Lecture 18

Instructor: Haipeng Luo

1 Bandit Convex Optimization

In this lecture we discuss the most general bandit problem: bandit convex optimization (BCO), which is basically the OCO setting with only bandit feedback. Specifically, for each $t = 1, \ldots, T$,

1. learner picks action $w_t \in \Omega \subset \mathbb{R}^d$ while simultaneously environment picks a convex loss function $f_t : \Omega \to [-1, 1]$;
2. learner suffers and observes $f_t(w_t)$.

Once again we assume that the environment is oblivious and aim to minimize expected regret:

$$\mathbb{E}[\mathcal{R}_T] = \mathbb{E}\left[\sum_{t=1}^{T} f_t(w_t)\right] - \sum_{t=1}^{T} f_t(w_\star).$$

where $w_\star = \operatorname{argmin}_{w\in\Omega} \sum_{t=1}^{T} f_t(w)$.

Recall that in the full information setting, there is no extra difficulty when f_t is convex compared to the case when f_t is linear, due to the so-called convexity trick: $f_t(w_t) - f_t(w_\star) \leq \nabla f_t(w_t)^\top (w_t - w_\star)$. In the bandit setting, however, this is no longer true because the only feedback is $f_t(w_t)$ while to apply a linear bandit algorithm one needs to observe $\nabla f_t(w_t)^\top w_t$. In fact, this is a very challenging problem and there are still many open problems unsolved.

The convexity trick above can still be helpful though. By now it is clear that one key technique to solve bandit problems is to come up with estimators. If we try to solve the problem directly, it appears that we need to construct an estimator for the function f_t given its value at only one point, which is very challenging. However, by the convexity trick it is clear that one only needs to construct an estimator $\widehat{g}_t$ for the gradient $\nabla f_t(w_t)$, which intuitively is much more manageable. Given such estimators, we can again execute FTRL

$$w_t = \operatorname*{argmin}_{w\in\Omega} \sum_{\tau=1}^{t-1} w^\top \widehat{g}_\tau + \frac{1}{\eta}\psi(w)$$

for some regularizer ψ and learning rate η. To construct these estimators, we need to make use of the following lemma:

Lemma 1. *Given a function f, an invertible matrix M and $\delta > 0$, define the smoothed version of f as $\widehat{f}(w) = \mathbb{E}_{b\sim\mathbb{B}^d}[f(w + \delta M b)]$ where b is a uniform sample of the d-dimensional unit ball $\mathbb{B}^d = \{b \in \mathbb{R}^d : \|b\|_2 \leq 1\}$. Then the following holds*

$$\nabla \widehat{f}(w) = \mathbb{E}_{s\sim\mathbb{S}^d}\left[\frac{d}{\delta} f(w + \delta M s) M^{-1} s\right] \tag{1}$$

where s is a uniform sample of the d-dimensional unit sphere $\mathbb{S}^d = \{s \in \mathbb{R}^d : \|s\|_2 = 1\}$.

We omit the proof here but one can simply verify this fact when $d = 1$ so that the unit ball is simply the segment $[-1, 1]$ and the unit sphere is simply two points -1 and 1. Indeed in this case, with F

being the antiderivative of f we have

$$\nabla \mathbb{E}_{b\sim\mathbb{B}^d}[f(w+\delta Mb)] = \frac{1}{2}\frac{d}{dw}\int_{-1}^{1} f(w+\delta Mb)db = \frac{1}{2\delta M}\frac{d}{dw}\left(F(w+\delta M) - F(w-\delta M)\right)$$
$$= \frac{1}{2\delta M}\left(f(w+\delta M) - f(w-\delta M)\right) = \mathbb{E}_{s\sim\mathbb{S}^d}\left[\frac{d}{\delta}f(w+\delta Ms)M^{-1}s\right].$$

This lemma directly implies a way to construct the gradient estimator $\widehat{g}_t$: draw a uniform sample s from the unit sphere, query the value of $f_t(w_t+\delta Ms)$ for some matrix M and δ by playing $\widetilde{w}_t = w_t + \delta Ms$, and then use $\widehat{g}_t = \frac{d}{\delta}f(w+\delta Ms)M^{-1}s$ as an unbiased estimator of the gradient of $\widehat{f}_t$.

Importantly, this is an unbiased estimator for the smoothed version of f_t but not f_t itself. This leads to one key difficulty in solving BCO using this approach: bias-variance tradeoff of the estimator, which is controlled by the parameter δ. When δ is close to 0, $\widehat{f}_t$ is very close to f_t but $\widehat{g}_t$ will have very large magnitude and large variance; on the other hand, when δ is large, the variance goes down while $\widehat{f}_t$ becomes very different from f_t. We will see how exactly one should tune δ later in the analysis.

The next step is to decide what M should be. The simplest choice is the identity. From a geometric viewpoint, this amounts to exploring the sphere centered at w_t with radius δ. Extra care needs to be taken to ensure that $\widetilde{w}_t$ is never outside the set Ω. Indeed, one of the first BCO algorithms [Flaxman et al., 2005] uses exactly this exploration scheme, together with $\psi(w) = \frac{1}{2}\|w\|_2^2$ so that FTRL is simply gradient descent.

However, as discussed last time, sampling uniformly in all directions might not be the best idea. We have seen that using a self-concordant barrier ψ together with a Dikin ellipsoid exploration scheme works well for linear bandit. Here we can in fact use the same idea (first proposed in [Saha and Tewari, 2011]). Specifically, at time t let $H_t = \nabla^2\psi(w_t)$ be the Hessian of ψ at w_t. If we let $M = H_t^{-\frac{1}{2}}$, then note that

$$\|\widetilde{w}_t - w_t\|_{w_t} = \delta\sqrt{s^\top H_t^{-\frac{1}{2}} H_t H_t^{-\frac{1}{2}} s} = \delta,$$

which means $\widetilde{w}_t$ is exactly on the surface of the Dikin ellipsoid $\mathcal{E}_\delta(w_t)$. Since $\mathcal{E}_1(w_t)$ is contained in Ω, we can safely choose any $\delta \in (0,1]$. See Algorithm 1 for the complete pseudocode.

2 Regret Analysis

We analyze the regret of Algorithm 1 in this section. Recall that with our choice of M, $\widehat{f}_t(w)$ is defined as $\mathbb{E}_{b\sim\mathbb{B}^d}\left[f_t\left(w+\delta H_t^{-\frac{1}{2}}b\right)\right]$. First note that FTRL guarantees a bound on the quantity $\mathbb{E}\left[\sum_{t=1}^T \widehat{f}_t(w_t) - \widehat{f}_t(u)\right] \le \mathbb{E}\left[\sum_{t=1}^T \nabla\widehat{f}_t(w_t)^\top(w_t-u)\right] = \mathbb{E}\left[\sum_{t=1}^T \widehat{g}_t^\top(w_t-u)\right]$ for any $u \in \Omega$. To connect this quantity to the actual regret $\mathbb{E}[\mathcal{R}_T]$, we decompose the regret into five terms and bound each of them separately:

$$\mathbb{E}[\mathcal{R}_T] = \underbrace{\mathbb{E}\left[\sum_{t=1}^T f_t(\widetilde{w}_t) - \widehat{f}_t(\widetilde{w}_t)\right]}_{A_1} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T \widehat{f}_t(\widetilde{w}_t) - \widehat{f}_t(w_t)\right]}_{A_2}$$
$$+ \underbrace{\mathbb{E}\left[\sum_{t=1}^T \widehat{f}_t(w_t) - \widehat{f}_t(u)\right]}_{A_3} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T \widehat{f}_t(u) - f_t(u)\right]}_{A_4} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T f_t(u) - f_t(w_\star)\right]}_{A_5},$$

First, A_1 is simply non-positive by Jensen's inequality:

$$\widehat{f}_t(\widetilde{w}_t) = \mathbb{E}_{b\sim\mathbb{B}^d}\left[f_t\left(\widetilde{w}_t + \delta H_t^{-\frac{1}{2}}b\right)\right] \ge f_t\left(\widetilde{w}_t + \delta H_t^{-\frac{1}{2}}\mathbb{E}_{b\sim\mathbb{B}^d}[b]\right) = f_t(\widetilde{w}_t)$$

Algorithm 1: Variant of SCRiBLe for BCO

Input: parameter $\delta \in (0, 1]$, learning rate $\eta > 0$, and a ν-self-concordant function ψ
for $t = 1, \ldots, T$ **do**
  compute $w_t = \operatorname{argmin}_{w \in \Omega} \sum_{\tau=1}^{t-1} w^\top \widehat{g}_\tau + \frac{1}{\eta}\psi(w)$
  compute Hessian $H_t = \nabla^2 \psi(w_t)$ and sample $s_t \in \mathbb{S}^d$ uniformly at random
  play $\widetilde{w}_t = w_t + \delta H_t^{-\frac{1}{2}} s_t$ and observe $f_t(\widetilde{w}_t)$
  construct estimator $\widehat{g}_t = \frac{d}{\delta} f_t(\widetilde{w}_t) H_t^{\frac{1}{2}} s_t$

Next we look at term A_3, which is bounded by $\mathbb{E}\left[\sum_{t=1}^T \widehat{g}_t^\top (w_t - u)\right]$ as mentioned. Using the analysis from SCRiBLe, it is clear that as long as $\eta \le \frac{1}{16\|\widehat{g}_t\|_{w_t}^\star}$,

$$A_3 \le \frac{\psi(u) - \psi(w_1)}{\eta} + 8\eta \sum_{t=1}^T \|\widehat{g}_t\|_{w_t}^{\star 2}.$$

Note that the dual norm term is bounded as

$$\|\widehat{g}_t\|_{w_t}^{\star 2} = \frac{d^2}{\delta^2} f_t(\widetilde{w}_t)^2 \left(s_t^T H^{\frac{1}{2}} H^{-1} H^{\frac{1}{2}} s_t\right) \le \frac{d^2}{\delta^2}.$$

Together with the discussions from last lecture which says u should be chosen as $\frac{1}{1+\epsilon} w_\star + \frac{\epsilon}{1+\epsilon} w_1$ for some $\epsilon > 0$ so that $\psi(u) - \psi(w_1)$ is bounded by $\nu \ln\left(\frac{1}{\epsilon} + 1\right)$, we have

$$A_3 \le \frac{\nu \ln\left(\frac{1}{\epsilon} + 1\right)}{\eta} + \frac{8\eta d^2 T}{\delta^2},$$

as long as $\eta \le \frac{\delta}{16d}$. For simplicity we assume that T is large enough so that $\eta = \frac{\delta}{d}\sqrt{\frac{\nu}{T}\ln\left(\frac{1}{\epsilon}+1\right)} \le \frac{\delta}{16d}$ and with this η we have $A_3 = \mathcal{O}\left(\frac{d}{\delta}\sqrt{T\nu \ln\left(\frac{1}{\epsilon}+1\right)}\right)$.

With this specific choice of u, we can also bound the term A_5 using Jensen's inequality:

$$f_t(u) - f_t(w_\star) \le \frac{1}{1+\epsilon} f_t(w_\star) + \frac{\epsilon}{1+\epsilon} f_t(w_1) - f_t(w_\star) \le \epsilon |f_t(w_1) - f_t(w_\star)| \le 2\epsilon,$$

and thus $A_5 \le 2T\epsilon$. We will simply pick $\epsilon = 1/T$ so that A_5 is negligible.

Finally, to bound A_2 and A_4, we will make one additional assumption on f_t (although even without any more assumptions one can still prove some weak bounds as shown in [Flaxman et al., 2005]). We will consider two different choices of the assumption, each of which leads to a different rate in the end.

2.1 Lipschitzness Assumption

Assume f_t's are L-Lipschitz, that is, for all $w, w' \in \Omega$, $|f_t(w) - f_t(w')| \le L\|w - w'\|_2$. Note that this implies that $\widehat{f}_t$ is L-Lipschitz too:

$$|\widehat{f}_t(w) - \widehat{f}_t(w')| = \left|\mathbb{E}_{b \in \mathbb{B}^d}\left[f_t\left(w + \delta H_t^{-\frac{1}{2}} b\right) - f_t\left(w' + \delta H_t^{-\frac{1}{2}} b\right)\right]\right| \le L\|w - w'\|_2.$$

Therefore, we have

$$\widehat{f}_t(\widetilde{w}_t) - \widehat{f}_t(w_t) \le \delta L \left\|H_t^{-\frac{1}{2}} s_t\right\|_2 = \delta L \left\|w_t + H_t^{-\frac{1}{2}} s_t - w_t\right\|_2 \le \delta L \max_{w, w' \in \Omega} \|w - w'\|_2$$

where the last inequality holds since $w_t + H_t^{-\frac{1}{2}} s_t \in \mathcal{E}_1(w_t)$ is indeed in Ω as discussed. Letting $D = \max_{w,w' \in \Omega} \|w - w'\|_2$ denote the diameter of Ω, we have $A_2 \le \delta LDT$.

Similarly, we also have

$$\widehat{f}_t(u) - f_t(u) = \mathbb{E}_{b \sim \mathbb{B}^d}\left[f_t\left(u + \delta H_t^{-\frac{1}{2}} b\right) - f_t(u)\right] \le \delta L \mathbb{E}_{b \sim \mathbb{B}^d}\left[\left\|H_t^{-\frac{1}{2}} b\right\|_2\right] \le \delta LD,$$

and thus A_4 is also bounded by δLDT. Putting everything together we have proven

$$\mathbb{E}[\mathcal{R}_T] = \mathcal{O}\left(\delta LDT + \frac{d}{\delta}\sqrt{T\nu \ln T}\right),$$

where one can clearly see the tradeoff between the bias (the first term) and the variance (the second term) controlled by δ. With the optimal tuning of δ (assuming again that T is large enough so that $\delta \leq 1$) this shows $\mathbb{E}[\mathcal{R}_T] = \widetilde{\mathcal{O}}\left(\sqrt{dLD}\nu^{\frac{1}{4}}T^{\frac{3}{4}}\right)$ where the notation $\widetilde{\mathcal{O}}$ hides the small $\ln T$ terms.

2.2 Smoothness Assumption

Next we forget about the Lipschitzness assumption and make a different assumption that f_t's are β-smooth, that is, for any $w, w' \in \Omega$, $f_t(w) - f_t(w') \leq \nabla f_t(w')^\top (w - w') + \frac{\beta}{2}\|w - w'\|_2^2$, which also means that the gradient of f_t exists and is β-Lipschitz. Similarly, we can show that $\widehat{f}_t$ is β-smooth too:

$$\begin{aligned}
\widehat{f}_t(w) - \widehat{f}_t(w') &= \mathbb{E}_{b\sim\mathbb{B}^d}\left[f_t\left(w + \delta H_t^{-\frac{1}{2}} b\right) - f_t\left(w' + \delta H_t^{-\frac{1}{2}} b\right)\right] \\
&\leq \mathbb{E}_{b\sim\mathbb{B}^d}\left[\nabla f_t\left(w' + \delta H_t^{-\frac{1}{2}} b\right)\right]^\top (w - w') + \frac{\beta}{2}\|w - w'\|_2^2 \\
&= \nabla \mathbb{E}_{b\sim\mathbb{B}^d}\left[f_t\left(w' + \delta H_t^{-\frac{1}{2}} b\right)\right]^\top (w - w') + \frac{\beta}{2}\|w - w'\|_2^2 \\
&= \nabla \widehat{f}_t(w')^\top (w - w') + \frac{\beta}{2}\|w - w'\|_2^2 .
\end{aligned}$$

With this fact, A_2 and A_4 can both be bounded by $\frac{1}{2}\beta\delta^2 D^2 T$ since

$$\mathbb{E}\left[\widehat{f}_t(\widetilde{w}_t) - \widehat{f}_t(w_t)\right] \leq \mathbb{E}\left[\delta \nabla \widehat{f}_t(w_t)^\top H_t^{-\frac{1}{2}} s_t + \frac{\beta\delta^2}{2}\left\|H_t^{-\frac{1}{2}} s_t\right\|_2^2\right] \leq \frac{\beta\delta^2 D^2}{2}$$

and

$$\begin{aligned}
\widehat{f}_t(u) - f_t(u) &= \mathbb{E}_{b\sim\mathbb{B}^d}\left[f_t\left(u + \delta H_t^{-\frac{1}{2}} b\right) - f_t(u)\right] \\
&\leq \mathbb{E}_{b\sim\mathbb{B}^d}\left[\delta \nabla f_t(u)^\top H_t^{-\frac{1}{2}} b + \frac{\beta\delta^2}{2}\left\|H_t^{-\frac{1}{2}} b\right\|_2^2\right] \leq \frac{\beta\delta^2 D^2}{2}.
\end{aligned}$$

Putting everything together we have proven $\mathbb{E}[\mathcal{R}_T] = \mathcal{O}\left(\beta\delta^2 D^2 T + \frac{d}{\delta}\sqrt{T\nu \ln T}\right)$. With the optimal tuning of δ this becomes $\mathbb{E}[\mathcal{R}_T] = \widetilde{\mathcal{O}}\left((\beta\nu)^{\frac{1}{3}}(TdD)^{\frac{2}{3}}\right)$, improving the dependence on T from $T^{\frac{3}{4}}$ to $T^{\frac{2}{3}}$ compared to the Lipschitz case. Also, note that linear functions are smooth with $\beta = 0$. Therefore if f_t's are linear we can simply set $\delta = 1$ and recover the SCRiBLe regret bound $\mathcal{O}\left(d\sqrt{T\nu \ln T}\right)$. The two algorithms are slightly different in terms of sampling scheme though.

It turns out that these are all suboptimal results and the optimal regret for this problem is still $\mathcal{O}(\sqrt{T})$ (ignoring dependence on other parameters). A polynomial time algorithm to achieve this regret was only discovered very recently [Bubeck et al., 2016]. However, the algorithm is very complicated and far from being practical. Obtaining simple and optimal algorithms (even with extra assumptions) is still an important open problem.

References

Sébastien Bubeck, Ronen Eldan, and Yin Tat Lee. Kernel-based methods for bandit convex optimization. *arXiv preprint arXiv:1607.03084*, 2016.

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*, 2005.

Ankan Saha and Ambuj Tewari. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *The 14th International Conference on Artificial Intelligence and Statistics*, 2011.