CSCI 659 Lecture 9

Spring 2026

Instructor: Haipeng Luo

1 Adversarial Bandit Linear/Convex Optimization

So far most of our discussions on adversarial bandit problems focused on a set of discrete actions. Put differently, the decision set of the learner is always a simplex. In this lecture, we go back to the general adversarial OCO setting with a general compact convex decision set $\Omega \subseteq \mathbb{R}^d$ and study bandit feedback in this setting. Concretely, consider the following adversarial Bandit Convex Optimization (BCO) problem: at each round $t = 1, \ldots, T$,

1. the learner decides an action $w_t \in \Omega$ while simultaneously the adversary decides a convex loss function $f_t : \Omega \to [-1, 1]$;
2. the learner suffers and observes (only) $f_t(w_t)$.

We again assume that the adversary is oblivious for simplicity. The learner's goal is to minimize her expected regret against the best fixed action: $\mathbb{E}[\mathcal{R}_T] = \mathbb{E}[\sum_{t=1}^T f_t(w_t)] - \sum_{t=1}^T f_t(w^\star)$ where $w^\star \in \operatorname{argmin}_{w\in\Omega} \sum_{t=1}^T f_t(w)$ and the expectation is with respect to the learner's internal randomness.

While in the full-information setting, we argue that considering only linear loss functions is without loss of generality via the linearization trick: $f_t(w_t) - f_t(w^\star) \leq \langle \nabla f_t(w_t), w_t - w^\star \rangle$, this is no longer true in the bandit setting since our observation is only $f_t(w_t)$, instead of $\langle \nabla f_t(w_t), w_t \rangle$. For this reason, BCO is an extremely challenging problem, even for the simpler case where all loss functions are drawn from a fixed distribution, in which case the problem is also known under many different names such as blackbox/zeroth-order/derivative-free optimization and has many applications in practice.

We will therefore start by considering the special case, Bandit Linear Optimization (BLO), where each loss function f_t is a linear function parameterized by a vector $\ell_t \in \mathbb{R}^d$, that is, $f_t(w) = \langle w, \ell_t \rangle$. This still captures many interesting and important applications. For example, consider the bandit version of the combinatorial problems discussed in Lecture 2 (also known as *combinatorial bandits*), where there is a set of combinatorial actions $A = \{a_1, \ldots, a_K\} \subseteq \{0, 1\}^d$ and picking action a at time t incurs loss $\langle a, \ell_t \rangle$ for some loss vector ℓ_t, which is also the only observation for the learner. We have discussed examples such as m-set where each action corresponds to picking exactly m out of d items (e.g. recommending m out of d products to the customer), or online shortest path where each action corresponds to picking one path of a given graph (e.g. deciding which route to commute to work each day). The bandit feedback fits particularly well for the online shortest path example since most often we only observe/record the total loss (travel time) of the selected path.

To solve combinatorial bandits using BLO, one can take Ω as the convex hull of A as we did in Lecture 2. The only extra subtlety is that after the BLO algorithm selects $w_t \in \Omega$, if w_t is not already one of the combinatorial actions, we need to sample $a_t \in A$ with expectation w_t (recall that w_t, being a point in the convex hull of A, exactly corresponds to a distribution over the elements in A). This makes the feedback to the learner $\langle a_t, \ell_t \rangle$, instead of $\langle w_t, \ell_t \rangle$ as the protocol of BLO specifies. As we will see, however, this will not be an issue for the algorithms we consider. This also makes MAB a special case of BLO with $A = \{e_1, \ldots, e_K\}$ (the set of standard basis vectors).

Compared to the stochastic linear bandit problem discussed in Lecture 7, the key difference here is that the parameter deciding the loss (i.e. ℓ_t) is changing over time arbitrarily, but the action set is fixed and we care about competing to the overall best fixed action, while in stochastic linear bandits we consider a fixed parameter θ, allow the action set to be changing over time arbitrarily, and compare to the best action at each time.

2 The Exp2 Algorithm for BLO

We start with an inefficient but optimal algorithm that operates over a discrete subset A of Ω of size K and in a sense treats BLO as a K-armed bandit problem with a linear structure. This subset A can be obtained by discretizing Ω so that any two points in A are $\frac{1}{T}$-close (say in terms of L_2 norm), in which case K is of order $\mathcal{O}(T^d)$ and the extra regret introduced by this discretization is only $\mathcal{O}(1)$. On the other hand, if Ω is itself already a convex hull of a discrete set, which is the case for combinational bandits for example, then we can directly take this set as A since in this case the best action $w^\star$ can always be selected from A (the minimum of a linear function over a polytope can always be achieved by one of its corners). Without loss of generality, we assume that A is full rank (since otherwise we can first project them onto a full-rank subspace with lower dimension).

If we simply treat this as a standard K-armed bandit problem, then the regret is $\mathcal{O}(\sqrt{TK})$, clearly unacceptable since K can be exponentially large as mentioned. The issue of this approach is that it completely ignores the linear structure of the losses. As the simplest example, if A contains two actions a and $2a$, then no matter what ℓ_t is, knowing one action's loss completely reveals the loss for the other. In other words, similar to the case for Exp4, bandit feedback here does not really mean only $1/K$ fraction of information is available. Instead, since there are at most d independent directions in $\mathbb{R}^d$, bandit feedback here should be intuitively viewed as having only $1/d$ fraction of information.

To make use of this structure, we will directly construct a loss estimator $\widehat{\ell}_t \in \mathbb{R}^d$ for ℓ_t, and then estimate the loss for each action $a \in A$ naturally as $\langle a, \widehat{\ell}_t\rangle$. Having these estimators for all actions, we use Hedge to come up with a distribution $p_{t+1} \in \Delta(A)$ based on $p_{t+1}(a) \propto \exp(-\eta \sum_{s\leq t}\langle a, \widehat{\ell}_s\rangle)$, and sample a_{t+1} from p_{t+1}. It remains to figure out how to construct $\widehat{\ell}_t$.

Unlike the stochastic linear bandit problem where ℓ_t is fixed and can be estimated by standard linear regression based on the past t observations, here, ℓ_t can be arbitrarily changing over time and we have only one sample $\langle a_t, \ell_t\rangle$. Thanks to the randomness in selecting a_t, however, it is in fact possible to do a "one-point regression" by imagining having K samples, each with probability $p_t(a)$. More specifically, we construct the estimator as:

$$\widehat{\ell}_t = V(p_t)^{-1} a_t a_t^\top \ell_t \quad \text{where} \quad V(p) = \sum_{a \in A} p(a) a a^\top = \mathbb{E}_{a\sim p}\left[a a^\top\right] \tag{1}$$

is the covariance matrix with respect to p. Note that 1) although ℓ_t appears in this formula, the dependence is only through $a_t^\top \ell_t$, a quantity that we indeed observe; and 2) $V(p_t)$ is indeed invertible since A is full rank and p_t, computed based on the exponential weight, has a full support. In fact, when $A = \{e_1, \ldots, e_K\}$, this exactly recovers the importance weighted estimator for MAB (verify it yourself). The following lemma shows that this estimator is not only unbiased, but also leads to a nice bound for the local-norm term of Hedge.

Lemma 1. *For any distribution $p_t \in \Delta(A)$ with a full support, let $\widehat{\ell}_t$ be the loss estimator defined in Eq. (1) where a_t is sampled from p_t. Then we have (expectations below are with respect to $a_t \sim p_t$)*

$$\mathbb{E}\left[\widehat{\ell}_t\right] = \ell_t \quad \textit{and} \quad \mathbb{E}\left[\sum_{a\in A} p_t(a)\left\langle a, \widehat{\ell}_t\right\rangle^2\right] \leq d.$$

Proof. Direct calculations show: $\mathbb{E}[\widehat{\ell}_t] = V(p_t)^{-1}\mathbb{E}\left[a_t a_t^\top\right]\ell_t = V(p_t)^{-1}V(p_t)\ell_t = \ell_t$, and

$$\mathbb{E}\left[\sum_{a\in A} p_t(a)(a^\top \widehat{\ell}_t)^2\right] = \sum_{a\in A} p_t(a)\mathbb{E}\left[(a_t^\top \ell_t)^2 a^\top V(p_t)^{-1} a_t a_t^\top V(p_t)^{-1} a\right]$$

$$\le \sum_{a\in A} p_t(a)a^\top V(p_t)^{-1}\mathbb{E}\left[a_t a_t^\top\right] V(p_t)^{-1}a = \sum_{a\in A} p_t(a)a^\top V(p_t)^{-1}a$$
$$= \left\langle \sum_{a\in A} p_t(a)aa^\top, V(p_t)^{-1} \right\rangle = \left\langle V(p_t), V(p_t)^{-1}\right\rangle = d,$$

where the inequality is by the fact $|a^\top \ell_t| \le 1$ for all a (coming from the assumption that the range of f_t is in $[-1,1]$ in the BCO problem description). □

Therefore, if we still have the following local-norm regret bound from Hedge: for any $a^\star$,

$$\sum_{t=1}^{T}\sum_{a\in A} p_t(a)\left\langle a, \widehat{\ell}_t\right\rangle - \sum_{t=1}^{T}\left\langle a^\star, \widehat{\ell}_t\right\rangle \le \frac{\ln K}{\eta} + \eta\sum_{t=1}^{T}\sum_{a\in A} p_t(a)\left\langle a, \widehat{\ell}_t\right\rangle^2, \tag{2}$$

then taking expectation on both sides would imply a regret bound of $\mathcal{O}(\sqrt{dT\ln K})$ after optimally tuning η, much better than the aforementioned $\mathcal{O}(\sqrt{TK})$ bound. However, there is one caveat: recall that Eq. (2) holds only when $\eta\langle a, \widehat{\ell}_t\rangle \ge -1$ holds for all a and t. This condition trivially holds for Exp3/Exp4, since in MAB or contextual bandits the loss estimators are always nonnegative. On the other hand, $\langle a, \widehat{\ell}_t\rangle = a^\top V(p_t)^{-1}a_t a_t^\top \ell_t$ can now be very negative and violate the condition if a is a direction that has small correlation with a_t with high probability (this is true even if we assume $\langle a, \ell_t\rangle \ge 0$ for all $a \in A$). Also recall that this is not just a technical requirement in the analysis, but is in fact related to the necessity of exploration as discussed in Lecture 6: Hedge with a loss estimator that is too negative will discourage exploration.

To address this issue, we modify the algorithm slightly and explicitly enforce a small amount of exploration. Specifically, let γ be the probability of performing explicit exploration and $q \in \Delta(A)$ be an exploration distribution over A to be specified later. We now redefined p_t as $(1-\gamma)p'_t + \gamma q$ where p'_t is the Hedge distribution with $p'_t(a) \propto \exp(-\eta\sum_{s<t}\langle a, \widehat{\ell}_s\rangle)$, and then sample a_t from p_t and construct estimator $\widehat{\ell}_t$ the same way as Eq. (1). The resulting algorithm is called by many names such as Exp2 (Expanded Exponential weight) or GeometricHedge [Dani et al., 2008, Cesa-Bianchi and Lugosi, 2012, Bubeck et al., 2012] and is summarized below.

Algorithm 1: Exp2

```
Input: learning rate $\eta > 0$, exploration probability $\gamma \in (0,1)$ and distribution $q \in \Delta(A)$
for $t = 1, \ldots, T$ do
    compute $p'_t \in \Delta(A)$ such that $p'_t(a) \propto \exp(-\eta\sum_{s<t}\langle a, \widehat{\ell}_s\rangle)$
    sample $a_t$ from $p_t = (1-\gamma)p'_t + \gamma q$
    observe $\langle a_t, \ell_t\rangle$ and construct $\widehat{\ell}_t = V(p_t)^{-1}a_t a_t^\top \ell_t$ where $V(p) = \sum_{a\in A} p(a)aa^\top$
```

The following lemma tells us what property we need from the exploration distribution q.

Lemma 2. *Let $G(q) = \max_{a\in A}\|a\|^2_{V(q)^{-1}}$. We have for any $a \in A$ and t, $|\langle a, \widehat{\ell}_t\rangle| \le \frac{G(q)}{\gamma}$ and thus $\eta|\langle a, \widehat{\ell}_t\rangle| \le 1$ as long as $\eta \le \frac{\gamma}{G(q)}$.*

Proof. By definition, we have

$$|\langle a, \widehat{\ell}_t\rangle| = |a^\top V(p_t)^{-1}a_t||a_t^\top \ell_t| \le |a^\top V(p_t)^{-1}a_t| = |a^\top V(p_t)^{-1/2}V(p_t)^{-1/2}a_t|$$
$$\le \sqrt{a^\top V(p_t)^{-1}a}\cdot\sqrt{a_t^\top V(p_t)^{-1}a_t} \le \max_{a\in A}\|a\|^2_{V(p_t)^{-1}} \le \max_{a\in A}\|a\|^2_{(\gamma V(q))^{-1}} = \frac{G(q)}{\gamma},$$

where the second inequality is by Cauchy-Schwarz inequality and the last inequality uses the fact $V(p_t) - \gamma V(q) = (1-\gamma)V(p'_t)$ is positive-semidefinite and thus so is $(\gamma V(q))^{-1} - V(p_t)^{-1}$. □

Note that the smaller the learning rate, the larger the term $\frac{\ln N}{\eta}$ in Eq. (2). This motivates us to pick q such that $G(q)$ is as large as possible. As a naive first attempt, if we simply let q to be the uniform distribution over A, then it can be verified that $G(q) \le K$, which is too large since it would eventually introduce poly(K)-dependence to the regret. This is because a uniform exploration is wasteful when actions all live in a d-dimensional space. Intuitively, it should be enough if we sufficiently explore some directions that span this d-dimensional space, making $G(q)$ order poly(d) instead of poly(K). It turns out that this is indeed always possible.

Lemma 3. *For any set A, we have $\min_{q\in\Delta(A)} G(q) = d$. Moreover, the minimizer (called the G-optimal design) is also a maximizer of the function $\ln\det V(q)$ (called the D-optimal design).*

For a complete proof, see e.g., Lattimore and Szepesvári [2020, Theorem 21.1]. Here, we only show one direction $\min_{q\in\Delta(A)} G(q) \le d$ that is already enough to tell us that a good exploration distribution always exists. Indeed, this can be shown via an application of the minimax theorem:

$$\begin{aligned}
\min_{q\in\Delta(A)} G(q) &= \min_{q\in\Delta(A)} \max_{a\in A} \|a\|^2_{V(q)^{-1}} \\
&= \min_{q\in\Delta(A)} \max_{p\in\Delta(A)} \mathbb{E}_{a\sim p}[\|a\|^2_{V(q)^{-1}}] \\
&= \max_{p\in\Delta(A)} \min_{q\in\Delta(A)} \mathbb{E}_{a\sim p}[\|a\|^2_{V(q)^{-1}}] \\
&\le \max_{p\in\Delta(A)} \mathbb{E}_{a\sim p}[\|a\|^2_{V(p)^{-1}}] \\
&= \max_{p\in\Delta(A)} \left\langle \mathbb{E}_{a\sim p}[aa^\top], V(p)^{-1} \right\rangle = d.
\end{aligned}$$

We note that the second statement of the lemma implies that there is a concrete way to find such a good exploration distribution, that is, by maximizing the concave function $\ln\det V(q)$, which can be done via standard optimization methods. In fact, it can also be shown that there always exists a solution with support size at most $d(d+1)/2$. Using such a good exploration distribution, we now conclude the regret bound of Exp2.

Theorem 1. *With $\eta \le \frac{\gamma}{d}$ and q being a G-optimal design of the action set A, Exp2 ensures $\mathbb{E}[\mathcal{R}_T] \le \frac{\ln K}{\eta} + 2\gamma T + \eta T d$. Thus, setting $\gamma = d\eta$ and $\eta = \sqrt{\frac{\ln K}{Td}}$ leads to $\mathbb{E}[\mathcal{R}_T] = \mathcal{O}\left(\sqrt{dT\ln K}\right)$.*

Proof. By Lemmas 2 and 3, we know that the condition $\eta\langle a, \widehat{\ell}_t\rangle \ge -1$ holds for all a and t, and thus we can apply the standard analysis of Hedge: for any $a^\star \in A$, we have

$$\sum_{t=1}^T \sum_{a\in A} p'_t(a) \left\langle a, \widehat{\ell}_t \right\rangle - \sum_{t=1}^T \left\langle a^\star, \widehat{\ell}_t \right\rangle \le \frac{\ln K}{\eta} + \eta \sum_{t=1}^T \sum_{a\in A} p'_t(a) \left\langle a, \widehat{\ell}_t \right\rangle^2 .$$

Plugging in $p'_t(a) = \frac{p_t(a) - \gamma q(a)}{1-\gamma}$, multiplying both sides by $1-\gamma$, and rearranging give

$$\begin{aligned}
&\sum_{t=1}^T \sum_{a\in A} p_t(a) \left\langle a, \widehat{\ell}_t \right\rangle - \sum_{t=1}^T \left\langle a^\star, \widehat{\ell}_t \right\rangle \\
&\le \frac{(1-\gamma)\ln K}{\eta} + \gamma \sum_{t=1}^T \sum_{a\in A} q(a) \left\langle a, \widehat{\ell}_t \right\rangle - \gamma \sum_{t=1}^T \left\langle a^\star, \widehat{\ell}_t \right\rangle + \eta \sum_{t=1}^T \sum_{a\in A} (p_t(a) - \gamma q(a)) \left\langle a, \widehat{\ell}_t \right\rangle^2 \\
&\le \frac{\ln K}{\eta} + \gamma \sum_{t=1}^T \sum_{a\in A} q(a) \left\langle a, \widehat{\ell}_t \right\rangle - \gamma \sum_{t=1}^T \left\langle a^\star, \widehat{\ell}_t \right\rangle + \eta \sum_{t=1}^T \sum_{a\in A} p_t(a) \left\langle a, \widehat{\ell}_t \right\rangle^2 .
\end{aligned}$$

Taking expectation on both sides and using Lemma 1 and the fact $|\langle a, \ell_t\rangle| \le 1$, we arrive at

$$\mathbb{E}[\mathcal{R}_T] = \mathbb{E}\left[\sum_{t=1}^T \langle a_t, \ell_t\rangle\right] - \min_{a\in A} \sum_{t=1}^T \langle a, \ell_t\rangle \le \frac{\ln K}{\eta} + 2\gamma T + \eta T d,$$

proving the first statement. □

Going back to an earlier statement that any bounded action set Ω can be discretized into a finite action set A with $K = \mathcal{O}(T^d)$, we see that in this case Exp2 enjoys $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret, which is known to be near optimal [Dani et al., 2008].

3 The SCRiBLe Algorithm for BLO

While Exp2 achieves the optimal regret, it is computational inefficient since it explicitly maintains a distribution over potentially exponentially many actions. Similar to our discussion for combinatorial problems in Lecture 2, to obtain an efficient algorithm, we need to directly perform FTRL over the d-dimensional decision space Ω, that is, at time t compute $w_t = \operatorname{argmin}_{w\in\Omega}\langle w, \sum_{s<t}\widehat{\ell}_s\rangle + \frac{1}{\eta}\psi(w)$ for some loss estimators $\widehat{\ell}_1, \ldots, \widehat{\ell}_{t-1}$ and regularizer ψ. Note, however, that we cannot directly play w_t as the final decision, since randomness is required to construct the loss estimators as it should have become clear by now after seeing so many examples. Thus, another thing we need to figure out is how to randomly decide the final decision, denoted by $\widetilde{w}_t \in \Omega$, based on the FTRL solution w_t. These there elements (random decision, loss estimators, and regularizer) are all tied together closely, and there happens to be a delicate combination of the three that makes thing work.

First, having w_t, we will randomly explore a local region centered at w_t. One possibility of this local region is just a small L_2 ball, but this does not take into account the "shape" of the decision set Ω at all. For example, if w_t is very close to the boundary of Ω, then this ball needs to be very small, limiting the exploration in all directions. Directly considering the shape of Ω, an arbitrary convex set, is indeed highly challenging. Instead, we will somehow let the regularizer take care of this and explore over the surface of an ellipsoid defined with respect to the local behavior of the regularizer. Specifically, we play $\widetilde{w}_t = w_t + H_t^{-1/2}s_t$, where $H_t = \nabla^2\psi(w_t)$ (invertible as long as ψ is strictly convex) and s_t is uniformly at random sampled from the d-dimensional sphere, denoted by $\mathbb{S}^d$. If H_t is the identity matrix, then $\|\widetilde{w}_t - w_t\|_2 = 1$ and thus $\widetilde{w}_t$ is exactly a uniform sample from the surface of a unit L_2 ball centered at w_t. More generally, for a positive definite H_t, we have $\|\widetilde{w}_t - w_t\|_{H_t} = 1$ and thus $\widetilde{w}_t$ is a uniform sample from the surface of an ellipsoid centered at w_t. The eigenvectors of H_t define the principal axes of this ellipsoid and the corresponding eigenvalues are the reciprocals of the square of the semi-axes. It is clear that $\mathbb{E}_t[\widetilde{w}_t] = w_t$.

Of course, for this scheme to be valid, we need to make sure that $\widetilde{w}_t$ is indeed within Ω, an issue that we will come back later. Assuming its validity, after playing $\widetilde{w}_t$ and observing $\widetilde{w}_t^\top \ell_t$ we construct the loss estimator as $\widehat{\ell}_t = dH_t^{1/2}s_t\widetilde{w}_t^\top\ell_t$. This is in fact closely related to the estimator used in Exp2, since (expectation below is with respect to the randomness of s_t)

$$\left(\mathbb{E}\left[(\widetilde{w}_t - w_t)(\widetilde{w}_t - w_t)^\top\right]\right)^{-1}(\widetilde{w}_t - w_t) = \left(\mathbb{E}\left[H_t^{-1/2}s_ts_t^\top H_t^{-1/2}\right]\right)^{-1}H_t^{-1/2}s_t = dH_t^{1/2}s_t,$$

where we use the fact $\mathbb{E}[s_ts_t^\top] = \frac{1}{d}I_d$ (I_d is the d by d identity matrix). The lemma below shows that the estimator enjoys not only unbiasedness but also a small local norm.

Lemma 4. *The estimator defined above satisfies:* $\mathbb{E}[\widehat{\ell}_t] = \ell_t$ *and* $\|\widehat{\ell}_t\|_{H_t^{-1}} \le d$ *where the expectation is with respect to the randomness of* s_t.

Proof. By direct calculations and the facts $\mathbb{E}[s_t] = \mathbf{0}$ and $\mathbb{E}[s_ts_t^\top] = \frac{1}{d}I_d$, we have

$$\mathbb{E}\left[\widehat{\ell}_t\right] = \mathbb{E}\left[dH_t^{1/2}s_t\widetilde{w}_t^\top\ell_t\right] + \mathbb{E}\left[dH_t^{1/2}s_ts_t^\top H_t^{-1/2}\ell_t\right] = \ell_t,$$

and

$$\|\widehat{\ell}_t\|_{H_t^{-1}}^2 = \widehat{\ell}_t^\top H_t^{-1}\widehat{\ell}_t = d^2(\widetilde{w}_t^\top\ell_t)^2 s_t^\top H_t^{1/2}H_t^{-1}H_t^{1/2}s_t = d^2(\widetilde{w}_t^\top\ell_t)^2 \le d^2,$$

where in the last step we further use $s_t^\top s_t = 1$ and the assumption $|\widetilde{w}_t^\top\ell_t| \le 1$. □

We point out that, unlike all other local norm calculations we have seen, this one is bounded *always*, instead of only in expectation, and it holds for any strictly convex regularizer. As before, however, we still need to argue that the stability term of FTRL is indeed related to the local norm. This, together with the earlier issue on the validity of $\widetilde{w}_t$, can be simultaneously addressed by using a special type of regularizers called *self-concordant barriers*. Self-concordant barriers play a fundamental role in optimization theory (in particular, the Interior Point Method), and its (somewhat surprising) role for BLO was discovered by the seminal work by Abernethy et al. [2008], who proposed the following SCRiBLe (Self-Concordant Regularization in Bandit Learning) algorithm.[1]

[1] This version is slightly different from their original algorithm which samples s_t from the eigenbasis of H_t instead, but this makes no real difference to the regret analysis.

```
Algorithm 2: SCRiBLe
Input: learning rate $\eta > 0$ and a self-concordant barrier $\psi$ for $\Omega$
for $t = 1, \ldots, T$ do
    compute $w_t = \operatorname{argmin}_{w\in\Omega} \left\langle w, \sum_{s<t} \widehat{\ell}_s \right\rangle + \frac{1}{\eta}\psi(w)$
    sample $s_t \in \mathbb{S}^d$ uniformly at random and play $\widetilde{w}_t = w_t + H_t^{-1/2} s_t$ where $H_t = \nabla^2\psi(w_t)$
    observe $\widetilde{w}_t^\top \ell_t$ and construct estimator $\widehat{\ell}_t = dH_t^{1/2} s_t \widetilde{w}_t^\top \ell_t$
```

A barrier on Ω is a function that approaches $+\infty$ on the boundary of Ω. We defer the formal definition of self-concordance to the end of the discussion and first list the following useful facts (all can be found in [Nesterov and Nemirovskii, 1994]).

Fact 1. *If ψ is a self-concordant barrier on Ω, then for any w in the interior of Ω, the ellipsoid $\{\widetilde{w} : \|\widetilde{w} - w\|_{\nabla^2\psi(w)} \le 1\}$, called the Dikin ellipsoid centered at w, is contained by Ω.*

Since ψ is a barrier, the FTRL solution w_t is always in the interior of Ω. This tells us that $\widetilde{w}_t$, being on the surface of the Dikin ellipsoid center at w_t, is indeed a valid decision. Compared to simply exploring a small ball centered at w_t, the Dikin ellipsoid can make much better use of the space and explore more adaptively and aggressively.

Fact 2. *Let ψ be a self-concordant barrier on Ω and $w^\star$ be its minimizer. For any $w \in \Omega$, if its Newton decrement $\lambda_\psi(w)$, defined as $\|\nabla\psi(w)\|_{\nabla^{-2}\psi(w)}$, is at most $1/2$, then $\|w - w^\star\|_{\nabla^2\psi(w)} \le 2\lambda_\psi(w)$.*

This fact says that by looking at the Newton decrement of w, which is the local norm of the gradient of w, one can tell how far away w is from the minimizer $w^\star$ (as long as this Newton decrement is not too vacuously large). This fact helps us relate the stability of FTRL to the local norm of the loss estimator, as shown in the following lemma.

Lemma 5. *If $\eta \le \frac{1}{2d}$, SCRiBLe ensures $\langle w_t - w_{t+1}, \widehat{\ell}_t \rangle \le 2\eta\|\widehat{\ell}_t\|^2_{H_t^{-1}}$ for all t.*

Proof. By Hölder's inequality, we first bound $\langle w_t - w_{t+1}, \widehat{\ell}_t \rangle$ by $\|w_t - w_{t+1}\|_{H_t}\|\widehat{\ell}_t\|_{H_t^{-1}}$. Then, note that w_{t+1} is the minimizer of the function $F_t(w) = \eta\langle w, \sum_{s\le t} \widehat{\ell}_s\rangle + \psi(w)$, which is a self-concordant barrier (the linear terms does not affect the self-concordance coming from ψ, as it will become clear once we see the definition). To apply Fact 2, we calculate the Newton decrement:

$$\lambda_F(w_t) = \|\nabla F(w_t)\|_{\nabla^{-2}F(w_t)} = \left\|\eta\sum_{s\le t}\widehat{\ell}_s + \nabla\psi(w_t)\right\|_{H_t^{-1}} = \eta\|\widehat{\ell}_t\|_{H_t^{-1}}$$

where the last step uses the first-order condition: $\eta\sum_{s<t}\widehat{\ell}_s + \nabla\psi(w_t) = \mathbf{0}$, since w_t minimizes the barrier function $F_{t-1}(w) = \eta\langle w, \sum_{s<t}\widehat{\ell}_s\rangle + \psi(w)$. By Lemma 4 and the condition $\eta \le \frac{1}{2d}$, we know $\lambda_F(w_t) \le 1/2$ and thus Fact 2 implies $\|w_t - w_{t+1}\|_{H_t} \le 2\lambda_F(w_t) = 2\eta\|\widehat{\ell}_t\|_{H_t^{-1}}$, which finishes the proof. □

Note that the proof crucially relies on one fact mentioned earlier: the local-norm of the estimator is bounded always, not just in expectation. It remains to deal with the penalty term of FTRL, which is a bit trickier than what we have seen for other regularizers — in the past we have always bounded the penalty term by the range of the regularizer, but now the range of a barrier is by definition $+\infty$! To deal with this issue, we require an additional property from the regularizer, making it a so-called ν-self-concordant barrier for some parameter $\nu > 0$. We again defer the formal definition and first mention the following useful fact.

Fact 3. *If ψ is a ν-self-concordant barrier on Ω, then for any $\epsilon > 0$, we have $\psi(u) - \psi(w_1) \le \nu\ln(\frac{1}{\epsilon}+1)$ for any u from a shrunk (towards w_1) version of Ω defined as $\{\frac{1}{1+\epsilon}w + \frac{\epsilon}{1+\epsilon}w_1 : w \in \Omega\}$.*

Therefore, even though ψ has an infinite range on Ω, its range becomes only $\nu\ln(\frac{1}{\epsilon}+1)$ if one looks at a slightly shrunk version of it, since a ν-self-concordant barrier changes its value rapidly close to the boundary. Based on all these discussions, we are now ready to prove the following regret bound for SCRiBLe.

Theorem 2. *With a ν-self-concordant barrier regularizer and $\eta = \min\left\{\frac{1}{2d}, \sqrt{\frac{\nu \ln T}{Td^2}}\right\}$, SCRiBLe ensures $\mathbb{E}[\mathcal{R}_T] = \mathcal{O}(d\sqrt{\nu T \ln T} + d\nu \ln T)$.*

Proof. Based on Lemma 3 of Lecture 2, FTRL ensures for any $u \in \Omega$:

$$\sum_{t=1}^{T} \left\langle w_t - u, \widehat{\ell}_t \right\rangle \leq \frac{\psi(u) - \psi(w_1)}{\eta} + \sum_{t=1}^{T} \left\langle w_t - w_{t+1}, \widehat{\ell}_t \right\rangle.$$

Pick $u = \frac{1}{1+\epsilon}w^\star + \frac{\epsilon}{1+\epsilon}w_1$ for $\epsilon = 1/T$. Then in expectation the left-hand side is almost the regret of the learner due to the unbiasedness of the loss estimators:

$$\begin{aligned}\mathbb{E}_t[\langle w_t - u, \widehat{\ell}_t\rangle] &= \langle w_t - u, \ell_t\rangle = \mathbb{E}_t[\langle \widetilde{w}_t - u, \ell_t\rangle] = \mathbb{E}_t[\langle \widetilde{w}_t - w^\star, \ell_t\rangle] + \mathbb{E}_t[\langle w^\star - u, \ell_t\rangle] \\ &= \mathbb{E}_t[\langle \widetilde{w}_t - w^\star, \ell_t\rangle] + \frac{\epsilon}{1+\epsilon}\langle w^\star - w_1, \ell_t\rangle \geq \mathbb{E}_t[\langle \widetilde{w}_t - w^\star, \ell_t\rangle] - \frac{2}{T}.\end{aligned}$$

For the right-hand side, the penalty term is at most $\frac{\nu \ln(T+1)}{\eta}$ based on Fact 3, and the stability term is at most $2\eta d^2 T$ based on Lemmas 4 and 5. Combining everything shows $\mathbb{E}[\mathcal{R}_T] \leq 2 + \frac{\nu \ln(T+1)}{\eta} + 2\eta d^2 T$, and plugging in the (optimal) value of the learning rate finishes the proof. □

Finally, to get a sense of how good this regret bound is, we point out one last important fact.

Fact 4. *For any closed convex set in $\mathbb{R}^d$, there exists a ν-self-concordant barrier with $\nu = \mathcal{O}(d)$.*

Therefore, using such an $\mathcal{O}(d)$-self-concordant barrier, SCRiBLe achieves $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ regret, slightly worse than the $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret of Exp2. The advantage of SCRiBLe, however, is that it can be implemented efficiently for most problems that we care about, since it only requires solving a d-dimensional convex problem to find w_t. In fact, Abernethy et al. [2008] even showed that it suffices to do one Damped Newton step at each round to achieve the same regret, that is, instead of computing w_{t+1} as exactly the minimizer of $F_t(w) = \eta\langle w, \sum_{s\leq t}\widehat{\ell}_s\rangle + \psi(w)$, we do the following:

$$w_{t+1} = w_t - \frac{1}{1+\lambda_{F_t}(w_t)}\nabla^{-2}F_t(w_t)\nabla F_t(w_t).$$

Thus, the bottleneck is only in computing the Hessian inverse of F_t (or ψ equivalently).

Definition and Examples of Self-concordant Barriers. For completeness we now give the formal definition of ν-self-concordant barriers. First, consider the one dimensional case ($d = 1$). A function $\psi : \Omega \to \mathbb{R}$ is self-concordant if it is third-order differentiable, strictly convex, and satisfies the following Lipschitz Hessian condition: $|\psi(w)'''| \leq 2(\psi(w)'')^{3/2}$ for all w in the interior of Ω, and it is ν-self-concordant if in addition it satisfies the Lipschitz condition: $|\psi(w)'| \leq \sqrt{\nu\psi(w)''}$ again for all w in the interior of Ω. For the general d-dimensional case, ψ is ν-self-concordant if restricting it onto any direction gives a ν-self-concordant one-dimensional function.

These conditions say that both the Hessian and the function value move slowly relative to the movement of the gradient. Importantly, unlike common Lipschitz conditions (such as $|\psi(w)'| \leq C$ for some constant $C > 0$), these two conditions are both *affine-invariant*, meaning that if ψ satisfies them, then so does $\psi(Mw + u)$ for any affine transformation defined via M and u. Canonical examples include the following (try to verify them at least for $d = 1$ to convince yourself):

- $\psi(w) = -\sum_{i=1}^{d} \ln w_i$ (the log-barrier) is a d-self-concordant barrier for $\Omega = \mathbb{R}_+^d$;
- $\psi(w) = -\sum_{j=1}^{m} \ln(\alpha_j^\top w - \beta_j)$ is an m-self-concordant barrier for the polytope $\Omega = \{w \in \mathbb{R}^d : \alpha_j^\top w \geq \beta_j$ for $j = 1, \ldots, m\}$;
- $\psi(w) = -\ln(1 - \|w\|_2^2)$ is a 1-self-concordant barrier for the unit ball $\Omega = \{w \in \mathbb{R}^d : \|w\|_2 \leq 1\}$ (note that the self-concordant parameter here is 1 instead of d).

Question 1. *How would you use SCRiBLe to solve a combinatorial bandit problem? Think about what regularizer you will use and also what the final decision at each round you will play.*

4 One-Point Gradient Estimate for BCO

We now go back to the general BCO problem where the loss function f_t is not necessarily linear. As mentioned, the linearizion trick $f_t(w_t) - f_t(w^\star) \le \langle \nabla f_t(w_t), w_t - w^\star \rangle$ does not reduce BCO to BLO, but it is still useful since it suggests that we do not need to estimate the entire function f_t based on just one value $f_t(w_t)$, a problem that sounds extremely challenging, but instead only need to estimate one gradient $\nabla f_t(w_t)$ based on $f_t(w_t)$. Such estimate is called a one-point gradient estimate. While this is still highly non-trivial, one can at least estimate the gradient of a smoothed version of f_t based on the following lemma.

Lemma 6. *Given a function f and an invertible matrix M, define the smoothed version of f as $\widehat{f}(w) = \mathbb{E}_{b \sim \mathbb{B}^d}[f(w + Mb)]$ where b is a uniform sample of the d-dimensional unit ball $\mathbb{B}^d = \{b \in \mathbb{R}^d : \|b\|_2 \le 1\}$. Then the following holds*

$$\nabla \widehat{f}(w) = \mathbb{E}_{s \sim \mathbb{S}^d}\left[df(w + Ms)M^{-1}s \right] \tag{3}$$

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

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

This lemma directly implies a way to construct the gradient estimator $\widehat{\ell}_t$: draw a uniform sample s from the unit sphere, query the value of $f_t(w_t + Ms)$ for some M by playing $\widetilde{w}_t = w_t + Ms$, and then use $\widehat{\ell}_t = df(w + Ms)M^{-1}s$ as an unbiased estimator of the gradient $\nabla \widehat{f}_t(w_t)$ where $\widehat{f}_t(w) = \mathbb{E}_{b \sim \mathbb{B}^d}[f_t(w + Mb)]$ is a smoothed version of f_t. Inspired by SCRiBLe, we pick M based on the Hessian of a self-concordant regularizer together with an extra scaling parameter $\delta \in (0, 1]$, leading to the following algorithm proposed by [Saha and Tewari, 2011]. Note that when $\delta = 1$, this exactly recovers SCRiBLe.

Algorithm 3: A Generalization of SCRiBLe for BCO

```
Input: parameter $\delta \in (0,1]$, learning rate $\eta > 0$, and a $\nu$-self-concordant function $\psi$
for $t = 1, \ldots, T$ do
    compute $w_t = \mathrm{argmin}_{w \in \Omega} \left\langle w, \sum_{s<t} \widehat{\ell}_s \right\rangle + \frac{1}{\eta}\psi(w)$
    sample $s_t \in \mathbb{S}^d$ uniformly at random and play $\widetilde{w}_t = w_t + \delta H_t^{-1/2} s_t$ where $H_t = \nabla^2 \psi(w_t)$
    observe $f_t(\widetilde{w}_t)$ and construct gradient estimator $\widehat{\ell}_t = \frac{d}{\delta} f_t(\widetilde{w}_t) H_t^{\frac{1}{2}} s_t$
```

Note that $\widetilde{w}_t$ is a valid point as it is within the Dikin ellipsoid centered at w_t. Importantly, since $\widehat{\ell}_t$ is not exactly an unbiased estimator for f_t itself, this leads to one key issue in this approach: bias-variance trade-off 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 $\widehat{\ell}_t$ has a very large variance; on the other hand, when δ is large, the variance goes down while $\widehat{f}_t$ becomes very different from f_t. Due to this trade-off, this algorithm at best achieves $\widetilde{\mathcal{O}}(T^{3/4})$ regret for Lipschitz loss functions or $\widetilde{\mathcal{O}}(T^{2/3})$ regret for smooth loss functions; see [Luo, 2017] for the analysis. The best upper bound for this problem is $\widetilde{\mathcal{O}}(d^{2.5}\sqrt{T})$ [Fokkema et al., 2024]; while the best existing lower bound is (somewhat surprisingly) still $\Omega(d\sqrt{T})$ coming from the linear case [Dani et al., 2008]. Closing this gap with an efficient algorithm is still a key open problem in the bandit literature. See [Lattimore, 2024] for a thorough discussion on this topic.

References

Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *21st Annual Conference on Learning Theory*, 2008.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Sham Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *25th Annual Conference on Learning Theory*, 2012.

Nicolò Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

Varsha Dani, Sham M Kakade, and Thomas P Hayes. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems 21*, 2008.

Hidde Fokkema, Dirk van der Hoeven, Tor Lattimore, and Jack J Mayo. Online newton method for bandit convex optimisation. *Conference on Learning Theory*, 2024.

Tor Lattimore. Bandit convex optimisation. *arXiv preprint arXiv:2402.06535*, 2024.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Haipeng Luo. Lecture notes 18, introduction to online learning, 2017. URL `https://haipeng-luo.net/courses/CSCI699/lecture18.pdf`.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

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.