Proceedings of Machine Learning Research vol 75:1–29, 2018 31st Annual Conference on Learning Theory

More Adaptive Algorithms for Adversarial Bandits

Chen-Yu Wei CHENYU.WEI@USC.EDU
University of Southern California

Haipeng Luo HAIPENGL@USC.EDU
University of Southern California

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We develop a novel and generic algorithm for the adversarial multi-armed bandit problem (or more generally the combinatorial semi-bandit problem). When instantiated differently, our algorithm achieves various new data-dependent regret bounds improving previous work. Examples include: 1) a regret bound depending on the variance of only the best arm; 2) a regret bound depending on the first-order path-length of only the best arm; 3) a regret bound depending on the sum of the first-order path-lengths of all arms as well as an important negative term, which together lead to faster convergence rates for some normal form games with partial feedback; 4) a regret bound that simultaneously implies small regret when the best arm has small loss *and* logarithmic regret when there exists an arm whose expected loss is always smaller than those of other arms by a fixed gap (e.g. the classic i.i.d. setting). In some cases, such as the last two results, our algorithm is completely parameter-free.

The main idea of our algorithm is to apply the optimism and adaptivity techniques to the well-known Online Mirror Descent framework with a special log-barrier regularizer. The challenges are to come up with appropriate optimistic predictions and correction terms in this framework. Some of our results also crucially rely on using a sophisticated increasing learning rate schedule.

Keywords: multi-armed bandit, semi-bandit, adaptive regret bounds, optimistic online mirror descent, increasing learning rate

1. Introduction

The adversarial Multi-Armed Bandits (MAB) problem (Auer et al., 2002) is a classic online learning problem with partial information feedback. In this problem, at each round the learner selects one of the K arms while simultaneously the adversary decides the loss of each arm, then the learner suffers and observes (only) the loss of the picked arm. The goal of the learner is to minimize the regret, that is, the difference between her total loss and the total loss of the best fixed arm. The classic Exp3 algorithm (Auer et al., 2002) achieves a regret bound of order $\tilde{\mathcal{O}}(\sqrt{TK})$ after T rounds,[1] which is worst-case optimal up to logarithmic factors.

There are several existing works on deriving more adaptive bandit algorithms, replacing the dependence on T in the regret bound by some data-dependent quantity that is $\mathcal{O}(T)$ in the worst-case but could be potentially much smaller in benign environments. Examples of such data-dependent quantities include the loss of the best arm (Allenberg et al., 2006; Foster et al., 2016) or the empirical variance of all arms (Hazan and Kale, 2011a; Bubeck et al., 2017). Extensions to more general

1. Throughout the paper we use the notation $\tilde{\mathcal{O}}(\cdot)$ to suppress factors that are poly-logarithmic in T and K.

settings such as semi-bandit, two-point bandit, and graph bandit have also been studied (Neu, 2015; Chiang et al., 2013; Lykouris et al., 2017). These adaptive algorithms not only enjoy better performance guarantees, but also have important applications for other areas such as game theory (Foster et al., 2016).

In this work, we propose a novel and generic bandit algorithm in the more general semi-bandit setting (formally defined in Section 2). By instantiating this generic algorithm differently, we obtain various adaptive algorithms with new data-dependent expected regret bounds that improve previous work. When specified to the MAB setting with $\ell_{t,i} \in [-1,1]$ denoting the loss of arm i at time t (and $\ell_{0,i} \triangleq 0$), these bounds replace the dependence on T by (also see Table 1 for a summary):

- $\sum_{t=1}^{T}(\ell_{t,i^\star} - \frac{1}{T}\sum_{s=1}^{T}\ell_{s,i^\star})^2$, that is, the (unnormalized) variance of the best arm $i^\star$. Similar existing bounds of (Hazan and Kale, 2011a,b; Bubeck et al., 2017) replace T by the average of the variances of all arms. In general these two are incomparable. However, note that the variance of the best arm is always bounded by K times the average variance, while it is possible that the latter is of order $\Theta(T)$ and the former is only $\mathcal{O}(1)$. (Section 3.1)

- $K\sum_{t=1}^{T}|\ell_{t,i^\star} - \ell_{t-1,i^\star}|$, that is, ($K$ times) the first-order path-length of the best arm. (Section 3.2)

- $\sum_{i=1}^{K}\sum_{t=1}^{T}|\ell_{t,i} - \ell_{t-1,i}|$, that is, the sum of the first-order path-lengths of all arms. Importantly, there is also an additional negative term in the regret similar to the one of (Syrgkanis et al., 2015) for the full information setting. This implies a fast convergence rate of order $1/T^{\frac{3}{4}}$ for several game playing settings with bandit feedback. (Sections 4.1)

- A new quantity in terms of some second-order excess loss (see Eq. (9) for the exact form). While the bound is not easy to interpret on it own, it in fact automatically and simultaneously implies the so-called "small-loss" bound $\tilde{\mathcal{O}}\left(\sqrt{K\sum_{t=1}^{T}\ell_{t,i^\star}}\right)$,[2] *and* logarithmic regret $\mathcal{O}(\frac{K\ln T}{\Delta})$ if there is an arm whose expected loss is always smaller than those of other arms by a fixed gap Δ (e.g. the classic i.i.d. MAB setting (Lai and Robbins, 1985)). (Section 4.2)

These bounds are incomparable in general. All of them have known counterparts in the full information setting (see for example (Steinhardt and Liang, 2014) and (De Rooij et al., 2014)), but are novel in the bandit setting to the best of our knowledge. Note that for the first two results that depend on some quantities of only the best arm, we require tuning a learning rate parameter in terms of these (unknown) quantities. Obtaining the same results with parameter-free algorithms remains open, even for the full information setting. However, for the other results, we indeed provide parameter-free algorithms based on a variant of the doubling trick.

Our general algorithm falls into the Online Mirror Descent (OMD) framework (see for example (Hazan et al., 2016)) with the "log-barrier" as the regularizer, originally proposed in (Foster et al., 2016). However, to obtain our results, two extra crucial ingredients are needed:

- First, we adopt the ideas of optimism and adaptivity from (Steinhardt and Liang, 2014), which roughly speaking amounts to incorporating a correction term as well as an optimistic prediction into the loss vectors. In (Steinhardt and Liang, 2014), this technique was developed in the Follow-the-Regularized-Leader (FTRL) framework,[3] but it is in fact crucial here to re-derive

2. Assuming that losses are non-negative in this case as it is common for small-loss bounds.
3. Although it was confusingly referred as OMD in (Steinhardt and Liang, 2014).

it in the OMD framework (due to the next ingredient). The challenges here are to come up with the right correction terms and optimistic predictions.

- Second, we apply an individual and increasing learning rate schedule for one of the path-length results. Such increasing learning rate schedule was originally proposed in (Bubeck et al., 2016) and also recently used in (Agarwal et al., 2017), but for different purposes.

Although most algorithmic techniques we use in this work have been studied before, combining all of them, in the general semi-bandit setting, requires novel and non-trivial analysis. The use of log-barrier in the semi-bandit setting is also new as far as we know.

Related work. There is a rich literature in deriving adaptive algorithms and regret bounds for online learning with full information feedback (see recent work (Luo and Schapire, 2015; Koolen and Van Erven, 2015; van Erven and Koolen, 2016; Orabona and Pál, 2016; Cutkosky and Boahen, 2017) and references therein), as well as the stochastic bandit setting (such as (Garivier and Cappé, 2011; Lattimore, 2015; Degenne and Perchet, 2016)). Similar results for the adversarial bandit setting, however, are relatively sparse and have been mentioned above. While obtaining regret bounds that depend on the quality of the best action is common in the full information setting, it is in fact much more challenging in the bandit setting, and the only existing result of this kind is the "small-loss" bound (Allenberg et al., 2006; Foster et al., 2016). We hope that our work opens up more possibilities in obtaining these results, despite some recent negative results discovered by Gerchinovitz and Lattimore (2016).

Chiang et al. (2013) proposed bandit algorithms with second-order path-length bounds, but their work requires stronger two-point feedback. The implication of path-length regret bounds on faster convergence rate for computing equilibriums was studied in (Syrgkanis et al., 2015). Other examples of adaptive online learning leading to faster convergence in game theory include (Rakhlin and Sridharan, 2013b; Daskalakis et al., 2015; Foster et al., 2016).

There exist several bandit algorithms that achieve almost optimal regret in both the adversarial setting ($\mathcal{O}(\sqrt{TK})$) and the i.i.d. setting ($\mathcal{O}(\sum_{i:\Delta_i\neq 0}\frac{\ln T}{\Delta_i})$ where Δ_i is the gap between the expected loss of arm i and the one of the optimal arm) (Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Auer and Chiang, 2016; Seldin and Lugosi, 2017). Our results in Section 4.2 have slightly weaker guarantee for the i.i.d. setting (at most K times worse specifically) since it essentially replaces all Δ_i by $\min_{i:\Delta_i\neq 0}\Delta_i$. On the other hand, however, our results have several advantages compared to previous work. First, our guarantee for the adversarial setting is stronger since it replaces the dependence on T by the loss of the best arm. Second, our logarithmic regret result applies to not just the simple i.i.d. setting, but the more general setting mentioned above where neither independence nor identical distributions is required. Our dependence on $\ln T$ is also better than previous works, resolving an open problem raised by Seldin and Lugosi (2017). Finally, our algorithm and analysis are also arguably much simpler, without performing any stationarity detection or gap estimation. Indeed, the result is in some sense algorithm-independent and solely through a new adaptive regret bound Eq. (9), similar to the results in the full-information setting such as (Gaillard et al., 2014).

Using a self-concordant barrier as regularizer was proposed in the seminal work of (Abernethy et al., 2008) for general linear bandit problems. The log-barrier is technically not a barrier for the decision set of the semi-bandit problem, but still it exhibits many similar properties as shown in our proofs. Optimistic FTRL/OMD was developed in (Chiang et al., 2012; Rakhlin and Sridharan, 2013a). As pointed out in (Steinhardt and Liang, 2014), incorporating correction terms in the loss

vectors can also be viewed as using adaptive regularizers, which was studied in several previous works, mostly for the full information setting (see (McMahan, 2017) for a survey).

2. Problem Setup and Algorithm Overview

We consider the combinatorial bandit problem with semi-bandit feedback, which subsumes the classic multi-armed bandit problem. The learning process proceeds for T rounds. In each round, the learner selects a subset of arms, denoted by a binary vector b_t from a predefined action set $\mathcal{X} \subseteq \{0,1\}^K$, and suffers loss $b_t^\top \ell_t$, where $\ell_t \in [-1,1]^K$ is a loss vector decided by an adversary. The feedback received by the learner is the vector $(b_{t,1}\ell_{t,1}, \ldots, b_{t,K}\ell_{t,K})$, or in other words, the loss of each chosen arm. For simplicity, we assume that the adversary is oblivious and the loss vectors $\ell_1, \ldots, \ell_T$ are decided ahead of time independent of the learner's actions.

The learner's goal is to minimize the *regret*, which is the gap between her accumulated loss and that of the best fixed action $b^* \in \mathcal{X}$. Formally the regret is defined as

$$\mathrm{Reg}_T \triangleq \sum_{t=1}^T b_t^\top \ell_t - \sum_{t=1}^T b^{*\top} \ell_t, \text{ where } b^* \triangleq \min_{b \in \mathcal{X}} \sum_{t=1}^T b^\top \ell_t.$$

In the special case of multi-armed bandit, the action set $\mathcal{X}$ is $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K\}$ where $\mathbf{e}_i$ denotes the i-th standard basis vector. In other words, in each round the learner picks one arm $i_t \in [K] \triangleq \{1, 2, \ldots, K\}$ (corresponding to $b_t = \mathbf{e}_{i_t}$), and receives the loss ℓ_{t,i_t}. We denote the best arm by $i^* \triangleq \min_{i \in [K]} \sum_{t=1}^T \ell_{t,i}$.

Notation. For a convex function ψ defined on a convex set Ω, the Bregman divergence of two points $u, v \in \Omega$ with respect to ψ is defined as $D_\psi(u,v) \triangleq \psi(u) - \psi(v) - \langle \nabla\psi(v), u - v\rangle$. The log-barrier used in this work is of the form $\psi(u) = \sum_{i=1}^K \frac{1}{\eta_i} \ln \frac{1}{u_i}$ for some learning rates $\eta_1, \ldots, \eta_K \geq 0$ and $u \in \mathrm{conv}(\mathcal{X})$, the convex hull of $\mathcal{X}$. With $h(y) \triangleq y - 1 - \ln y$, the Bregman divergence with respect to the log-barrier is: $D_\psi(u,v) = \sum_{i=1}^K \frac{1}{\eta_i}\left(\ln \frac{v_i}{u_i} + \frac{u_i - v_i}{v_i}\right) = \sum_{i=1}^K \frac{1}{\eta_i} h\left(\frac{u_i}{v_i}\right)$.

The all-zero and all-one vector are denoted by $\mathbf{0}$ and $\mathbf{1}$ respectively. Δ_K represents the $(K-1)$-dimensional simplex. For a binary vector b we write $i \in b$ if $b_i = 1$. Denote by $K_0 = \max_{b \in \mathcal{X}} \|b\|_0$ the maximum number of arms an action in $\mathcal{X}$ can pick. Note that for MAB, K_0 is simply 1.

We define $\ell_0 = \mathbf{0}$ for notational convenience. At round t, for an arm i we denote its accumulated loss by $L_{t,i} \triangleq \sum_{s=1}^t \ell_{s,i}$, its average loss by $\mu_{t,i} \triangleq \frac{1}{t} L_{t,i}$, its (unnormalized) variance by $Q_{t,i} \triangleq \sum_{s=1}^t (\ell_{s,i} - \mu_{t,i})^2$, and its first-order path-length by $V_{t,i} \triangleq \sum_{s=1}^t |\ell_{s,i} - \ell_{s-1,i}|$. For MAB, we define $\alpha_i(t)$ to be the most recent time when arm i is picked prior to round t , that is, $\alpha_i(t) = \max\{s < t : i_s = i\}$ (or 0 if the set is empty).

2.1. Algorithm Overview

As mentioned our algorithm falls into the OMD framework that operates on the set $\Omega = \mathrm{conv}(\mathcal{X})$. The vanilla OMD formula for the bandit setting is $w_t = \mathrm{argmin}_{w \in \Omega}\{\langle w, \hat{\ell}_{t-1}\rangle + D_\psi(w, w_{t-1})\}$ for some regularizer ψ and some (unbiased) estimator $\hat{\ell}_{t-1}$ of the true loss ℓ_{t-1}. The learner then picks an action b_t randomly such that $\mathbb{E}[b_t] = w_t$, and constructs the next loss estimator $\hat{\ell}_t$ based on the bandit feedback. Our algorithm, however, requires several extra ingredients. The generic update

Algorithm 1 **B**arrier-**R**egularized with **O**ptimism and **AD**aptivity **O**nline **M**irror **D**escent (**BROAD-OMD**)

Define: $\Omega = \mathrm{conv}(\mathcal{X})$, $\psi_t(w) = \sum_{i=1}^K \frac{1}{\eta_{t,i}} \ln \frac{1}{w_i}$.
Initialize: $w'_1 = \mathrm{argmin}_{w\in\Omega} \psi_1(w)$.
for $t = 1, 2, \ldots, T$ **do**
  $w_t = \mathrm{argmin}_{w\in\Omega} \left\{ \langle w, m_t\rangle + D_{\psi_t}(w, w'_t)\right\}$.
  Draw $b_t \sim w_t$, suffer loss $b_t^\top \ell_t$, and observe $\{b_{t,i}\ell_{t,i}\}_{i=1}^K$.
  Construct $\hat{\ell}_t$ as an unbiased estimator of ℓ_t.
  Let $a_{t,i} = \begin{cases} 6\eta_{t,i} w_{t,i} (\hat{\ell}_{t,i} - m_{t,i})^2, & \text{(Option I)} \\ 0. & \text{(Option II)} \end{cases}$
  $w'_{t+1} = \mathrm{argmin}_{w\in\Omega} \left\{ \langle w, \hat{\ell}_t + a_t\rangle + D_{\psi_t}(w, w'_t)\right\}$.
end

Table 1: Different configurations of BROAD-OMD and regret bounds for MAB. See Section 2 and the corresponding sections for the meaning of notation. For the last two rows, to obtain parameter-free algorithms one needs to apply a doubling trick to decrease the learning rate.

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

rule is

$$w_t = \operatorname*{argmin}_{w\in\Omega} \left\{ \langle w, m_t\rangle + D_{\psi_t}(w, w'_t)\right\}, \tag{1}$$

$$w'_{t+1} = \operatorname*{argmin}_{w\in\Omega} \left\{ \langle w, \hat{\ell}_t + a_t\rangle + D_{\psi_t}(w, w'_t)\right\}. \tag{2}$$

Here, we still play randomly according to w_t, which is now updated to minimize its loss with respect to $m_t \in [-1, 1]^K$, an *optimistic prediction* of the true loss vector ℓ_t, penalized by a Bregman divergence term associated with a *time-varying* regularizer ψ_t. In addition, we maintain a sequence of auxiliary points w'_t that is updated using the loss estimator $\hat{\ell}_t$ and an extra *correction term* a_t.

When $a_t = \mathbf{0}$, this is studied in (Rakhlin and Sridharan, 2013a) under the name optimistic OMD. When $a_t \neq \mathbf{0}$, the closest algorithm to this variant of OMD is its FTRL version studied by Steinhardt and Liang (2014). However, while ψ_t is fixed for all t in (Steinhardt and Liang, 2014),[4] some of our results crucially rely on using time-varying ψ_t (which corresponds to time-varying learning rate) and also the OMD update form instead of FTRL.

It is well known that the classic Exp3 algorithm falls into this framework with $m_t = a_t = \mathbf{0}$ and ψ_t being the (negative) entropy. To obtain our results, first, it is crucial to use the log-barrier

4. Steinhardt and Liang (2014) also uses the notation ψ_t, but it corresponds to putting a_t into a fixed regularizer.

as the regularizer instead, that is, $\psi_t(w) = \sum_{i=1}^K \frac{1}{\eta_{t,i}} \ln \frac{1}{w_i}$ for some individual and time-varying learning rates $\eta_{t,i}$. Second, we focus on two options of a_t. For results that depend on some quantity of only the best arm, we use a sophisticated choice of a_t that we explain in details in Section 3. For the other results we simply set $a_t = \mathbf{0}$. With the choices of $m_t, \hat{\ell}_t$, and η_t open, we present this generic framework in Algorithm 1 and name it BROAD-OMD (short for Barrier-Regularized with Optimism and ADaptivity Online Mirror Descent).

In Section 3 and 4 respectively, we prove general regret bounds for BROAD-OMD with Option I and Option II, followed by specific applications in the MAB setting achieved via specific choices of $m_t, \hat{\ell}_t$, and η_t. The results and the corresponding configurations of the algorithm are summarized in Table 1.

Computational efficiency. The sampling step $b_t \sim w_t$ can be done efficiently as long as Ω can be described by a polynomial number of constraints. The optimization problems in the update rules of w_t and w'_t are convex and can be solved by general optimization methods. For many special cases, however, these two computational bottlenecks have simple solutions. Take MAB as an example, w_t directly specifies the probability of picking each arm, and the optimization problems can be solved via a simple binary search (Agarwal et al., 2017).

3. BROAD-OMD with Option I

In this section we focus on BROAD-OMD with Option I. We first show a general lemma that update rules (1) and (2) guarantee, no matter what regularizer ψ_t is used and what a_t, m_t, and $\hat{\ell}_t$ are.

Lemma 1 *For the update rules (1) and (2), if the following condition holds:*

$$\langle w_t - w'_{t+1}, \hat{\ell}_t - m_t + a_t \rangle \le \langle w_t, a_t \rangle, \tag{3}$$

then for all $u \in \Omega$, we have

$$\langle w_t - u, \hat{\ell}_t \rangle \le D_{\psi_t}(u, w'_t) - D_{\psi_t}(u, w'_{t+1}) + \langle u, a_t \rangle - A_t, \tag{4}$$

where $A_t \triangleq D_{\psi_t}(w'_{t+1}, w_t) + D_{\psi_t}(w_t, w'_t) \ge 0$.

The important part of bound (4) is the term $\langle u, a_t \rangle$, which allows us to derive regret bounds that depend on only the comparator u. The key is now how to configure the algorithm such that condition (3) holds, while leading to a reasonable bound (4) at the same time.

In the work of (Steinhardt and Liang, 2014) for full-information problems, a_t can be defined as $a_{t,i} = \eta_{t,i}(\ell_{t,i} - m_{t,i})^2$, which suffices to derive many interesting results. However, in the bandit setting this is not applicable since ℓ_t is unknown. The natural first attempt is to replace ℓ_t by $\hat{\ell}_t$, but one would quickly realize the common issue in the bandit literature: $\hat{\ell}_{t,i}$ is often constructed via inverse propensity weighting, and thus $(\hat{\ell}_{t,i} - m_{t,i})^2$ can be of order $1/w_{t,i}^2$, which is too large.

Based on this observation, our choice for a_t is $a_{t,i} = 6\eta_{t,i} w_{t,i} (\hat{\ell}_{t,i} - m_{t,i})^2$ (the constant 6 is merely for technical reasons). The extra term $w_{t,i}$ can then cancel the aforementioned large term $1/w_{t,i}^2$ in expectation, similar to the classic trick done in the analysis of Exp3 (Auer et al., 2002).

Note that with a smaller a_t, condition (3) becomes more stringent. The entropy regularizer used in (Steinhardt and Liang, 2014) no longer suffices to maintain such a condition. Instead, it turns out that the log-barrier regularizer used by BROAD-OMD addresses the issue, as shown below.

Theorem 2 *If the following three conditions hold for all t, i: (i) $\eta_{t,i} \leq \frac{1}{162}$, (ii) $w_{t,i}|\hat{\ell}_{t,i} - m_{t,i}| \leq 3$, (iii) $\sum_{i=1}^{K} \eta_{t,i} w_{t,i}^2 (\hat{\ell}_{t,i} - m_{t,i})^2 \leq \frac{1}{18}$, then* BROAD-OMD *with $a_{t,i} = 6\eta_{t,i} w_{t,i} (\hat{\ell}_{t,i} - m_{t,i})^2$ guarantees condition (3). Moreover, it guarantees for any $u \in \Omega$ (recall $h(y) = y - 1 - \ln y \geq 0$),*

$$\sum_{t=1}^{T} \langle w_t - u, \hat{\ell}_t \rangle \leq \sum_{i=1}^{K} \left(\frac{\ln \frac{w'_{1,i}}{u_i}}{\eta_{1,i}} + \sum_{t=1}^{T} \left(\frac{1}{\eta_{t+1,i}} - \frac{1}{\eta_{t,i}} \right) h\left(\frac{u_i}{w'_{t+1,i}} \right) \right) + \sum_{t=1}^{T} \langle u, a_t \rangle. \tag{5}$$

The three conditions of the theorem are usually trivially satisfied as we will show. Note that $h(\cdot)$ is always non-negative. Therefore, if the sequence $\{\eta_{t,i}\}_{t=1}^{T+1}$ is non-decreasing for all i,[5] the term $\sum_{t=1}^{T} \left(\frac{1}{\eta_{t+1,i}} - \frac{1}{\eta_{t,i}} \right) h\left(\frac{u_i}{w'_{t+1,i}} \right)$ in bound (5) is non-positive. For some results we can simply discard this term, while for others, this term becomes critical. On the other hand, the term $\ln \frac{w'_{1,i}}{u_i}$ appears to be infinity if we want to compare with the best fixed action (where $u_i = 0$ for some i). However, this can be simply resolved by comparing with some close neighbor of the best action in Ω instead, similar to (Foster et al., 2016; Agarwal et al., 2017).

One can now derive different results using Theorem 2 with specific choices of $\hat{\ell}_t$ and m_t. As an example, we state the following corollary by using a variance-reduced importance-weighted estimator $\hat{\ell}_t$ as in (Rakhlin and Sridharan, 2013a).

Corollary 3 BROAD-OMD *with $a_{t,i} = 6\eta_{t,i} w_{t,i} (\hat{\ell}_{t,i} - m_{t,i})^2$, any $m_{t,i} \in [-1, 1]$, $\hat{\ell}_{t,i} = \frac{(\ell_{t,i} - m_{t,i}) \mathbb{1}\{i \in b_t\}}{w_{t,i}} + m_{t,i}$, and $\eta_{t,i} = \eta \leq \frac{1}{162 K_0}$ enjoys the following regret bound:*

$$\mathbb{E}\left[\mathrm{Reg}_T\right] = \mathbb{E}\left[\sum_{t=1}^{T} \langle b_t - b^*, \ell_t \rangle \right] \leq \frac{K \ln T}{\eta} + 6\eta \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i : i \in b^*} (\ell_{t,i} - m_{t,i})^2 \right] + \mathcal{O}(K_0).$$

One can see that the expected regret in Corollary 3 only depends on the squared estimation error of m_t for the actions that b^* chooses! This is exactly the counterpart of results in (Steinhardt and Liang, 2014), but for the more challenging combinatorial semi-bandit problem. Note that our dependence on K_0 is also optimal (Audibert et al., 2013).

In the following subsections, we invoke Theorem 2 with different choices of $\hat{\ell}_t$ and m_t to obtain various more concrete adaptive bounds. For simplicity, we state these results only in the MAB setting, but they can be straightforwardly generalized to the semi-bandit case.

3.1. Variance Bound

Our first application of BROAD-OMD is an adaptive bound that depends on the variance of the best arm, that is, a bound of order $\tilde{\mathcal{O}}\left(\sqrt{K Q_{T,i^*}}\right) = \tilde{\mathcal{O}}\left(\sqrt{K \sum_{t=1}^{T} (\ell_{t,i^*} - \mu_{T,i^*})^2}\right)$. According to Corollary 3, if we were able to use $m_t = \mu_T$, with a best-tuned η the bound is obtained immediately. The issue is of course that μ_T is unknown ahead of time. In fact, even setting $m_t = \mu_{t-1}$ is infeasible due to the bandit feedback.

Fortunately this issue was already solved by Hazan and Kale (2011a) via the "reservoir sampling" technique. The high level idea is that one can spend a small portion of time on estimating

5. One might notice that $\eta_{T+1,i}$ is not defined here. Indeed this term is artificially added only to make the analysis of Section 3.2 more concise, and $\eta_{T+1,i}$ can be any positive number. In Algorithm 2 we give it a concrete definition.

μ_t on the fly. More precisely, by performing uniform exploration with probability $\min\left\{1, \frac{MK}{t}\right\}$ at time t for some parameter M, one can obtain an estimator $\tilde{\mu}_t$ of μ_t such that $\mathbb{E}[\tilde{\mu}_t] = \mu_t$ and $\mathrm{Var}[\tilde{\mu}_{t,i}] \le \frac{Q_{t,i}}{Mt}$ (see (Hazan and Kale, 2011a) for details). Then we can simply pick $m_t = \tilde{\mu}_{t-1}$ and prove the following result.

Theorem 4 BROAD-OMD *with reservoir sampling (Hazan and Kale, 2011a),* $a_{t,i} = 6\eta_{t,i} w_{t,i} (\hat{\ell}_{t,i} - m_{t,i})^2$, $m_{t,i} = \tilde{\mu}_{t-1,i}$, $\hat{\ell}_{t,i} = \frac{(\ell_{t,i} - m_{t,i}) \mathbb{1}\{i_t = i\}}{w_{t,i}} + m_{t,i}$, *and* $\eta_{t,i} = \eta \le \frac{1}{162}$ *guarantees*

$$\mathbb{E}\left[\mathrm{Reg}_T\right] = \mathcal{O}\left(\frac{K \ln T}{\eta} + \eta Q_{T,i^*} + K (\ln T)^2\right).$$

With the optimal tuning of η*, the regret is thus of order* $\tilde{\mathcal{O}}\left(\sqrt{K Q_{T,i^*}} + K\right)$.

3.2. Path-length Bound

Our second application is to obtain path-length bounds. The counterpart in the full-information setting is a bound in terms of the second-order path-length $\sum_{t=1}^T (\ell_{t,i^*} - \ell_{t-1,i^*})^2$ (Steinhardt and Liang, 2014). Again, in light of Corollary 3, if we were able to pick $m_t = \ell_{t-1}$ the problem would be solved. The difficulty is again that ℓ_{t-1} is not fully observable.

While it is still not clear how to achieve such a second-order path-length bound or whether it is possible at all, we propose a way to obtain a slightly weaker first-order path-length bound $\tilde{\mathcal{O}}\left(K\sqrt{V_{T,i^*}}\right) = \tilde{\mathcal{O}}\left(K\sqrt{\sum_{t=1}^T |\ell_{t,i^*} - \ell_{t-1,i^*}|}\right)$. Note that in the worst case this is $\sqrt{K}$ times worse than the optimal regret $\tilde{\mathcal{O}}(\sqrt{TK})$.

The idea is to set $m_{t,i}$ to be the most recent observed loss of arm i, that is, $m_{t,i} = \ell_{\alpha_i(t),i}$, where $\alpha_i(t)$ is defined in Section 2. While the estimation error $(\ell_{t,i} - \ell_{\alpha_i(t),i})^2$ could be much larger than $(\ell_{t,i} - \ell_{t-1,i})^2$, the quantity we aim for, observe that if $t - \alpha_i(t)$ is large, it means that arm i has bad performance before time t so that the learner seldom draws arm i. In this case, the learner might have accumulated negative regret with respect to arm i, which can potentially be used to compensate the large estimation error.

To formalize this intuition, we go back to the bound in Theorem 2 and examine the key term $\sum_{t=1}^T \langle u, a_t \rangle$ after plugging in $u = \mathbf{e}_i$ for some arm i, $m_{t,i} = \ell_{\alpha_i(t),i}$, and $\hat{\ell}_{t,i} = \frac{(\ell_{t,i} - m_{t,i}) \mathbb{1}\{i_t = i\}}{w_{t,i}} + m_{t,i}$. We assume $\eta_{t,i} = \eta$ for simplicity and also use the fact $w_{t,i} |\hat{\ell}_{t,i} - m_{t,i}| \le 2$. We then have

$$\begin{aligned}
\sum_{t=1}^T \langle u, a_t \rangle &= 6\eta \sum_{t=1}^T w_{t,i} (\hat{\ell}_{t,i} - \ell_{\alpha_i(t),i})^2 \le 12\eta \sum_{t=1}^T |\hat{\ell}_{t,i} - \ell_{\alpha_i(t),i}| = 12\eta \sum_{t: i_t = i} \frac{|\ell_{t,i} - \ell_{\alpha_i(t),i}|}{w_{t,i}} \\
&\le 12\eta \sum_{t: i_t = i} \frac{\sum_{s=\alpha_i(t)+1}^t |\ell_{s,i} - \ell_{s-1,i}|}{w_{t,i}} \le 12\eta \left(\max_{t \in [T]} \frac{1}{w_{t,i}}\right) V_{T,i}. \qquad (6)
\end{aligned}$$

Therefore, the term $\sum_{t=1}^T \langle u, a_t \rangle$ is close to the first-order path-length but with an extra factor $\max_{t \in [T]} \frac{1}{w_{t,i}}$. To cancel this potentially large factor, we adopt the increasing learning rate schedule recently used in (Agarwal et al., 2017). The idea is that the term $h\left(\frac{u_i}{w'_{t+1,i}}\right)$ in Eq. (5) is close to $\frac{1}{w_{t+1,i}}$ if u_i is close to 1. If we increase the learning rate whenever we encounter a large $\frac{1}{w_{t+1,i}}$,

Algorithm 2 BROAD-OMD+ (specialized for MAB)

Define: $\kappa = e^{\frac{1}{\ln T}}$, $\psi_t(w) = \sum_{i=1}^K \frac{1}{\eta_{t,i}} \ln \frac{1}{w_i}$.
Initialize: $w'_{1,i} = 1/K$, $\rho_{1,i} = 2K$ for all $i \in [K]$.
for $t = 1, 2, \ldots, T$ **do**
  $w_t = \operatorname{argmin}_{w \in \Delta_K} \left\{ \langle w, m_t \rangle + D_{\psi_t}(w, w'_t) \right\}$.
  $\bar{w}_t = (1 - \frac{1}{T}) w_t + \frac{1}{KT} \mathbf{1}$.
  Draw $i_t \sim \bar{w}_t$, suffer loss ℓ_{t,i_t}, and let $\hat{\ell}_{t,i} = \frac{(\ell_{t,i} - m_{t,i}) \mathbb{1}\{i_t = i\}}{\bar{w}_{t,i}} + m_{t,i}$.
  Let $a_{t,i} = 6\eta_{t,i} w_{t,i} (\hat{\ell}_{t,i} - m_{t,i})^2$.
  $w'_{t+1} = \operatorname{argmin}_{w \in \Delta_K} \left\{ \langle w, \hat{\ell}_t + a_t \rangle + D_{\psi_t}(w, w'_t) \right\}$.
  for $i = 1, \ldots, K$ **do**
    if $\frac{1}{\bar{w}_{t,i}} > \rho_{t,i}$ **then** $\rho_{t+1,i} = \frac{2}{\bar{w}_{t,i}}$, $\eta_{t+1,i} = \kappa \eta_{t,i}$.
    else $\rho_{t+1,i} = \rho_{t,i}$, $\eta_{t+1,i} = \eta_{t,i}$.
  end
end

then $\left(\frac{1}{\eta_{t+1,i}} - \frac{1}{\eta_{t,i}}\right) h\left(\frac{u_i}{w'_{t+1,i}}\right)$ becomes a large negative term in terms of $\frac{-1}{w_{t+1,i}}$, which exactly compensates the term $\sum_{t=1}^T \langle u, a_t \rangle$.

To avoid the learning rates increased by too much, similarly to (Agarwal et al., 2017) we use some individual threshold ($\rho_{t,i}$) to decide when to increase the learning rate and update these thresholds in some doubling manner. Also, we mix w_t with a small amount of uniform exploration to further ensure that it cannot be too small. The final algorithm, call BROAD-OMD+, is presented in Algorithm 2 (only for the MAB setting for simplicity). We prove the following theorem.

Theorem 5 BROAD-OMD+ *with* $m_{t,i} = \ell_{\alpha_i(t),i}$ *and* $\eta_{1,i} = \eta \le \frac{1}{810}$ *guarantees*

$$\mathbb{E}\left[\mathrm{Reg}_T\right] \le \frac{2K \ln T}{\eta} + \mathbb{E}[\rho_{T+1,i^*}] \left(\frac{-1}{40\eta \ln T} + 90\eta V_{T,i^*} \right) + \mathcal{O}(1)$$

when $T \ge 3$. *Picking* $\eta = \min\left\{ \frac{1}{810}, \frac{1}{60\sqrt{V_{T,i^*} \ln T}} \right\}$ *so that the second term is non-positive leads to* $\mathbb{E}\left[\mathrm{Reg}_T\right] = \tilde{\mathcal{O}}\left(K\sqrt{V_{T,i^*}} + K\right)$.

4. BROAD-OMD with Option II

In this section, we move on to discuss BROAD-OMD with Option II, that is, $a_t = \mathbf{0}$. We also fix $\eta_{t,i} = \eta$, although in the doubling trick discussed later, different values of η will be used for different runs of BROAD-OMD. Again we start with a general lemma that holds no matter what regularizer ψ_t is used and what m_t and $\hat{\ell}_t$ are.

Lemma 6 *For the update rules* (1) *and* (2) *with* $a_t = \mathbf{0}$, *we have for all* $u \in \Omega$,

$$\langle w_t - u, \hat{\ell}_t \rangle \le D_{\psi_t}(u, w'_t) - D_{\psi_t}(u, w'_{t+1}) + \langle w_t - w'_{t+1}, \hat{\ell}_t - m_t \rangle - A_t,$$

where $A_t \triangleq D_{\psi_t}(w'_{t+1}, w_t) + D_{\psi_t}(w_t, w'_t) \ge 0$.

The proof is standard as in typical OMD analysis. The next theorem then shows how the term $\langle w_t - w'_{t+1}, \hat{\ell}_t - m_t \rangle$ is further bounded when ψ_t is the log-barrier as in BROAD-OMD.

Theorem 7 *If the following three conditions hold for all t, i: (i) $\eta \leq \frac{1}{162}$, (ii) $w_{t,i}|\hat{\ell}_{t,i} - m_{t,i}| \leq 3$, (iii) $\eta \sum_{i=1}^{K} w_{t,i}^2 (\hat{\ell}_{t,i} - m_{t,i})^2 \leq \frac{1}{18}$ (same as those in Theorem 2), then* BROAD-OMD *with $a_t = \mathbf{0}$ guarantees for any $u \in \Omega$,*

$$\sum_{t=1}^{T} \langle w_t - u, \hat{\ell}_t \rangle \leq \sum_{i=1}^{K} \frac{\ln \frac{w'_{1,i}}{u_i}}{\eta} + 3\eta \sum_{t=1}^{T} \sum_{i=1}^{K} w_{t,i}^2 (\hat{\ell}_{t,i} - m_{t,i})^2 - \sum_{t=1}^{T} A_t. \tag{7}$$

For MAB, the last term can further be lower bounded by $\sum_{t=1}^{T} A_t \geq \frac{1}{48\eta} \sum_{t=2}^{T} \sum_{i=1}^{K} \frac{(w_{t,i} - w_{t-1,i})^2}{w_{t-1,i}^2}$.

In bound (7), the first term can again be bounded by $\frac{K \ln T}{\eta}$ via picking an appropriate u. The last negative term is useful when we use the algorithm to play games, which is discussed in Section 4.1.1. The second term is the key term, which, compared to the key term $\sum_{t=1}^{T} \langle u, a_t \rangle$ in Eq. (5) for BROAD-OMD with Option I, has an extra $w_{t,i}$ and is in terms of all arms instead of the arms that u picks. As a comparison to Corollary 3, if we pick $\hat{\ell}_{t,i} = \frac{(\ell_{t,i} - m_{t,i}) \mathbb{1}\{i \in b_t\}}{w_{t,i}} + m_{t,i}$, we obtain an expected regret bound in terms of $\mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \in b_t} (\ell_{t,i} - m_{t,i})^2\right] = \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i=1}^{K} w_{t,i} (\ell_{t,i} - m_{t,i})^2\right]$, which is not as easy to interpret as the bound in Corollary 3. However, in the following subsections we will discuss in details how to apply bound (7) to obtain more concrete results.

Before that, we point out that since the bound is now in terms of all arms, we can in fact apply a doubling trick to make the algorithm parameter-free! The idea is that as long as the observable term $3\eta \sum_{s=1}^{t} \sum_{i=1}^{K} w_{s,i}^2 (\hat{\ell}_{s,i} - m_{s,i})^2$ becomes larger than $\frac{K \ln T}{\eta}$ at some round t, we half the learning rate η and restart the algorithm. This avoids the need for optimal tuning done in Section 3. We formally present the algorithm in Algorithm 3 (in Appendix G) and show its regret bound below.

Theorem 8 *If conditions (ii) and (iii) in Theorem 7 hold, then Algorithm 3 guarantees*

$$\mathbb{E}[\mathrm{Reg}_T] = \mathcal{O}\left(\sqrt{(K \ln T) \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i=1}^{K} w_{t,i}^2 (\hat{\ell}_{t,i} - m_{t,i})^2\right]} + K_0 K \ln T\right).$$

In the following subsections, we instantiate Theorem 7 or 8 with different m_t and $\hat{\ell}_t$. Again, for simplicity we only focus on the MAB setting.

4.1. Another Path-length Bound

If we configure BROAD-OMD with Option II in the same way as in Section 3.2, that is, $m_{t,i} = \ell_{\alpha_i(t),i}$ and $\hat{\ell}_{t,i} = \frac{(\ell_{t,i} - m_{t,i}) \mathbb{1}\{i_t = i\}}{w_{t,i}} + m_{t,i}$. Then the key term in Eq. (7) can be bounded as follows:

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{i=1}^{K} w_{t,i}^2 (\hat{\ell}_{t,i} - m_{t,i})^2 = \sum_{t=1}^{T} \sum_{i=1}^{K} (\ell_{t,i} - \ell_{\alpha_i(t),i})^2 \mathbb{1}\{i_t = i\} = \sum_{i=1}^{K} \sum_{t: i_t = i} (\ell_{t,i} - \ell_{\alpha_i(t),i})^2 \\
&\leq 2 \sum_{i=1}^{K} \sum_{t: i_t = i} |\ell_{t,i} - \ell_{\alpha_i(t),i}| \leq 2 \sum_{i=1}^{K} \sum_{t: i_t = i} \sum_{s = \alpha_i(t)+1}^{t} |\ell_{s,i} - \ell_{s-1,i}| \leq 2 \sum_{i=1}^{K} V_{T,i}.
\end{aligned} \tag{8}$$

Unlike Eq. (6), this is bounded even without the help of negative regret, but the price is that now the regret depends on the sum of all arms' path-length. With this calculation, we obtain the following corollary.

Corollary 9 BROAD-OMD *with* $a_{t,i} = 0$, $m_{t,i} = \ell_{\alpha_i(t),i}$, $\hat{\ell}_{t,i} = \frac{(\ell_{t,i}-m_{t,i})\mathbb{1}\{i_t=i\}}{w_{t,i}} + m_{t,i}$, *and* $\eta_{t,i} = \eta \le \frac{1}{162}$ *guarantees*

$$\mathbb{E}\left[\mathrm{Reg}_T\right] \le \mathcal{O}\left(\frac{K\ln T}{\eta}\right) + 6\eta\sum_{i=1}^{K} V_{T,i} - \mathbb{E}\left[\sum_{t=2}^{T}\sum_{i=1}^{K}\frac{(w_{t,i}-w_{t-1,i})^2}{48\eta w_{t-1,i}^2}\right] \le \mathcal{O}\left(\frac{K\ln T}{\eta} + \eta\sum_{i=1}^{K} V_{T,i}\right).$$

Using the doubling trick (Algorithm 3), we achieve expected regret $\tilde{\mathcal{O}}\left(\sqrt{K\sum_{i=1}^{K} V_{T,i}} + K\right)$.

This new path-length bound could be $\sqrt{K}$ times better than the one in Section 3.2 in some cases, but $\sqrt{T}$ times larger in others. The extra advantage, however, is the negative term in the regret,[6] explicitly spelled out in Corollary 9, which we discuss next.

4.1.1. FAST CONVERGENCE IN BANDIT GAMES

It is well-known that in a repeated two-player zero-sum game, if both players play according to some no-regret algorithms, then their average strategies converge to a Nash equilibrium (Freund and Schapire, 1999). Similar results for general multi-player games have also been discovered. The convergence rate of these results is governed by the regret bounds of the learning algorithms, and several recent works (such as those mentioned in the introduction) have developed adaptive algorithms with regret much smaller than the worst case $\mathcal{O}(\sqrt{T})$ by exploiting the special structure in this setup, which translates to convergence rates faster than $1/\sqrt{T}$ in computing equilibriums.

One way to obtain such fast rates is exactly via path-length regret bounds as shown in (Rakhlin and Sridharan, 2013b; Syrgkanis et al., 2015). In these works, the convergence rate $1/T$ is achieved when the players have full-information feedback. We generalize their results to the bandit setting, and show that convergence rate of $1/T^{\frac{3}{4}}$ can be obtained. Though faster than $1/\sqrt{T}$, it is still slower than $1/T$ compared to the full-information setting, which is due to the fact that in bandit we only have first-order instead of second-order path-length bound. We detail the proofs and the remaining open problems in Appendix I.

4.2. Adapting to Stochastic Bandits

Our last application is to obtain an algorithm that simultaneously enjoys near optimal regret in both adversarial and stochastic setting. Specifically, the stochastic setting we consider here is as follows: there exists an arm a^* and some fixed gap $\Delta > 0$ such that $\mathbb{E}_{\ell_t}\left[\ell_{t,i} - \ell_{t,a^*} | \ell_1, \ldots, \ell_{t-1}\right] \ge \Delta$ for all $i \ne a^*$ and $t \in [T]$. In other words, arm a^*'s expected loss is always smaller than those of other arms by a fixed amount. The classic i.i.d. MAB (Lai and Robbins, 1985) is clearly a special case of ours. Unlike the i.i.d. setting, however, we require neither independence nor identical distributions.

6. In fact, similar negative term, coming from the term A_t in Lemma 1, also exists (but is omitted) in the bound of Theorem 5. However, it is not clear to us how to utilize it in the same way as in Section 4.1.1 if we also want to exploit the other negative term coming from increasing learning rates.

Note that a^* can be different from the empirically best arm i^* defined in Section 2. The expected regret in this setting is still with respect to i^* and further takes into consideration the randomness over losses. In other words, we care about $\mathbb{E}_{\ell_1,\dots,\ell_T}\left[\mathbb{E}_{i_1,\dots,i_T}[\mathrm{Reg}_T]\right]$, abbreviated as $\mathbb{E}[\mathrm{Reg}_T]$ still.

We invoke BROAD-OMD with $a_t = \mathbf{0}$, $\hat{\ell}_{t,i} = \frac{\ell_{t,i}\mathbb{1}\{i_t=i\}}{w_{t,i}}$ being the typical importance-weighted unbiased estimator, and a somewhat special choice of m_t: $m_{t,i} = \ell_{t,i_t}$ for all i. This choice of m_t is seemingly invalid since it depends on i_t, which is drawn after we have constructed w_t based on m_t itself. However, note that because m_t now has identical coordinates, we have $w_t = \operatorname{argmin}_{w\in\Delta_K}\left\{\langle w, m_t\rangle + D_{\psi_t}(w, w'_t)\right\} = \operatorname{argmin}_{w\in\Delta_K}\left\{D_{\psi_t}(w, w'_t)\right\} = w'_t$, independent of the actual value of m_t. Therefore, the algorithm is still valid and is in fact equivalent to the vanilla log-barrier OMD of (Foster et al., 2016). Also note that we cannot define $\hat{\ell}_t$ as in previous sections (in terms of m_t) since it is not an unbiased estimator of ℓ_t anymore (due to the randomness of m_t).

Although the algorithm is the same, using our analysis framework we actually derive a tighter bound in terms of the following quantity based on Theorem 7: $\sum_{t=1}^T\sum_{i=1}^K w_{t,i}^2(\hat{\ell}_{t,i} - \ell_{t,i_t})^2 = \sum_{t=1}^T\sum_{i=1}^K(\ell_{t,i}\mathbb{1}\{i_t = i\} - w_{t,i}\ell_{t,i_t})^2$. It turns out that based on this quantity alone, one can derive both a "small-loss" bound for the adversarial setting and a logarithmic bound for the stochastic setting as shown below. We emphasize that the doubling trick of Algorithm 3 is essential to make the algorithm parameter-free, which is another key difference from (Foster et al., 2016).

Theorem 10 BROAD-OMD *with* $a_t = 0$, $m_{t,i} = \ell_{t,i_t}$, $\hat{\ell}_{t,i} = \frac{\ell_{t,i}\mathbb{1}\{i_t=i\}}{w_{t,i}}$, *and the doubling trick (Algorithm 3), guarantees*

$$\mathbb{E}\left[\mathrm{Reg}_T\right] = \mathcal{O}\left(\sqrt{(K\ln T)\mathbb{E}\left[\sum_{t=1}^T\sum_{i=1}^K(\ell_{t,i}\mathbb{1}\{i_t = i\} - w_{t,i}\ell_{t,i_t})^2\right]} + K\ln T\right). \tag{9}$$

This bound implies that in the stochastic setting, we have $\mathbb{E}\left[\mathrm{Reg}_T\right] = \mathcal{O}\left(\frac{K\ln T}{\Delta}\right)$, *while in the adversarial setting, we have* $\mathbb{E}\left[\mathrm{Reg}_T\right] = \mathcal{O}\left(\sqrt{KL_{T,i^*}\ln T} + K\ln T\right)$ *assuming non-negative losses.*

5. Conclusions and Discussions

In this work we develop and analyze a general bandit algorithm using techniques such as optimistic mirror descent, log-barrier regularizer, increasing learning rate, and so on. We show various applications of this general framework, obtaining several more adaptive algorithms that improve previous works. Future directions include 1) improving the dependence on K for the path-length results; 2) obtaining second-order path-length bounds; 3) generalizing the results to the linear bandit problem.

Acknowledgement. CYW is grateful for the support of NSF Grant #1755781. The authors would like to thank Chi-Jen Lu for posing the problem of bandit path-length, and to thank Chi-Jen Lu and Yi-Te Hong for helpful discussions in this direction.

References

Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Conference on Learning Theory*, pages 263–274, 2008.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38, 2017.

Chamy Allenberg, Peter Auer, Laszlo Gyorfi, and György Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *Algorithmic Learning Theory*, volume 4264, pages 229–243. Springer, 2006.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2013.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 116–120, 2016.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1, 2012.

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

Sébastien Bubeck, Michael B. Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. *arXiv preprint arXiv:1711.01037*, 2017.

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Conference on Learning Theory*, 2012.

Chao-Kai Chiang, Chia-Jung Lee, and Chi-Jen Lu. Beating bandits in gradually evolving worlds. In *Conference on Learning Theory*, pages 210–227, 2013.

Ashok Cutkosky and Kwabena Boahen. Online learning without prior information. In *Conference on Learning Theory*, 2017.

Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. *Games and Economic Behavior*, 92:327–348, 2015.

Steven De Rooij, Tim Van Erven, Peter D Grünwald, and Wouter M Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15(1):1281–1316, 2014.

Rémy Degenne and Vianney Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In *International Conference on Machine Learning*, pages 1587–1595, 2016.

Dylan J Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Eva Tardos. Learning in games: Robustness of fast convergence. In *Advances in Neural Information Processing Systems*, pages 4734–4742, 2016.

Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

Pierre Gaillard, Gilles Stoltz, and Tim Van Erven. A second-order bound with excess losses. In *Conference on Learning Theory*, pages 176–196, 2014.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Conference On Learning Theory*, pages 359–376, 2011.

Sébastien Gerchinovitz and Tor Lattimore. Refined lower bounds for adversarial bandits. In *Advances in Neural Information Processing Systems*, pages 1198–1206, 2016.

Elad Hazan and Satyen Kale. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12(Apr):1287–1311, 2011a.

Elad Hazan and Satyen Kale. A simple multi-armed bandit algorithm with optimal variation-bounded regret. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 817–820, 2011b.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Wouter M Koolen and Tim Van Erven. Second-order quantile methods for experts and combinatorial games. In *Conference on Learning Theory*, pages 1155–1175, 2015.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tor Lattimore. Optimally confident ucb: Improved regret for finite-armed bandits. *arXiv preprint arXiv:1507.07880*, 2015.

Haipeng Luo and Robert E Schapire. Achieving all with no parameters: Adanormalhedge. In *Conference on Learning Theory*, pages 1286–1304, 2015.

Thodoris Lykouris, Karthik Sridharan, and Eva Tardos. Small-loss bounds for online learning with partial information. *arXiv preprint arXiv:1711.03639*, 2017.

H Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *Journal of Machine Learning Research*, 18(90):1–50, 2017.

Gergely Neu. First-order regret bounds for combinatorial semi-bandits. In *Conference on Learning Theory*, pages 1360–1375, 2015.

Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. In *Advances in Neural Information Processing Systems*, pages 577–585, 2016.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019, 2013a.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013b.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, 2017.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pages 1287–1295, 2014.

Jacob Steinhardt and Percy Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *International Conference on Machine Learning*, pages 1593–1601, 2014.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, pages 2989–2997, 2015.

Tim van Erven and Wouter M Koolen. Metagrad: Multiple learning rates in online learning. In *Advances in Neural Information Processing Systems*, pages 3666–3674, 2016.

Appendix A. Proof of Lemma 1

Proof of Lemma 1. We first state a useful property used in typical OMD analysis. Let Ω be a convex compact set in $\mathbb{R}^K$, ψ be a convex function on Ω, w' be an arbitrary point in Ω, and $x \in \mathbb{R}^K$. If $w^* = \operatorname{argmin}_{w\in\Omega}\{\langle w, x\rangle + D_\psi(w, w')\}$, then for any $u \in \Omega$,

$$\langle w^* - u, x\rangle \le D_\psi(u, w') - D_\psi(u, w^*) - D_\psi(w^*, w').$$

This is by the first-order optimality condition of w^* and direct calculations. Applying this to update rule (2) we have

$$\langle w'_{t+1} - u, \hat{\ell}_t + a_t\rangle \le D_{\psi_t}(u, w'_t) - D_{\psi_t}(u, w'_{t+1}) - D_{\psi_t}(w'_{t+1}, w'_t); \tag{10}$$

while applying it to update rule (1) and picking $u = w'_{t+1}$ we have

$$\langle w_t - w'_{t+1}, m_t\rangle \le D_{\psi_t}(w'_{t+1}, w'_t) - D_{\psi_t}(w'_{t+1}, w_t) - D_{\psi_t}(w_t, w'_t). \tag{11}$$

Now we bound the instantaneous regret as follows:

$$\begin{aligned}
&\langle w_t - u, \hat{\ell}_t\rangle \\
&= \langle w_t - u, \hat{\ell}_t + a_t\rangle - \langle w_t, a_t\rangle + \langle u, a_t\rangle \\
&= \langle w_t - w'_{t+1}, \hat{\ell}_t + a_t\rangle - \langle w_t, a_t\rangle + \langle w'_{t+1} - u, \hat{\ell}_t + a_t\rangle + \langle u, a_t\rangle \\
&= \langle w_t - w'_{t+1}, \hat{\ell}_t + a_t - m_t\rangle - \langle w_t, a_t\rangle + \langle w'_{t+1} - u, \hat{\ell}_t + a_t\rangle + \langle w_t - w'_{t+1}, m_t\rangle + \langle u, a_t\rangle \\
&\le D_{\psi_t}(u, w'_t) - D_{\psi_t}(u, w'_{t+1}) - D_{\psi_t}(w'_{t+1}, w_t) - D_{\psi_t}(w_t, w'_t) + \langle u, a_t\rangle,
\end{aligned} \tag{12}$$

where last inequality is by the condition $\langle w_t - w'_{t+1}, \hat{\ell}_t + a_t - m_t\rangle - \langle w_t, a_t\rangle \le 0$, Eq. (10), and Eq. (11). ■

Appendix B. Lemmas for Log-barrier OMD

In this section we establish some useful lemmas for update rules (1) and (2) with log-barrier regularizer, which are used in the proofs of other theorems. We start with some definitions.

Definition 11 *For any $h \in \mathbb{R}^K$, define norm $\|h\|_{t,w} = \sqrt{h^\top \nabla^2 \psi_t(w) h} = \sqrt{\sum_{i=1}^K \frac{1}{\eta_{t,i}} \frac{h_i^2}{w_i^2}}$ and its dual norm $\|h\|_{t,w}^* = \sqrt{h^\top \nabla^{-2} \psi_t(w) h} = \sqrt{\sum_{i=1}^K \eta_{t,i} w_i^2 h_i^2}$. For some radius $r > 0$, define ellipsoid $\mathcal{E}_{t,w}(r) = \left\{ u \in \mathbb{R}^K : \|u - w\|_{t,w} \le r \right\}$.*

Lemma 12 *If $w' \in \mathcal{E}_{t,w}(1)$ and $\eta_{t,i} \le \frac{1}{81}$ for all i, then $w'_i \in \left[\frac{1}{2} w_i, \frac{3}{2} w_i\right]$ for all i, and also $0.9 \|h\|_{t,w} \le \|h\|_{t,w'} \le 1.2 \|h\|_{t,w}$ for any $h \in \mathbb{R}^K$.*

Proof $w' \in \mathcal{E}_{t,w}(1)$ implies $\sum_{i=1}^K \frac{1}{\eta_{t,i}} \frac{(w'_i - w_i)^2}{w_i^2} \le 1$. Thus for every i, we have $\frac{|w'_i - w_i|}{w_i} \le \sqrt{\eta_{t,i}} \le \frac{1}{9}$, implying $w'_i \in \left[\frac{8}{9} w_i, \frac{10}{9} w_i\right] \subset \left[\frac{1}{2} w_i, \frac{3}{2} w_i\right]$. Therefore, $\|h\|_{t,w'} = \sqrt{\sum_{i=1}^K \frac{1}{\eta_{t,i}} \frac{h_i^2}{w_i'^2}} \ge \sqrt{\sum_{i=1}^K \frac{1}{\eta_{t,i}} \frac{h_i^2}{\left(\frac{10}{9} w_i\right)^2}} = 0.9 \|h\|_{t,w}$. Similarly, we have $\|h\|_{t,w'} \le 1.2 \|h\|_{t,w}$. ■

Lemma 13 *Let w_t, w'_{t+1} follow (1) and (2) where ψ_t is the log-barrier with $\eta_{t,i} \le \frac{1}{81}$ for all i. If $\left\| \hat{\ell}_t - m_t + a_t \right\|_{t,w_t}^* \le \frac{1}{3}$, then $w'_{t+1} \in \mathcal{E}_{t,w_t}(1)$.*

Proof Define $F_t(w) = \langle w, m_t \rangle + D_{\psi_t}(w, w'_t)$ and $F'_{t+1}(w) = \langle w, \hat{\ell}_t + a_t \rangle + D_{\psi_t}(w, w'_t)$. Then by definition we have $w_t = \operatorname{argmin}_{w \in \Omega} F_t(w)$ and $w'_{t+1} = \operatorname{argmin}_{w \in \Omega} F'_{t+1}(w)$. To show $w'_{t+1} \in \mathcal{E}_{t,w_t}(1)$, it suffices to show that for all u on the boundary of $\mathcal{E}_{t,w_t}(1)$, $F'_{t+1}(u) \ge F'_{t+1}(w_t)$.

Indeed, using Taylor's theorem, for any $u \in \partial \mathcal{E}_{t,w_t}(1)$, there is an ξ on the line segment between w_t and u such that (let $h \triangleq u - w_t$)

$$
\begin{aligned}
F'_{t+1}(u) &= F'_{t+1}(w_t) + \nabla F'_{t+1}(w_t)^\top h + \frac{1}{2} h^\top \nabla^2 F'_{t+1}(\xi) h \\
&= F'_{t+1}(w_t) + (\hat{\ell}_t - m_t + a_t)^\top h + \nabla F_t(w_t)^\top h + \frac{1}{2} h^\top \nabla^2 \psi_t(\xi) h \\
&\ge F'_{t+1}(w_t) + (\hat{\ell}_t - m_t + a_t)^\top h + \frac{1}{2} \|h\|_{t,\xi}^2 && \text{(by the optimality of } w_t\text{)} \\
&\ge F'_{t+1}(w_t) + (\hat{\ell}_t - m_t + a_t)^\top h + \frac{1}{2} \times 0.9^2 \|h\|_{t,w_t}^2 && \text{(by Lemma 12)} \\
&\ge F'_{t+1}(w_t) - \left\| \hat{\ell}_t - m_t + a_t \right\|_{t,w_t}^* \|h\|_{t,w_t} + \frac{1}{3} \|h\|_{t,w_t}^2 \\
&= F'_{t+1}(w_t) - \left\| \hat{\ell}_t - m_t + a_t \right\|_{t,w_t}^* + \frac{1}{3} && (\|h\|_{t,w_t} = 1) \\
&\ge F'_{t+1}(w_t). && \text{(by the assumption)}
\end{aligned}
$$

■

Lemma 14 *Let w_t, w'_{t+1} follow (1) and (2) where ψ_t is the log-barrier with $\eta_{t,i} \leq \frac{1}{81}$ for all i. If $\left\|\hat{\ell}_t - m_t + a_t\right\|^*_{t,w_t} \leq \frac{1}{3}$, then $\left\|w'_{t+1} - w_t\right\|_{t,w_t} \leq 3\left\|\hat{\ell}_t - m_t + a_t\right\|^*_{t,w_t}$.*

Proof Define $F_t(w)$ and $F'_{t+1}(w)$ to be the same as in Lemma 13. Then we have

$$\begin{aligned} F'_{t+1}(w_t) - F'_{t+1}(w'_{t+1}) &= (w_t - w'_{t+1})^\top(\hat{\ell}_t - m_t + a_t) + F_t(w_t) - F_t(w'_{t+1}) \\ &\leq (w_t - w'_{t+1})^\top(\hat{\ell}_t - m_t + a_t) \qquad \text{(optimality of } w_t\text{)} \\ &\leq \left\|w_t - w'_{t+1}\right\|_{t,w_t}\left\|\hat{\ell}_t - m_t + a_t\right\|^*_{t,w_t}. \end{aligned} \tag{13}$$

On the other hand, for some ξ on the line segment between w_t and w'_{t+1}, we have by Taylor's theorem and the optimality of w'_{t+1},

$$\begin{aligned} F'_{t+1}(w_t) - F'_{t+1}(w'_{t+1}) &= \nabla F'_{t+1}(w'_{t+1})^\top(w_t - w'_{t+1}) + \frac{1}{2}(w_t - w'_{t+1})^\top\nabla^2 F'_{t+1}(\xi)(w_t - w'_{t+1}) \\ &\geq \frac{1}{2}\left\|w_t - w'_{t+1}\right\|^2_{t,\xi}. \end{aligned} \tag{14}$$

Since the condition in Lemma 13 holds, $w'_{t+1} \in \mathcal{E}_{t,w_t}(1)$, and thus $\xi \in \mathcal{E}_{t,w_t}(1)$. Using again Lemma 12, we have

$$\frac{1}{2}\left\|w_t - w'_{t+1}\right\|^2_{t,\xi} \geq \frac{1}{3}\left\|w_t - w'_{t+1}\right\|^2_{t,w_t}. \tag{15}$$

Combining (13), (14), and (15), we have $\left\|w_t - w'_{t+1}\right\|_{t,w_t}\left\|\hat{\ell}_t - m_t + a_t\right\|^*_{t,w_t} \geq \frac{1}{3}\left\|w_t - w'_{t+1}\right\|^2_{t,w_t}$, which leads to the stated inequality. ∎

Lemma 15 *When the three conditions in Theorem 2 hold, we have $\left\|\hat{\ell}_t - m_t + a_t\right\|^*_{t,w_t} \leq \frac{1}{3}$ for either $a_{t,i} = 6\eta_{t,i}w_{t,i}(\hat{\ell}_{t,i} - m_{t,i})^2$ or $a_{t,i} = 0$.*

Proof For $a_{t,i} = 6\eta_{t,i}w_{t,i}(\hat{\ell}_{t,i} - m_{t,i})^2$, we have

$$\begin{aligned} \left\|\hat{\ell}_t - m_t + a_t\right\|^{*2}_{t,w_t} &= \sum_{i=1}^K \eta_{t,i}w_{t,i}^2\left(\hat{\ell}_{t,i} - m_{t,i} + 6\eta_{t,i}w_{t,i}(\hat{\ell}_{t,i} - m_{t,i})^2\right)^2 \\ &= \sum_{i=1}^K \eta_{t,i}w_{t,i}^2(\hat{\ell}_{t,i} - m_{t,i})^2 + 12\eta_{t,i}^2w_{t,i}^3(\hat{\ell}_{t,i} - m_{t,i})^3 + 36\eta_{t,i}^3w_{t,i}^4(\hat{\ell}_{t,i} - m_{t,i})^4 \\ &\leq \sum_{i=1}^K \eta_{t,i}w_{t,i}^2(\hat{\ell}_{t,i} - m_{t,i})^2(1 + 36\eta_{t,i} + 324\eta_{t,i}^2) \qquad \text{(condition (ii))} \\ &\leq 2\sum_{i=1}^K \eta_{t,i}w_{t,i}^2(\hat{\ell}_{t,i} - m_{t,i})^2 \qquad \text{(condition (i))} \\ &\leq 2 \times \frac{1}{18} = \frac{1}{9}. \qquad \text{(condition (iii))} \end{aligned}$$

For $a_{t,i} = 0$, we have

$$\left\|\hat{\ell}_t - m_t + a_t\right\|_{t,w_t}^{*2} = \left\|\hat{\ell}_t - m_t\right\|_{t,w_t}^{*2} = \sum_{i=1}^{K} \eta_{t,i} w_{t,i}^2 (\hat{\ell}_{t,i} - m_{t,i})^2 \le \frac{1}{18} < \frac{1}{9}. \quad \text{(condition (iii))}$$

■

Lemma 16 *If the three conditions in Theorem 2 hold,* BROAD-OMD *(with either Option I or II) satisfies* $\frac{1}{2} w_{t,i} \le w'_{t+1,i} \le \frac{3}{2} w_{t,i}$.

Proof This is a direct application of Lemmas 15, 13, and 12.

■

Lemma 17 *For the MAB problem, if the three conditions in Theorem 2 hold,* BROAD-OMD *(with either Option I or II) satisfies* $\frac{1}{2} w_{t,i} \le w'_{t,i} \le \frac{3}{2} w_{t,i}$.

Proof It suffices to prove $w'_t \in \mathcal{E}_{t,w_t}(1)$ by Lemma 12. Since we assume that the three conditions in Theorem 2 hold and $w_t \in \Delta_K$, we have $\|m_t\|_{t,w_t}^* = \sqrt{\sum_{i=1}^K \eta_{t,i} w_{t,i}^2 m_{t,i}^2} \le \sqrt{\frac{1}{162} \sum_{i=1}^K w_{t,i}^2} \le \sqrt{\frac{1}{162}} < \frac{1}{3}$. This implies $w'_t \in \mathcal{E}_{t,w_t}(1)$ by a similar arguments as in the proof of Lemma 13 (one only needs to replace $F'_{t+1}(w)$ there by $G(w) \triangleq D_{\psi_t}(w, w'_t)$ and note that $w'_t = \operatorname{argmin}_{w \in \Delta_K} G(w)$). ■

Appendix C. Proof of Theorem 2 and Corollary 3

Proof of Theorem 2. We first prove Eq. (3) holds: by Lemmas 15 and 14, we have

$$\begin{aligned}
\langle w_t - w'_{t+1}, \hat{\ell}_t - m_t + a_t \rangle &\le \left\|w_t - w'_{t+1}\right\|_{t,w_t} \left\|\hat{\ell}_t - m_t + a_t\right\|_{t,w_t}^* \\
&\le 3 \left\|\hat{\ell}_t - m_t + a_t\right\|_{t,w_t}^{*2} \\
&\le 3 \sum_{i=1}^{K} \eta_{t,i} w_{t,i}^2 (\hat{\ell}_{t,i} - m_{t,i})^2 (1 + 36\eta_{t,i} + 324\eta_{t,i}^2) \\
&\le 6 \sum_{i=1}^{K} \eta_{t,i} w_{t,i}^2 (\hat{\ell}_{t,i} - m_{t,i})^2 = \langle w_t, a_t \rangle,
\end{aligned}$$

where the last two inequalities are by the same calculations done in the proof of Lemma 15.

Since Eq. (3) holds, using Lemma 1 we have (ignoring non-positive terms $-A_t$'s),

$$\sum_{t=1}^{T} \langle w_t - u, \hat{\ell}_t \rangle \le \sum_{t=1}^{T} \left(D_{\psi_t}(u, w'_t) - D_{\psi_t}(u, w'_{t+1}) \right) + \sum_{t=1}^{T} \langle u, a_t \rangle$$

$$\leq D_{\psi_1}(u, w'_1) + \sum_{t=1}^{T} \left(D_{\psi_{t+1}}(u, w'_{t+1}) - D_{\psi_t}(u, w'_{t+1}) \right) + \sum_{t=1}^{T} \langle u, a_t \rangle. \tag{16}$$

In the last inequality, we add a term $D_{\psi_{T+1}}(u, w'_{T+1}) \geq 0$ artificially. As mentioned, ψ_{T+1}, defined in terms of $\eta_{T+1,i}$, never appears in the BROAD-OMD algorithm. We can simply pick any $\eta_{T+1,i} > 0$ for all i here. This is just to simplify some analysis later.

The first term in (16) can be bounded by the optimality of w'_1:

$$\begin{aligned} D_{\psi_1}(u, w'_1) &= \psi_1(u) - \psi_1(w'_1) - \langle \nabla \psi_1(w'_1), u - w'_1 \rangle \\ &\leq \psi_1(u) - \psi_1(w'_1) = \sum_{i=1}^{K} \frac{1}{\eta_{1,i}} \ln \frac{w'_{1,i}}{u_i}. \end{aligned}$$

The second term, by definition, is

$$\sum_{t=1}^{T} \sum_{i=1}^{K} \left(\frac{1}{\eta_{t+1,i}} - \frac{1}{\eta_{t,i}} \right) h \left(\frac{u_i}{w'_{t+1,i}} \right).$$

Plugging the above two terms into (16) finishes the proof. ∎

Proof of Corollary 3. We first check the three conditions in Theorem 2 under our choice of $\eta_{t,i}$ and $\hat{\ell}_{t,i}$: $\eta_{t,i} = \eta = \frac{1}{162K_0} \leq \frac{1}{162}$; $w_{t,i}|\hat{\ell}_{t,i} - m_{t,i}| = |\ell_{t,i} - m_{t,i}| \mathbb{1}\{i \in b_t\} \leq 2 < 3$; $\sum_{i=1}^{K} \eta_{t,i} w_{t,i}^2 (\hat{\ell}_{t,i} - m_{t,i})^2 = \frac{1}{162K_0} \sum_{i=1}^{K} (\ell_{t,i} - m_{t,i})^2 \mathbb{1}\{i \in b_t\} \leq \frac{4}{162} < \frac{1}{18}$. Applying Theorem 2 we then have

$$\sum_{t=1}^{T} \langle w_t - u, \hat{\ell}_t \rangle \leq \sum_{i=1}^{K} \frac{\ln \frac{w'_{1,i}}{u_i}}{\eta} + \sum_{t=1}^{T} \langle u, a_t \rangle.$$

As mentioned, if we let $u = b^*$, then $\ln \frac{w'_{1,i}}{u_i}$ becomes infinity for those $i \notin b^*$. Instead, we let $u = \left(1 - \frac{1}{T}\right) b^* + \frac{1}{T} w'_1$. With this choice of u, we have $\frac{w'_{1,i}}{u_i} \leq \frac{w'_{1,i}}{\frac{1}{T} w'_{1,i}} = T$. Plugging u into the above inequality and rearranging, we get

$$\sum_{t=1}^{T} \langle w_t - b^*, \hat{\ell}_t \rangle \leq \frac{K \ln T}{\eta} + \sum_{t=1}^{T} \langle b^*, a_t \rangle + B, \tag{17}$$

where $B \triangleq \frac{1}{T} \sum_{t=1}^{T} \langle -b^* + w'_1, \hat{\ell}_t + a_t \rangle$.

Now note that $\mathbb{E}_{b_t}[a_{t,i}] = 6\eta(\ell_{t,i} - m_{t,i})^2 = \mathcal{O}(\eta)$ and $\mathbb{E}_{b_t}[\hat{\ell}_{t,i}] = \ell_{t,i} = \mathcal{O}(1)$ for all i. Thus, $\mathbb{E}[B] = \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \langle -b^* + w'_1, \mathbb{E}_{b_t}[\hat{\ell}_t + a_t] \rangle \right] \leq \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \| -b^* + w'_1 \|_1 \left\| \mathbb{E}_{b_t}[\hat{\ell}_t + a_t] \right\|_\infty \right] = \mathcal{O}(K_0)$. Taking expectation on both sides of (17), we have

$$\mathbb{E}\left[\sum_{t=1}^{T} b_t^\top \ell_t - \sum_{t=1}^{T} b^{*\top} \ell_t \right] \leq \frac{K \ln T}{\eta} + 6\eta \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \in b^*}^{K} (\ell_{t,i} - m_{t,i})^2 \right] + \mathcal{O}(K_0).$$

∎

Appendix D. Proof of Theorem 4

Proof of Theorem 4. As in Hazan and Kale (2011a), for the rounds we perform uniform sampling we do not update w'_t. Let $\mathcal{S}$ be the set of rounds of uniform sampling. Then for the other rounds we can apply Corollary 3 to arrive at

$$\mathbb{E}\left[\sum_{t\in[T]\setminus\mathcal{S}} \ell_{t,i_t} - \ell_{t,i^*}\right] \leq \frac{K\ln T}{\eta} + 6\eta\mathbb{E}\left[\sum_{t\in[T]\setminus\mathcal{S}} (\ell_{t,i^*} - \tilde{\mu}_{t-1,i^*})^2\right] + \mathcal{O}(1). \tag{18}$$

The second term can be bounded as follows:

$$\begin{aligned}\mathbb{E}\left[\sum_{t\in[T]\setminus\mathcal{S}} (\ell_{t,i^*} - \tilde{\mu}_{t-1,i^*})^2\right] &\leq \mathbb{E}\left[\sum_{t=2}^{T} (\ell_{t,i^*} - \tilde{\mu}_{t-1,i^*})^2\right] \\ &\leq 3\sum_{t=2}^{T}(\ell_{t,i^*} - \mu_{t,i^*})^2 + 3\sum_{t=2}^{T}(\mu_{t,i^*} - \mu_{t-1,i^*})^2 + 3\mathbb{E}\left[\sum_{t=2}^{T}(\mu_{t-1,i^*} - \tilde{\mu}_{t-1,i^*})^2\right]. \end{aligned} \tag{19}$$

The first and the third terms in (19) can be bounded using Lemma 10 and 11 of (Hazan and Kale, 2011a) respectively, and they are both of order $\mathcal{O}(Q_{T,i^*}+1)$ if we pick $M = \Theta(\ln T)$. The second term in (19) can be bounded by a constant by Lemma 18. Thus second term in (18) can be bounded by $\mathcal{O}\left(\eta(Q_{T,i^*}+1)\right)$. Finally, note that $\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i_t} - \ell_{t,i^*}\right] \leq \mathbb{E}\left[\sum_{t\in[T]\setminus\mathcal{S}} \ell_{t,i_t} - \ell_{t,i^*}\right] + 2\mathbb{E}[|\mathcal{S}|]$ and that $\mathbb{E}[|\mathcal{S}|] = \mathcal{O}\left(\sum_{t=1}^{T}\frac{MK}{t}\right) = \mathcal{O}\left(MK\ln T\right) = \mathcal{O}\left(K(\ln T)^2\right)$. Combining everything, we get

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i_t} - \ell_{t,i^*}\right] = \mathcal{O}\left(\frac{K\ln T}{\eta} + \eta Q_{T,i^*} + K(\ln T)^2\right).$$

■

Lemma 18 *For any i, $\sum_{t=2}^{T}(\mu_{t,i} - \mu_{t-1,i})^2 = \mathcal{O}(1)$.*

Proof By definition, $|\mu_{t,i} - \mu_{t-1,i}| = \left|\frac{1}{t}\sum_{s=1}^{t}\ell_{s,i} - \frac{1}{t-1}\sum_{s=1}^{t-1}\ell_{s,i}\right| = \left|\frac{1}{t}\ell_{t,i} - \frac{1}{t(t-1)}\sum_{s=1}^{t-1}\ell_{s,i}\right| \leq \left|\frac{1}{t}\ell_{t,i}\right| + \left|\frac{1}{t(t-1)}\sum_{s=1}^{t-1}\ell_{s,i}\right| \leq \frac{2}{t}$. Therefore, $\sum_{t=2}^{T}(\mu_{t,i} - \mu_{t-1,i})^2 \leq \sum_{t=2}^{T}\frac{4}{t^2} = \mathcal{O}(1)$. ■

Appendix E. Proof of Theorem 5

We first state a useful lemma.

Lemma 19 *Let n_i be such that $\eta_{T+1,i} = \kappa^{n_i}\eta_{1,i}$, i.e., the number of times the learning rate of arm i changes in* BROAD-OMD+. *Then $n_i \leq \log_2 T$, and $\eta_{t,i} \leq 5\eta_{1,i}$ for all t, i.*

Proof Let $t_1, t_2, \ldots, t_{n_i} \in [T]$ be the rounds the learning rate for arm i changes (i.e., $\eta_{t+1,i} = \kappa\eta_{t,i}$ for $t = t_1, \ldots, t_{n_i}$). By the algorithm, we have

$$KT \geq \frac{1}{\bar{w}_{t_{n_i},i}} > \rho_{t_{n_i},i} > 2\rho_{t_{n_i-1},i} > \cdots > 2^{n_i-1}\rho_{t_1,i} = 2^{n_i}K.$$

Therefore, $n_i \leq \log_2 T$. And we have $\eta_{t,i} \leq \kappa^{\log_2 T}\eta_{1,i} = e^{\frac{\log_2 T}{\ln T}}\eta_{1,i} \leq 5\eta_{1,i}$. ■

Proof of Theorem 5. Again, we verify the three conditions stated in Theorem 2. By Lemma 19, $\eta_{t,i} \leq 5\eta \leq 5 \times \frac{1}{810} = \frac{1}{162}$; also, $w_{t,j}\left|\hat{\ell}_{t,j} - m_{t,j}\right| = w_{t,j}\left|\frac{(\ell_{t,j}-m_{t,j})\mathbb{1}\{i_t=j\}}{\bar{w}_{t,j}}\right| \leq w_{t,j}\left|\frac{2}{w_{t,j}\left(1-\frac{1}{T}\right)}\right| \leq 3$ because we assume $T \geq 3$; finally, $\sum_{j=1}^{K}\eta_{t,j}w_{t,j}^2(\hat{\ell}_{t,j} - m_{t,j})^2 = \eta_{t,i_t}w_{t,i_t}^2(\hat{\ell}_{t,i_t} - m_{t,i_t})^2 \leq \frac{1}{162} \times 3^2 = \frac{1}{18}$.

Let τ_j denote the last round the learning rate for arm j is updated, that is, $\tau_j \triangleq \max\{t \in [T] : \eta_{t+1,j} = \kappa\eta_{t,j}\}$. We assume that the learning rate is updated at least once so that τ_j is well defined, otherwise one can verify that the bound is trivial. For any arm i to compete with, let $u = \left(1-\frac{1}{T}\right)\mathbf{e}_i + \frac{1}{T}w'_1 = \left(1-\frac{1}{T}\right)\mathbf{e}_i + \frac{1}{KT}\mathbf{1}$, which guarantees $\frac{w'_{1,i}}{u_i} \leq T$. Applying Theorem 2, with $B \triangleq \frac{1}{T}\sum_{t=1}^{T}\langle -\mathbf{e}_i + w'_1, \hat{\ell}_t + a_t\rangle$ we have

$$\begin{aligned}
\sum_{t=1}^{T}\langle w_t, \hat{\ell}_t\rangle - \hat{\ell}_{t,i} &\leq \frac{K\ln T}{\eta} + \sum_{t=1}^{T}\sum_{j=1}^{K}\left(\frac{1}{\eta_{t+1,j}} - \frac{1}{\eta_{t,j}}\right)h\left(\frac{u_j}{w'_{t+1,j}}\right) + \sum_{t=1}^{T}a_{t,i} + B \\
&\leq \frac{K\ln T}{\eta} + \left(\frac{1}{\eta_{\tau_i+1,i}} - \frac{1}{\eta_{\tau_i,i}}\right)h\left(\frac{u_i}{w'_{\tau_i+1,i}}\right) + \sum_{t=1}^{T}a_{t,i} + B \\
&\leq \frac{K\ln T}{\eta} + \frac{1-\kappa}{\eta_{\tau_i+1,i}}h\left(\frac{u_i}{w'_{\tau_i+1,i}}\right) + \sum_{t=1}^{T}a_{t,i} + B \\
&\leq \frac{K\ln T}{\eta} - \frac{1}{5\eta\ln T}h\left(\frac{u_i}{w'_{\tau_i+1,i}}\right) + \sum_{t=1}^{T}a_{t,i} + B, \qquad (20)
\end{aligned}$$

where the last inequality is by Lemma 19 and the fact $\kappa - 1 \geq \frac{1}{\ln T}$. Now we bound the second and the third term in (20) separately.

1. For the second term, by Lemma 16 and $T \geq 3$ we have

$$\frac{u_i}{w'_{\tau_i+1,i}} \geq \frac{1-\frac{1}{T}}{\frac{3}{2}w_{\tau_i,i}} \geq \frac{\left(1-\frac{1}{T}\right)^2}{\frac{3}{2}\bar{w}_{\tau_i,i}} = \frac{\left(1-\frac{1}{T}\right)^2}{\frac{3}{2}} \times \frac{\rho_{T+1,i}}{2} \geq \frac{\rho_{T+1,i}}{8} \geq \frac{4K}{8} \geq 1.$$

Noting that $h(y)$ is an increasing function when $y \geq 1$, we thus have

$$h\left(\frac{u_i}{w'_{\tau_i+1,i}}\right) \geq h\left(\frac{\rho_{T+1,i}}{8}\right) = \frac{\rho_{T+1,i}}{8} - 1 - \ln\left(\frac{\rho_{T+1,i}}{8}\right) \geq \frac{\rho_{T+1,i}}{8} - 1 - \ln\left(\frac{KT}{4}\right). \qquad (21)$$

2. For the third term, we proceed as

$$\sum_{t=1}^{T} a_{t,i} = 6\sum_{t=1}^{T} \eta_{t,i} w_{t,i} (\hat{\ell}_{t,i} - m_{t,i})^2 \le 90\eta \sum_{t=1}^{T} |\hat{\ell}_{t,i} - m_{t,i}|$$
$$\le 90\eta \left(\max_{t\in[T]} \frac{1}{\bar{w}_{t,i}} \right) \sum_{t=1}^{T} |\ell_{t,i} - \ell_{t-1,i}| \le 90\eta \rho_{T+1,i} V_{T,i}, \tag{22}$$

where in the first inequality, we use $w_{t,i}|\hat{\ell}_{t,i} - m_{t,i}| \le 3$ and $\eta_{t,i} \le 5\eta$; in the second inequality, we do a similar calculation as in Eq. (6) (only replacing $w_{t,i}$ by $\bar{w}_{t,i}$); and in the last inequality, we use the fact $\frac{1}{\bar{w}_{t,i}} \le \rho_{T+1,i}$ for all $t \in [T]$ by the algorithm.

Combining Eq. (21) and Eq. (22) and using the fact $\frac{1+\ln\left(\frac{KT}{4}\right)}{5\ln T} \le K \ln T$, we continue from Eq. (20) to arrive at

$$\sum_{t=1}^{T} \langle w_t, \hat{\ell}_t \rangle - \hat{\ell}_{t,i} \le \frac{2K\ln T}{\eta} + \rho_{T+1,i} \left(\frac{-1}{40\eta \ln T} + 90\eta V_{T,i} \right) + B, \tag{23}$$

We are almost done here, but note that the left-hand side of (23) is not the desired regret. What we would like to bound is

$$\sum_{t=1}^{T} \langle \bar{w}_t, \hat{\ell}_t \rangle - \sum_{t=1}^{T} \hat{\ell}_{t,i} = \sum_{t=1}^{T} \langle \bar{w}_t - w_t, \hat{\ell}_t \rangle + \sum_{t=1}^{T} \left(\langle w_t, \hat{\ell}_t \rangle - \hat{\ell}_{t,i} \right), \tag{24}$$

where the second summation on the right-hand side is bounded by Eq. (23). The first term can be written as $\sum_{t=1}^{T} \langle -\frac{1}{T} w_t + \frac{1}{KT}\mathbf{1}, \hat{\ell}_t \rangle$. Note that $\frac{1}{T}\sum_{t=1}^{T} \langle -w_t, \hat{\ell}_t \rangle \le \frac{1}{T}\sum_{t=1}^{T} |\langle w_t, \hat{\ell}_t - m_t \rangle| + \frac{1}{T}\sum_{t=1}^{T} |\langle w_t, m_t \rangle| \le 3 + 1 = 4$, and $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \langle \frac{1}{K}\mathbf{1}, \hat{\ell}_t \rangle \right] = \frac{1}{T}\sum_{t=1}^{T} \langle \frac{1}{K}\mathbf{1}, \ell_t \rangle \le 1$. Therefore, taking expectation on both sides of (24), we get

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i_t} \right] - \sum_{t=1}^{T} \ell_{t,i} \le \frac{2K\ln T}{\eta} + \mathbb{E}[\rho_{T+1,i}] \left(\frac{-1}{40\eta \ln T} + 90\eta V_{T,i} \right) + \mathcal{O}(1),$$

because $\mathbb{E}[B]$ is also $\mathcal{O}(1)$ as proved in Corollary 3. ■

Appendix F. Proofs of Lemma 6 and Theorem 7

Proof of Lemma 6. By the same arguments as in the proof of Lemma 1, we have

$$\langle w'_{t+1} - u, \hat{\ell}_t \rangle \le D_{\psi_t}(u, w'_t) - D_{\psi_t}(u, w'_{t+1}) - D_{\psi_t}(w'_{t+1}, w'_t);$$

and

$$\langle w_t - w'_{t+1}, m_t \rangle \le D_{\psi_t}(w'_{t+1}, w'_t) - D_{\psi_t}(w'_{t+1}, w_t) - D_{\psi_t}(w_t, w'_t).$$

Therefore, by expanding the instantaneous regret, we have

$$\langle w_t - u, \hat{\ell}_t \rangle$$

$$= \langle w_t - w'_{t+1}, \hat{\ell}_t - m_t \rangle + \langle w'_{t+1} - u, \hat{\ell}_t \rangle + \langle w_t - w'_{t+1}, m_t \rangle$$
$$\le \langle w_t - w'_{t+1}, \hat{\ell}_t - m_t \rangle + D_{\psi_t}(u, w'_t) - D_{\psi_t}(u, w'_{t+1}) - D_{\psi_t}(w'_{t+1}, w_t) - D_{\psi_t}(w_t, w'_t).$$

■

Proof of Theorem 7. Applying Lemma 6, we have

$$\sum_{t=1}^{T} \langle w_t - u, \hat{\ell}_t \rangle \le \sum_{t=1}^{T} \left(D_{\psi_t}(u, w'_t) - D_{\psi_t}(u, w'_{t+1}) + \langle w_t - w'_{t+1}, \hat{\ell}_t - m_t \rangle - A_t \right)$$
$$\le \sum_{i=1}^{K} \frac{\ln \frac{w'_{1,i}}{u_i}}{\eta} + \sum_{t=1}^{T} \langle w_t - w'_{t+1}, \hat{\ell}_t - m_t \rangle - A_t.$$

For the second term, using Lemma 15 and 14 we bound $\langle w_t - w'_{t+1}, \hat{\ell}_t - m_t \rangle$ by

$$\left\| w_t - w'_{t+1} \right\|_{t,w_t} \left\| \hat{\ell}_t - m_t \right\|^*_{t,w_t} \le 3 \left\| \hat{\ell}_t - m_t \right\|^{*2}_{t,w_t} = 3\eta \sum_{i=1}^{K} w_{t,i}^2 (\hat{\ell}_{t,i} - m_{t,i})^2$$

Finally we lower bound A_t for the MAB case. Note $h(y) = y - 1 - \ln y \ge \frac{(y-1)^2}{6}$ for $y \in [\frac{1}{2}, 2]$. By Lemma 16 and 17, $\frac{w'_{t+1,i}}{w_{t,i}}$ and $\frac{w_{t,i}}{w'_{t,i}}$ both belong to $[\frac{1}{2}, 2]$. Therefore,

$$A_t = D_{\psi_t}(w'_{t+1}, w_t) + D_{\psi_t}(w_t, w'_t) = \frac{1}{\eta} \sum_{i=1}^{K} \left(h\left(\frac{w'_{t+1,i}}{w_{t,i}} \right) + h\left(\frac{w_{t,i}}{w'_{t,i}} \right) \right)$$
$$\ge \frac{1}{6\eta} \sum_{i=1}^{K} \left(\frac{(w'_{t+1,i} - w_{t,i})^2}{w_{t,i}^2} + \frac{(w_{t,i} - w'_{t,i})^2}{w_{t,i}^{\prime 2}} \right)$$
$$\ge \frac{1}{24\eta} \sum_{i=1}^{K} \left(\frac{(w'_{t+1,i} - w_{t,i})^2}{w_{t,i}^2} + \frac{(w_{t,i} - w'_{t,i})^2}{w_{t-1,i}^2} \right),$$

and

$$\sum_{t=1}^{T} A_t \ge \frac{1}{24\eta} \sum_{t=2}^{T} \sum_{i=1}^{K} \frac{(w'_{t,i} - w_{t-1,i})^2}{w_{t-1,i}^2} + \sum_{t=2}^{T} \sum_{i=1}^{K} \frac{(w_{t,i} - w'_{t,i})^2}{w_{t-1,i}^2} \ge \frac{1}{48\eta} \sum_{t=2}^{T} \sum_{i=1}^{K} \frac{(w_{t,i} - w_{t-1,i})^2}{w_{t-1,i}^2}.$$

■

Appendix G. Doubling Trick

We include the version of our algorithm with the doubling trick in Algorithm 3. For simplicity we still assume the time horizon T is known; the extension to unknown horizon is straightforward.

Proof of Theorem 8. Let $u = \left(1 - \frac{1}{T}\right) b^* + \frac{1}{T} w'_1$ so that $\ln \frac{w'_{1,i}}{u_i} \le \ln T$. At some epoch β, by Theorem 7, the break condition, and condition (iii) we have with $\eta_\beta \triangleq \frac{2^{-\beta}}{162 K_0}$,

$$\sum_{t=T_\beta+1}^{T_{\beta+1}} \langle w_t - u, \hat{\ell}_t \rangle \le \frac{K \ln T}{\eta_\beta} + 3\eta_\beta \sum_{t=T_\beta+1}^{T_{\beta+1}} \sum_{i=1}^{K} w_{t,i}^2 (\hat{\ell}_{t,i} - m_{t,i})^2$$

Algorithm 3 Doubling trick for BROAD-OMD with $a_t = \mathbf{0}$

Initialize: $\eta = \frac{1}{162K_0}, T_0 = 0, t = 1$.
for $\beta = 0, 1, \ldots$ **do**
  $w'_t = \operatorname{argmin}_{w\in\Omega} \psi_1(w)$ (restart BROAD-OMD).
  while $t \le T$ **do**
    Update w_t, sample $b_t \sim w_t$, and update w'_{t+1} as in BROAD-OMD with Option II.
    if $\sum_{s=T_\beta+1}^{t}\sum_{i=1}^{K} w_{s,i}^2(\hat{\ell}_{s,i} - m_{s,i})^2 \ge \frac{K\ln T}{3\eta^2}$ **then**
      $\eta \leftarrow \eta/2, T_{\beta+1} \leftarrow t, t \leftarrow t+1$.
      break.
    end
    $t \leftarrow t+1$.
  end
end

$$\le \frac{2K\ln T}{\eta_\beta} + 3\eta_\beta \sum_{i=1}^{K} w_{T_{\beta+1},i}^2(\hat{\ell}_{T_{\beta+1},i} - m_{T_{\beta+1},i})^2 = \mathcal{O}\left(\frac{K\ln T}{\eta_\beta}\right).$$

Suppose that at time T, the algorithm is at epoch $\beta = \beta^*$. Then we have

$$\sum_{t=1}^{T}\langle w_t - u, \hat{\ell}_t\rangle \le \sum_{\beta=0}^{\beta^*}\mathcal{O}\left(\frac{K\ln T}{\eta_\beta}\right) \le \sum_{\beta=0}^{\beta^*}\mathcal{O}\left(2^\beta K_0 K\ln T\right) \le \mathcal{O}\left(2^{\beta^*}K_0K\ln T\right).$$

It remains to bound β^*. If $\beta^* = 0$ (no restart ever happened), then trivially $\sum_{t=1}^{T}\langle w_t - u, \hat{\ell}_t\rangle = \mathcal{O}(K_0K\ln T)$. Otherwise, because epoch $\beta^* - 1$ finishes, we have

$$\sum_{t=T_{\beta^*-1}+1}^{T_{\beta^*}}\sum_{i=1}^{K} w_{t,i}^2(\hat{\ell}_{t,i} - m_{t,i})^2 \ge \frac{K\ln T}{3(\eta_{\beta^*-1})^2} = \Omega(2^{2\beta^*}K_0^2K\ln T).$$

Combining them, we have

$$\sum_{t=1}^{T}\langle w_t - u, \hat{\ell}_t\rangle \le \mathcal{O}\left(2^{\beta^*}K_0K\ln T\right) \le \mathcal{O}\left(\sqrt{(K\ln T)\sum_{t=T_{\beta^*-1}+1}^{T_{\beta^*}}\sum_{i=1}^{K} w_{t,i}^2(\hat{\ell}_{t,i} - m_{t,i})^2}\right)$$
$$\le \mathcal{O}\left(\sqrt{(K\ln T)\sum_{t=1}^{T}\sum_{i=1}^{K} w_{t,i}^2(\hat{\ell}_{t,i} - m_{t,i})^2}\right), \tag{25}$$

Combining both cases we have

$$\sum_{t=1}^{T}\langle w_t - u, \hat{\ell}_t\rangle \le \mathcal{O}\left(\sqrt{K\ln T\sum_{t=1}^{T}\sum_{i=1}^{K} w_{t,i}^2(\hat{\ell}_{t,i} - m_{t,i})^2} + K_0K\ln T\right). \tag{26}$$

Now substituting u by its definition and taking expectations, with $B \triangleq \frac{1}{T}\sum_{t=1}^{T}\langle -b^* + w_1', \hat{\ell}_t\rangle$ we arrive at

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle b_t - b^*, \ell_t\rangle\right] \le \mathcal{O}\left(\mathbb{E}\left[\sqrt{K\ln T\sum_{t=1}^{T}\sum_{i=1}^{K} w_{t,i}^2(\hat{\ell}_{t,i} - m_{t,i})^2}\right] + K_0 K\ln T\right) + \mathbb{E}[B]$$
$$\le \mathcal{O}\left(\sqrt{K\ln T\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} w_{t,i}^2(\hat{\ell}_{t,i} - m_{t,i})^2\right]} + K_0 K\ln T\right),$$

where the last inequality uses the fact $\mathbb{E}[B] = \mathcal{O}(K)$ and Jensen's inequality. ■

Appendix H. Proofs of Corollary 9 and Theorem 20

Proof of Corollary 9. We first verify the three conditions in Theorem 7: $\eta \le \frac{1}{162}$ by assumption; $w_{t,i}\left|\hat{\ell}_{t,i} - m_{t,i}\right| = \left|(\ell_{t,i} - \ell_{\alpha_i(t),i})\mathbb{1}\{i_t = i\}\right| \le 2 < 3$; $\eta\sum_{i=1}^{K} w_{t,i}^2(\hat{\ell}_{t,i} - m_{t,i})^2 = \eta w_{t,i_t}^2(\hat{\ell}_{t,i_t} - m_{t,i_t})^2 \le \frac{9}{162} = \frac{1}{18}$. Let $u = \left(1 - \frac{1}{T}\right)\mathbf{e}_{i^*} + \frac{1}{T}w_1'$, which guarantees $\frac{w_{1,i}'}{u_i} \le T$. By Theorem 7 and some rearrangement, we have

$$\sum_{t=1}^{T}\langle w_t - \mathbf{e}_{i^*}, \hat{\ell}_t\rangle \le \frac{K\ln T}{\eta} + 3\eta\sum_{t=1}^{T}\sum_{i=1}^{K} w_{t,i}^2(\hat{\ell}_{t,i} - m_{t,i})^2 - \sum_{t=1}^{T} A_t + B,$$

where $B \triangleq \frac{1}{T}\sum_{t=1}^{T}\langle -\mathbf{e}_{i^*} + w_1', \hat{\ell}_t\rangle$. To get the stated bound, just note that $\mathbb{E}[B] = \mathcal{O}(1)$, and replace $\sum_{t=1}^{T}\sum_{i=1}^{K} w_{t,i}^2(\hat{\ell}_{t,i} - m_{t,i})^2$ by the upper bound at (8) and A_t by the lower bound in Theorem 7. ■

Appendix I. Omitted Details in Section 4.1.1

Although the generalization to multi-player games is straightforward, for simplicity we only consider two-player zero-sum games.

We first describe the protocol of the game. The game is defined by an unknown matrix $G \in [-1,1]^{M\times N}$ where entry $G(i,j)$ specifies the loss (or reward) for Player 1 (or Player 2) if Player 1 picks row i while Player 2 picks column j. The players play the game repeatedly for T rounds. At round t, Player 1 randomly picks a row $i_t \sim x_t$ for some $x_t \in \Delta_M$ while Player 2 randomly picks a column $j_t \sim y_t$ for some $y_t \in \Delta_N$. In (Syrgkanis et al., 2015), the feedbacks they receive are the vectors Gy_t and $x_t^\top G$ respectively. As a natural extension to the bandit setting, we consider a setting where the feedbacks are the scalar values $\mathbf{e}_{i_t}^\top G y_t$ and $x_t^\top G\mathbf{e}_{j_t}$ respectively, that is, the expected loss/reward for the players' own realized actions (over the opponent's randomness).

It is clear that each player is essentially facing an MAB problem and thus can employ an MAB algorithm. Specifically, if both players apply Exp3 for example, their expected average strategies converge to a Nash equilibrium at rate $1/\sqrt{T}$. However, if instead Player 1 applies BROAD-OMD configured as in Corollary 9, then her regret has a path-length term that can be bounded as follows:

$$\sum_{i=1}^{K}\sum_{t=2}^{T}\left|\mathbf{e}_i^\top G y_t - \mathbf{e}_i^\top G y_{t-1}\right| \le \sum_{i=1}^{K}\sum_{t=2}^{T}\left\|\mathbf{e}_i^\top G\right\|_\infty \|y_t - y_{t-1}\|_1 \le K\sum_{t=2}^{T}\|y_t - y_{t-1}\|_1,$$

which is closely related to the negative regret term in Corollary 9 for Player 2 if she also employs the same BROAD-OMD. The cancellation of these terms then lead to faster convergence rate.

Theorem 20 *For the setting described above, if both players run* BROAD-OMD *configured as in Corollary 9 except that* $\eta_{t,i} = \eta = (M+N)^{-\frac{1}{4}} T^{-\frac{1}{4}}$*, then their expected average strategies converge to Nash equilibriums at the rate of* $\tilde{\mathcal{O}}\left((M+N)^{\frac{5}{4}}/T^{\frac{3}{4}}\right)$*, that is,*

$$\max_{y\in\Delta_N} \mathbb{E}[\bar{x}]^\top G y \le \text{Val} + \tilde{\mathcal{O}}((M+N)^{\frac{5}{4}}/T^{\frac{3}{4}}) \quad \textit{and} \quad \min_{x\in\Delta_M} x^\top G \mathbb{E}[\bar{y}] \ge \text{Val} - \tilde{\mathcal{O}}((M+N)^{\frac{5}{4}}/T^{\frac{3}{4}}),$$

where $\bar{x} = \frac{1}{T}\sum_{t=1}^T x_t, \bar{y} = \frac{1}{T}\sum_{t=1}^T y_t$ *and* $\text{Val} = \min_{x\in\Delta_M}\max_{y\in\Delta_N} x^\top G y = \max_{y\in\Delta_N}\min_{x\in\Delta_M} x^\top G y$.

Proof As mentioned, Player 1's $V_{T,i}$ is

$$\sum_{t=1}^T |\ell_{t,i} - \ell_{t-1,i}| = \sum_{t=1}^T |\mathbf{e}_i^\top G y_t - \mathbf{e}_i^\top G y_{t-1}| \le \sum_{t=1}^T \left\|\mathbf{e}_i^\top G\right\|_\infty \|y_t - y_{t-1}\|_1 \le \sum_{t=1}^T \|y_t - y_{t-1}\|_1$$

due to the assumption $|G(i,j)| \le 1$. Therefore, by Corollary 9, Player 1's (pseudo) regret is

$$\begin{aligned}
&\max_{x\in\Delta_M} \mathbb{E}\left[\sum_{t=1}^T x_t^\top G y_t - \sum_{t=1}^T x^\top G y_t\right] \\
&\le \mathcal{O}\left(\frac{M\ln T}{\eta}\right) + \mathbb{E}\left[6\eta M \sum_{t=1}^T \|y_t - y_{t-1}\|_1 - \frac{1}{48\eta}\sum_{t=2}^T\sum_{i=1}^M \frac{(x_{t,i} - x_{t-1,i})^2}{x_{t-1,i}^2}\right],
\end{aligned}$$

while Player 2's (pseudo) regret is

$$\begin{aligned}
&\max_{y\in\Delta_N} \mathbb{E}\left[\sum_{t=1}^T x_T^\top G y - \sum_{t=1}^T x_t^\top G y_t\right] \\
&\le \mathcal{O}\left(\frac{N\ln T}{\eta}\right) + \mathbb{E}\left[6\eta N \sum_{t=1}^T \|x_t - x_{t-1}\|_1 - \frac{1}{48\eta}\sum_{t=2}^T\sum_{i=1}^N \frac{(y_{t,i} - y_{t-1,i})^2}{y_{t-1,i}^2}\right].
\end{aligned}$$

Summing up the above two bounds, and using the following fact (by the inequality $a - b \le \frac{a^2}{4b}$):

$$\sum_{i=1}^N \left(6\eta M |y_{t,i} - y_{t-1,i}| - \frac{(y_{t,i} - y_{t-1,i})^2}{48\eta y_{t-1,i}^2}\right) \le 432\eta^3 M^2 \sum_{i=1}^N y_{t-1,i}^2 \le 432\eta^3 M^2,$$

we get

$$\max_{y\in\Delta_N} \mathbb{E}[\bar{x}]^\top G y - \min_{x\in\Delta_M} x^\top G \mathbb{E}[\bar{y}] = \mathcal{O}\left(\frac{(M+N)\ln T}{T\eta} + \eta^3(M^2+N^2)\right).$$

With $\eta = \tilde{\Theta}\left((M+N)^{-\frac{1}{4}} T^{-\frac{1}{4}}\right)$ the above bound becomes $\tilde{\mathcal{O}}\left((M+N)^{\frac{5}{4}} T^{-\frac{3}{4}}\right)$. Rearranging then gives

$$\max_{y\in\Delta_N} \mathbb{E}[\bar{x}]^\top G y \le \min_{x\in\Delta_M} x^\top G \mathbb{E}[\bar{y}] + \tilde{\mathcal{O}}((M+N)^{\frac{5}{4}} T^{-\frac{3}{4}}),$$

$$\leq \min_{x\in\Delta_M}\max_{y\in\Delta_N} x^\top Gy + \tilde{\mathcal{O}}((M+N)^{\frac{5}{4}}T^{-\frac{3}{4}}) = \text{Val} + \tilde{\mathcal{O}}((M+N)^{\frac{5}{4}}T^{-\frac{3}{4}}),$$

and similarly

$$\begin{aligned}\min_{x\in\Delta_M} x^\top G\mathbb{E}[\bar{y}] &\geq \max_{y\in\Delta_N} \mathbb{E}[\bar{x}]^\top Gy - \tilde{\mathcal{O}}((M+N)^{\frac{5}{4}}T^{-\frac{3}{4}}) \\ &\geq \max_{y\in\Delta_N}\min_{x\in\Delta_M} x^\top Gy - \tilde{\mathcal{O}}((M+N)^{\frac{5}{4}}T^{-\frac{3}{4}}) = \text{Val} - \tilde{\mathcal{O}}((M+N)^{\frac{5}{4}}T^{-\frac{3}{4}}),\end{aligned}$$

completing the proof. ■

As shown by the theorem, we obtain convergence rate faster than $1/\sqrt{T}$, but still slower than the $1/T$ rate compared to the full-information setup of (Rakhlin and Sridharan, 2013b; Syrgkanis et al., 2015), due to the fact that we only have first-order instead of second-order path-length bound.

Note that Rakhlin and Sridharan (2013b) also studies two-player zero-sum games with bandit feedback but with an unnatural restriction that in each round the players play the same strategy for four times. Foster et al. (2016) greatly weakened the restriction, but their algorithm only converges to some approximation of Val. For further comparisons, the readers are referred to the comparisons to (Syrgkanis et al., 2015) in (Foster et al., 2016). We also point out that the question raised in (Rakhlin and Sridharan, 2013b) remains open: if the players only receive the realized loss/reward $\mathbf{e}_{i_t}^\top G\mathbf{e}_{j_t}$ as feedback (a more natural setup), can the convergence rate to Val be faster than $1/\sqrt{T}$?

Appendix J. Proof of Theorem 10

Proof of Theorem 10. We first verify conditions (ii) and (iii) in Theorem 8 hold for $\hat{\ell}_{t,i} = \frac{\ell_{t,i}\mathbb{1}\{i_t=i\}}{w_{t,i}}$ and $m_{t,i} = \ell_{t,i_t}$. Indeed, condition (ii) holds since $w_{t,i}|\hat{\ell}_{t,i} - m_{t,i}| = |\ell_{t,i}\mathbb{1}\{i_t = i\} - w_{t,i}\ell_{t,i_t}| \leq 2 < 3$. Other the other hand, condition (iii) also holds because

$$\begin{aligned}\eta\sum_{i=1}^K w_{t,i}^2(\hat{\ell}_{t,i} - m_{t,i})^2 &= \eta\sum_{i=1}^K(\ell_{t,i}\mathbb{1}\{i_t=i\} - w_{t,i}\ell_{t,i_t})^2 \\ &= \eta\sum_{i=1}^K(\ell_{t,i}^2\mathbb{1}\{i_t=i\} - 2\ell_{t,i}w_{t,i}\ell_{t,i_t}\mathbb{1}\{i_t=i\} + w_{t,i}^2\ell_{t,i_t}^2) \\ &\leq \frac{1}{162}\left(\ell_{t,i_t}^2 - 2w_{t,i_t}\ell_{t,i_t}^2 + \left(\sum_{i=1}^K w_{t,i}^2\right)\ell_{t,i_t}^2\right) \\ &\leq \frac{1}{162}(1+0+1) < \frac{1}{18}.\end{aligned}$$

Thus, by Theorem 8, we have

$$\mathbb{E}\left[\sum_{t=1}^T \ell_{t,i_t} - \ell_{t,i^*}\right] = \mathcal{O}\left(\sqrt{(K\ln T)\mathbb{E}\left[\sum_{t=1}^T\sum_{i=1}^K w_{t,i}^2(\hat{\ell}_{t,i} - \ell_{t,i_t})^2\right]} + K\ln T\right). \tag{27}$$

Now we consider the stochastic setting. In this case, we further take expectations over $\ell_1, \ldots, \ell_T$ on both sides of (27). The left-hand side of (27) can be lower bounded by

$$\mathbb{E}\left[\sum_{t=1}^T \ell_{t,i_t} - \ell_{t,i^*}\right] = \mathbb{E}\left[\sum_{t=1}^T \ell_{t,i_t} - \min_j\sum_{t=1}^T \ell_{t,j}\right] \geq \mathbb{E}\left[\sum_{t=1}^T \ell_{t,i_t} - \sum_{t=1}^T \ell_{t,a^*}\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} w_{t,i}(\ell_{t,i} - \ell_{t,a^*})\right] \geq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\neq a^*} w_{t,i}\Delta\right] = \Delta\mathbb{E}\left[\sum_{t=1}^{T}(1 - w_{t,a^*})\right]. \tag{28}$$

On the other hand,

$$\begin{aligned}
&\mathbb{E}_{i_t\sim w_t}\left[\sum_{i=1}^{K} w_{t,i}^2(\hat{\ell}_{t,i} - \ell_{t,i_t})^2\right] = \mathbb{E}_{i_t\sim w_t}\left[\sum_{i=1}^{K} w_{t,i}^2\left(\frac{\ell_{t,i}\mathbb{1}\{i_t = i\}}{w_{t,i}} - \ell_{t,i_t}\right)^2\right] \\
&= \mathbb{E}_{i_t\sim w_t}\left[\sum_{i=1}^{K}\left(\ell_{t,i}\mathbb{1}\{i_t = i\} - w_{t,i}\ell_{t,i_t}\right)^2\right] \\
&= \sum_{i=1}^{K}\left(w_{t,i}\left(\ell_{t,i} - w_{t,i}\ell_{t,i}\right)^2 + \sum_{j\neq i} w_{t,j}(w_{t,i}\ell_{t,j})^2\right) \\
&\leq \sum_{i=1}^{K}\left(w_{t,i}\left(1 - w_{t,i}\right)^2 + \sum_{j\neq i} w_{t,j}w_{t,i}^2\right) = \sum_{i=1}^{K} w_{t,i}(1 - w_{t,i}) \\
&\leq (1 - w_{t,a^*}) + \sum_{i\neq a^*} w_{t,i} = 2(1 - w_{t,a^*}).
\end{aligned} \tag{29}$$

Therefore, the first term on the right-hand side of (27) can be upper bounded by

$$\sqrt{(K\ln T)\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} w_{t,i}^2(\hat{\ell}_{t,i} - \ell_{t,i_t})^2\right]} \leq \sqrt{(K\ln T)\mathbb{E}\left[\sum_{t=1}^{T} 2(1 - w_{t,a^*})\right]}. \tag{30}$$

Let $H = \mathbb{E}\left[\sum_{t=1}^{T}(1 - w_{t,a^*})\right]$. Combining (28), (30), and (27), we have

$$H\Delta \leq \mathcal{O}\left(\sqrt{(K\ln T)H} + K\ln T\right),$$

which implies $H = \mathcal{O}\left(\frac{K\ln T}{\Delta^2}\right)$. Therefore, the expected regret is upper bounded by

$$\mathcal{O}\left(\sqrt{(K\ln T)H} + K\ln T\right) = \mathcal{O}\left(\frac{K\ln T}{\Delta}\right).$$

For the adversarial setting, we continue from an intermediate step of (29):

$$\begin{aligned}
&\mathbb{E}_{i_t\sim w_t}\left[\sum_{i=1}^{K} w_{t,i}^2(\hat{\ell}_{t,i} - \ell_{t,i_t})^2\right] = \sum_{i=1}^{K}\left(w_{t,i}(1 - w_{t,i})^2\ell_{t,i}^2 + \sum_{j\neq i} w_{t,j}w_{t,i}^2\ell_{t,j}^2\right) \\
&\leq \sum_{i=1}^{K} w_{t,i}\ell_{t,i}^2 + \sum_{j=1}^{K}\sum_{i\neq j} w_{t,j}w_{t,i}^2\ell_{t,j}^2 \leq \sum_{i=1}^{K} w_{t,i}\ell_{t,i}^2 + \sum_{j=1}^{K} w_{t,j}\ell_{t,j}^2 = 2\mathbb{E}_{i_t\sim w_t}\left[\ell_{t,i_t}^2\right]
\end{aligned}$$

Assuming $\ell_{t,i} \in [0,1]$, we thus have $\ell_{t,i_t}^2 \leq \ell_{t,i_t}$ and

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i_t}\right] - \sum_{t=1}^{T} \ell_{t,i^*} = \mathcal{O}\left(\sqrt{(K \ln T)\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i_t}\right]} + K \ln T\right).$$

Solving for $\sqrt{\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i_t}\right]}$ and rearranging then give

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i_t}\right] - \sum_{t=1}^{T} \ell_{t,i^*} = \mathcal{O}\left(\sqrt{(K \ln T)\sum_{t=1}^{T} \ell_{t,i^*}} + K \ln T\right) = \mathcal{O}\left(\sqrt{K L_{T,i^*} \ln T} + K \ln T\right).$$

■