1 Outline

In this lecture, we cover

- online mirror descent,
- bandit convex optimization.

2 Online Mirror Descent

We learned the online gradient descent (OGD) algorithm that proceeds with the update rule

$$x_{t+1} = \mathrm{proj}_C \{x_t - \eta_t g_t\}$$

where g_t is a subgradient of f_t at x_t. Note that

$$\|x - x_t + \eta_t g_t\|_2^2 = 2\eta_t \left(g_t^\top x + \frac{1}{2\eta_t} \|x - x_t\|_2^2 \right) - 2\eta_t g_t^\top x_t + \eta_t^2 \|g_t\|_2^2.$$

This implies that

$$\begin{aligned} x_{t+1} &= \mathrm{proj}_C \{x_t - \eta_t g_t\} \\ &= \mathrm{argmin}_{x \in C} \|x - x_t + \eta_t g_t\|_2^2 \\ &= \mathrm{argmin}_{x \in C} \left\{ g_t^\top x + \frac{1}{2\eta_t} \|x - x_t\|_2^2 \right\}. \end{aligned}$$

Recall that $f_t(x_t) + g_t^\top x$ is a first-order approximation of f_t at x_t. Moreover, the quadratic term $\|x - x_t\|_2^2 / 2\eta_t$ encourages to choose a solution nearby x_t. Hence, the choice of x_{t+1} is given by a tradeoff between minimizing the first-order approximation and choosing a solution nearby x_t. Here, the distance between x and x_t is measured by the ℓ_2 norm. Then the regret upper bound is given by

$$\frac{3}{2} LR\sqrt{T}$$

where L is a global upper bound on the Lipschitz constants of the loss functions and $R^2 = \sup_{x,y \in C} \|x - y\|_2^2$.

Let us consider the case where

- loss functions $f_1, \ldots, f_T$ are L-Lipschitz in the ℓ_1 norm,
- the domain C is given by

$$C = \left\{ x \in \mathbb{R}_+^d : \; x_1 + \cdots + x_d = \frac{R}{2} \right\}.$$

Note that

$$\sup_{x,y\in C} \|x-y\|_1^2 = R^2 \quad \text{and} \quad \sup_{x,y\in C} \|x-y\|_2^2 = \frac{1}{2}R^2.$$

As the loss functions are L-Lipschitz in the ℓ_1 norm, it follows that

$$\|\nabla f_t(x)\|_\infty \le L$$

for $x \in C$ and $t = 1, \dots, T$. Here, we have

$$\|\nabla f_t(x)\|_2 \le \sqrt{d}\|\nabla f_t(x)\|_\infty \le L\sqrt{d}$$

for $x \in C$ and $t = 1, \dots, T$. In fact, it can be that the Lipschitz constant of f_t in the ℓ_2 norm can blow up by a factor of $\sqrt{d}$. Then, online gradient descent may incur a regret of order

$$O(LR\sqrt{dT}).$$

Here, we have the additional factor of $\sqrt{d}$. Again, this is due to the observation that the upper bound on the Lipschitz constants of loss functions has become $L\sqrt{d}$, not L. Perhaps, measuring the Lipschitz constants of loss functions in the ℓ_2-norm is not the best idea.

Recall that the update rule of OGD is

$$x_{t+1} = \text{argmin}_{x\in C}\left\{g_t^\top x + \frac{1}{2\eta_t}\|x - x_t\|_2^2\right\}.$$

Here, the distance between x and x_t is captured by the ℓ_2 distance between them. Instead, we may consider the notion of **Bregman divergence**. To define it, we take a strongly convex function ψ with respect to a norm $\|\cdot\|$. Then the Bregman divergence of p and q with respect to ψ is given by

$$D_\psi(p,q) = \psi(p) - \psi(q) - \nabla\psi(q)^\top(p-q).$$

Example 22.1. Note that

$$\psi(x) = \frac{1}{2}\|x\|_2^2$$

is strongly convex in the ℓ_2 norm, and the corresponding Bregman divergence is given by

$$D_\psi(p,q) = \frac{1}{2}\|p-q\|_2^2.$$

Example 22.2. Consider

$$\psi(x) = \sum_{i=1}^d x_i \log x_i,$$

which is strongly convex over

$$C = \{x \in \mathbb{R}_+^d : \ \|x\|_1 = 1\}$$

in the ℓ_1 norm. Moreover, the corresponding Bregman divergence is given by

$$D_\psi(p,q) = \sum_{i=1}^d p_i \log\frac{p_i}{q_i} = KL(p,q)$$

for $p, q \in C$. **Pinsker's inequality** states that

$$KL(p,q) \ge \frac{1}{2}\|p-q\|_1^2.$$

The **Online Mirror Descent (OMD)** algorithm runs with the update rule

$$x_{t+1} = \operatorname{argmin}_{x \in C} \left\{ g_t^\top x + \frac{1}{\eta_t} D_\psi(x, x_t) \right\}.$$

Theorem 22.3. *Let $f_1, \dots, f_T$ be an arbitrary sequence of convex loss functions that are L-Lipschitz in a norm $\|\cdot\|$. Assume that the Bregman divergence D_ψ satisfies*

$$D_\psi(x, y) \geq \frac{1}{2}\|x - y\|^2$$

for any $x, y \in C$. Moreover, $R^2 = \sup_{x,y \in C} D_\psi(x, y)$. Then online mirror descent with step sizes $\eta_t = R/(L\sqrt{t})$ guarantees that

$$\sum_{t=1}^{T} f_t(x_t) - \min_{x \in C} \sum_{t=1}^{T} f_t(x) = O\left(LR\sqrt{T}\right).$$

3 Bandit Convex Optimization

So far, we have learned online convex optimization (OCO) and its applications in sequential decision making. Platforms like streaming services, e-commerce websites, or news aggregators use OCO to recommend items (movies, products, articles) to users. In medical studies, one may apply OCO to assign treatments to patients with the goal of identifying the most effective ones. Moreover, retailers and service providers adjust prices in real-time to maximize revenue or market share.

Recall that the OCO framework considers a sequence of loss functions $f_1, \dots, f_T$ and aims to sequantially generate solutions $x_1, \dots, x_T$ so that the regret

$$\sum_{t=1}^{T} f_t(x_t) - \min_{x \in C} \sum_{t=1}^{T} f_t(x)$$

can be minimized. We learned that online gradient descent (OGD) and online mirror descent (OMD) guarantee a sublinear regret. Here, OMD as well as OGD proceeds with the update rule

$$x_{t+1} = \operatorname{argmin}_{x \in C} \left\{ g_t^\top x + \frac{1}{\eta_t} D_\psi(x, x_t) \right\}$$

where g_t is the gradient $\nabla f_t(x_t)$ of f_t at x_t. Hence, solving the OCO framework with the algorithms requires knowledge of the gradient.

In some real-world applications, however, it may not be feasible to assume full knowledge of the gradient. For recommender systems, the system learns from user interactions (e.g., clicks, purchases) to improve recommendations, but it only receives feedback on the items shown to the user. For clinical trials, feedback is limited to the outcomes observed in the patients who received the treatments. Furthermore, for dynamic pricing, the feedback comes from customer purchases at specific price points, so a pricing strategy is found without full knowledge of the demand curve. In these settings, the feedback is typically just the loss corresponding to the chosen action.

The **bandit feedback** means that for a chosen decision x_t, the decision-maker receives its associated loss $f_t(x_t)$. The bandit feedback is a **zeroth-order** feedback, whereas the gradient information is a **first-order** feedback. To construct the gradient $\nabla f_t(x_t)$ at x_t, we require knowledge of the loss function f_t, which is not feasible in many applications such as recommender systems, medical

trials, and dynamic pricing. For these scenarios, the decision-maker is asked to generate a sequence of decisions based on the bandit feedback.

Bandit Convex Optimization (BCO) is basically an extension of OCO where the decision-maker generates a sequence of decisions based on the history of bandit feedback. To be more specific, the decision-maker prepares a decision x_{t+1} for the $(t+1)$th time slot based on the zeroth-order information $f_1(x_1), \ldots, f_t(x_t)$ up to the first t time slots. Then the performance is measured based on the regret as in the OCO framework. As BCO works with more limited information than OCO, one would expect a worse regret for BCO. In this lecture, we will cover two algorithmic frameworks; the first algorithm guarantees a regret of $O(T^{3/4})$, and the second one guarantees a regret of $O(\sqrt{T})$ while it requries more information than the first one.