CSE599M **Spring 2021**

Data-Dependent Regret Bounds

Omid Sadeghi, Max Gray, Tanner Fiez

1 Introduction

In this scribe note, we provide an overview of recent work deriving data-dependent bounds for adversarial bandits. This line of work asks the question of whether improved regret bounds can be obtained when the loss sequence is not completely adversarial. In other words, while classical algorithms such as Exp3 obtain worst-case optimal regret of $\widetilde{O}(\sqrt{KT})$ where K is the number of actions and T is the time-horizon, this line of work explores algorithms that can adapt to easier loss sequences and provides regret bounds in terms of meaningful characteristics of the loss sequence. Deriving results of this nature requires novel algorithms and analysis methods. In particular, variants of optimistic mirror descent framework have emerged as the primary algorithmic method that can naturally adapt to easy data with careful choices of the parameters and the regularization function. We give a high level overview of these methods and highlight main results.

Problem Setup and Regret Definition. For the majority of this document, we consider the standard adversarial bandit problem with K arms and a time horizon of T. For reasons that become clear later on, we often represent any arm i by the i-th canonical basis vector e_i and the collection of these vectors as the set $\mathcal{X}$. The convex hull of $\mathcal{X}$ denoted by $\Omega = \text{conv}(\mathcal{X})$ corresponds to the standard simplex. In general, the losses at each time are assumed to be bounded. That is, $\ell_{t,i} \in [-1, 1]$ or $\ell_{t,i} \in [0, 1]$ for all times t and actions $i \in [K]$. The goal is to minimize the regret during the course of the algorithm. The regret is defined as

$$R(T) = \sum_{t=1}^{T} \ell_{t,i_t} - \sum_{t=1}^{T} \ell_{t,i^*}$$

where i_t denotes the arm selection by an algorithm at time t and the $i^* = \min_{i\in[K]} \sum_{t=1}^{T} \ell_{t,i}$ denotes the best fixed arm in hindsight. The purpose of the line of work reviewed in this document is to obtain expected regret bounds that replace the usual dependence on T with a data-dependent quantity.

Data-Dependent Quantities. Toward defining key data-dependent quantities, we present notation for problem and arm-dependent quantities. For any arm $i \in [K]$ and round $t \in [T]$, we denote the accumulated loss by $L_{t,i} := \sum_{s=1}^{t} \ell_{s,i}$, the mean loss by $\mu_{t,i} := \frac{L_{t,i}}{t}$, the unnormalized variance by $Q_{t,i} := \sum_{s=1}^{t} (\ell_{s,i} - \mu_{t,i})^2$, and the "first-order path length" $V_{t,i} := \sum_{s=1}^{t} |\ell_{s,i} - \ell_{s-1,i}|$. Existing data-dependent regret bounds for adversarial bandit problems replace the usual dependence on T with one of the following quantities:

- Variance (unnormalized) of the best arm (non-negative losses):

$$Q_{T,i^*} = \sum_{t=1}^{T} (\ell_{t,i^*} - \mu_{t,i^*})^2$$

- K times the path length of the best arm:

$$KV_{T,i^*} = K \sum_{t=1}^{T} |\ell_{t,i^*} - \ell_{t-1,i^*}|$$

- Sum of the path lengths of all arms:

$$\sum_{i=1}^{K} V_{T,i} = \sum_{i=1}^{K} \sum_{t=1}^{T} |\ell_{t,i} - \ell_{t-1,i}|$$

- Best-of-both worlds type of guarantee:

$$\min\left\{\sqrt{KL_{T,i^*}}, \frac{K}{\Delta}\right\} = \min\left\{\sqrt{K\sum_{t=1}^{T}\ell_{t,i^*}}, \frac{K}{\Delta}\right\}$$

  where the latter term is applicable when there is an arm whose expected loss is always smaller than that of all other arms by a fixed gap Δ.

Each of the data-dependent quantities can be $O(T)$ in the worst-case, but can potentially be much smaller. In general, the quantities improve over the usual worst-case dependence on T when the loss sequences (potentially that of the optimal arm) either have low variance or are slowly changing. In the next section, we present an algorithmic framework which dependent on the parameter choices can obtain regret guarantees for each of the quantities above. We will focus our presentation on the path-length regret bounds. Following that, we go into an improved path-length regret bound with slightly different algorithmic techniques. Finally, we cover application of these methods to the linear bandit problem in the final section.

2 BROAD-OMD Algorithm Framework

In this section, we present the BROAD-OMD algorithmic framework [4]. The BROAD-OMD framework is a general algorithmic framework based on optimistic mirror descent that can be tweaked to optimize for regret bounds in terms of particular data-dependent quantities.

The BROAD-OMD framework is based on a randomized algorithm that computes a distribution ω_t on the simplex $\Omega = \text{conv}(\mathcal{X})$ over the action set $\mathcal{X} = \{e_1, \ldots, e_K\}$ and then samples an action $i_t \sim \omega_t$ to play at the given round $t \in [T]$. Following the action selection the loss of the action taken i_t given by $\ell_{t,i}$ is observed and realized, while the losses of all other actions are not observed. This setup follows the usual approach for adversarial bandits, where the key distinctions will come from how the distribution ω_t is chosen at each time $t \in [T]$ and also how the losses of actions not taken are estimated which we denote by $\hat{\ell}_{t,i}$.

Toward fully introducing the BROAD-OMD framework, we need to recall the definition of the Bregman divergence with respect to a regularization function. Let Ψ be a convex function defined on a convex set Ω. The *Bregman divergence* between any $x, y \in \Omega$ with respect to Ψ is defined as

$$D_\Psi(x, y) = \Psi(x) - \Psi(y) - \langle \nabla\Psi(y), x - y\rangle.$$

A useful property of the Bregman divergence is the generalized Pythagorean theorem, which says for all $x, y, z \in \Omega$,

$$\langle x - z, \nabla\Psi(x) - \nabla\Psi(y)\rangle = D_\Psi(x, y) + D_\Psi(z, x) - D_\Psi(z, y).$$

The generic mirror descent algorithm computes the distribution over actions as follows

$$\omega_t = \arg\min_{\omega\in\Omega}\{\langle \omega, \hat{\ell}_{t-1}\rangle + D_\Psi(\omega, \omega_{t-1})\}$$

where $\hat{\ell}_{t-1}$ denotes the estimates of the last loss for each of the actions. In plain words, the mirror descent update selects a distribution over actions that minimizes the expected loss over the estimated last loss with the Bregman divergence term penalizing shifts from the last distribution over actions. Toward analyzing mirror descent type algorithms, near directly from first-order optimality conditions and the generalized Pythagorean theorem for Bregman divergence, the generic mirror descent update ensures that for all $u \in \Omega$,

$$\langle \omega_t - u, \hat{\ell}_{t-1}\rangle \leq D_\Psi(u, \omega_{t-1}) - D_\Psi(u, \omega_t) - D_\Psi(\omega_t, \omega_{t-1}).$$

This fact is important to the analysis of mirror descent algorithms, and also plays a key role in the analysis of the BROAD-OMD framework. The BROAD-OMD framework uses a variant of the usual mirror descent framework that is known as optimistic mirror descent. The optimistic mirror descent algorithm differs the usual mirror descent framework by maintaining two separate sequences $\{\omega_t\}$ and $\{\omega'_t\}$. The sequence $\{\omega_t\}$

Algorithm 1 BROAD-OMD

Parameters: $\Omega = \text{conv}(\mathcal{X})$ and $\Psi_t(\omega) = -\sum_{i=1}^{K} \eta_{t,i}^{-1} \ln \omega_i$.
Initialization: Set $\omega_1'(i) = \frac{1}{K}$ $\forall i \in [K]$.
for $t = 1, 2, \ldots, T$ **do**
  Set $\omega_t = \arg\min_{\omega \in \Omega} \{\langle \omega, m_t \rangle + D_{\Psi_t}(\omega, \omega_t')\}$.
  Draw arm $i_t \sim \omega_t$.
  Observe and suffer the loss ℓ_{t,i_t}.
  Construct $\hat{\ell}_t$ as an unbiased estimator of ℓ_t
  Let

$$a_{t,i} = \begin{cases} 6\eta_{i,t}\omega_{t,i}(\hat{\ell}_{t,i} - m_{t,i})^2 & \text{(Option I)} \\ 0 & \text{(Option II)} \end{cases}$$

  $\omega_{t+1}' = \arg\min_{\omega \in \Omega} \{\langle \omega, m_t \rangle + D_{\Psi_t}(\omega, \omega_t')\}$
end for

Option	$m_{t,i}$	$\hat{\ell}_{t,i}$	$\eta_{t,i}$	$\mathbb{E}[\text{Reg}_T]$ in $\tilde{\mathcal{O}}$
I	$\tilde{\mu}_{t-1,i}$	$\frac{(\ell_{t,i} - m_{t,i})\mathbb{1}\{i_t = i\}}{w_{t,i}} + m_{t,i}$	fixed	$\sqrt{KQ_{T,i^*}}$
I	$\ell_{\alpha_i(t),i}$	$\frac{(\ell_{t,i} - m_{t,i})\mathbb{1}\{i_t = i\}}{\tilde{w}_{t,i}} + m_{t,i}$	increasing	$K\sqrt{V_{T,i^*}}$
II	$\ell_{\alpha_i(t),i}$	$\frac{(\ell_{t,i} - m_{t,i})\mathbb{1}\{i_t = i\}}{w_{t,i}} + m_{t,i}$	fixed	$\sqrt{K\sum_{i=1}^{K} V_{T,i}}$
II	ℓ_{t,i_t}	$\frac{\ell_{t,i}\mathbb{1}\{i_t = i\}}{w_{t,i}}$	fixed	$\min\{\sqrt{KL_{T,i^*}}, \frac{K}{\Delta}\}$

Figure 1: Regret bounds obtained by the BROAD-OMD algorithm with particular parameter choices in the algorithm. Note that $\alpha_i(t)$ denotes the last time that the algorithm selected action $i \in [K]$ and $\tilde{\mu}_{t-1,i}$ is an unbiased estimate of $\mu_{t-1,i}$.

is selected at each time to minimize the expected loss over a prediction of the current loss and a penalty term, while the sequence $\{\omega_t'\}$ is selected at each time to minimize the expected loss over an estimate of the current loss (potentially with a correction term) and a penalty term. In particular, at the beginning of each round, optimistic mirror descent takes in a prediction m_t of the current loss vector, and computes the distribution over actions according to

$$\omega_t = \arg\min_{\omega \in \Omega} \{\langle \omega, m_t \rangle + D_{\Psi_t}(\omega, \omega_t')\}.$$

Then, after playing an action $i_t \sim \omega_t$, observing a loss ℓ_{t,i_t}, forming an estimated loss vector $\hat{\ell}_t$, taking in a correction term a_t, the auxillary sequence is updated according to the update

$$\omega_{t+1}' = \arg\min_{\omega \in \Omega} \{\langle \omega, \hat{\ell}_t + a_t \rangle + D_{\Psi_t}(\omega, \omega_t')\}.$$

Observe here that the regularization function in the Bregman divergence may be time-dependent.

The key choices in parameterizing the algorithm are the regularization function Ψ, and the methods to obtain the prediction m_t, the loss estimate $\hat{\ell}_t$, and the correction term a_t. These choices will be detailed shortly and they are informed by the following property of optimistic mirror descent. For optimistic mirror descent, if

$$\langle \omega_t - \omega_{t+1}', \hat{\ell}_t - m_t + a_t \rangle \leq \langle \omega_t, a_t \rangle,$$

then for all $u \in \Omega$,

$$\langle \omega_t - u, \hat{\ell}_t \rangle \leq D_{\Psi_t}(u, \omega_t') - D_{\Psi_t}(u, \omega_{t+1}') + \langle u, a_t \rangle - D_{\Psi_t}(\omega_{t+1}', \omega_t) - D_{\Psi_t}(\omega_{t+1}', \omega_t). \tag{1}$$

In the special case, $a_t = 0$, then for all $u \in \Omega$,

$$\langle \omega_t - u, \hat{\ell}_t \rangle \leq D_{\Psi_t}(u, \omega'_t) - D_{\Psi_t}(u, \omega'_{t+1}) + \langle \omega_t - \omega'_{t+1}, \hat{\ell}_t - m_t \rangle - D_{\Psi_t}(\omega'_{t+1}, \omega_t) - D_{\Psi_t}(\omega'_{t+1}, \omega_t).$$

This result by [4] is key to the analysis of optimistic mirror descent and leads to the basic path toward deriving regret bounds. In particular, summing this inequality over all time gives a bound on the estimate loss of the algorithm compared to any fixed comparator and then taking the expectation gives a bound on the expected regret. Typically, the terms $D_{\Psi_t}(u, \omega'_t) - D_{\Psi_t}(u, \omega'_{t+1})$ are not the dominating components and simply telescope with a time-independent regularizer. Moreover, the final terms $-D_{\Psi_t}(\omega'_{t+1}, \omega_t) - D_{\Psi_t}(\omega'_{t+1}, \omega_t)$ are negative and may or not be useful in the regret bound or could be simply dropped. Finally, controlling $\langle u, a_t \rangle$ or $\langle \omega_t - \omega'_{t+1}, \hat{\ell}_t - m_t \rangle$ is the key to obtaining desirable regret bounds.

The BROAD-OMD algorithmic framework is presented in Algorithm 1 along with the results obtained for various parameter choices in Figure 1. A key component of the BROAD-OMD framework is the choice of the log barrier regularization function. This regularization function may also be time-dependent. In particular, the log barrier regularizer for any $x \in \Omega$ is given by $\Psi_t(x) = -\sum_{i=1}^K \eta_{t,i}^{-1} \ln x_i$ and the Bregman divergence between $x, y \in \Omega$ is given by $D_{\Psi_t}(x, y) = -\sum_{i=1}^K \eta_{t,i}^{-1} h\left(\frac{u_i}{v_i}\right)$ where $h(y) = y - 1 - \ln y$. This regularization function is key since it makes it so that with proper choice of learning rate the optimistic mirror guarantee given in (1) holds with appropriate choice of correction term a_t.

3 Path Length Bound: BROAD-OMD with Option I

In this section, we outline how the regret path-length bound with Option I is obtained with the BROAD-OMD framework described in Algorithm 1. As described in Figure 1, this regret bound is chosen by selecting $m_{t,i} = \ell_{\alpha_i(t),i}$ and $\hat{\ell}_{t,i} = \frac{(\ell_{t,i} - m_{t,i}) 1\{i_t = i\}}{w_{t,i}} + m_{t,i}$. Thus the prediction is simply the last observed loss for each action, the estimator is chosen so it is unbiased. The choice of $a_{t,i}$ is more complicated and is based on the following reasoning. In the full information setting, it is typically $a_{t,i} = \eta_{t,i}(\ell_{t,i} - m_{t,i})^2$, but this is unknown with bandit feedback. We also cannot replace $\ell_{t,i}$ directly with $\hat{\ell}_{t,i}$ since then $(\ell_{t,i} - m_{t,i})^2$ can be as big as $1/\omega_{t,i}^2$. The fix for this is to select $a_{t,i} = 6\eta_{t,i}\omega_{t,i}(\hat{\ell}_{t,i} - m_{t,i})^2$ so that the extra $\omega_{t,i}$ term can cancel the term $1/\omega_{t,i}^2$ in expectation.

Then if we consider bounding the term $\sum_{t=1}^T \langle u, a_t \rangle$ that arises from (1) with a constant learning rate of $\eta_{t,i} = \eta$, it is possible to show that

$$\sum_{t=1}^T \langle u, a_t \rangle \leq 12\eta \max_t \omega_{t,i}^{-1} V_{T,i}.$$

Hence, $\sum_{t=1}^T \langle u, a_t \rangle$ is close to the desired path-length quantity, except for the leading $\max_t \omega_{t,i}^{-1}$. In [4] the fix proposed for this issue is an increasing learning rate schedule. In particular, the term $\sum_{t=1}^T D_{\Psi_t}(u, \omega'_t) - D_{\Psi_t}(u, \omega'_{t+1})$ that arises from (1) can be bounded by $\sum_{t=1}^T \sum_{i=1}^K (\frac{1}{\eta_{t+1,i}} - \frac{1}{\eta_{t,i}}) h(\frac{u_i}{\omega'_{t+1,i}})$ which is negative for an increasing learning rate. Moreover, $h(\frac{u_i}{\omega'_{t+1,i}})$ is close to $\frac{1}{\omega_{t+1,i}}$ if u_i close to 1. Thus, by increasing the learning rate when $\frac{1}{\omega_{t+1,i}}$ is big, then $\left(\frac{1}{\eta_{t+1,i}} - \frac{1}{\eta_{t,i}}\right) h\left(\frac{u_i}{\omega'_{t+1,i}}\right)$ is a negative in terms of $\frac{-1}{\omega_{t+1,i}}$ that compensates for the positive term coming from $\sum_{t=1}^T \langle u, a_t \rangle$ This idea results in expected regret bound that is $\tilde{\mathcal{O}}(K\sqrt{V_{T,i^*}})$. Full details can be found in Section 3 of [4].

4 Path Length Bound: BROAD-OMD with Option II

Compared to the BROAD-OMD algorithm with option I, BROAD-OMD algorithm with option II does not have the correction terms $\{a_t\}_{t=1}^T$ and they have been set equal to zero. Also, for each $i \in [K]$ and $t \in [T]$, the optimistic prediction $m_{t,i} = \ell_{\alpha_i(t),i}$ is chosen to be the observed loss at the last time when arm i was picked. Otherwise, the algorithm is identical to the BROAD-OMD with option I. [4] used Algorithm 1 to obtain an

$\mathcal{O}(\sqrt{K\sum_{t=2}^{T}\|\ell_t-\ell_{t-1}\|_1})$ regret bound. Compared to the $\mathcal{O}(K\sqrt{\sum_{t=2}^{T}|\ell_{t,i^*}-\ell_{t-1,i^*}|})$ path length bound obtained earlier using the BROAD-OMD algorithm with option I, the new bound could be $\mathcal{O}(\sqrt{K})$ smaller or $\mathcal{O}(\sqrt{T})$ larger than the earlier result depending on the loss functions.

5 Minimax-optimality of algorithms with $\tilde{\mathcal{O}}(\sqrt{KV_1})$ regret bounds

One might wonder after seeing the adaptive regret bounds of [4] whether they're optimal. To address this, Theorem 1 of [3] constructs an adaptive loss sequence that forces any algorithm playing against it to obtain $\tilde{\Omega}(\sqrt{KV_1})$ regret. The idea comes from the works of Auer and colleagues - for example, a proof sketch is given in Theorem 6.1 of [2] of the fact that one can always force $\Omega(\sqrt{KT})$ regret. The idea of [3]'s lower bound is to use the bounds in, e.g., [2], and rephrase them in terms of the path length: if you create a loss sequence of the form in [2], then ensure that for this sequence you have

$$V_1=\sum_{t=2}^{T}\|l_t-l_{t-1}\|_1=\mathcal{O}(T), \tag{2}$$

then you can use the same argument to get

$$R_T=\Omega(\sqrt{KT})=\Omega(\sqrt{KV_1}), \tag{3}$$

for any algorithm run against these losses.

6 Algorithm 1 of [3]: bounds in terms of V_∞

Algorithm 2 Algorithm 1 of [3]

Define: $\Psi(x)=\frac{1}{\eta}\sum_{i=1}^{K}\ln\frac{1}{x_i}$ for some learning rate η; parameter $\alpha\in(0,1)$.
Initialize: w_1 is the uniform distribution, $c_0=0$.
for $t=1,2,...,T$ **do**
  Play $i_t\sim w_t$ and observe $c_t=l_{t,i_t}$.
  Construct unbiased estimator $\hat{l}_t$ s.t. $\hat{l}_{t,i}=\frac{l_{t,i}-c_{t-1}}{w_{t,i}}\mathbf{1}\{i_t=i\}+c_{t-1}$ for all i.
  Update $x_{t+1}=\arg\min_{x\in\Delta_K}\langle x,\hat{l}_t\rangle+D_\Psi(x,x_t)$.
  $w_{t+1}=(1-\alpha_{t+1})x_{t+1}+\alpha_{t+1}\mathbf{e}_{i_t}$, where $\alpha_{t+1}=\frac{\alpha(1-c_t)}{1+\alpha(1-c_t)}$.
end

First, let's address the similarities and differences between this algorithm and BROAD-OMD [4]. Looking at Algorithm 2, you'll notice that the regularizer is of the same form as BROAD-OMD, with a simple constant learning rate. A main difference is the way in which OMD itself is used: rather than generating a complicated "optimistic" (though it's worth noting that it isn't necessary that this be an accurate descriptor) loss predictor to affect actions, the adjustment is made by directly biasing the next arm choice toward the most-recently-chosen arm.

Briefly, the initial weights x_t are decided by optimistic mirror descent, and then the actual sampling weights w_t are

$$w_{ti}=\begin{cases}(1-\alpha_t)x_{ti}+\alpha_t\cdot 1, & i=i_t,\\ (1-\alpha_t)x_{ti}, & \text{else}\end{cases}, \tag{4}$$

where α_t is a quantity that's negatively, though nonlinearly, correlated to the loss just observed, c_{t-1}:

$$\alpha_t=\frac{(1-c_{t-1})\alpha}{1+(1-c_{t-1})\alpha}. \tag{5}$$

To understand why α_t is defined the way that it is, it helps to take a look at the proof of Theorem 2. At some point in the proof, it's important to have the following identity hold:

$$\frac{\alpha_t}{1-\alpha_t} = (1 - c_{t-1}). \tag{6}$$

This is what ultimately allows us to cancel the term $-\frac{1}{2}\sum_{t=2}^{T}(c_t - c_{t-1})^2$ that comes out of the application of [4]'s Theorem 7 in the proof of Theorem 2.

7 Algorithm 2 of [3]: getting close to $\mathcal{O}(\sqrt{V_1})$ (oblivious)

Algorithm 2 is pretty confusing. It's in some ways analogous to Algorithm 1, but the analysis of the regret is a bit more complicated. The important quantities are listed below.

- The "minority set", $\mathcal{S}_t = \{i \in [K] : x_{ti} < \beta\}$ where β is a parameter.
- $\tau(t) = \max\{\tau \leq T : i_\tau \in \mathcal{S}_{\tau-1}\}$, i.e. the last time step where a minority arm was chosen.
- Along with $\tau(t)$ we also have $c_{\tau(t)} = l_{\tau(t),i_\tau(t)}$. A counterintuitive fact is that they use $c_{\tau(t-1)}$ as the loss estimate at round t, even though it is in general *not true* that $i_{\tau(t-1)} \in S_{\tau(t-1)}$. The authors' justification is essentially that this is still (maybe) a decent estimate of the loss for arms in the minority set, or at least (maybe) better than c_{t-1} itself, since arms in the minority set won't have been chosen as recently - these are assumptions meant to make sense in a situation where the relative ordering of arm rewards doesn't change quickly, which is one situation where the path length makes sense to worry about.
- $\alpha_t = \frac{(1-c_{\tau(t-1)})\alpha}{1+(1-c_{\tau(t-1)})\alpha}$, which is an analog to the quantity of the same name in Algorithm 1. Here we can kind of see that the idea of this algorithm is to play Algorithm 1 on a subset of arms, using a sequence of times $\tau(t)$ rather than the times t to compute reweighted probabilities w_t.
- The regularizer Ψ now includes a negative Shannon entropy term. This is used in the analysis, and is discussed surrounding the proof of Theorem 3 of [3].

There's a fair amount of similarity to Algorithm 1 in the analysis after the first bit of the proof of Theorem 3; once the mirror descent analysis is complete, it's mostly left to deal with the minority arms and show that playing with respect to w_t attains reasonable regret (this is Lemma 12 of [3]). For completeness, the regret bound they achieve is

$$\tilde{\mathcal{O}}\left(K^{1/3}\sqrt{V_1^{2/3}T^{1/3}} + K^2\right), \tag{7}$$

which is smaller than $\sqrt{KV_1}$ when $V_1 \geq T/K$. The authors also note that they believe they can achieve $\sqrt{V_1}$ regret for an oblivious adversary, though they don't currently have a simple algorithm.

8 Path Length Bounds for Linear Bandits

We first introduce the linear bandits framework: The learner's decision set is $\Omega \subseteq \mathcal{B} = \{z \in \mathbb{R}^d : \|z\| \leq 1\}$. At each round $t \in [T]$, the learner picks an action $\omega_t \in \Omega$ and simultaneously, the adaptive adversary picks a linear loss function parametrized by $\ell_t \in \mathcal{L} \subseteq \mathcal{B}_* = \{z \in \mathbb{R}^d : \|z\|_* \leq 1\}$. The learner then incurs and observes the loss $\langle \omega_t, \ell_t \rangle$. The goal is to minimize the (pseudo) regret defined as $\max_{\omega^* \in \Omega} \mathbb{E}\big[\sum_{t=1}^{T}\langle \omega_t - \omega^*, \ell_t\rangle\big]$. [3] introduced Algorithm 3 as the meta-algorithm for obtaining regret bounds that are parametrized by the path length of the sequence of loss functions. The algorithm is based on the optimistic OMD framework, but its choice of the regularizer Ψ, unbiased loss estimators $\{\hat{\ell}_t\}_{t=1}^T$ and the optimistic predictions $\|m_t\|_{t=1}^T$ are quite different from the BROAD-OMD algorithm of [4]. We explain these choices in more detail below:

Algorithm 3 Meta-Algorithm for Path Length Bounds for Linear Bandits

Parameters: $\Psi(x)$ is a ν-self-concordant barrier, learning rate η
Initialization: Set $x_1 = x'_1 = \arg\min_{x\in\Omega} \Psi(x)$ and m_1=0.
for $t = 1, 2, \ldots, T$ **do**
Compute eigendecomposition $\nabla^2\Psi(x_t) = \sum_{i=1}^{d} \lambda_{t,i} v_{t,i} v_{t,i}^T$.
Sample $i_t \in [d]$ and $\sigma_t \in \{-1, +1\}$ uniformly at random.
Play $\omega_t = x_t + \frac{\sigma_t}{\sqrt{\lambda_{t,i_t}}} v_{t,i_t}$.
Observe and suffer the loss $c_t = \langle \omega_t, \ell_t \rangle$.
Set $\hat{\ell}_t = d(c_t - \langle \omega_t, m_t \rangle)\sigma_t \sqrt{\lambda_{t,i_t}} v_{t,i_t} + m_t$.
Set

$$m_{t+1} = \begin{cases} \Pi_{\mathcal{B}}\big(m_t - \frac{1}{4}(\langle \omega_t, m_t \rangle - c_t)\omega_t\big) & \text{Option I} \\ \Pi_{\mathcal{K}_{t+1}}(m_t) \text{ where } \mathcal{K}_{t+1} = \{m \in \mathcal{B} : \langle \omega_t, m \rangle = c_t\} & \text{Option II} \\ \text{via the convex body chasing algorithm} & \text{Option III} \end{cases}$$

Update $x'_{t+1} = \arg\min_{x\in\Omega} \eta\langle x, \hat{\ell}_t \rangle + D_\Psi(x, x'_t)$.
Update $x_{t+1} = \arg\min_{x\in\Omega} \eta\langle x, m_{t+1} \rangle + D_\Psi(x, x'_{t+1})$.
end for

- Instead of a log-barrier or negative entropy function, Algorithm 3 employs a ν-self-concordant barrier function as the regularizer. More details on the definition of these classes of functions and the intuition behind choosing them could be found in [1].

- For all $t \in [T]$, the chosen action ω_t is an unbiased estimator of the output of the optimistic OMD x_t.

$$\mathbb{E}[\omega_t] = \mathbb{E}[x_t + \frac{\sigma_t}{\sqrt{\lambda_{t,i_t}}} v_{t,i_t}] = x_t + \mathbb{E}[\frac{\sigma_t}{\sqrt{\lambda_{t,i_t}}} v_{t,i_t}] = x_t + \mathbb{E}[\frac{1}{2}\frac{1}{\sqrt{\lambda_{t,i_t}}} v_{t,i_t} - \frac{1}{2}\frac{1}{\sqrt{\lambda_{t,i_t}}} v_{t,i_t}] = x_t.$$

  The choice of ω_t could be thought of as performing simultaneous exploration and exploitation where the latter corresponds to choosing the best action according to the prediction of the optimistic OMD framework x_t and the former is aimed toward best estimating ℓ_t through sampling in a wide region in the space of loss functions.

- $\hat{\ell}_t$ is an unbiased estimator of ℓ_t.

$$\begin{aligned} \mathbb{E}[d\langle \omega_t, \ell_t - m_t \rangle \sigma_t \sqrt{\lambda_{t,i_t}} v_{t,i_t} + m_t] &= \mathbb{E}[\frac{1}{2} d\langle \ell_t - m_t, x_t + \frac{1}{\sqrt{\lambda_{t,i_t}}} v_{t,i_t} \rangle \sqrt{\lambda_{t,i_t}} v_{t,i_t} \\ &\quad - \frac{1}{2} d\langle \ell_t - m_t, x_t - \frac{1}{\sqrt{\lambda_{t,i_t}}} v_{t,i_t} \rangle \sqrt{\lambda_{t,i_t}} v_{t,i_t}] + m_t \\ &= \mathbb{E}[d\langle \ell_t - m_t, v_{t,i_t} \rangle v_{t,i_t}] + m_t \\ &= (\sum_{i=1}^{d} v_{t,i} v_{t,i}^T)(\ell_t - m_t) + m_t \\ &= \ell_t. \end{aligned}$$

- To better understand the three different options for the optimistic predictions $\{m_t\}_{t=1}^T$, we first note that the regret bound of Algorithm 3 is $\mathcal{O}(\frac{\nu \ln T}{\eta} + \eta d^2 \mathbb{E}[\sum_{t=1}^{T} \langle \omega_t, \ell_t - m_t \rangle^2])$. Thus, the predictions should be chosen such that the term $\mathbb{E}[\sum_{t=1}^{T} \langle \omega_t, \ell_t - m_t \rangle^2]$ is close to the path length of the loss

functions. If we were in the full information setting, we could choose $m_t = \ell_{t-1}$ and use the Cauchy-Schwarz inequality to obtain $\sum_{t=1}^{T} \langle \omega_t, \ell_t - m_t \rangle^2 \leq \mathcal{O}(\sum_{t=2}^{T} \|\ell_t - \ell_{t-1}\|)$. However, since we only have access to $\langle \omega_t, \ell_t \rangle$, this is not possible in the bandit setting. Each of the three options for choosing m_t in Algorithm 3 corresponds to using a certain online algorithm whose regret or competitive bound is parametrized by the path length.

References

[1] Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. "Competing in the dark: An efficient algorithm for bandit linear optimization". In: (2009).

[2] Peter Auer et al. "Gambling in a rigged casino: The adversarial multi-armed bandit problem". In: *Proceedings of IEEE 36th Annual Foundations of Computer Science*. IEEE. 1995, pp. 322–331.

[3] Sébastien Bubeck et al. "Improved path-length regret bounds for bandits". In: *Conference On Learning Theory*. PMLR. 2019, pp. 508–528.

[4] Chen-Yu Wei and Haipeng Luo. "More adaptive algorithms for adversarial bandits". In: *Conference On Learning Theory*. PMLR. 2018, pp. 1263–1291.