Proceedings of Machine Learning Research vol 99:1–21, 2019 32nd Annual Conference on Learning Theory

Improved Path-length Regret Bounds for Bandits

Sébastien Bubeck SEBUBECK@MICROSOFT.COM
Microsoft Research, Redmond

Yuanzhi Li YUANZHIL@STANFORD.EDU
Stanford University

Haipeng Luo HAIPENGL@USC.EDU
University of Southern California

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

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We study adaptive regret bounds in terms of the variation of the losses (the so-called path-length bounds) for both multi-armed bandit and more generally linear bandit. We first show that the seemingly suboptimal path-length bound of (Wei and Luo, 2018) is in fact not improvable for adaptive adversary. Despite this negative result, we then develop two new algorithms, one that strictly improves over (Wei and Luo, 2018) with a smaller path-length measure, and the other which improves over (Wei and Luo, 2018) for oblivious adversary when the path-length is large. Our algorithms are based on the well-studied optimistic mirror descent framework, but importantly with several novel techniques, including new optimistic predictions, a slight bias towards recently selected arms, and the use of a hybrid regularizer similar to that of (Bubeck et al., 2018).

Furthermore, we extend our results to linear bandit by showing a reduction to obtaining dynamic regret for a full-information problem, followed by a further reduction to convex body chasing. As a consequence we obtain new dynamic regret results as well as the first path-length regret bounds for general linear bandit.

Keywords: multi-armed bandit, linear bandit, path-length regret bound, optimistic mirror descent, dynamic regret, convex body chasing

1. Introduction

The multi-armed bandit (MAB) problem (Auer et al., 2002) is a classic online learning problem with partial information feedback. In the general adversarial environment, it is well known that $\Theta(\sqrt{KT})$ is the worst-case optimal regret bound where T is the number of rounds and K is the number of arms. Linear bandit generalizes MAB to learning linear loss functions with an arbitrary bounded convex set in $\mathbb{R}^d$, and it is also known that $\Theta(d\sqrt{T})$ is the worst-case optimal regret (Dani et al., 2008; Bubeck et al., 2012).

Despite these worst-case bounds, several works have studied more adaptive algorithms with data-dependent regret bounds that can be much smaller than the worst-case bounds under reasonable conditions. For example, recent work (Wei and Luo, 2018) proposes several such data-dependent regret bounds for MAB, including those that replace the dependence on T by the actual losses of the arms, the variance of the losses, or the variation of the losses measured by the so-called path-length.

Table 1: Main results and comparisons with previous works (see Section 1.2 for notation definition). For linear bandit, the upper bound with V_2 holds when the decision set is a 2-norm ball, and the lower bound with V_1 holds when the decision set is a max-norm ball.

		Oblivious Adversary	Adaptive Adversary
MAB	upper bound	$\sqrt{KV_\infty}$ (Theorem 2)	
	lower bound	$\sqrt{KV_\infty}$	
MAB	upper bound	$K^{\frac{1}{3}}\sqrt{V_1^{2/3}T^{1/3}}$ (Theorem 3)	$\sqrt{KV_1}$ (Wei and Luo, 2018)
	lower bound	$\sqrt{V_1}$	$\sqrt{KV_1}$ (Theorem 1)
Linear Bandit	upper bound	$d^{3/2}\sqrt{V_2}$ (Corollary 4 or 8) ; $d^2\sqrt{V_*}$ (Corollary 6)	
	lower bound	$d\sqrt{V_1}$ (Dani et al., 2008)	

In particular, since path-length is the smallest among these different measures, in this work we focus on extending and improving the existing path-length bounds for bandits. We start from a curious investigation on whether the bound $\widetilde{\mathcal{O}}(\sqrt{KV_1})$[1] of (Wei and Luo, 2018) can be improved, where $V_1 = \mathbb{E}[\sum_{t=2}^{T}\|\ell_t - \ell_{t-1}\|_1]$ is the 1-norm path-length and $\ell_1, \ldots, \ell_T$ are the loss vectors chosen by the adversary. Indeed, since V_1 can be as large as KT and $\Omega(\sqrt{KT})$ is a lower bound for MAB, it is very natural to ask whether one can improve the bound $\widetilde{\mathcal{O}}(\sqrt{KV_1})$ to $\widetilde{\mathcal{O}}(\sqrt{V_1})$.

Surprisingly, we show (in Theorem 1) that the bound $\widetilde{\mathcal{O}}(\sqrt{KV_1})$ is not improvable in general, at least not for an *adaptive adversary* who can pick the loss vectors based on the learner's previous actions. Despite this negative result, however, we also show the following two improvements:

- First, in Section 2.1 we propose a new algorithm with regret bound $\widetilde{\mathcal{O}}(\sqrt{KV_\infty})$ where $V_\infty = \mathbb{E}\big[\sum_{t=2}^{T}\|\ell_t - \ell_{t-1}\|_\infty\big]$ is the max-norm path-length. This is a strict improvement over (Wei and Luo, 2018) since $V_\infty \le V_1$, and moreover it is optimal even for oblivious adversary (that is, adversary who picks loss vectors independently of the learner's actions).
- Second, building on top of our first algorithm, in Section 2.2 we propose a more sophisticated algorithm with regret bound $\widetilde{\mathcal{O}}\Big(K^{\frac{1}{3}}\sqrt{V_1^{2/3}T^{1/3}}\Big)$ for oblivious adversary. This improves over (Wei and Luo, 2018) whenever $V_1 \ge T/K$. For example when $V_1 = T$, our bound becomes $\widetilde{\mathcal{O}}\Big(K^{\frac{1}{3}}\sqrt{T}\Big)$ while the one of (Wei and Luo, 2018) becomes the worst-case bound $\widetilde{\mathcal{O}}(\sqrt{TK})$. Note that in light of our aforementioned lower bound, this also shows a strict distinction between oblivious and adaptive adversary, which is uncommon in online learning.

Our algorithms are based on the optimistic mirror descent framework (Chiang et al., 2012; Rakhlin and Sridharan, 2013). However, several novel techniques are needed to achieve our results, including new optimistic predictions, a slight bias towards recently selected arms, and also a hybrid

1. We use the notation $\widetilde{\mathcal{O}}(\cdot)$ to suppress poly-logarithmic dependence on T.

regularizer. In particular, our second algorithm dynamically partitions the arms into two groups based on their probabilities of being selected, and applies different optimistic predictions and essentially different regularizers to these two groups. This new technique might be of independent interest.

Moreover, in Section 3 we further extend our results to general linear bandit and achieve a regret bound $\widetilde{\mathcal{O}}(d^2\sqrt{V_*})$ where $V_* = \mathbb{E}\big[\sum_{t=2}^{T} \|\ell_t - \ell_{t-1}\|_*\big]$ is the path-length measured by some arbitrary dual norm. When the decision set of the learner is a 2-norm ball, we also obtain an improved bound of order $\widetilde{\mathcal{O}}(d^{\frac{3}{2}}\sqrt{V_2})$ where $V_2 = \mathbb{E}\big[\sum_{t=2}^{T} \|\ell_t - \ell_{t-1}\|_2\big]$. Our algorithm is based on optimistic SCRiBLe (Abernethy et al., 2008; Hazan and Kale, 2011; Rakhlin and Sridharan, 2013) and the key challenge is to come up with the optimistic prediction under partial information feedback. We reduce this problem to obtaining dynamic regret (Zinkevich, 2003) for a full-information online learning problem, and then further reduce the latter to an instance of convex body chasing (Friedman and Linial, 1993). We discuss the implications of existing results through our reduction chain, and also propose a simple greedy approach for chasing convex sets with squared 2-norm, leading to the stated path-length bound $\widetilde{\mathcal{O}}(d^{\frac{3}{2}}\sqrt{V_2})$ for linear bandit.

Our main results are summarized in Table 1, where the two lower bounds without references are direct implications from known results as discussed in Section 2.

1.1. Related work

Path-length regret bounds were studied in (Chiang et al., 2012; Steinhardt and Liang, 2014) for full information problems, and in (Wei and Luo, 2018) for MAB and semi-bandit. Chiang et al. (2013) and Yang et al. (2016) also studies path-length bounds for a partial information setting under the easier two-point bandit feedback. Our result in Section 3 is the first path-length bound for general linear bandit as far as we know.

Dynamic regret bounds of (Besbes et al., 2015; Wei et al., 2016) for MAB are also expressed in terms of some path-length measure. However, the bound is much weaker compared to ours since dynamic regret is a stronger benchmark. For example, results of (Wei et al., 2016) only imply a bound $\mathcal{O}(\sqrt{KTV_\infty})$ for our problem, which is linear in T in the worst case.

Hybrid regularizer was first proposed by Bubeck et al. (2018) for sparse bandit and bandit with variance bound, and was also recently used in (Luo et al., 2018) for online portfolio. Our hybrid regularizer is similar to the one of (Bubeck et al., 2018) which is a combination of Shannon entropy and log-barrier, but importantly the weight for log-barrier is much larger than that of (Bubeck et al., 2018). The purpose of the hybrid regularizer and the role it plays in the analysis are also very different in all these three works.

As mentioned we show a strict distinction between oblivious and adaptive adversary, which is uncommon in online learning. The other two examples are online learning with switching costs (Cesa-Bianchi et al., 2013) and best-of-both-worlds results for MAB (Auer and Chiang, 2016).

1.2. Problem setup and notation

The multi-armed bandit problem proceeds for T rounds with $K \leq T$ fixed arms. In each round t, the learner selects one arm $i_t \in [K] \triangleq \{1, 2, \ldots, K\}$, and simultaneously the adversary decides the loss vector $\ell_t \in [0,1]^K$ where $\ell_{t,i}$ is the loss for arm i at time t. If ℓ_t is selected independently of the learner's previous actions $i_1, \ldots, i_{t-1}$, then the adversary is said to be *oblivious*; otherwise the

adversary is *adaptive*. In the end of round t, the learner suffers and observes the loss of the selected arm ℓ_{t,i_t}.

The learner's goal is to minimize her (pseudo) regret, which is the gap between her total loss and the loss of the best fixed arm, formally defined as

$$\mathrm{Reg} \triangleq \max_{i^\star \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i_t} - \sum_{t=1}^{T} \ell_{t,i^\star}\right]$$

where the expectation is with respect to the randomness of both the learner and the adversary.

We also consider the more general linear bandit problem, where the learner's decision set is an arbitrary convex compact set $\Omega \subset \mathbb{R}^d$. At each time t, the learner picks an action $w_t \in \Omega$ and simultaneously the adversary picks a linear loss function parametrized by $\ell_t \in \mathcal{L} \subset \mathbb{R}^d$. The learner suffers and observes the linear loss $\langle w_t, \ell_t \rangle$. Without loss of generality, we assume that Ω is contained in a unit ball $\mathcal{B} \triangleq \{z \in \mathbb{R}^d : \|z\| \leq 1\}$ for some arbitrary norm $\|\cdot\|$ and $\mathcal{L}$ is contained in the dual norm ball $\mathcal{B}_* \triangleq \{z \in \mathbb{R}^d : \|z\|_* \leq 1\}$ (thus the magnitude of the loss for any action is always bounded by 1). For linear bandit we consider general adaptive adversary so ℓ_t can depend on $w_1, \ldots, w_{t-1}$. The (pseudo) regret is defined in a similar way

$$\mathrm{Reg} \triangleq \max_{w^\star \in \Omega} \mathbb{E}\left[\sum_{t=1}^{T} \langle w_t, \ell_t \rangle - \sum_{t=1}^{T} \langle w^\star, \ell_t \rangle\right]$$

where again the expectation is with respect to the randomness of both the learner and the adversary.

As mentioned we study adaptive regret bounds that depend on the variation of the loss sequence $\ell_1, \ldots, \ell_T$, measured by it path-length $V_p = \mathbb{E}\left[\sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|_p\right]$ for some $p \geq 1$, where we define $\ell_0 = \mathbf{0}$ to be the all-zero vector and the expectation is taken with respect to the randomness of the adversary as well as the randomness of the learner in the case of adaptive adversary. In particular, we consider path-length V_1 and V_∞ for MAB and V_2 for linear bandit. Note $V_\infty \leq V_2 \leq V_1$ and also $V_1 \leq K V_\infty$. For linear bandit, we also consider path-length measured by a general dual norm and denote it by $V_* = \mathbb{E}\left[\sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|_*\right]$. For simplicity we assume that these quantities are known when tuning the optimal learning rate.

For MAB, we define $\rho_i(t) \triangleq \max\{s \leq t : i_s = i\}$ (or 0 if the set is empty) as the most recent time arm i is selected (prior to round $t+1$). We use $\boldsymbol{e}_i$ to denote the standard basis vector in K dimension with coordinate i being 1 and others being 0.

2. Path-length Bounds for Multi-armed Bandit

In this section we first show path-length lower bounds for MAB, followed by our proposed algorithms with new upper bounds.

First note that $\Omega(\sqrt{KV_\infty})$ is a trivial lower bound for oblivious adversary (and thus also for the more powerful adaptive adversary) in light of the standard $\sqrt{KT}$ lower bound construction for MAB. Indeed, for any $\gamma \in [K/T, 1]$ and any MAB algorithm, one can find a loss sequence with $V_\infty = \mathcal{O}(T\gamma)$ and $\mathrm{Reg} = \Omega(\sqrt{KT\gamma}) = \Omega(\sqrt{KV_\infty})$, just by using the standard lower bound construction (Auer et al., 2002) for a game with $T\gamma$ rounds as the first $T\gamma$ rounds, and setting all losses to be zero for the rest. Since $KV_\infty \geq V_1$, $\Omega(\sqrt{V_1})$ is clearly also a lower bound.

However, it turns out that for adaptive adversary, one can prove a stronger lower bound in terms of V_1, as shown in the following theorem.

Theorem 1 *For any $\gamma \in [K/T, 1]$ and any MAB algorithm, there exists an adaptively chosen sequence of ℓ_t such that $V_1 = \mathcal{O}(T\gamma)$ and* $\mathrm{Reg} = \Omega(\sqrt{KT\gamma}) = \Omega(\sqrt{KV_1})$.

Proof The construction of the loss sequence is as follows. First uniformly at random pick an arm $i^\star$ as the "good" arm. Then for each time t and each $i \in [K]$, set

$$\ell_{t,i} = \begin{cases} 0, & \text{if } t > T\gamma \\ \ell_{t-1,i}, & \text{else if } t > 1 \text{ and } i \neq i_{t-1}, \\ \text{a fresh sample drawn from } \mathrm{Ber}(0.5), & \text{else if } i \neq i^\star, \\ \text{a fresh sample drawn from } \mathrm{Ber}\left(0.5 - \frac{1}{4}\sqrt{\frac{K}{T\alpha}}\right), & \text{else.} \end{cases}$$

Note that the construction is so that only the first $T\gamma$ rounds matter clearly, and more importantly for these rounds the regret of the algorithm is exactly the same as in the case where fresh samples are drawn every time according to $\mathrm{Ber}\left(0.5 - \frac{1}{4}\sqrt{\frac{K}{T\gamma}}\right)$ for the good arm and $\mathrm{Ber}(0.5)$ for the others, because the loss of each arm is a fresh sample from the algorithm's perspective until this arm is picked and the loss is observed (in which case a new sample is drawn). Standard MAB lower bound proofs (see (Auer et al., 2002)) show that the regret in this case is $\Omega(\sqrt{KT\gamma})$. On the other hand, it is also clear that under this construction we have $V_1 = \mathcal{O}(T\gamma)$ since only one coordinate of the loss vector changes for each time $t \leq T\gamma$, which finishes the proof. ■

Theorem 1 shows that the algorithm of Wei and Luo (2018, Section 4.1) is optimal in terms of V_1 path-length bound for adaptive adversary. In the next two subsections we respectively show improvements in terms of V_∞ path-length and oblivious adversary.

2.1. Improved bounds in terms of V_∞

We propose a new algorithm that improves the result of Wei and Luo (2018) from $\widetilde{\mathcal{O}}(\sqrt{KV_1})$ to $\widetilde{\mathcal{O}}(\sqrt{KV_\infty})$ for both oblivious and adaptive adversary. Similar to (Wei and Luo, 2018), our algorithm is also based on the optimistic mirror descent framework (Chiang et al., 2012; Rakhlin and Sridharan, 2013). Specifically, optimistic mirror descent for general linear bandit over a decision set Ω maintains two sequences $x_1, \ldots, x_T$ and $x'_1, \ldots, x'_T$ based on the following update rules:

$$x_t = \operatorname*{argmin}_{x \in \Omega} \langle x, m_t \rangle + D_\psi(x, x'_t),$$
$$x'_{t+1} = \operatorname*{argmin}_{x \in \Omega} \langle x, \hat{\ell}_t \rangle + D_\psi(x, x'_t),$$

where m_t and $\hat{\ell}_t$ are respectively some optimistic prediction and unbiased estimator for the true loss vector ℓ_t, ψ is some convex differentiable regularizer and $D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle$ is the Bregman divergence with respect to ψ.

For MAB, $\Omega = \Delta_K$ is the simplex of distributions over K arms and $\hat{\ell}_t$ is usually set to the unbiased estimator with $\hat{\ell}_{t,i} = \frac{\ell_{t,i} - m_{t,i}}{w_{t,i}} \mathbf{1}\{i_t = i\} + m_{t,i}$ where the selected arm i_t is drawn from some final sample distribution w_t (computed based on x_t). Wei and Luo (2018) use the log-barrier $\psi(x) = \frac{1}{\eta}\sum_{i=1}^K \ln \frac{1}{x_i}$ with some learning rate η as the regularizer and $w_t = x_t$, and prove that the regret is bounded by (ignoring constants): $\frac{K \ln T}{\eta} + \eta \sum_{t=1}^T (\ell_{t,i_t} - m_{t,i_t})^2$. With $m_{t,i} = \ell_{\rho_i(t-1),i}$

Algorithm 1

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 $t = 1, 2, \ldots, T$ **do**
1 Play $i_t \sim w_t$ and observe $c_t = \ell_{t,i_t}$.
2 Construct unbiased estimator $\hat{\ell}_t$ s.t. $\hat{\ell}_{t,i} = \frac{\ell_{t,i} - c_{t-1}}{w_{t,i}} \mathbf{1}\{i_t = i\} + c_{t-1}$ for all i.
3 Update $x_{t+1} = \operatorname{argmin}_{x \in \Delta_K} \langle x, \hat{\ell}_t \rangle + D_\psi(x, x_t)$.
4 $w_{t+1} = (1 - \alpha_{t+1}) x_{t+1} + \alpha_{t+1} \boldsymbol{e}_{i_t}$, where $\alpha_{t+1} = \frac{\alpha(1-c_t)}{1+\alpha(1-c_t)}$.
end

(that is, the most recently observed loss for arm i), it is further shown that the regret bound above is at most $\widetilde{\mathcal{O}}(\sqrt{KV_1})$ with the optimal learning rate η.

Our algorithm makes the following two modifications (see Algorithm 1 for pseudocode). First, we simply use the observed loss at time t as the optimistic prediction for *all* arms at time $t+1$. Formally, we set $m_{t+1,i} = c_t \triangleq \ell_{t,i_t}$ for all i. Note that in this case $\langle x, m_t \rangle = c_{t-1}$ for any $x \in \Delta_K$ and thus $x_t = \operatorname{argmin}_{x \in \Delta_K} \langle x, m_t \rangle + D_\psi(x, x'_t) = x'_t$, meaning that we only need to maintain one sequence (line 3). Second, instead of using x_{t+1} to sample i_{t+1}, we slightly bias towards the most recently picked arm by moving a small fraction α_{t+1} of each arm's weight to arm i_t, where $\alpha_{t+1} = \frac{\alpha(1-c_t)}{1+\alpha(1-c_t)}$ for some fixed parameter α (line 4). Note that the smaller the loss of arm i_t is, the more we bias towards this arm, but the correlation is in some nonlinear form. Such bias is intuitive in a slowly changing environment where we expect a good arm remains reasonably good for a while. In the next theorem we formally prove the improved regret bound of our algorithm.

Theorem 2 *Algorithm 1 with* $\eta \le \frac{1}{162}$ *and* $\alpha = 8\eta$ *ensures*

$$\mathrm{Reg} = \mathcal{O}\left(\frac{K \ln T}{\eta} + \eta \mathbb{E}\left[\sum_{t=1}^{T-1} |\ell_{t+1,i_t} - \ell_{t,i_t}|\right]\right) = \mathcal{O}\left(\frac{K \ln T}{\eta} + \eta V_\infty\right)$$

for adaptive adversary. Picking the optimal η *leads to regret bound* $\mathcal{O}(\sqrt{KV_\infty \ln T} + K \ln T)$.

Proof We first analyze the regret of the sequence $x_1, \ldots, x_T$ using the analysis of (Wei and Luo, 2018). Specifically by their Theorem 7 and our choice of m_t and $\hat{\ell}_t$, we have for any arm i,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \langle x_t - \boldsymbol{e}_i, \ell_t \rangle\right] &\le \mathcal{O}\left(\frac{K \ln T}{\eta}\right) + 3\eta \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} x_{t,i}^2 (\hat{\ell}_{t,i} - c_{t-1})^2\right] \\
&= \mathcal{O}\left(\frac{K \ln T}{\eta}\right) + 3\eta \mathbb{E}\left[\sum_{t=1}^{T} \frac{x_{t,i_t}^2}{w_{t,i_t}^2} (c_t - c_{t-1})^2\right] \\
&\le \mathcal{O}\left(\frac{K \ln T}{\eta}\right) + 4\eta \mathbb{E}\left[\sum_{t=1}^{T} (c_t - c_{t-1})^2\right], \qquad (1)
\end{aligned}$$

where in the last step we use $\frac{x_{t,i}}{w_{t,i}} \leq \frac{1}{1-\alpha_t} \leq 1+\alpha = 1+8\eta \leq \sqrt{4/3}$ by our choice of η. In the rest of the proof we analyze the difference between using x_t and w_t. Specifically we prove

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle w_t - x_t, \ell_t\rangle\right] \leq \mathcal{O}(1) + \alpha\mathbb{E}\left[\sum_{t=2}^{T}|\ell_{t,i_{t-1}} - \ell_{t-1,i_{t-1}}| - \frac{1}{2}\sum_{t=2}^{T}(c_t - c_{t-1})^2\right], \tag{2}$$

which finishes the proof by combining the two inequalities above and using $\alpha = 8\eta$. Indeed, observe that for each time $t > 1$, we have by the definition of w_t

$$\mathbb{E}\left[\langle w_t - x_t, \ell_t\rangle\right] = \mathbb{E}\left[\alpha_t\langle \mathbf{e}_{i_{t-1}} - x_t, \ell_t\rangle\right] = \mathbb{E}\left[\alpha_t\langle \mathbf{e}_{i_{t-1}} - w_t, \ell_t\rangle\right] + \mathbb{E}\left[\alpha_t\langle w_t - x_t, \ell_t\rangle\right].$$

Rearranging and plugging the definition of α_t gives

$$\begin{aligned}
\mathbb{E}\left[\langle w_t - x_t, \ell_t\rangle\right] &= \mathbb{E}\left[\frac{\alpha_t}{1-\alpha_t}\langle \mathbf{e}_{i_{t-1}} - w_t, \ell_t\rangle\right] \\
&= \alpha\mathbb{E}\left[(1-c_{t-1})\langle \mathbf{e}_{i_{t-1}} - w_t, \ell_t\rangle\right] \\
&= \alpha\mathbb{E}\left[(1-c_{t-1})(\ell_{t,i_{t-1}} - c_t)\right] \\
&= \alpha\mathbb{E}\left[(1-c_{t-1})(\ell_{t,i_{t-1}} - c_{t-1} + c_{t-1} - c_t)\right] \\
&\leq \alpha\mathbb{E}\left[|\ell_{t,i_{t-1}} - c_{t-1}| + (1-c_{t-1})(c_{t-1} - c_t)\right] \\
&= \alpha\mathbb{E}\left[|\ell_{t,i_{t-1}} - c_{t-1}| + (c_{t-1} - c_t - c_{t-1}^2 + c_{t-1}c_t)\right].
\end{aligned}$$

Summing over t, and combining Eq. (1), we can bound the regret by (recall $c_{t-1} = \ell_{t-1,i_{t-1}}$)

$$\begin{aligned}
&\mathcal{O}\left(\frac{K\ln T}{\eta}\right) + 8\eta\mathbb{E}\left[\sum_{t=2}^{T}|\ell_{t,i_{t-1}} - \ell_{t-1,i_{t-1}}|\right] + \mathbb{E}\left[4\eta\sum_{t=2}^{T}(c_t - c_{t-1})^2 + 8\eta\sum_{t=2}^{T}(-c_{t-1}^2 + c_{t-1}c_t)\right] \\
&= \mathcal{O}\left(\frac{K\ln T}{\eta}\right) + 8\eta\mathbb{E}\left[\sum_{t=2}^{T}|\ell_{t,i_{t-1}} - \ell_{t-1,i_{t-1}}|\right], \qquad \text{(telescoping)}
\end{aligned}$$

which finishes the proof. ■

2.2. Improved bounds in terms of V_1 for oblivious adversary

Next we come back to V_1 path-length bound and show that despite the lower bound of Theorem 1 for adaptive adversary, one can still improve the regret for oblivious adversary when the path-length is large.

Our algorithm (see Algorithm 2 for the pseudocode) still follows the general optimistic mirror descent framework (line 5 and line 6). The novelty is that we divide all arms into two groups: the minority group $\mathcal{S}_t$ consisting of arms with weight $x_{t,i}$ smaller than some parameter β, and the majority group $[K]\backslash\mathcal{S}_t$.[2] At a high level our algorithm uses the same strategy as Algorithm 1 for the minority group and a different strategy for the majority group, discussed in detail below.

2. The concept of majority and minority groups is reminiscent of the recent work (Allen-Zhu et al., 2018) on first-order regret bounds for contextual bandits.

Algorithm 2

Define: $\psi(x) = \frac{1}{\eta}\sum_{i=1}^{K} \ln \frac{1}{x_i} + \frac{K}{\eta}\sum_{i=1}^{K} x_i \ln x_i$ for some learning rate η; parameters $\alpha, \beta \in (0,1)$.
Initialize: x'_1, x_1, w_1 are uniform distributions, $m_1 = \mathbf{0}$, and $\mathcal{S}_0 = [K]$.
for $t = 1, 2, \ldots, T$ **do**

1 Play $i_t \sim w_t$ and observe $c_t = \ell_{t,i_t}$.

2 Let $\tau(t) = \max\{\tau \le t : i_\tau \in \mathcal{S}_{\tau-1}\}$ and $\mathcal{S}_t = \{i \in [K] : x_{t,i} < \beta\}$.

3 Construct unbiased estimator $\hat{\ell}_t$ s.t. $\hat{\ell}_{t,i} = \frac{\ell_{t,i} - m_{t,i}}{w_{t,i}} \mathbf{1}\{i_t = i\} + m_{t,i}$ for all i.

4 Construct optimistic prediction $m_{t+1,i} = \begin{cases} c_{\tau(t)} & \text{if } i \in \mathcal{S}_t, \\ \ell_{\rho_i(t),i} & \text{else.} \end{cases}$

5 Update $x'_{t+1} = \operatorname{argmin}_{x \in \Delta_K} \langle x, \hat{\ell}_t \rangle + D_\psi(x, x'_t)$.

6 Update $x_{t+1} = \operatorname{argmin}_{x \in \Delta_K} \langle x, m_{t+1} \rangle + D_\psi(x, x'_{t+1})$.

7 Let $\alpha_{t+1} = \frac{\alpha(1-c_{\tau(t)})}{1+\alpha(1-c_{\tau(t)})}$ and w_{t+1} be such that

$$w_{t+1,i} = \begin{cases} x_{t+1,i}(1-\alpha_{t+1}) & \text{if } i \in \mathcal{S}_t \backslash \{i_{\tau(t)}\}, \\ x_{t+1,i} + \alpha_{t+1} \sum_{j \in \mathcal{S}_t \backslash \{i_\tau\}} x_{t+1,j} & \text{else if } i = i_{\tau(t)}, \\ x_{t+1,i} & \text{else.} \end{cases}$$

end

Optimistic prediction (line 4). For arms in the minority group $\mathcal{S}_t$, we use $m_{t+1,i} = c_{\tau(t)}$ as the optimistic prediction for time $t+1$, where $\tau(t) \triangleq \max\{\tau \le t : i_\tau \in \mathcal{S}_{\tau-1}\}$ is basically the most recent time we selected a minority arm, and $c_t \triangleq \ell_{t,i_t}$ is the loss of the algorithm at time t (same as Algorithm 1). On the other hand, for arms in the majority group $[K]\backslash\mathcal{S}_t$, just like (Wei and Luo, 2018) we use their most recently observed loss as the optimistic prediction, that is, $m_{t+1,i} = \ell_{\rho_i(t),i}$. This is very natural since intuitively majority arms are selected more often by definition and therefore their last observed loss could be a good proxy for the current loss.

Slight bias (line 7). Among the minority, we also bias towards the most recently picked one $i_{\tau(t)}$ by moving a fraction α_{t+1} of the weights of all arms in $\mathcal{S}_t \backslash \{i_{\tau(t)}\}$ to arm $i_{\tau(t)}$, just in the same way as Algorithm 1.[3] For the rest of the arms we simply set $w_{t+1,i} = x_{t+1,i}$.

Regularizer. Our algorithm uses a hybrid regularizer $\psi(x) = \frac{1}{\eta}\sum_{i=1}^{K} \ln \frac{1}{x_i} + \frac{K}{\eta}\sum_{i=1}^{K} x_i \ln x_i$. Roughly speaking, in our analysis we apply the log-barrier $\frac{1}{\eta}\sum_{i=1}^{K} \ln \frac{1}{x_i}$ to the minority group and the (negative) Shannon entropy $\frac{K}{\eta}\sum_{i=1}^{K} x_i \ln x_i$ to the majority group (see Lemma 11). This regularizer is similar to the one used by Bubeck et al. (2018). The difference is that in (Bubeck et al., 2018), the purpose of the hybrid regularizer is to ensure that the algorithm is stable (in the sense of Lemma 9), and for that purpose it suffices to set the coefficient of the log-barrier to be as small as K. On the other hand, we use a much larger coefficient $\frac{1}{\eta}$ for the log-barrier. In fact, this is in some sense the most natural choice since it leads to the smallest variance term for mirror descent while keeping the same regularization overhead as the entropy part (see Lemma 11 for details). Our analysis exactly exploits the smaller variance for the minority group due to this hybrid regularizer.

3. It is possible that $i_{\tau(t)}$ is not in the current minority group $\mathcal{S}_t$ though.

The next theorem shows the improved regret of Algorithm 2 (see Appendix B for the proof).

Theorem 3 *Algorithm 2 with* $\eta \le \min\left\{\frac{1}{K}, \frac{1}{162}\right\}$ *and* $\alpha = 8\eta$ *ensures*

$$\mathrm{Reg} = \mathcal{O}\left(\frac{K \ln T}{\eta} + \frac{\eta V_1}{K\beta} + \eta\sqrt{\beta T V_1}\right)$$

for oblivious adversary. Picking the optimal parameters leads to regret $\widetilde{\mathcal{O}}\left(K^{\frac{1}{3}}\sqrt{V_1^{2/3}T^{1/3}} + K^2\right)$.

The new regret bound is smaller than $\widetilde{\mathcal{O}}(\sqrt{KV_1})$ (Wei and Luo, 2018) as long as $V_1 \ge T/K$. This serves as a proof of concept that improved regret in terms of V_1 is possible for oblivious adversary, and we expect that even $\widetilde{\mathcal{O}}(\sqrt{V_1})$ is achievable using our techniques, although we do not have a simple algorithm achieving it yet.

3. Path-length Bounds for Linear Bandit

In this section we move on to the more general linear bandit problem. Recall that the decision sets of the learner and the adversary are assumed to be contained in a unit norm ball $\mathcal{B}$ and the dual norm ball $\mathcal{B}_*$ respectively. Our algorithm (see Algorithm 3 for the pseudocode) is based on the optimistic SCRiBLe algorithm (Abernethy et al., 2008; Hazan and Kale, 2011; Rakhlin and Sridharan, 2013) with new optimistic predictions.

Specifically, optimistic SCRiBLe is again an instance of the general optimistic mirror descent reviewed in Section 2.1. The regularizer is any ν-self-concordant barrier of the decision set Ω for some $\nu > 0$. Having the point x_t, the algorithm uniformly at random selects one of the $2d$ endpoints of the principal axes of the unit Dikin ellipsoid centered at x_t, as the final action w_t (Lines 1, 2 and 3). After observing the loss $c_t = \langle w_t, \ell_t\rangle$, the algorithm then constructs an unbiased loss estimator (line 4) and uses it in the next optimistic mirror descent update (line 6 and line 7, note that the learning rate η is explicitly spelled out here). We refer the readers to (Abernethy et al., 2008; Rakhlin and Sridharan, 2013) for more detailed explanation of the (optimistic) SCRiBLe algorithm.

For any optimistic prediction sequence $m_1, \ldots, m_T$, Rakhlin and Sridharan (2013) shows that the regret of optimistic SCRiBLe is bounded as

$$\mathrm{Reg} = \mathcal{O}\left(\frac{\nu \ln T}{\eta} + \eta d^2 \mathbb{E}\left[\sum_{t=1}^{T}\langle w_t, \ell_t - m_t\rangle^2\right]\right). \tag{3}$$

It remains to specify how to pick the optimistic predictions $m_1, \ldots, m_T$ such that the last term above $\sum_{t=1}^{T}\langle w_t, \ell_t - m_t\rangle^2$ is close to the path-length of the loss sequence $\ell_1, \ldots, \ell_T$. This is trivial in the full information setting where one observes ℓ_t at the end of round t and can simply set $m_t = \ell_{t-1}$. In the bandit setting, however, only $c_t = \langle w_t, \ell_t\rangle$ is observed and the problem becomes more challenging. In the next subsections we propose two approaches, one through a reduction to obtaining dynamic regret in an online learning problem with full information, and another via a further reduction to an instance of convex body chasing. As a side result, we obtain new dynamic regret bounds that may be of independent interest.

Algorithm 3

Define: $\psi(x)$ is a ν-self-concordant barrier; learning rate η.
Initialize: $x_1 = x'_1 = \operatorname{argmin}_{x\in\Omega} \psi(x)$ and $m_1 = \mathbf{0}$.
for $t = 1, 2, \ldots, T$ **do**

Compute eigendecomposition $\nabla^2\psi(x_t) = \sum_{i=1}^{d} \lambda_{t,i} v_{t,i} v_{t,i}^\top$.

Sample $i_t \in [d]$ and $\sigma_t \in \{-1, +1\}$ uniformly at random.

Play $w_t = x_t + \frac{\sigma_t}{\sqrt{\lambda_{t,i_t}}} v_{t,i_t}$ and observe $c_t = \langle w_t, \ell_t \rangle$.

Construct unbiased estimator $\hat{\ell}_t = d(c_t - \langle w_t, m_t \rangle)\sigma_t \sqrt{\lambda_{t,i_t}} v_{t,i_t} + m_t$.

Set m_{t+1} $\begin{cases} = \operatorname{Proj}_{\mathcal{B}}\left(m_t - \frac{1}{4}(\langle w_t, m_t \rangle - c_t) w_t\right) & \text{(Option I)} \\ = \operatorname{Proj}_{\mathcal{K}_{t+1}}(m_t) \text{ where } \mathcal{K}_{t+1} \triangleq \{m \in \mathcal{B} : \langle w_t, m \rangle = c_t\} & \text{(Option II)} \\ \text{via the convex body chasing algorithm of (Sellke, 2019)} & \text{(Option III)} \end{cases}$

Update $x'_{t+1} = \operatorname{argmin}_{x\in\Omega} \eta\langle x, \hat{\ell}_t \rangle + D_\psi(x, x'_t)$.

Update $x_{t+1} = \operatorname{argmin}_{x\in\Omega} \eta\langle x, m_{t+1} \rangle + D_\psi(x, x'_{t+1})$.

end

3.1. Reduction to dynamic regret

Rakhlin and Sridharan (2013) suggest treating the problem of selecting m_t as another online learning problem. Specifically, consider the following online learning formulation: at each time t the algorithm selects $m_t \in \mathcal{B}_*$ and then observes the loss function $f_t(m) = \langle w_t, \ell_t - m \rangle^2 = (c_t - \langle w_t, m \rangle)^2$. Note that this is a full information problem even though ℓ_t is unknown and is in fact the standard problem of online linear regression with squared loss. Further observe that applying Online Newton Step (Hazan et al., 2007) to learn m_t ensures

$$\sum_{t=1}^{T} f_t(m_t) \le \min_{m^\star \in \mathcal{B}_*} \sum_{t=1}^{T} f_t(m^\star) + \mathcal{O}(d \ln T) \le \min_{m^\star \in \mathcal{B}_*} \sum_{t=1}^{T} \|\ell_t - m^\star\|_*^2 + \mathcal{O}(d \ln T)$$

Picking $m^\star = \frac{1}{T}\sum_{s=1}^{T} \ell_s$ and combining the above with Eq. (3) immediately recover the main result of (Hazan and Kale, 2011) with a different approach. (This observation was not made in (Rakhlin and Sridharan, 2013) though.)

However, competing with a fixed $m^\star$ is not adequate for getting path-length bound. Instead in this case we need a *dynamic regret* bound (Zinkevich, 2003) that allows the algorithm to compete with some sequence $m_1^\star, \ldots, m_T^\star$ instead of a fixed $m^\star$. Typical dynamic regret bounds depend on either the variation of the loss functions or the competitor sequence (Jadbabaie et al., 2015; Mokhtari et al., 2016; Yang et al., 2016; Zhang et al., 2017, 2018), and here we need the latter one. Specifically, when $\mathcal{B}_*$ is the unit 2-norm ball, Yang et al. (2016) discover that projected gradient descent with a constant learning rate ensures for any minimizer sequence $m_1^\star \in \operatorname{argmin}_{m\in\mathcal{B}_*} f_1(m), \ldots, m_T^\star \in \operatorname{argmin}_{m\in\mathcal{B}_*} f_T(m)$,

$$\sum_{t=1}^{T} f_t(m_t) - \sum_{t=1}^{T} f_t(m_t^\star) \le \mathcal{O}\left(L \sum_{t=2}^{T} \left\|m_t^\star - m_{t-1}^\star\right\|_2\right) \tag{4}$$

as long as the following assumption holds:

Assumption 1 *Each f_t is convex and L-smooth (that is, for any $m, m' \in \mathcal{B}_*$, $f_t(m) \le f_t(m') + \langle \nabla f_t(m'), m - m' \rangle + \frac{L}{2} \|m - m'\|_2^2$). Additionally, $\nabla f_t(m^\star) = 0$ for any $m^\star \in \operatorname{argmin}_{m \in \mathcal{B}_*} f_t(m)$.*

It is clear that $f_t(m) = (c_t - \langle w_t, m \rangle)^2$ satisfies Assumption 1 with $L = 4$. Also note that Option I in line 5 is exactly doing projected gradient descent with f_t (we define $\operatorname{Proj}_{\mathcal{K}}(m) = \operatorname{argmin}_{m' \in \mathcal{K}} \|m - m'\|$). Therefore picking $m_t^\star = \ell_t$ and combining Eq. (4) and Eq. (3) immediately imply the following.

Corollary 4 *When $\mathcal{B}$ and $\mathcal{B}_*$ are unit 2-norm balls, Algorithm 3 with Option I ensures* $\mathrm{Reg} = \mathcal{O}\left(\frac{\nu \ln T}{\eta} + \eta d^2 \mathbb{E}\left[\sum_{t=1}^T \|\ell_t - \ell_{t-1}\|_2\right]\right)$, *which is of order $\widetilde{\mathcal{O}}\left(d\sqrt{\nu V_2}\right)$ with the optimal η.*

To deal with the case when $\mathcal{B}$ and $\mathcal{B}_*$ are arbitrary primal-dual norm balls, we require dynamic regret bounds that are similar to Eq. (4) but hold for an arbitrary norm. We are not aware of any such existing results. Instead, in the next section we provide a solution via a further reduction to convex body chasing.

3.2. Further reduction to convex body chasing

Next we provide an alternative approach to obtain dynamic regret similar to Eq. (4), via a reduction to convex body chasing (Friedman and Linial, 1993), which in turn leads to a different approach for obtaining path-length bound for linear bandit.

We first describe the general convex body chasing problem (overloading some of our notations for convenience). For each time $t = 1, \ldots, T$, the algorithm is presented with a convex set $\mathcal{K}_t$ in some metric space and then needs to select a point $m_t \in \mathcal{K}_t$. The algorithm performance is measured by the total movement cost $\sum_{t=1}^{T-1} \operatorname{dist}(m_t, m_{t+1})$ where $\operatorname{dist}(\cdot, \cdot)$ is some distance function (usually a metric). The algorithm is said to be ω-competitive if its total movement cost is at most $\omega \sum_{t=1}^{T-1} \operatorname{dist}(m_t^\star, m_{t+1}^\star)$ for any sequence $m_1^\star \in \mathcal{K}_1, \ldots, m_T^\star \in \mathcal{K}_T$.

Now for a sequence of convex functions $f_1, \ldots, f_T$ defined on $\mathcal{B}_*$ that are G-Lipschitz with respect to norm $\|\cdot\|_*$, suppose we have a ω-competitive algorithm for $\mathcal{K}_t = \operatorname{argmin}_{m \in \mathcal{B}_*} f_{t-1}(m)$ ($\mathcal{K}_1 = \mathcal{B}_*$) and $\operatorname{dist}(m, m') = \|m - m'\|_*$, to produce a sequence $m_1, \ldots, m_T$, then it holds for any minimizer sequence $m_1^\star \in \mathcal{K}_2, \ldots, m_T^\star \in \mathcal{K}_{T+1}$,

$$\sum_{t=1}^{T-1} f_t(m_t) - f_t(m_t^\star) = \sum_{t=1}^{T-1} f_t(m_t) - f_t(m_{t+1}) \le G \sum_{t=1}^{T-1} \|m_t - m_{t+1}\|_* \le G\omega \sum_{t=1}^{T-1} \left\|m_t^\star - m_{t+1}^\star\right\|_*,$$

where the first step is by the fact $m_t^\star, m_{t+1} \in \mathcal{K}_{t+1}$, the second step is due to G-Lipschitzness, and the last step is by ω-competitiveness. This is exactly a dynamic regret bound of the form Eq. (4), thus showing a reduction from dynamic regret to convex body chasing, under *only the Lipschitzness assumption* on f_t. Furthermore, recent work of (Sellke, 2019) shows that for any norm $\|\cdot\|_*$, there exists an algorithm with competitive ratio $\omega = d$. This immediately implies the following new dynamic regret result.

Proposition 5 *For an online convex optimization problem with convex loss functions $f_1, \ldots, f_T$ that are defined over a subset of $\mathcal{B}_*$ and are G-Lipschitz with respect to some norm $\|\cdot\|_*$, there exists an online learning algorithm that selects $m_1, \ldots, m_T$ such that*

$$\sum_{t=1}^{T} f_t(m_t) - \sum_{t=1}^{T} f_t(m_t^\star) \le \mathcal{O}\left(Gd \sum_{t=1}^{T-1} \left\|m_t^\star - m_{t+1}^\star\right\|_*\right)$$

for any minimizer sequence $m_1^\star \in \operatorname{argmin}_m f_1(m), \ldots, m_T^\star \in \operatorname{argmin}_m f_T(m)$.[4]

We note that this is the first dynamic regret bound in terms of the variation $\sum_{t=1}^{T-1} \left\|m_t^\star - m_{t+1}^\star\right\|_*$ for an arbitrary norm, without any explicit dependence on T, and under only the Lipschitzness assumption. Combining this result with Eq. (3) and picking $m_t^\star = \ell_t$ immediately imply the following path-length regret bound.

Corollary 6 *Algorithm 3 with Option III ensures* $\mathrm{Reg} = \mathcal{O}\left(\frac{\nu \ln T}{\eta} + \eta d^3 \mathbb{E}\left[\sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|_*\right]\right)$, *which is of order* $\widetilde{\mathcal{O}}\left(d^{3/2}\sqrt{\nu V_*}\right)$ *with the optimal* η.

While the result above holds for an arbitrary norm, it is $\sqrt{d}$ worse than that of Corollary 4 when the norm is 2-norm. Below we make another observation that when $\mathcal{B}_*$ is the 2-norm ball and when Assumption 1 holds, the problem in fact reduces to a slightly different convex body chasing problem that admits a constant competitive ratio in some sense. In particular, note that if $m_t^\star, m_{t+1} \in \mathcal{K}_{t+1}$ for all t, then by smoothness and $\nabla f_t(m_{t+1}) = 0$ it holds

$$\begin{aligned}
&\sum_{t=1}^{T-1} f_t(m_t) - f_t(m_t^\star) = \sum_{t=1}^{T-1} f_t(m_t) - f_t(m_{t+1}) \\
&\leq \sum_{t=1}^{T-1} \langle \nabla f_t(m_{t+1}), m_t - m_{t+1} \rangle + \frac{L}{2}\|m_t - m_{t+1}\|_2^2 = \frac{L}{2}\sum_{t=1}^{T-1} \|m_t - m_{t+1}\|_2^2 . \qquad (5)
\end{aligned}$$

Therefore, if we had a chasing algorithm with $\mathcal{K}_t$'s as the sets and *squared* 2-norm as the distance function, we would have a dynamic regret in terms of $\sum_{t=1}^{T-1} \left\|m_t^\star - m_{t+1}^\star\right\|_2^2$. It turns out that this is not possible in general. However, we propose a very natural greedy approach that is "competitive" in a slightly weaker sense where we measure the movement cost of the algorithm by squared 2-norm and the movement cost of the benchmark by 2-norm. More concretely we prove the following (see Appendix C for the proof):

Theorem 7 (Convex body chasing with squared 2-norm) *Suppose* $\mathcal{B}_*$ *is the unit 2-norm ball. Let* $\mathcal{K}_1, \ldots, \mathcal{K}_T \subset \mathcal{B}_*$ *be a sequence of convex sets and* $m_t = \mathrm{Proj}_{\mathcal{K}_t}(m_{t-1})$ *(with* $m_0 \in \mathcal{B}_*$ *being arbitrary). Then the following competitive guarantee holds for any sequence* $m_1^\star \in \mathcal{K}_1, \ldots, m_T^\star \in \mathcal{K}_T$*:* $\sum_{t=1}^{T-1} \|m_t - m_{t+1}\|_2^2 \leq 4 + 6\sum_{t=1}^{T-1} \left\|m_t^\star - m_{t+1}^\star\right\|_2$.

Combining Eq. (5) and Theorem 7 then recovers the dynamic regret bound of Eq. (4) with a different approach compared to (Yang et al., 2016) (under the same Assumption 1). Note that Option II of line 5 exactly implements the greedy projection approach of Theorem 7. Therefore according to the discussions in Section 3.1 we have:

Corollary 8 *When* $\mathcal{B}$ *and* $\mathcal{B}_*$ *are unit 2-norm balls, Algorithm 3 with Option II ensures* $\mathrm{Reg} = \mathcal{O}\left(\frac{\nu \ln T}{\eta} + \eta d^2 \mathbb{E}\left[\sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|_2\right]\right)$, *which is of order* $\widetilde{\mathcal{O}}\left(d\sqrt{\nu V_2}\right)$ *with the optimal* η.

It is well known that any convex body in d dimension admits an $\mathcal{O}(d)$-self-concordant barrier, and therefore Algorithm 3 admits a regret bound $\widetilde{\mathcal{O}}\left(d^{3/2}\sqrt{V_2}\right)$ or more generally $\widetilde{\mathcal{O}}\left(d^2\sqrt{V_*}\right)$.

4. According to (Sellke, 2019), when $\mathcal{B}_*$ is the 2-norm ball, the dependence on d can be improved to $\sqrt{d \ln T}$.

Acknowledgments

The authors would like to thank all the anonymous reviewers for their valuable comments. HL and CYW are supported by NSF Grant #1755781.

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.

Zeyuan Allen-Zhu, Sébastien Bubeck, and Yuanzhi Li. Make the minority great again: First-order regret bound for contextual bandits. In *International Conference on Machine Learning*, 2018.

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.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Sébastien Bubeck and Ronen Eldan. The entropic barrier: a simple and optimal universal self-concordant barrier. In *Conference on Learning Theory*, 2015.

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

Sébastien Bubeck, Michael B. Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, 2018.

Nicolo Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. Online learning with switching costs and other adaptive adversaries. In *Advances in Neural Information Processing Systems*, pages 1160–1168, 2013.

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 Francesco Orabona. Black-box reductions for parameter-free online learning in banach spaces. In *Conference on Learning Theory*, 2018.

Varsha Dani, Sham M Kakade, and Thomas P Hayes. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, pages 345–352, 2008.

Joel Friedman and Nathan Linial. On convex body chasing. *Discrete & Computational Geometry*, 9(3):293–321, 1993.

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

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Ali Jadbabaie, Alexander Rakhlin, Shahin Shahrampour, and Karthik Sridharan. Online optimization: Competing with dynamic comparators. In *Artificial Intelligence and Statistics*, pages 398–406, 2015.

Haipeng Luo, Chen-Yu Wei, and Kai Zheng. Efficient online portfolio with logarithmic regret. In *Advances in Neural Information Processing Systems*, 2018.

Aryan Mokhtari, Shahin Shahrampour, Ali Jadbabaie, and Alejandro Ribeiro. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. In *55th IEEE Conference on Decision and Control*, pages 7195–7201, 2016.

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

Mark Sellke. Chasing convex bodies optimally. *arXiv preprint arXiv:1905.11968*, 2019.

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

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference on Learning Theory*, 2018.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Tracking the best expert in non-stationary stochastic environments. In *Advances in Neural Information Processing Systems*, pages 3972–3980, 2016.

Tianbao Yang, Lijun Zhang, Rong Jin, and Jinfeng Yi. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *International Conference on Machine Learning*, pages 449–457, 2016.

Lijun Zhang, Tianbao Yang, Jinfeng Yi, Jing Rong, and Zhi-Hua Zhou. Improved dynamic regret for non-degenerate functions. In *Advances in Neural Information Processing Systems*, pages 732–741, 2017.

Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Adaptive online learning in dynamic environments. In *Advances in Neural Information Processing Systems*, pages 1330–1340, 2018.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, 2003.

Appendix A. Open Problems

In this work we provide several improvements on path-length bounds for bandit. There are several open problems in this direction and we hope that the techniques we develop here are useful for solving these problems. First, our upper bound for MAB with oblivious adversary has some dependence on T. Whether one can remove this dependence, and in particular, whether $\mathcal{O}(\sqrt{V_1})$ is achievable are clear open problems.

Second, all existing path-length bounds for bandit are "first-order", while smaller bounds in terms of "second-order" path-length $\mathbb{E}\left[\sum_{t=1}^{T}\|\ell_t-\ell_{t-1}\|_p^2\right]$ for some $p\geq 1$ are only known for full information problems. It is therefore natural to ask whether second-order path-length bounds are achievable for bandits, or there is a distinction here due to the partial information feedback.

Third, in light of the bound $\sqrt{d\sum_{t=1}^{T}\langle w^\star,\ell_t\rangle^2}$ of (Cutkosky and Orabona, 2018) for full information setting with linear losses (where $w^\star$ is the competitor the regret is with respect to), it is also very natural to ask if path-length bounds of the form $\mathrm{poly}(d)\sqrt{\sum_{t=1}^{T}|\langle w^\star,\ell_t-\ell_{t-1}\rangle|}$ are possible (for either full-information or bandit feedback).

Appendix B. Proof of Theorem 3

We outline the proof below and defer several technical lemmas to the next subsection.

Proof [of Theorem 3] Define $B_t=\mathbf{1}\{i_t\in\mathcal{S}_{t-1}\}$. Standard optimistic mirror descent analysis (see Lemma 11) shows that the hybrid regularizer ensures the following regret bound for the x_t sequence: for any arm i,

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle x_t-\boldsymbol{e}_i,\ell_t\rangle\right]\leq\mathcal{O}\left(\frac{K\ln T}{\eta}\right)+4\eta\mathbb{E}\left[\sum_{t:B_t=0}\frac{(\ell_{t,i_t}-m_{t,i_t})^2}{Kx_{t,i_t}}+\sum_{t:B_t=1}(\ell_{t,i_t}-m_{t,i_t})^2\right]. \tag{6}$$

Since $x_{t,i_t}\geq\frac{x_{t-1,i_t}}{2}$ (see Lemma 9) and $x_{t-1,i_t}\geq\beta$ when $B_t=0$, we have by the definition of m_t

$$\begin{aligned}\mathbb{E}\left[\sum_{t:B_t=0}\frac{(\ell_{t,i_t}-m_{t,i_t})^2}{Kx_{t,i_t}}\right]&\leq\mathbb{E}\left[\frac{2}{K\beta}\sum_{t:B_t=0}(\ell_{t,i_t}-m_{t,i_t})^2\right]=\mathbb{E}\left[\frac{2}{K\beta}\sum_{t:B_t=0}(\ell_{t,i_t}-\ell_{\rho_{i_t}(t-1),i_t})^2\right]\\&\leq\mathbb{E}\left[\frac{2}{K\beta}\sum_{t=1}^{T}|\ell_{t,i_t}-\ell_{\rho_{i_t}(t-1),i_t}|\right]=\mathbb{E}\left[\frac{2}{K\beta}\sum_{i=1}^{K}\sum_{t:i_t=i}|\ell_{t,i}-\ell_{\rho_i(t-1),i}|\right]\leq\frac{2V_1}{K\beta}.\end{aligned} \tag{7}$$

On the other hand we have by the definition of m_t, c_t and $\tau(t)$,

$$\mathbb{E}\left[\sum_{t:B_t=1}(\ell_{t,i_t}-m_{t,i_t})^2\right]=\mathbb{E}\left[\sum_{t:B_t=1}\left(c_t-c_{\tau(t-1)}\right)^2\right]=\mathbb{E}\left[\sum_{t=2}^{T}\left(c_{\tau(t)}-c_{\tau(t-1)}\right)^2\right]. \tag{8}$$

Next in Lemma 12 we show an analogue of Eq. (2) which bounds $\mathbb{E}\left[\sum_{t=1}^{T}\langle w_t-x_t,\ell_t\rangle\right]$, the difference between playing x_t and w_t, by

$$\mathcal{O}(1)+\alpha\mathbb{E}\left[\sum_{t:B_{t+1}=1}|\ell_{t+1,i_{\tau(t)}}-\ell_{\tau(t),i_{\tau(t)}}|\right]-\frac{\alpha}{2}\mathbb{E}\left[\sum_{t=2}^{T}(c_{\tau(t)}-c_{\tau(t-1)})^2\right]. \tag{9}$$

It remains to bound the second term above. Note that for any integer L between 1 and T, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t:B_{t+1}=1}|\ell_{t+1,i_{\tau(t)}}-\ell_{\tau(t),i_{\tau(t)}}|\right]\\
&=\mathbb{E}\left[\sum_{t:B_{t+1}=1,t+1-\tau(t)\geq L}|\ell_{t+1,i_{\tau(t)}}-\ell_{\tau(t),i_{\tau(t)}}|\right]+\mathbb{E}\left[\sum_{t:B_{t+1}=1,t+1-\tau(t)<L}|\ell_{t+1,i_{\tau(t)}}-\ell_{\tau(t),i_{\tau(t)}}|\right]\\
&\leq\frac{T}{L}+\mathbb{E}\left[\sum_{t:B_t=1}\sum_{s=t}^{\min\{t+L-1,T\}}|\ell_{s+1,i_t}-\ell_{s,i_t}|\right]\\
&\leq\frac{T}{L}+\mathbb{E}\left[\sum_{t}\sum_{s=t}^{\min\{t+L-1,T\}}\sum_{i\in\mathcal{S}_{t-1}}w_{t,i}|\ell_{s+1,i}-\ell_{s,i}|\right]\\
&\leq\frac{T}{L}+4\beta\mathbb{E}\left[\sum_{t}\sum_{s=t}^{\min\{t+L-1,T\}}\sum_{i\in\mathcal{S}_{t-1}}|\ell_{s+1,i}-\ell_{s,i}|\right]\\
&\leq\frac{T}{L}+4\beta LV_1.
\end{aligned}$$

Here, the first inequality is by the fact that there are at most T/L non-overlapping intervals with length at least L (for the first term) and triangle inequality (for the second term); the second inequality holds by taking the expectation of the indicator $B_t=1$ and the *obliviousness of the adversary* (the only place obliviousness is required); and the third inequality holds since by Lemma 9 $x_{t,i}\leq 2x_{t-1,i}$ which is at most 2β for all $i\in\mathcal{S}_{t-1}$ and thus $w_{t,i}\leq x_{t,i}+\alpha_t\sum_{j\in\mathcal{S}_{t-1}}x_{t,j}\leq 2\beta+2K\alpha_t\beta\leq 4\beta$ by the condition $\alpha_t\leq\alpha\leq\frac{1}{K}$. Picking the optimal L gives

$$\mathbb{E}\left[\sum_{t:B_{t+1}=1}|\ell_{t+1,i_{\tau(t)}}-\ell_{\tau(t),i_{\tau(t)}}|\right]=\mathcal{O}\left(\sqrt{\beta TV_1}\right). \tag{10}$$

Finally, combining Eq. (6), (7), (8), (9), and (10), and using $\alpha=8\eta$ proves the theorem. ■

B.1. Technical lemmas

Lemma 9 (Multiplicative Stability) *If $\eta\leq\min\left\{\frac{1}{K},\frac{1}{162}\right\}$, line 5 and line 6 of Algorithm 2 ensure $\max\left(\frac{x_{t+1,i}}{x_{t,i}},\frac{x_{t,i}}{x_{t+1,i}}\right)\leq 2$ for all t and i.*

To prove this lemma we make use of the following auxiliary result, where we use the notation $\|a\|_M=\sqrt{a^\top Ma}$ for a vector $a\in\mathbb{R}^K$ and a positive semi-definite matrix $M\in\mathbb{R}^{K\times K}$.

Lemma 10 *For some arbitrary $b_1,b_2\in\mathbb{R}^K$, $a_0\in\Delta_K$ and ψ defined as in Algorithm 2 with $\eta\leq\frac{1}{162}$, define*

$$\begin{cases}a_1=\operatorname{argmin}_{a\in\Delta_K}F_1(a), & \text{where } F_1(a)\triangleq\langle a,b_1\rangle+D_\psi(a,a_0),\\ a_2=\operatorname{argmin}_{a\in\Delta_K}F_2(a), & \text{where } F_2(a)\triangleq\langle a,b_2\rangle+D_\psi(a,a_0).\end{cases}$$

Then as long as $\|b_1 - b_2\|_{\nabla^{-2}\psi(a_1)} \le \frac{1}{9}$*, we have for all* $i \in [K]$, $\max\left\{\frac{a_{2,i}}{a_{1,i}}, \frac{a_{1,i}}{a_{2,i}}\right\} \le \frac{27}{26}$.

Proof [of Lemma 10] First, we prove $\|a_1 - a_2\|_{\nabla^2\psi(a_1)} \le \frac{1}{3}$ by contradiction. Assume $\|a_1 - a_2\|_{\nabla^2\psi(a_1)} > \frac{1}{3}$. Then there exists some a_2' lying in the line segment between a_1 and a_2 such that $\|a_1 - a_2'\|_{\nabla^2\psi(a_1)} = \frac{1}{3}$. By Taylor's theorem, there exists $\overline{a}$ that lies in the line segment between a_1 and a_2' such that

$$\begin{aligned}
F_2(a_2') &= F_2(a_1) + \langle \nabla F_2(a_1), a_2' - a_1\rangle + \frac{1}{2}\|a_2' - a_1\|^2_{\nabla^2 F_2(\overline{a})} \\
&= F_2(a_1) + \langle b_2 - b_1, a_2' - a_1\rangle + \langle \nabla F_1(a_1), a_2' - a_1\rangle + \frac{1}{2}\|a_2' - a_1\|^2_{\nabla^2\psi(\overline{a})} \\
&\ge F_2(a_1) - \|b_2 - b_1\|_{\nabla^{-2}\psi(a_1)}\|a_2' - a_1\|_{\nabla^2\psi(a_1)} + \frac{1}{2}\|a_2' - a_1\|^2_{\nabla^2\psi(\overline{a})} \\
&\ge F_2(a_1) - \frac{1}{9}\times\frac{1}{3} + \frac{1}{2}\|a_2' - a_1\|^2_{\nabla^2\psi(\overline{a})} \qquad (11)
\end{aligned}$$

where in the first inequality we use Hölder inequality and the first-order optimality condition, and in the last inequality we use the conditions $\|b_1 - b_2\|_{\nabla^{-2}\psi(a_1)} \le \frac{1}{9}$ and $\|a_1 - a_2'\|_{\nabla^2\psi(a_1)} = \frac{1}{3}$. Note that $\nabla^2\psi(x)$ is a diagonal matrix and $\nabla^2\psi(x)_{ii} = \frac{1}{\eta}\frac{1}{x_i^2} + \frac{K}{\eta}\frac{1}{x_i} \ge \frac{1}{\eta}\frac{1}{x_i^2}$. Therefore for any $i \in [K]$,

$$\frac{1}{3} = \|a_2' - a_1\|_{\nabla^2\psi(a_1)} \ge \sqrt{\sum_{j=1}^K \frac{(a_{2,j}' - a_{1,j})^2}{\eta a_{1,j}^2}} \ge \frac{|a_{2,i}' - a_{1,i}|}{\sqrt{\eta} a_{1,i}}$$

and thus $\frac{|a_{2,i}' - a_{1,i}|}{a_{1,i}} \le \frac{\sqrt{\eta}}{3} \le \frac{1}{27}$, which implies $\max\left\{\frac{a_{2,i}'}{a_{1,i}}, \frac{a_{1,i}}{a_{2,i}'}\right\} \le \frac{27}{26}$. Thus the last term in Eq. (11) can be lower bounded by

$$\begin{aligned}
\|a_2' - a_1\|^2_{\nabla^2\psi(\overline{a})} &= \frac{1}{\eta}\sum_{i=1}^K \left(\frac{1}{\overline{a}_i^2} + \frac{K}{\overline{a}_i}\right)(a_{2,i}' - a_{1,i})^2 \ge \frac{1}{\eta}\left(\frac{26}{27}\right)^2 \sum_{i=1}^K \left(\frac{1}{a_{1,i}^2} + \frac{K}{a_{1,i}}\right)(a_{2,i}' - a_{1,i})^2 \\
&\ge 0.9\|a_2' - a_1\|^2_{\nabla^2\psi(a_1)} = 0.9 \times \left(\frac{1}{3}\right)^2 = 0.1.
\end{aligned}$$

Combining with Eq. (11) gives

$$F_2(a_2') \ge F_2(a_1) - \frac{1}{27} + \frac{1}{2}\times 0.1 > F_2(a_1).$$

Recall that a_2' is a point in the line segment between a_1 and a_2. By the convexity of F_2, the above inequality implies $F_2(a_1) < F_2(a_2)$, contradicting the optimality of a_2.

Thus we conclude $\|a_1 - a_2\|_{\nabla^2\psi(a_1)} \le \frac{1}{3}$. Since $\|a_1 - a_2\|_{\nabla^2\psi(a_1)} \ge \frac{|a_{2,i} - a_{1,i}|}{\sqrt{\eta} a_{1,i}}$ for all i according to previous discussions, we get $\frac{|a_{2,i} - a_{1,i}|}{\sqrt{\eta} a_{1,i}} \le \frac{\sqrt{\eta}}{3} \le \frac{1}{27}$, which implies $\max\left\{\frac{a_{2,i}}{a_{1,i}}, \frac{a_{1,i}}{a_{2,i}}\right\} \le \frac{27}{26}$. ■

Proof [of Lemma 9] We prove the following two stability inequalities

$$\max\left\{\frac{x_{t,i}}{x_{t+1,i}'}, \frac{x_{t+1,i}'}{x_{t,i}}\right\} \le \frac{27}{26} \qquad (12)$$

$$\max\left\{\frac{x_{t+1,i}}{x'_{t+1,i}}, \frac{x'_{t+1,i}}{x_{t+1,i}}\right\} \leq \frac{27}{26}, \tag{13}$$

which clearly implies the lemma since then $\max\left\{\frac{x_{t,i}}{x_{t+1,i}}, \frac{x_{t+1,i}}{x_{t,i}}\right\} \leq \frac{27}{26} \times \frac{27}{26} \leq 2$. To prove Eq. (12), observe that

$$\begin{cases} x_t = \operatorname{argmin}_{x\in\Delta_K} \langle x, m_t\rangle + D_\psi(x, x'_t), \\ x'_{t+1} = \operatorname{argmin}_{x\in\Delta_K} \langle x, \hat{\ell}_t\rangle + D_\psi(x, x'_t). \end{cases} \tag{14}$$

To apply Lemma 10 and obtain Eq. (12), we only need to show $\|\hat{\ell}_t - m_t\|_{\nabla^{-2}\psi(x_t)} \leq \frac{1}{9}$. Recall $\nabla^2\psi(u)_{ii} = \frac{1}{\eta}\frac{1}{u_i^2} + \frac{K}{\eta}\frac{1}{u_i}$. By our algorithm we have

$$\begin{aligned} \|\hat{\ell}_t - m_t\|^2_{\nabla^{-2}\psi(x_t)} &\leq \sum_{i=1}^K \eta x_{t,i}^2 \left(\frac{(\ell_{t,i} - m_{t,i})\mathbf{1}\{i_t = i\}}{w_{t,i}}\right)^2 \\ &\leq \sum_{i=1}^K \frac{\eta}{(1-\alpha)^2} x_{t,i}^2 \left(\frac{(\ell_{t,i} - m_{t,i})\mathbf{1}\{i_t = i\}}{x_{t,i}}\right)^2 \qquad (\tfrac{x_{t,i}}{w_{t,i}} \leq \tfrac{1}{1-\alpha_t} \leq \tfrac{1}{1-\alpha}\ \forall i) \\ &\leq \frac{\eta}{(1-\alpha)^2} \leq \frac{\frac{1}{162}}{(1-\frac{8}{162})^2} < \frac{1}{81}, \end{aligned}$$

finishing the proof for Eq. (12). To prove Eq. (13), we observe:

$$\begin{cases} x'_{t+1} = \operatorname{argmin}_{x\in\Delta_K} D_\psi(x, x'_{t+1}), \\ x_{t+1} = \operatorname{argmin}_{x\in\Delta_K} \langle x, m_{t+1}\rangle + D_\psi(x, x'_{t+1}). \end{cases} \tag{15}$$

Similarly, with the help of Lemma 10, we only need to show $\|m_{t+1}\|_{\nabla^{-2}\psi(x'_{t+1})} \leq \frac{1}{9}$. This can be seen by

$$\|m_{t+1}\|^2_{\nabla^{-2}\psi(x'_{t+1})} \leq \sum_{i=1}^K \eta x'^2_{t+1,i} m^2_{t+1,i} \leq \eta \leq \frac{1}{81}.$$

This finishes the proof. ∎

Lemma 11 *If $\eta \leq \min\left\{\frac{1}{K}, \frac{1}{162}\right\}$, line 5 and line 6 of Algorithm 2 ensure for any arm i^*,*

$$\mathbb{E}\left[\sum_{t=1}^T \langle x_t - \boldsymbol{e}_{i^*}, \ell_t\rangle\right] \leq \mathcal{O}\left(\frac{K\ln T}{\eta}\right) + 4\eta\mathbb{E}\left[\sum_{t:i_t\notin\mathcal{S}_{t-1}} \frac{(\ell_{t,i_t} - m_{t,i_t})^2}{K x_{t,i_t}} + \sum_{t:i_t\in\mathcal{S}_{t-1}} (\ell_{t,i_t} - m_{t,i_t})^2\right].$$

Proof We will in fact prove a stronger statement

$$\mathbb{E}\left[\sum_{t=1}^T \langle x_t - \boldsymbol{e}_{i^*}, \ell_t\rangle\right] \leq \mathcal{O}\left(\frac{K\ln T}{\eta}\right) + 4\eta\mathbb{E}\left[\sum_{t=1}^T \min\left\{\frac{(\ell_{t,i_t} - m_{t,i_t})^2}{K x_{t,i_t}}, (\ell_{t,i_t} - m_{t,i_t})^2\right\}\right],$$

which clearly implies the stated bound.

By standard analysis of optimistic mirror descent (e.g., Lemma 6 in (Wei and Luo, 2018), Lemma 5 in (Chiang et al., 2012)), we have

$$\langle x_t - u, \hat{\ell}_t \rangle \le D_\psi(u, x'_t) - D_\psi(u, x'_{t+1}) + \langle x_t - x'_{t+1}, \hat{\ell}_t - m_t \rangle \tag{16}$$

for all $u \in \Delta_K$. Specifically, we pick $u = \left(1 - \frac{1}{T}\right) \boldsymbol{e}_{i^*} + \frac{1}{KT}\mathbf{1}$. The first two terms on the right hand side of Eq. (16) telescope when summing over t. The non-negative remaining term is

$$\begin{aligned} D_\psi(u, x'_1) &= \psi(u) - \psi(x'_1) - \langle \nabla\psi(x'_1), u - x'_1 \rangle \\ &= \psi(u) - \psi(x'_1) \le \frac{K \ln T}{\eta} + \frac{K \ln K}{\eta} = \mathcal{O}\left(\frac{K \ln T}{\eta}\right), \end{aligned}$$

where in the first equality we use $x'_1 = \frac{1}{K}\mathbf{1}$ and ψ's definition. Below we focus on the last term in Eq. (16). Acoording to line 5 and line 6 of Algorithm 2, we can write

$$\begin{cases} x_t = \operatorname{argmin}_{x \in \Delta_K} F(x), & \text{where } F(x) \triangleq \langle x, m_t \rangle + D_\psi(x, x'_t), \\ x'_{t+1} = \operatorname{argmin}_{x \in \Delta_K} F'(x), & \text{where } F'(x) \triangleq \langle x, \hat{\ell}_t \rangle + D_\psi(x, x'_t). \end{cases} \tag{17}$$

By Taylor's theorem, there exists some $\overline{x}$ that lies in the line segment between x_t and x'_{t+1}, such that

$$F'(x_t) - F'(x'_{t+1}) = \langle \nabla F'(x'_{t+1}), x_t - x'_{t+1} \rangle + \frac{1}{2}\|x_t - x'_{t+1}\|^2_{\nabla^2 F'(\overline{x})} \ge \frac{1}{2}\|x_t - x'_{t+1}\|^2_{\nabla^2\psi(\overline{x})}, \tag{18}$$

where the last inequality uses the optimality of x'_{t+1} and that $\nabla^2 F' = \nabla^2\psi$. As shown in the proof of Lemma 9, $\max\left\{\frac{x'_{t+1,i}}{x_{t,i}}, \frac{x_{t,i}}{x'_{t+1,i}}\right\} \le \frac{27}{26}$, so

$$\begin{aligned} \frac{1}{2}\|x_t - x'_{t+1}\|^2_{\nabla^2\psi(\overline{x})} &= \frac{1}{2}\sum_{i=1}^K \frac{1}{\eta}\left(\frac{1}{\overline{x}_i^2} + \frac{K}{\overline{x}_i}\right)(x_{t,i} - x'_{t+1,i})^2 \\ &\ge \frac{1}{2}\left(\frac{26}{27}\right)^2 \sum_{i=1}^K \frac{1}{\eta}\left(\frac{1}{x_{t,i}^2} + \frac{K}{x_{t,i}}\right)(x_{t,i} - x'_{t+1,i})^2 \\ &\ge \frac{0.9}{2}\|x_t - x'_{t+1}\|^2_{\nabla^2\psi(x_t)} \end{aligned} \tag{19}$$

On the other hand,

$$\begin{aligned} F'(x_t) - F'(x'_{t+1}) &= \langle x_t - x'_{t+1}, \hat{\ell}_t - m_t \rangle + F(x_t) - F(x'_{t+1}) \\ &\le \langle x_t - x'_{t+1}, \hat{\ell}_t - m_t \rangle \qquad \text{(by the optimality of } x_t\text{)} \\ &\le \|x_t - x'_{t+1}\|_{\nabla^2\psi(x_t)} \|\hat{\ell}_t - m_t\|_{\nabla^{-2}\psi(x_t)}. \end{aligned} \tag{20}$$

Combining Eq. (18), Eq. (19) and Eq. (20) we get

$$\|x_t - x'_{t+1}\|_{\nabla^2\psi(x_t)} \le \frac{2}{0.9}\|\hat{\ell}_t - m_t\|_{\nabla^{-2}\psi(x_t)},$$

and thus

$$\begin{aligned}
&\langle x_t - x'_{t+1}, \hat{\ell}_t - m_t \rangle \\
&\le \|x_t - x'_{t+1}\|_{\nabla^2 \psi(x_t)} \|\hat{\ell}_t - m_t\|_{\nabla^{-2} \psi(x_t)} \le 3 \|\hat{\ell}_t - m_t\|^2_{\nabla^{-2} \psi(x_t)} \\
&\le 3\eta \sum_{i=1}^K \frac{1}{\frac{1}{x_{t,i}^2} + \frac{K}{x_{t,i}}} (\hat{\ell}_{t,i} - m_{t,i})^2 \\
&\le 3\eta \sum_{i=1}^K \min\left\{ x_{t,i}^2, \frac{x_{t,i}}{K} \right\} (\hat{\ell}_{t,i} - m_{t,i})^2 \\
&\le 3 \frac{\eta}{(1-\alpha)^2} \sum_{i=1}^K \min\left\{ x_{t,i}^2, \frac{x_{t,i}}{K} \right\} \left(\frac{(\ell_{t,i} - m_{t,i}) \mathbf{1}\{i_t = i\}}{x_{t,i}} \right)^2 && (\tfrac{x_{t,i}}{w_{t,i}} \le \tfrac{1}{1-\alpha_t} \le \tfrac{1}{1-\alpha} \ \forall i) \\
&\le 4\eta \min\left\{ (\ell_{t,i_t} - m_{t,i_t})^2, \frac{\ell_{t,i_t} - m_{t,i_t})^2}{K x_{t,i_t}} \right\}. && (\alpha = 8\eta \le \tfrac{8}{162})
\end{aligned}$$

We have thus showed

$$\sum_{t=1}^T \langle w_t - u, \hat{\ell}_t \rangle \le \mathcal{O}\left(\frac{K \ln T}{\eta} \right) + 4\eta \sum_{t=1}^T \min\left\{ (\ell_{t,i_t} - m_{t,i_t})^2, \frac{\ell_{t,i_t} - m_{t,i_t})^2}{K x_{t,i_t}} \right\}.$$

Finally realizing by u's definition,

$$\sum_{t=1}^T \langle u - \boldsymbol{e}_{i^*}, \hat{\ell}_t \rangle = \frac{1}{T} \sum_{t=1}^T \left\langle -\boldsymbol{e}_{i^*} + \frac{1}{K}\mathbf{1}, \hat{\ell}_t \right\rangle,$$

combining the two inequalities above and taking expectation finish the proof. ■

Lemma 12 *line 7 of Algorithm 2 ensures*

$$\mathbb{E}\left[\sum_{t=1}^T \langle w_t - x_t, \ell_t \rangle \right] \le \mathcal{O}(1) + \alpha \mathbb{E}\left[\sum_{t : i_{t+1} \in \mathcal{S}_t} |\ell_{t+1, i_{\tau(t)}} - \ell_{\tau(t), i_{\tau(t)}}| \right] - \frac{\alpha}{2} \mathbb{E}\left[\sum_{t=2}^T (c_{\tau(t)} - c_{\tau(t-1)})^2 \right].$$

Proof First fix any $t > 1$ and denote $\tau(t-1)$ by τ for notational convenience (note that τ is thus fixed at the beginning of round t). Note that by the construction of w_t, we have $w_{t,i} = x_{t,i}$ for any $i \notin \mathcal{S}_{t-1} \cup \{i_\tau\}$ and also $\sum_{i \in \mathcal{S}_{t-1} \cup \{i_\tau\}} x_{t,i} = \sum_{i \in \mathcal{S}_{t-1} \cup \{i_\tau\}} w_{t,i}$. Therefore

$$\begin{aligned}
\langle w_t - x_t, \ell_t \rangle &= \alpha_t \left(\left(\sum_{i \in \mathcal{S}_{t-1} \cup \{i_\tau\}} x_{t,i} \right) \ell_{t,i_\tau} - \sum_{i \in \mathcal{S}_{t-1} \cup \{i_\tau\}} x_{t,i} \ell_{t,i} \right) \\
&= \alpha_t \left(\left(\sum_{i \in \mathcal{S}_{t-1} \cup \{i_\tau\}} w_{t,i} \right) \ell_{t,i_\tau} - \sum_{i \in \mathcal{S}_{t-1} \cup \{i_\tau\}} x_{t,i} \ell_{t,i} \right) \\
&= \alpha_t \left(\sum_{i \in \mathcal{S}_{t-1} \cup \{i_\tau\}} w_{t,i} \ell_{t,i_\tau} - \sum_{i \in \mathcal{S}_{t-1} \cup \{i_\tau\}} w_{t,i} \ell_{t,i} \right) + \alpha_t \langle w_t - x_t, \ell_t \rangle
\end{aligned}$$

$$= \alpha_t \left(\sum_{i \in \mathcal{S}_{t-1}} w_{t,i} \ell_{t,i_\tau} - \sum_{i \in \mathcal{S}_{t-1}} w_{t,i} \ell_{t,i} \right) + \alpha_t \langle w_t - x_t, \ell_t \rangle.$$

Rearranging and using the definition of α_t gives:

$$\langle w_t - x_t, \ell_t \rangle = \frac{\alpha_t}{1-\alpha_t} \left(\sum_{i \in \mathcal{S}_{t-1}} w_{t,i} \ell_{t,i_\tau} - \sum_{i \in \mathcal{S}_{t-1}} w_{t,i} \ell_{t,i} \right) = \alpha(1-c_\tau) \left(\sum_{i \in \mathcal{S}_{t-1}} w_{t,i} (\ell_{t,i_\tau} - \ell_{t,i}) \right).$$

Taking expectation gives:

$$\begin{aligned}
&\mathbb{E}[\langle w_t - x_t, \ell_t \rangle] = \alpha \mathbb{E} \left[(1-c_\tau) \left(\sum_{i \in \mathcal{S}_{t-1}} w_{t,i} (\ell_{t,i_\tau} - \ell_{t,i}) \right) \right] \\
&= \alpha \mathbb{E} \left[\mathbf{1}\{i_t \in \mathcal{S}_{t-1}\} (1-c_\tau)(\ell_{t,i_\tau} - \ell_{t,i_t}) \right] \\
&= \alpha \mathbb{E} \left[\mathbf{1}\{i_t \in \mathcal{S}_{t-1}\} (1-c_\tau)(\ell_{t,i_\tau} - c_\tau + c_\tau - c_t)) \right] \\
&\leq \alpha \mathbb{E} \left[\mathbf{1}\{i_t \in \mathcal{S}_{t-1}\} |\ell_{t,i_\tau} - c_\tau| \right] + \alpha \mathbb{E} \left[\mathbf{1}\{i_t \in \mathcal{S}_{t-1}\} (c_\tau - c_t - c_\tau^2 + c_t c_\tau) \right] \\
&= \alpha \mathbb{E} \left[\mathbf{1}\{i_t \in \mathcal{S}_{t-1}\} |\ell_{t,i_\tau} - \ell_{\tau,i_\tau}| \right] + \alpha \mathbb{E} \left[c_{\tau(t-1)} - c_{\tau(t)} - c_{\tau(t-1)}^2 + c_{\tau(t)} c_{\tau(t-1)} \right],
\end{aligned}$$

where in the last step we use the fact that $\tau(t)$ is t if $i_t \in S_{t-1}$ and is $\tau(t-1)$ otherwise. Finally summing over t and telescoping finish the proof. ■

Appendix C. Proof of Theorem 7

Since m_{t+1} is the projection of m_t on $\mathcal{K}_{t+1}$ and also $m^\star_{t+1} \in \mathcal{K}_{t+1}$, by the generalized Pythagorean theorem we have

$$\left\| m_{t+1} - m^\star_{t+1} \right\|_2^2 + \left\| m_{t+1} - m_t \right\|_2^2 \leq \left\| m_t - m^\star_{t+1} \right\|_2^2.$$

On the other hand, by triangle inequality we also have

$$\begin{aligned}
\left\| m_t - m^\star_{t+1} \right\|_2^2 &\leq \left(\left\| m_t - m^\star_t \right\|_2 + \left\| m^\star_t - m^\star_{t+1} \right\|_2 \right)^2 \\
&= \left\| m_t - m^\star_t \right\|_2^2 + 2 \left\| m_t - m^\star_t \right\|_2 \left\| m^\star_t - m^\star_{t+1} \right\|_2 + \left\| m^\star_t - m^\star_{t+1} \right\|_2^2 \\
&\leq \left\| m_t - m^\star_t \right\|_2^2 + 6 \left\| m^\star_t - m^\star_{t+1} \right\|_2. && (\mathcal{K}_t \subset \mathcal{B})
\end{aligned}$$

Combining the two inequalities above, summing over t, and telescoping give

$$\sum_{t=1}^{T-1} \left\| m_{t+1} - m_t \right\|_2^2 \leq \left\| m_1 - m^\star_1 \right\|_2^2 + 6 \sum_{t=1}^{T-1} \left\| m^\star_t - m^\star_{t+1} \right\|_2 \leq 4 + 6 \sum_{t=1}^{T-1} \left\| m^\star_t - m^\star_{t+1} \right\|_2,$$

finishing the proof.