ECE 273 Convex Optimization and Applications Instructor: Jun-Kun Wang
Scribe: Animesh Singhal, Shantanu Shukla, Rohan Nagrik May 14, 2024
Editor/TA: Marialena Sfyraki

Lecture 13: Mirror Descent (Continued) and Online Convex Optimization

1 Review

1.1 Mirror Descent

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

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

Algorithm 1 Mirror Descent

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

1.2 Example: Probability Simplex

Suppose $C := \{x \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1, x_i \geq 0\}$ is the probability simplex. The mirror descent update is:

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

The update for each coordinate $i \in [d]$ is:

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

1.3 Theorem: Mirror Descent Convergence

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

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

1.4 Comparison: Mirror Descent vs. Projected Gradient Descent

For the problem $\min_{x \in C} f(x)$, where C is the simplex $\Delta_d := \{x \in \mathbb{R}^d : \sum_{i=1}^d x[i] = 1, x[i] \geq 0, \forall i\}$:

- Projected Gradient Descent: $\epsilon = O\left(\sqrt{\frac{d}{K}}\right)$, where K is the number of iterations.
- Mirror Descent: $\epsilon = O\left(\sqrt{\frac{\log d}{K}}\right)$.

2 Alternative View of Mirror Descent

We've previously formulated the optimization of a function using mirror descent. Now, let's explore the geometric intuition behind the term 'Mirror' by discussing an alternate representation of the mirror descent process.

2.1 Mirror Descent Update

The Mirror Descent update can be expressed as:

$$x_{k+1} = \arg\min_{x \in C} \left(\langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\eta} D_{x_k}^{\phi}(x) \right).$$

For a distance-generating function $\phi(\cdot)$ that is closed and convex, with a differentiable conjugate $\phi^*(\cdot)$, this can be equivalently written as:

$$\begin{aligned} \nabla \phi(y_{k+1}) &= \nabla \phi(x_k) - \eta \nabla f(x_k), \\ y_{k+1} &= \nabla \phi^* \left(\nabla \phi(x_k) - \eta \nabla f(x_k) \right), \\ x_{k+1} &= \arg\min_{x \in C} D_{y_{k+1}}^{\phi}(x). \end{aligned}$$

We will use the Fenchel Inequality to demonstrate the equivalence between these two representations.

2.2 Fenchel Conjugate and Inequality

Given a function $f(\cdot)$, its Fenchel Conjugate is defined as:

$$f^*(y) = \sup_{x \in \mathrm{dom}(f)} \left(y^\top x - f(x) \right).$$

Theorem 2 (Fenchel Inequality)**.** *For any x and y:*

$$f^*(y) \geq y^\top x - f(x).$$

Remark: Intuitively, this inequality holds because the supremum of an expression is always at least as large as the expression itself.

Question: When do we have the equality, i.e. $f^*(y) = y^\top x - f(x)$?
Answer: When the supremum is attained.

2.2.1 Closed functions

A function is closed if its sublevel set is a closed set, i.e.,

$$\{x \in \text{dom}(f) : f(x) \leq \alpha\} \text{ is a closed set.}$$

2.2.2 Equality in Fenchel Inequality

Theorem 3. *If $f(\cdot)$ is closed and convex, then:*

$$f^*(y) + f(x) = y^\top x \iff y \in \partial f(x) \iff x \in \partial f^*(y).$$

Remark: When $f(\cdot)$ is differentiable, Theorem 2 implies:

$$f^*(y) + f(x) = y^\top x \iff y = \nabla f(x) \iff x = \nabla f^*(y).$$

We have two important results follow from these definitions:

1. For the maximization problem:

$$\arg\max_x \langle x, y\rangle - f(x),$$

the solution is $x \in \partial f^*(y)$. This is because when $\langle x, y\rangle - f(x)$ is maximized, Fenchel's inequality holds as an equality, implying:

$$\langle x, y\rangle - f(x) = f^*(y) \implies x \in \partial f^*(y).$$

2. For the maximization problem:

$$\arg\max_y \langle y, x\rangle - f^*(y),$$

the solution is $y \in \partial f(x)$. Applying Fenchel's inequality in reverse gives:

$$f(x) = \sup_{y \in \text{dom}(f^*)} \left(y^T x - f^*(y) \right).$$

Therefore, when $\langle y, x\rangle - f^*(y)$ is maximized, Fenchel's inequality holds as an equality, implying:

$$\langle y, x\rangle - f^*(y) = f(x) \implies y \in \partial f(x).$$

Thus, the Fenchel conjugate of the Fenchel conjugate of x is x itself, confirming that $y \in \partial f(x)$.

2.2.3 Inverse Map

Theorem 4. *Suppose that $f(\cdot)$ is closed and convex. Then, $y \in \partial f(x)$ if and only if $x \in \partial f^*(y)$.*

Remark: This result implies that the gradient map has an inverse when the function is differentiable, and the inverse is the gradient of the conjugate function. In fact, when the function and its conjugate are differentiable, this yields the fact that $y = \nabla f(x)$ if and only if $x = \nabla f^*(y)$. Observe that if we have

$$y = \nabla f(x),$$

then if we let $x = \nabla f^*(y)$, we get

$$y = \nabla f\left(f^*(y)\right).$$

Similarly, if we have

$$x = \nabla f^*(y),$$

then if we let $y = \nabla f(x)$, we get

$$x = \nabla f^*\left(\nabla f(x)\right).$$

Since f is convex and closed, ∇f and ∇f^* are inverses of each other:

$$\nabla f(\nabla f^*(x)) = x,$$
$$\nabla f^*(\nabla f(y)) = y.$$

This result is given by the following theorem.

Theorem 5. *If a differentiable function $\phi(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is closed and convex, then:*

$$y = \nabla\phi(x) \quad \iff \quad x = \nabla\phi^*(y).$$

2.3 Geometric Picture of Mirror Descent

The first step of alternative mirror descent representation can be written as:

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

The second step involves:

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

To understand the geometric intuition of mirror descent, consider these steps:

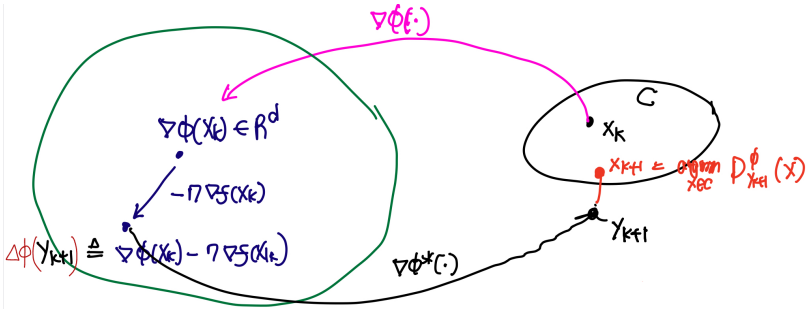

Figure 1: Mirror descent: Geometric picture

1. Mapping to Another Space: Use $\nabla\phi(\cdot)$ to map x_k to a new space, leveraging the geometry induced by $\phi(\cdot)$.

2. Gradient Step in the New Space: Take a step in the direction of the negative gradient of f, yielding $\nabla\phi(y_{k+1})$.

3. Inverse Mapping: Map back to the original space using $\nabla\phi^*(\cdot)$, bringing us to y_{k+1}.

4. Projection Back to the Constraint Set: Project y_{k+1} back onto C by minimizing the Bregman divergence $D^{\phi}_{y_{k+1}}(x)$:

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

Proof of Equivalency

$$
\begin{aligned}
x_{k+1} &= \arg\min_{x\in C} \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{\eta} D^{\phi}_{x_k}(x) && \text{(primal view)} \\
&= \arg\min_{x\in C} \eta \nabla f(x_k)^\top x + \phi(x) - \phi(x_k) - \langle \nabla\phi(x_k), x - x_k\rangle && \text{(def. of Bregman divergence)} \\
&= \arg\min_{x\in C} \phi(x) - (\nabla\phi(x_k) - \eta\nabla f(x_k))^\top x && \text{(keeping terms depending on } x\text{)} \\
&= \arg\min_{x\in C} \phi(x) - (\nabla\phi(y_{k+1}))^\top x && \text{(first step of alternative view)} \\
&= \arg\min_{x\in C} \phi(x) - \phi(y_{k+1}) - \langle \nabla\phi(y_{k+1}), x - y_{k+1}\rangle && \text{(adding terms not depending on } x\text{)} \\
&= \arg\min_{x\in C} D^{\phi}_{y_{k+1}}(x) && \text{(def. of Bregman divergence)}
\end{aligned}
$$

3 Online Convex Optimization

Online convex optimization (OCO) lies at the intersection between learning and convex optimization. Most Online Convex Optimization scenarios follow the following generic protocol:

Algorithm 2 Protocol/Setting

1: **for** $t = 1, 2, \ldots, T$ **do**
2: Commit a point z_t with its convex decision space $Z \subset \mathbb{R}^d$.
3: Receive a convex loss function $\ell_t(\cdot) : Z \to \mathbb{R}$ and incur a loss $\ell_t(z_t)$.
4: **end for**

The objective of Online Convex Optimization (OCO) is to perform comparably to any preset benchmark or reference within a convex decision space, denoted as Z. Specifically, the algorithm's regret compared to any fixed benchmark or reference point $z^* \in Z$ is calculated as follows:

$$
\text{Regret}_T(z_*) := \underbrace{\sum_{t=1}^{T} \ell_t(z_t)}_{\text{cumulative loss of learner}} - \underbrace{\sum_{t=1}^{T} \ell_t(z_*)}_{\text{cumulative loss of benchmark}},
$$

where T is the number of total rounds. We wish to achieve sub-linear regret i.e.

$$
\text{Regret}_T(z_*) = o(T).
$$

That is, we wish to have

$$\underbrace{\frac{\text{Regret}_T(z_*)}{T}}_{\text{average regret}} \to 0 \quad \text{as} \quad T \to \infty.$$

Thus, we need the average regret to be vanishing over time.

Question: Why not compete with the best action at each time step ?

$$\sum_{t=1}^{T} \ell_t(z_t) - \sum_{t=1}^{T} \min_z \ell_t(z) = \Omega(T).$$

Answer: The reason is that we only learn the loss $\ell_t(\cdot)$ after we have already made our decision z_t within the decision space Z. Based on the metric in the above equation, if we choose anything apart from the point that minimizes the loss function $\ell_t(\cdot)$ for each time-step, we'll inevitably incur some loss. As a result, the value of the above equation will increase linearly with T. This issue leads us to focus on "regret" relative to a consistent benchmark action z^*, instead of always trying to minimize the loss at each step.

3.1 Online Linear Optimization

This is a special case of Online Convex Optimization where the cost function l_t is linear in z at each time-step i.e. $l_t(z) = c_t^T z_t$

Algorithm 3 Protocol/Setting

1: **for** $t = 1, 2, \ldots$ **do**
2: Commit a point z_t with its convex decision space $Z \subseteq \mathbb{R}^d$.
3: Receive a loss function $\ell_t(z) = c_t^\top z$ and incur a loss $\ell_t(z_t)$.
4: **end for**

The regret of OCO can be bounded by the regret of OLO using the property of sub-gradients. If g_x is a sub-gradient of $f(x)$ at x then :

$$f(y) \geq f(x) + \langle g_x, y - x \rangle, \quad \forall y \in \mathcal{C}$$

$$\underbrace{\ell_t(z_t) - \ell_t(z_*)}_{\text{per-round regret of OCO}} \leq \underbrace{\langle z_t - z_*, c_t \rangle}_{\text{per-round regret of OLO}},$$

where c_t is the sub-gradient of $l_t(\cdot)$ at z_t.

3.2 Follow-the-Leader (FTL)

In this algorithm, at the first time step we select some initial vector to start with and in the subsequent time-steps, we choose the vector with the minimum loss across all preceding time-steps (or rounds). The algorithm is given below:

Algorithm 4 Follow-the-Leader (FTL)

Input a convex decision space $Z \subseteq \mathbb{R}^d$ and $z_{\text{init}} \in Z$.
for $t = 1, 2, \ldots$ **do**
  if $t = 1$ **then**
    Commit $z_t = z_{\text{init}}$
  else
    Commit $z_t = \arg\min_{z \in Z} \sum_{s=1}^{t-1} \ell_s(z)$.
  end if
  Receive a convex loss function $\ell_t(\cdot) : Z \to \mathbb{R}$ and incur a loss $\ell_t(z_t)$.
end for

Lemma 1. *Let $z_1, z_2, \ldots$ be the sequences of points generated by FTL. Then for any reference $z_* \in Z$:*

$$\textit{Regret}_T(z_*) = \sum_{t=1}^{T} \ell_t(z_t) - \sum_{t=1}^{T} \ell_t(z_*) \leq \sum_{t=1}^{T} \ell_t(z_t) - \sum_{t=1}^{T} \ell_t(z_{t+1})$$

Remark: Before moving on to the proof of the Lemma, we can see that subtracting $\sum_{t=1}^{T} \ell_t(z_t)$ would yield

$$\sum_{t=1}^{T} \ell_t(z_{t+1}) \leq \sum_{t=1}^{T} \ell_t(z_*).$$

Proof. We proceed by induction. We prove this inequality by induction. The base case of $T = 1$ follows directly from the definition of z_{t+1} i.e.

$$z_2 = \arg\min_{z \in Z} \ell_1(z) \;\Rightarrow\; \ell_1(z_2) \leq \ell_1(z),\ \forall z \in Z.$$

Thus, if we choose $z = z_*$, we have that

$$\ell_1(z_2) \leq \ell_1(z_*)$$

Assume the inequality holds for $t = T - 1$. Then for all $z_* \in Z$, we have

$$\sum_{t=1}^{T-1} \ell_t(z_{t+1}) \leq \sum_{t=1}^{T-1} \ell_t(z_*)$$

Now adding $l_T(z_{T+1})$ to both sides gives that for all $z_* \in Z$, we have

$$\ell_T(z_{T+1}) + \sum_{t=1}^{T-1} \ell_t(z_{t+1}) \leq \ell_T(z_{T+1}) + \sum_{t=1}^{T-1} \ell_t(z_*).$$
$$\iff \sum_{t=1}^{T} \ell_t(z_{t+1}) \leq \ell_T(z_{T+1}) + \sum_{t=1}^{T-1} \ell_t(z_*).$$

Since the above holds for all $z_* \in Z$ we can choose $z_* = z_{T+1}$, which implies

$$\begin{aligned}\sum_{t=1}^{T} \ell_t(z_{t+1}) &\leq \sum_{t=1}^{T} \ell_T(z_{T+1}) \\ &\leq \sum_{t=1}^{T} \ell_T(z_*), \quad \forall z_* \in Z,\end{aligned}$$

where we used the fact that $z_{T+1} = \arg\min_{z \in Z} \sum_{t=1}^{T} l_T(z)$. Thus, the inductive argument is complete. □

Theorem 6. *Consider $\ell_t(z) = \frac{1}{2}||z - c_t||_2^2$. Assume $max_t||c_t|| \leq G$. Then FTL has a regret at most :*

$$Regret_T(z^*) \leq 4G^2(\log(T) + 1), \quad \textit{for any } z^* \in \mathbb{R}^d.$$

for any $z^* \in \mathbb{R}^d$.

Proof. We assume $Z = \mathbb{R}^d$. Using the FTL rule:

$$\begin{aligned}z_t &= \arg\min_{z \in Z} \sum_{s=1}^{t-1} \ell_s(z) \\ &= \arg\min_{z \in Z} \sum_{s=1}^{t-1} \|z - c_s\|_2^2 \\ &= \arg\min_{z \in Z} F(z)\end{aligned}$$

Since $F(z)$ is convex, we have:

$$\frac{\partial F(z_t)}{\partial z} = \sum_{s=1}^{t-1} 2(z_t - c_s) = 0 \implies z_t = \frac{1}{t-1} \sum_{s=1}^{t-1} c_s$$

where z_t is the average of $c_1, c_2, \ldots c_{t-1}$. This can now be rewritten as:

$$z_{t+1} = \underbrace{\frac{1}{t}\left(c_t + (t-1)z_t\right)}_{\sum_{s=1}^{t} c_s} = \left(1 - \frac{1}{t}\right) z_t + \frac{1}{t} c_t.$$

By subtracting c_t from both sides, we get:

$$z_{t+1} - c_t = \left(1 - \frac{1}{t}\right) z_t + \left(\frac{1}{t} - 1\right) c_t = \left(1 - \frac{1}{t}\right)(z_t - c_t).$$

Therefore,

$$\begin{aligned}
\ell_t(z_t) - \ell_t(z_{t+1}) &= \frac{1}{2}\|z_t - c_t\|_2^2 - \frac{1}{2}\|z_{t+1} - c_t\|_2^2 \\
&= \frac{1}{2}\|z_t - c_t\|_2^2 - \frac{1}{2}\left(1 - \frac{1}{t}\right)^2 \|z_t - c_t\|_2^2 \\
&= \frac{1}{2}\left(1 - \left(1 - \frac{1}{t}\right)^2\right) \|z_t - c_t\|_2^2 \\
&= \frac{1}{2}\left(1 - 1 + \frac{2}{t} - \frac{1}{t^2}\right) \|z_t - c_t\|_2^2 \\
&\leq \frac{1}{2}\left(\frac{2}{t}\right) \|z_t - c_t\|_2^2 \\
&= \frac{1}{t}\|z_t - c_t\|_2^2
\end{aligned}$$

Let $G = \max_t ||c_t||$. We have $z_t \leq G$ since z_t is the average of $c_1, \ldots, c_{t-1}$ and hence, by the Triangle Inequality we get $||z_t - c_t||_2 \leq 2G$. We therefore obtain:

$$\begin{aligned}
\ell_t(z_t) - \ell_t(z_{t+1}) &= \frac{1}{t}\|z_t - c_t\|_2^2 \\
&= \frac{1}{t}\left(\|z_t - c_t\|_2\right)^2 \\
&\leq \frac{1}{t}\left(\|z_t\|_2 + \|c_t\|_2\right)^2 \\
&\leq \frac{1}{t}(2G)^2.
\end{aligned}$$

Summing over $t = 1, \ldots, T$ we obtain:

$$\sum_{t=1}^{T} (\ell_t(z_t) - \ell_t(z_{t+1})) \leq (2G)^2 \sum_{t=1}^{T} \frac{1}{t}.$$

Now using induction, we prove the inequality:

$$\sum_{t=1}^{T} \frac{1}{t} \leq \log(T) + 1,$$

For the base case $T = 1$ it is obviously true that:

$$\sum_{t=1}^{1} \frac{1}{t} = 1 \leq \log(1) + 1 = 1.$$

Let's assume that the inequality holds for $T - 1$. Thus we have:

$$\sum_{t=1}^{T-1} \frac{1}{t} \leq \log(T-1) + 1.$$

Observe that we have:

$$\begin{aligned}
\log(T) + 1 &= \log\left((T-1)\frac{T}{T-1}\right) + 1 \\
&= \log(T-1) + \log(\frac{T}{T-1}) + 1 \\
&\geq \sum_{t=1}^{T-1} \frac{1}{t} + \log(\frac{T}{T-1})
\end{aligned}$$

Using the fact that $\log(x) \geq 1 - \frac{1}{x}$, we get $\log(\frac{T}{T-1}) \geq \frac{1}{T}$, which after substitution in the above equation gives:

$$\log(T) + 1 \geq \sum_{t=1}^{T-1} \frac{1}{t} + \frac{1}{T} = \sum_{t=1}^{T} \frac{1}{t}.$$

Combining the above inequality with Lemma 1 and using the fact that:

$$\sum_{t=1}^{T} \frac{1}{t} \leq \log(T) + 1,$$

we derive:

$$\text{Regret}_T(z^*) \leq \sum_{t=1}^{T} (\ell_t(z_t) - \ell_t(z_{t+1})) \leq (2L)^2 \sum_{t=1}^{T} \frac{1}{t} \leq 4L^2(\log(T) + 1),$$

which completes the proof. □