Two different views of (Online) Mirror Descent

ComComX • 13 Jun 2025

Online Convex Optimization

The online convex optimization is stated as a general repeated game between a player and an adversary as follows:

On a given set $\mathcal{X} \subset \mathbb{R}^n$, at each round $t = 1, 2, \ldots, T$:

1. The adversary chooses a loss function $l_t : \mathcal{X} \to \mathbb{R}^n$ and keeps it in secret;
2. The player plays an action $\mathbf{x}_t \in X$;
3. The adversary reveals l_t;
4. The player incurs the loss $l_t(\mathbf{x}_t)$;

Goal: The player tries to minimize the cumulated loss $\sum_{t=1}^{T} l_t(\mathbf{x}_t)$.

The performance of algorithms for the above problem is often measured by the regret, which is the difference between the cumulated loss derived by the algorithm and the best *fixed* action in hindsight:

$$R_T := \sum_{t=1}^{T} l_t(\mathbf{x}_t) - \min_{\mathbf{x} \in \mathcal{X}} \sum_{t=1}^{T} l_t(\mathbf{x}). \tag{1}$$

The goal of the player now is to keep the growth of R_T sublinear in T, so that the average regret per step goes to zero as $T \to \infty$.

Without assumptions on the loss function l_t, the above problem is impossible to solve. Therefore, we need to restrict the adversary's power in choosing the loss functions and also provide the player with some additional hints:

Assumption 1.

1. *$\mathcal{X}$ is closed and convex, and* $\operatorname{int}(\mathcal{X}) \neq \varnothing$*;*
2. *For all t, l_t is convex and L_t-Lipschitz continuous w.r.t. some fixed norm $\|\cdot\|$;*
3. *For all t, a subgradient $\mathbf{g}_t \in \partial l_t(\mathbf{x}_t)$ of l_t at $\mathbf{x}_t$ is accessible by the player.*

View 1: Subgradient algorithm with projection

Definition 2 (Legendre function)**.** *A function $\psi : \mathcal{X} \to \mathbb{R}$ is a Legendre function if it is:*

1. *Essentially strictly convex: strictly convex and continuously differentiable on $\operatorname{int}(\mathcal{X})$;*
2. *Essentially smooth: for $\mathbf{x} \in \operatorname{int}(\mathcal{X})$, $\lim_{\mathbf{x} \to \operatorname{bdr}(\mathcal{X})} \|\nabla\psi(\mathbf{x})\| = +\infty$.*

Definition 3 (Bregman divergence)**.** *Given a Legendre function ψ, the Bregman divergence w.r.t. ψ, denoted by $B_\psi : \mathcal{X} \times \operatorname{int}(\mathcal{X}) \to \mathbb{R}$, is defined as:*

$$B_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle. \tag{2}$$

Since ψ is strictly convex, we know that $B_\psi(\mathbf{x}, \mathbf{y})$ is nonnegative and $B_\psi(\mathbf{x}, \mathbf{y}) = 0$ iff $\mathbf{x} = \mathbf{y}$.

Proposition 4. *Given a Legendre function ψ, the Bregman projection $\Pi_\mathcal{X}(\mathbf{z}) = \arg\min_{\mathbf{x} \in \mathcal{X}} B_\psi(\mathbf{x}, \mathbf{z})$ exists and unique. Moreover, $\Pi_\mathcal{X}(\mathbf{z}) \in \operatorname{int}(\mathcal{X})$.*

Let's break down the projection operator above to get a handy result we will use later. For that purpose, define the indicator function

$$\delta_\mathcal{X}(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in \mathcal{X} \\ +\infty & \text{otherwise.} \end{cases} \tag{3}$$

We also have that the subdifferential $\partial\delta_\mathcal{X}(\mathbf{x})$ is the normal cone to $\mathcal{X}$ at $\mathbf{x}$, denoted by $N_\mathcal{X}(\mathbf{x})$. Now, expanding the projection, we have

$$\begin{aligned} \Pi_\mathcal{X}(\mathbf{z}) &= \arg\min_{\mathbf{x} \in \mathcal{X}} \Big\{ \psi(\mathbf{x}) - \psi(\mathbf{z}) - \langle \nabla\psi(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle \Big\} \\ &= \arg\min_{\mathbf{x} \in \mathbb{R}^n} \Big\{ \psi(\mathbf{x}) - \psi(\mathbf{z}) - \langle \nabla\psi(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle + \delta_\mathcal{X}(\mathbf{x}) \Big\}. \end{aligned}$$

By first-order optimality condition:

$$0 \in \nabla\psi(\Pi_\mathcal{X}(\mathbf{z})) - \nabla\psi(\mathbf{z}) + N_\mathcal{X}(\Pi_\mathcal{X}(\mathbf{z})). \tag{4}$$

Rearranging, it follows that

$$\Pi_\mathcal{X}(\mathbf{z}) \in (\nabla\psi + N_\mathcal{X})^{-1}(\nabla\psi(\mathbf{z})). \tag{5}$$

Since the projection exists and unique, we can write, without any ambiguous,

$$\Pi_\mathcal{X}(\mathbf{z}) = (\nabla\psi + N_\mathcal{X})^{-1}(\nabla\psi(\mathbf{z})). \tag{6}$$

We are now at the position to present our first view of the (Online) Mirror Descent algorithm. The algorithm is often written in the form of *subgradient algorithm with projection* as follows:

View 1: subgradient algorithm with projection

$$\mathbf{x}_1 \in \text{int}(\mathcal{X}) \tag{7}$$

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \mathbf{x}, \mathbf{g}_t \rangle + \frac{1}{\eta_t} B_\psi(\mathbf{x}, \mathbf{x}_t) \right\}. \tag{8}$$

This is a generalization of the subgradient algorithm, where the projection can be based on any Bregman divergence rather than the Euclidean distance. Therefore, it captures the geometric properties of $\mathcal{X}$ more effectively.

Let's explain why this update procedure is called a subgradient algorithm with projection, i.e., where the "projection" comes from. We rewrite our algorithm as

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \langle \mathbf{x}, \mathbf{g}_t \rangle + \frac{1}{\eta_t} B_\psi(\mathbf{x}, \mathbf{x}_t) + \delta_\mathcal{X}(\mathbf{x}) \right\}. \tag{9}$$

By the first-order optimality condition, we have

$$0 \in \eta_t \mathbf{g}_t + \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t) + N_\mathcal{X}(\mathbf{x}_{t+1}) \tag{10}$$

Rearranging, we get

$$\mathbf{x}_{t+1} \in (\nabla\psi + N_\mathcal{X})^{-1}(\nabla\psi(\mathbf{x}_t) - \eta_t \mathbf{g}_t). \tag{11}$$

Now, let $\mathbf{y}_{t+1}$ be a point such that $\nabla\psi(\mathbf{y}_{t+1}) = \nabla\psi(\mathbf{x}_t) - \eta_t \mathbf{g}_t$. We can then rewrite (11) as

$$\mathbf{x}_{t+1} \in (\nabla\psi + N_\mathcal{X})^{-1}(\nabla\psi(\mathbf{y}_{t+1})). \tag{12}$$

By (6), $(\nabla\psi + N_\mathcal{X})^{-1}(\nabla\psi(\mathbf{y}_{t+1}))$ is exactly the Bregman projection of $\mathbf{y}_{t+1}$ onto $\mathcal{X}$. Moreover, this projection is guaranteed to exist and be unique, thus we can write

$$\mathbf{x}_{t+1} = \Pi_\mathcal{X}(\mathbf{y}_{t+1}). \tag{13}$$

In addition, by proposition 4, $\mathbf{x}_{t+1}$ is guaranteed to be in the interior of $\mathcal{X}$. Therefore, our algorithm is well-defined.

View 2: Mirror descent

Now we move to the second view of the algorithm, the view of Mirror Descent introduced by Nemirovski and Yudin for convex optimization. Let first define the convex conjugate (or Legendre-Fenchel transformation) of the function ψ, denoted by ψ^*:

$$\psi^*(\mathbf{z}) = \max_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \mathbf{x}, \mathbf{z} \rangle - \psi(\mathbf{x}) \right\}. \tag{14}$$

Proposition 5. *If ψ is proper, lsc, and σ-strongly convex. Then*

1. *$(\partial\psi)^{-1}$ is everywhere single-valued and $1/\sigma$-Lipschitz continuous;*
2. *ψ^* is finite everywhere and differentiable.*

View 2: Online Mirror Descent

$$\mathbf{x}_1 \in \operatorname{int}(\mathcal{X}) \tag{15}$$
$$\mathbf{z}_{t+1} = \nabla\psi(\mathbf{x}_t) - \eta_t \mathbf{g}_t \tag{16}$$
$$\mathbf{x}_{t+1} = \nabla\psi^*(\mathbf{z}_{t+1}). \tag{17}$$

Intuitively, first $\nabla\psi$ maps $\mathbf{x}_t$ from the primal space $\mathcal{X}$ into its dual space, and the update is done in the dual space. Then, $\nabla\psi^*$ maps the new point from the dual space back into $\mathcal{X}$.

Now we show that this second view is equivalent to our first view, i.e. the sequences $(x_t)_{t=1\ldots T}$ produced by the two algorithms are identical. Applying the first-order optimality condition to the definition of $\psi^*(z)$, we have $0 \in \mathbf{z} - \nabla\psi(\mathbf{x}) - N_{\mathcal{X}}(\mathbf{x})$, or $\mathbf{x} \in (\nabla\psi + N_{\mathcal{X}})^{-1}(\mathbf{z})$. Moreover, by proposition 5 and by the conjugacy, we have

$$\mathbf{x} = (\nabla\psi + N_{\mathcal{X}})^{-1}(\mathbf{z}) = \nabla\psi^*(\mathbf{z}) = (\partial\psi)^{-1}(\mathbf{z}). \tag{18}$$

We can then rewrite our update as

$$\begin{aligned}\mathbf{x}_{t+1} &= \nabla\psi^*(\mathbf{z}_{t+1}) \\ &= (\nabla\psi + N_{\mathcal{X}})^{-1}(\nabla\psi(\mathbf{x}_t) - \eta_t \mathbf{g}_t) \\ &= \Pi_{\mathcal{X}}(\mathbf{y}_{t+1}),\end{aligned}$$

where $\mathbf{y}_{t+1}$ is the point with $\nabla\psi(\mathbf{y}_{t+1}) = \nabla\psi(\mathbf{x}_t) - \eta_t \mathbf{g}_t$. This is exactly the interpretation of our first view as showed in (12).