x) = \sum_{i=1}^{d} x_i \log x_i$, which is the negative entropy function and is 1-strongly convex with respect to $\|\cdot\|_1$.

4.1 Solving the Problem Using Mirror Descent

Let $x_1 = \frac{1}{d}\mathbf{1}_d$ (the uniform discrete distribution). Then, the following inequality holds:

$$\begin{aligned} D^{\phi}_{x_1}(x_*) &= \sum_{i=1}^{d} x_{*,i} \log \frac{x_{*,i}}{1/d} \\ &= \underbrace{\sum_{i=1}^{d} x_{*,i} \log x_{*,i}}_{\leq 0} + \log d \underbrace{\sum_{i=1}^{d} x_{*,i}}_{=1 \text{ as } x_* \in C} \\ &\leq \log d. \end{aligned}$$

Suppose that $\|g_k\|_\infty^2 \leq 1$. Denoting the number of iterations by K, we have:

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

The tightest bound is achieved with parameter tuning when the following holds:

$$\frac{1}{\eta} \log d = \frac{\eta}{2} K \Leftrightarrow \eta = \sqrt{\frac{2 \log d}{K}}.$$

Therefore,

$$\sum_{k=1}^{K} f(x_k) - f(x_*) \; \leq \; \sqrt{\frac{K}{2 \log d}} \cdot \log d + \frac{1}{2} \sqrt{\frac{2 \log d}{K}} K \; = \sqrt{2K \log d} = \mathcal{O}(\sqrt{K \log d}).$$

Since the algorithm returns $\bar{x}$, we apply Jensen's inequality to obtain:

$$\begin{aligned} f(\bar{x}) - f(x_*) &\leq \frac{1}{K} \sum_{k=1}^{K} f(x_k) - f(x_*) \\ &= \mathcal{O}\left(\sqrt{\frac{\log d}{K}} \right). \end{aligned}$$

From the above discussion, we see that after K iterations, MD achieves an ϵ-optimality gap, where $\epsilon = \mathcal{O}\left(\sqrt{\frac{\log d}{K}}\right)$.

4.2 Solving the Same Problem Using PGD

We now apply Projected Gradient Descent (PGD) on the same problem, where the Bregman divergence $D^{\phi}_{x_1}(x_*) = \frac{1}{2}\|x_1 - x_*\|_2^2 \leq B$, where B is a bound on the initial distance. This setting corresponds to the quadratic form $\frac{1}{2}\|\cdot\|_2^2$, which is strongly convex with respect to the ℓ_2-norm, which is a self-dual norm.

Considering the norm inequality $\|z\|_\infty \leq \|z\|_2 \leq \sqrt{d}\|z\|_\infty$ for all $z \in \mathbb{R}^d$, it follows that $\|g_k\|_2^2 \leq d\|g_k\|_\infty^2 \leq d$.

Using Theorem 2, the cumulative error bound over K iterations is given by:

$$\begin{aligned}\sum_{k=1}^{K} f(x_k) - f(x_*) &\leq \frac{1}{\eta} D^{\phi}_{x_1}(x_*) + \sum_{k=1}^{K} \frac{\eta}{2}\|g_k\|_*^2 \\ &= \frac{1}{2\eta}\|x_1 - x_*\|_2^2 + \sum_{k=1}^{K} \frac{\eta}{2}\|g_k\|_*^2 \\ &\leq \frac{1}{\eta}B + \frac{\eta}{2}Kd.\end{aligned}$$

For the optimal choice of $\eta = \sqrt{\frac{2B}{Kd}}$ that gives the tightest bound, the cumulative error bound achieves the order $\mathcal{O}(\sqrt{BKd})$.

$$\begin{aligned}&\sum_{k=1}^{K} \big(f(x_k) - f(x_*)\big) = \mathcal{O}(\sqrt{BKd}) \\ \Leftrightarrow\ &\frac{1}{K}\sum_{k=1}^{K} \big(f(x_k) - f(x_*)\big) = \mathcal{O}\left(\sqrt{\frac{Bd}{K}}\right).\end{aligned}$$

Applying Jensen's inequality to the convex function f, we deduce:

$$f(\overline{x}_k) - f(x_*) \leq \frac{1}{K}\sum_{k=1}^{K} \big(f(x_k) - f(x_*)\big) = \mathcal{O}\left(\sqrt{\frac{Bd}{K}}\right).$$

Hence, the convergence rate of PGD is $\mathcal{O}\left(\sqrt{\frac{1}{K}}\right)$, similarly to MD. However, the constant factor is crucial, as it makes MD particularly more efficient for high-dimensional problems (d large). To achieve an ϵ-optimality gap, the required number of iterations for MD is approximated by $\sqrt{K} \approx \frac{\sqrt{\log d}}{\epsilon}$, while for PGD it scales as $\sqrt{K} \approx \frac{\sqrt{d}}{\epsilon}$, noting that $\log d \leq d$.

Table 1 summarizes the expressions for the ϵ-optimality gap and the required number of iterations to achieve this gap for the MD and PGD algorithms.

Method	ϵ	Approx. Required Iterations ($\sqrt{K}$)
MD	$\mathcal{O}\left(\sqrt{\frac{\log d}{K}}\right)$	$\frac{\sqrt{\log d}}{\epsilon}$
PGD	$\mathcal{O}\left(\sqrt{\frac{d}{K}}\right)$	$\frac{\sqrt{d}}{\epsilon}$

Table 1: Summary of ϵ for MD and PGD.

Bibliographic Notes

The Mirror Descent algorithm is covered in more detail in Chapter 7 of [Vishnoi (2021)] and in Chapter 5 of [Nemirovski (2022)].

References

[Vishnoi (2021)] Nisheeth K. Vishnoi. Algorithms for Convex Optimization. Cambridge University Press, 2021

[Nemirovski (2022)] Arkadi Nemirovsky. Lectures on Modern Convex Optimization. `https://urldefense.com/v3/__https://www2.isye.gatech.edu/*nemirovs/LMCOLN2022Fall.pdf__;fg!!Mih3wA!Ha3qLKdYzLUSCnY1cE_8WyLDp9VTholjqwrPAGLUz3duCx3nDdTqzhk8d6wcAWH9BOwkAjJAc9gD3yP0_zOWxQ$`