DS 801: Advanced Optimization for Data Science **KAIST, Fall 2024**
Lecture #21: Bandit Convex Optimization May 27, 2024
Lecturer: Dabeen Lee

1 Outline

In this lecture, we cover

- introduction to bandit convex optimization,
- the algorithm by Flaxman, Kalai, and McMahan,
- an algorithm by two-point feedback.

2 Bandit Convex Optimization

In the last lecture, we 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 $f_1, \ldots, f_T$ and aims to sequantially generate solutions $x_1, \ldots, x_T$ 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.

3 Algorithm by Flaxman, Kalai, and McMahan

In this section, we discuss an algorithm by Flaxman et al. [FKM05] for bandit convex optimization. Before we present the algorithm, let us provide some basic intuition behind it. The idea is basically to construct an estimator of the gradient based on bandit feedback. To elaborate, we consider a univariate function $f : \mathbb{R} \to \mathbb{R}$, and we want an estimator of the derivative $f'(x)$ at a point $x \in \mathbb{R}$. Then we take u from $\{-1, 1\}$ uniformly at random and evaluate $f(x + \delta u)$ for some $\delta > 0$. We claim that

$$\frac{1}{\delta} f(x + \delta u) u$$

is an approximation of $f'(x)$. To see this,

$$\mathbb{E}\left[\frac{1}{\delta} f(x + \delta u) u\right] = \frac{f(x+\delta) - f(x-\delta)}{2\delta},$$

which converges to $f'(x)$ as $\delta \to 0$. One may generalize this idea to the d-dimensional case as follows.

Let $\mathbb{B}$ and $\mathbb{S}$ be the unit ball and the unit sphere in $\mathbb{R}^d$ defined as

$$\mathbb{B} = \left\{x \in \mathbb{R}^d : \|x\|_2 \leq 1\right\} \quad \text{and} \quad \mathbb{S} = \left\{x \in \mathbb{R}^d : \|x\|_2 = 1\right\}.$$

For a function $f : \mathbb{R}^d \to \mathbb{R}$ and a point $x \in \mathbb{R}^d$, the procedure to deduce an estimator of the gradient $\nabla f(x)$ is as follows.

1. Sample a vector u from $\mathbb{S} = \left\{x \in \mathbb{R}^d : \|x\|_2 = 1\right\}$ uniformly at random.
2. Evaluate $f(x + \delta u)$.
3. Take vector g given by
$$g = \frac{d}{\delta} f(x + \delta u) u.$$

Then we consider

$$\hat{f}^\delta(x) = \mathbb{E}_{v \sim \mathbb{B}}\left[f(x + \delta v)\right]$$

where the expectation is taken over the random vector v sampled from the unit ball $\mathbb{B}$ uniformly at random. We will see that $\hat{f}_\delta(x)$ is an approximation of f.

Lemma 21.1. *If f is L-Lipschitz in the ℓ_2-norm,*

$$\left|f(x) - \hat{f}^\delta(x)\right| \leq \delta L.$$

Proof. Note that

$$\left|f(x) - \hat{f}^\delta(x)\right| = \mathbb{E}_{v\sim\mathbb{B}}\left[|f(x) - f(x+\delta v)|\right] \leq \mathbb{E}_{v\sim\mathbb{B}}\left[L\|\delta v\|_2\right] \leq L\delta,$$

as required. □

In fact, we can argue that g is an unbiased estimator of $\nabla\hat{f}^\delta(x)$.

Lemma 21.2 ([FKM05]). *For any $\delta > 0$,*

$$\mathbb{E}_{u\sim\mathbb{S}}\left[\frac{d}{\delta}f(x+\delta u)u\right] = \nabla\hat{f}^\delta(x)$$

where the expectation is taken over the random vector u sampled from the unit sphere $\mathbb{S}$ uniformly at random.

Hence, for any u sampled from $\mathbb{S}$ uniformly at random,

$$g = \frac{d}{\delta}f(x+\delta u)u$$

is an unbiased estimator of $\nabla\hat{f}^\delta(x)$. Given this, we are ready to explain the algorithm of Flaxman et al. that guarantees a sublinear regret.

As before, we consider the sequential optimization setting where we receive a sequence of loss functions $f_1,\ldots,f_T$. To simplify our presentation, we assume the following.

Assumption 21.3. The domain C contains the origin, i.e., $0 \in C$.

This assumption is without loss of generality because we may translate the coordinate space by a point x_1 in C.

Assumption 21.4. The diameter of the domain C is bounded above by R for so, i.e., $\sup_{x,y\in C}\|x-y\|_2 \leq R$ for some $R \geq 1$.

This assumption holds if C is bounded. In particular, one may take $R = \max\{1, \sup_{x,y\in C}\|x-y\|_2\}$. The last component is to define

$$(1-\delta)C := \left\{x \in \mathbb{R}^d : \frac{1}{1-\delta}x \in C\right\}.$$

Then the algorithm is as follows. As we play $x_t + \delta u_t$, not x_t, the regret of Algorithm 1 is given by

Algorithm 1 FKM Bandit Gradient Descent

```
Initialize $x_1 = 0$.
for $t = 1,\ldots,T$ do
    Sample $u_t$ from the unit sphere $\mathbb{S} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ uniformly at random.
    Evaluate $f_t(x_t + \delta u_t)$.
    Construct $g_t = (d/\delta)f_t(x_t + \delta u_t)u_t$.
    Obtain $x_{t+1} = \mathrm{proj}_{(1-\delta)C}\{x_t - \eta g_t\}$ for a step size $\eta > 0$.
end for
```

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

Another difference compared to online gradient descent is that we project solutions to $(1-\delta)C$, not to C.

Theorem 21.5. *Let $f_1, \ldots, f_T : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz convex loss functions. Setting*

$$\eta = \frac{R}{d} T^{-3/4} \quad \textit{and} \quad \delta = T^{-1/4},$$

the expected regret is bounded as

$$\mathbb{E}\left[\text{Regret}\right] = \mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t + \delta u_t) - \min_{x\in C}\sum_{t=1}^{T} f_t(x)\right] = O(LRdT^{3/4})$$

where the expectation is taken over the random vectors $u_1, \ldots, u_T$ sampled from the unit sphere $\mathbb{S}$ uniformly at random.

Proof. Let $x \in C$, and let $x_\delta = \text{proj}_{(1-\delta)C}(x)$. Since $(1-\delta)x \in (1-\delta)C$, the choice of x_δ states that

$$\|x_\delta - x\|_2 \le \|(1-\delta)x - x\|_2 = \delta\|x\|_2 = \delta\|x - 0\|_2 \le \delta R$$

where the last inequality holds because the diameter of C is bounded above by R and $0, x \in C$. Since each f_t is L-Lipschitz, it follows that

$$f_t(x_\delta) - f_t(x) \le L\|x_\delta - x\|_2 \le \delta LR.$$

Moreover,

$$f_t(x_t + \delta u_t) - f_t(x_t) \le L\|\delta u_t\|_2 = \delta L \le \delta LR$$

where the last inequality holds because $R \ge 1$. This implies that

$$\text{Regret} = \sum_{t=1}^{T} f_t(x_t + \delta u_t) - \sum_{t=1}^{T} f_t(x) \le \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x_\delta) + 2\delta LRT.$$

Next, we bound the right-hand side. Note that for any $x \in C$,

$$\left| f_t(x) - \hat{f}_t^\delta(x) \right| = \left|\mathbb{E}_{v\sim\mathbb{B}}\left[f_t(x) - f_t(x+\delta v)\right]\right| \le \mathbb{E}_{v\sim\mathbb{B}}\left[L\|\delta v\|_2\right] \le \delta LR.$$

This implies that

$$\text{Regret} \le \sum_{t=1}^{T} \hat{f}_t^\delta(x_t) - \sum_{t=1}^{T} \hat{f}_t^\delta(x_\delta) + 4\delta LRT.$$

Note that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \hat{f}_t^\delta(x_t) - \sum_{t=1}^{T} \hat{f}_t^\delta(x_\delta)\right] &\le \mathbb{E}\left[\sum_{t=1}^{T} \nabla \hat{f}_t^{\delta\top}(x_t - x_\delta)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} g_t^\top(x_t - x_\delta)\right] \\
&\le \mathbb{E}\left[\frac{R^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|_2^2\right]
\end{aligned}$$

where the first inequality holds due to the convexity of the loss functions, the equality holds because g_t is an unbiased estimator of $\nabla \hat{f}_t^\delta$ by Lemma 21.2, and the second inequality holds because Algorithm 1 works in the same as online gradient descent over domain $(1-\delta)C$ with linear functions $g_t^\top x$ for $t \in [T]$. Note that

$$\|g_t\|_2^2 = \frac{d^2}{\delta^2}(f_t(x_t+\delta u_t) - f_t(0) + f_t(0))^2\|u_t\|_2^2 \leq \frac{d^2}{\delta^2}(L((1-\delta)R+\delta) + f_t(0))^2 = \frac{d^2}{\delta^2}(LR + f_t(0))^2.$$

Hence, we have that

$$\mathbb{E}\left[\frac{R^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \|g_t\|_2^2\right] \leq \frac{R^2}{2\eta} + \frac{\eta T}{2\delta^2}d^2 \max_{t\in[T]}(LR + f_t(0))^2.$$

Therefore,

$$\mathbb{E}\left[\text{Regret}\right] \leq 4\delta LRT + \frac{R^2}{2\eta} + \frac{\eta T}{2\delta^2}d^2 \max_{t\in[T]}(LR + f_t(0))^2.$$

Setting

$$\eta = \frac{R}{d}T^{-3/4} \quad \text{and} \quad \delta = T^{-1/4},$$

we deduce that

$$\mathbb{E}\left[\text{Regret}\right] = O(dLRT^{3/4}),$$

as required. □

4 Algorithm with Two-Point Feedback

In the previous section, we explained an algorithm that guarantees a regret upper bound of $O(T^{3/4})$. Hence, there is still a gap from the lower bound of $\Omega(\sqrt{T})$. Recall that Algorithm 1 tests a single point at each iteration. Indeed, this might be too restrictive. Instead of the single-point feedback regime, [ADX10] proposed a relaxed setting where one can evaluate multiple points for an iteration while the number of tests is still less than or equal to a fixed constant.

In this section, we explain a framework due to Shamir [Sha17]. The framework provides a gradient estimation scheme based on two points. For a function $f : \mathbb{R}^d \to \mathbb{R}$ and a point $x \in \mathbb{R}^d$, we consider the following procedure.

1. Sample a vector u from the unit sphere $\mathbb{S} = \left\{x \in \mathbb{R}^d : \|x\|_2 = 1\right\}$ uniformly at random.
2. Evaluate $f(x+\delta u)$ and $f(x-\delta u)$.
3. Take vector g given by
$$g = \frac{d}{2\delta}(f(x+\delta u) - f(x-\delta u))u.$$

Here, we also define $\hat{f}^\delta$ as before:

$$\hat{f}^\delta(x) = \mathbb{E}_{v\sim\mathbb{B}}\left[f(x+\delta v)\right].$$

Lemma 21.6 ([Sha17])**.** *For any $\delta > 0$,*

$$\mathbb{E}_{u\sim\mathbb{S}}\left[\frac{d}{2\delta}(f(x+\delta u) - f(x-\delta u))u\right] = \nabla \hat{f}^\delta(x)$$

where the expectation is taken over the random vector u sampled from the unit sphere $\mathbb{S}$ uniformly at random.

Proof. Note that

$$\frac{d}{2\delta}(f(x+\delta u)-f(x-\delta u))u = \frac{d}{2\delta}f(x+\delta u)u - \frac{d}{2\delta}f(x-\delta u)u.$$

Then, the assertion follows from Lemma 21.2. □

Moreover, we have the following fact.

Lemma 21.7 ([Sha17])**.** *Let*

$$\delta = R\sqrt{\frac{2d}{T}},$$

and let f be a L-Lipschitz continuous convex function. Then there exists some constant $\kappa > 0$ such that for all x,

$$\mathbb{E}_{u\sim\mathbb{S}}\left[\left\|\frac{d}{2\delta}(f(x+\delta u)-f(x-\delta u))u\right\|^2\right] \le \kappa d L^2.$$

Algorithm 2 Algorithm with two-point feedback

```
Initialize $x_1 = 0$.
for $t = 1, \ldots, T$ do
    Sample $u_t$ from the unit sphere $\mathbb{S} = \left\{x \in \mathbb{R}^d : \|x\|_2 = 1\right\}$ uniformly at random.
    Evaluate $f_t(x_t + \delta u_t)$ and $f_t(x_t - \delta u_t)$
    Construct $g_t = (d/2\delta)\,(f_t(x_t + \delta u_t) - f_t(x_t - \delta u_t))\,u_t$.
    Obtain $x_{t+1} = \mathrm{proj}_C\,\{x_t - \eta g_t\}$ for a step size $\eta > 0$.
end for
```

Theorem 21.8 ([Sha17])**.** *Let $f_1, \ldots, f_T : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz convex loss functions. Setting*

$$\eta = \frac{R}{L\sqrt{dT}} \quad \text{and} \quad \delta = R\sqrt{\frac{2d}{T}},$$

the expected regret is bounded as

$$\mathbb{E}\left[\text{Regret}\right] = \mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t+\delta u_t) - \min_{x\in C}\sum_{t=1}^{T} f_t(x)\right] = O(LR\sqrt{dT})$$

where the expectation is taken over the random vectors $u_1, \ldots, u_T$ sampled from the unit sphere $\mathbb{S}$ uniformly at random.

References

[ADX10] Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, June 2010. 4

[FKM05] Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, page 385–394, USA, 2005. Society for Industrial and Applied Mathematics. 3, 21.2

[Sha17] Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *J. Mach. Learn. Res.*, 18(1):1703–1713, jan 2017. 4, 21.6, 21.7, 21.8