Tight First- and Second-Order Regret Bounds for Adversarial Linear Bandits

Shinji Ito*
NEC Corporation
i-shinji@nec.com

Shuichi Hirahara
National Institute of Informatics
s_hirahara@nii.ac.jp

Tasuku Soma
The University of Tokyo
tasuku_soma@mist.i.u-tokyo.ac.jp

Yuichi Yoshida
National Institute of Informatics
yyoshida@nii.ac.jp

Abstract

We propose novel algorithms with first- and second-order regret bounds for adversarial linear bandits. These regret bounds imply that our algorithms perform well when there is an action achieving a small cumulative loss or the loss has a small variance. In addition, we need only assumptions weaker than those of existing algorithms; our algorithms work on discrete action sets as well as continuous ones without a priori knowledge about losses, and they run efficiently if a linear optimization oracle for the action set is available. These results are obtained by combining optimistic online optimization, continuous multiplicative weight update methods, and a novel technique that we refer to as distribution truncation. We also show that the regret bounds of our algorithms are tight up to polylogarithmic factors.

1 Introduction

The *adversarial linear bandit problem* models sequential decision making with limited information, and it has been used in a wide range of applications, including combinatorial bandits [18; 22] and the adaptive routing problem [12]. In this problem, a player is given a set $\mathcal{A} \subseteq \mathbb{R}^d$ of actions represented by d-dimensional feature vectors, and the player is to choose an action $a_t \in \mathcal{A}$ in each round $t = 1, \ldots, T$ of the decision process. Just after choosing the action a_t, the player receives the *bandit feedback* $\langle \ell_t, a_t \rangle$ as the loss of the action, where $\ell_t \in \mathbb{R}^d$ is the *loss vector* of the t-th round chosen by an adversary.[1] We should note here that the loss vector ℓ_t is not revealed to the player even after choosing the action. The goal of the player is to minimize cumulative loss $\sum_{t=1}^{T} \langle \ell_t, a_t \rangle$. Player performance of the player is evaluated by means of the *regret* $R_T(a^*)$ defined as

$$R_T(a^*) = \sum_{t=1}^{T} \langle \ell_t, a_t \rangle - \sum_{t=1}^{T} \langle \ell_t, a^* \rangle \tag{1}$$

for $a^* \in \mathcal{A}$. A plethora of algorithms have been proposed with the expected regret $\mathbf{E}[R_T(a^*)] = \tilde{O}(d\sqrt{T})$ for any $a^* \in \mathcal{A}$ [18; 32; 40] under the assumption that $| \langle \ell_t, a_t \rangle | = O(1)$, where $\tilde{O}(\cdot)$ hides a logarithmic factor in d and T. These algorithms are *worst-case optimal* up to logarithmic factors:

*This work was done while Shinji Ito was at the University of Tokyo.

[1] In this paper, we are concerned with *adaptive adversaries*, i.e., an adversary can choose loss ℓ_t based on the past player's actions $a_1, \ldots, a_{t-1}$. We note that ℓ_t cannot depend on the t-th action a_t since otherwise the player would always suffer $\Omega(T)$ regret.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Table 1: Regret bounds for adversarial linear bandits. Here, we denote $\bar{\ell} = \frac{1}{T}\sum_{t=1}^{T} \ell_t$. The results with † require additional assumptions on the feasible set $\mathcal{A}$ and a priori knowledge of deviations.

	Upper Bound	Lower Bound
Worst case	$\tilde{O}(d\sqrt{T})$ [18; 32]	$\Omega(d\sqrt{T})$ [25; 35]
First order	$\tilde{O}\left(d\sqrt{\sum_{t=1}^{T}\langle \ell_t, a^* \rangle}\right)$ **[Theorem 3]**	$\Omega\left(d\sqrt{\sum_{t=1}^{T}\langle \ell_t, a^* \rangle}\right)$
Second order	$\tilde{O}\left(d\sqrt{\theta\sum_{t=1}^{T}\|\ell_t - \bar{\ell}\|_2^2}\right)$ † [30]	$\Omega\left(d\sqrt{\sum_{t=1}^{T}\|\ell_t - \bar{\ell}\|_*^2}\right)$
	$\tilde{O}\left(d\sqrt{\sum_{t=1}^{T}\|\ell_t - \bar{\ell}\|_*^2}\right)$ **[Theorem 2]**	
Predictable	$\tilde{O}\left(d\sqrt{\theta\sum_{t=1}^{T}\|\ell_t - m_t\|_2^2}\right)$ † [44]	$\Omega\left(d\sqrt{\sum_{t=1}^{T}\|\ell_t - m_t\|_*^2}\right)$
sequence	$\tilde{O}\left(d\sqrt{\sum_{t=1}^{T}\left(\langle \ell_t - m_t, a_t \rangle\right)^2}\right)$ **[Theorem 1]**	

Any algorithm suffers the expected regret of $\Omega(d\sqrt{T})$ in the worst case [25; 35]. These worst-case bounds are, however, too pessimistic in practical situations since we rarely encounter truly adversarial environments in real-world problems.

To get around the worst-case lower bound, algorithms with *first-order* and *second-order* regret bounds have been developed for some adversarial bandit problems [6; 15; 30; 48]. First-order regret bounds are those depending on the minimum cumulative loss $L_T^* = \min_{a^* \in \mathcal{A}} \sum_{t=1}^{T} \langle \ell_t, a^* \rangle$, rather than on the number T of rounds. For example, Allenberg et al. [6] proposed an algorithm with a first-order regret bound for the *adversarial multi-armed bandit (MAB) problem*, a special case of the adversarial linear bandit problem in which the action set $\mathcal{A}$ is just a finite set of size K.[2] For MAB, their algorithm achieves regret of $\tilde{O}(\sqrt{KL_T^*})$, which improves over the worst-case optimal bound of $O(\sqrt{KT})$ [8; 11], especially when L_T^* is much smaller than T, i.e., where there is an action with a small cumulative loss. It is worth noting that their algorithm achieves a nearly optimal *worst-case* bound, as well, since $L_T^* \leq T$ follows from a standard assumption. For more general linear bandits, however, such an algorithm was not known in the literature. In this paper, second-order regret bounds refer to those depending on the *second-order variation* $\sum_{t=1}^{T} \|\ell_t - \bar{\ell}\|^2$ rather than on T, where $\|\cdot\|$ is an arbitrary norm and $\bar{\ell}$ stands for the average of the loss vectors $\{\ell_t\}_{t=1}^{T}$.[3] For MAB, Bubeck et al. [15] have proposed an algorithm with a regret bound of $\tilde{O}(\sqrt{\sum_{t=1}^{T}\|\ell_t - \bar{\ell}\|_2^2})$. This algorithm is nearly worst-case optimal since $\|\ell_t - \bar{\ell}\|_2^2 \leq KT$, and it performs better when losses $\{\ell_t\}$ have small variation. For linear bandits, Hazan and Kale [30] proposed an algorithm achieving $\tilde{O}(d^{\frac{3}{2}}\sqrt{\sum_{t=1}^{T}\|\ell_t - \bar{\ell}\|_2^2})$ regret under certain assumptions. Though this bound is better than the worst-case optimal bound of $\tilde{O}(d\sqrt{T})$ when $\sum_{t=1}^{T}\|\ell_t - \bar{\ell}\|_2^2 = O(T/d)$, it is not worst-case optimal in general.

Another (and deeply relevant) line of work that tries to get around the worst-case lower bound is a framework called *predictable sequences* [44; 45; 48]. This framework assumes that the player is given *predicted loss vector* m_t, which is produced by an arbitrary process, before choosing actions. One would hope that the regret would get smaller when m_t predicts ℓ_t well. In fact, as shown in [44], we can achieve $\tilde{O}(d^{\frac{3}{2}}\sqrt{\sum_{t}^{T}\|\ell_t - m_t\|_2^2})$ regret for linear bandits, which can be smaller than the worst-case optimal bound if m_t are sufficiently close to ℓ_t so that $\sum_{t=1}^{T}\|\ell_t - m_t\|_2^2 = O(T/d)$.

[2] In fact, MAB is equivalent to the linear bandit problem with an action set being the standard basis of a K-dimensional space $\mathcal{A} = \{e_1, \ldots, e_K\} \subseteq \{0,1\}^K$: Each element ℓ_{ti} of the loss vector $\ell_t \in [0,1]^K$ stands for the loss incurred by choosing the i-th action e_i.

[3] The term "second-order regret" has been used to mean various bounds in the literature. In this work, we adopt the one stated here.

Our contributions

In this paper, we present newly-devised, efficient algorithms with improved first- and second-order regret bounds for the adversarial linear bandit problem. Our algorithms not only yield better regret bounds but also make only fewer assumptions than have been seen in previous studies. Previous results and our contributions are summarized in Table 1. In the table, $\bar{\ell}$ denotes the average of the loss vectors, i.e., $\bar{\ell} = \frac{1}{T}\sum_{t=1}^{T} \ell_t$. The bounds with † in Table 1 require additional prior knowledge w.r.t. the loss vectors. For example, the results by [30] and [44] in Table 1 are based on the assumption that, respectively, (approximated values of) the quantities $\sum_{t=1}^{T} \|\ell_t - \bar{\ell}\|^2$ and $\sum_{t=1}^{T} \|\ell_t - m_t\|_2^2$ are given before the game starts. In addition, they assume the following conditions: (i) The action set $\mathcal{A}$ is convex, (ii) $\max_{a \in \mathcal{A}} \|a\|_2 = O(1)$, and (iii) $\mathcal{A}$ has a self-concordance barrier with parameter $\theta \geq 1$. Note that the self-concordance parameter θ can be $\Omega(d)$, e.g., when $\mathcal{A} = \{a \in \mathbb{R}^d \mid \|a\|_\infty \leq 1\}$.

We provide two algorithms with the guarantees described below:

1. The first one achieves $\mathbf{E}[R_T(a^*)] = \tilde{O}\left(\mathbf{E}\left[d\sqrt{\sum_{t=1}^{T} \left(\langle \ell_t - m_t, a_t\rangle\right)^2}\right]\right)$ for predictable sequences. If $\|a_t\| = O(1)$ holds for a fixed norm $\|\cdot\|$, our bound implies a regret bound of $\tilde{O}\left(\mathbf{E}\left[d\sqrt{\sum_{t=1}^{T} \|\ell_t - m_t\|_*^2}\right]\right)$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$. This result encompasses the result by [44] in Table 1 since $\max_{a \in \mathcal{A}} \|a\|_2 = O(1)$ is assumed in their work and $\theta \geq 1$ in general. Further, this algorithm does not require the above-mentioned assumptions. In particular, the action set can be discrete.
2. The second achieves the following second-order regret bound: $\mathbf{E}[R_T(a^*)] = \tilde{O}\left(d\sqrt{\sum_{t=1}^{T} \|\ell_t - \bar{\ell}\|_*^2}\right)$ for arbitrary $\bar{\ell}$, as shown in Theorem 2. This result encompasses the regret bound by [30] in Table 1, and again, as before, this algorithm does not require the above-mentioned assumptions. When losses $\langle \ell_t, a_t\rangle$ are non-negative, this algorithm achieves the first-order regret bound of $\mathbf{E}[R_T(a^*)] = \tilde{O}\left(d\sqrt{\sum_{t=1}^{T} \langle \ell_t, a^*\rangle}\right)$ simultaneously, as shown in Theorem 3.

Each regret bound shown in Theorems 1, 2 and 3 is of $\tilde{O}(d\sqrt{T})$, and hence, enjoys worst-case optimality up to logarithmic factors, in contrast to existing algorithms [30; 44]. We should note, however, that our regret bounds are not tight for MAB since the worst-case optimal regret bound is known to be $\Theta(\sqrt{dT})$ for this special case. Similarly, in the special case where the action set is the unit ball, algorithms proposed in [30; 44] achieves regret bounds comparable to ours (up to logarithmic factors), as there is a self-concordant barrier of parameter $\theta = O(1)$.

Our algorithms are based on the multiplicative weight update (MWU) method [7; 34] with an unbiased estimator $\hat{\ell}_t$ of the loss vector ℓ_t. As with existing algorithms [18; 30; 32] for the adversarial linear bandit problem, we construct an unbiased estimator from a single observation $\langle \ell_t, a_t\rangle$, where a_t follows a distribution p_t maintained by the MWU method. The regret strongly depends on the *stability* of the unbiased estimators $\hat{\ell}_t$; we want the norm and variance of $\hat{\ell}_t$ to be small enough. In order to make an unbiased estimator stable, previous studies [18; 32] have mixed p_t with another probability distribution. This approach, however, requires the mixing rate of $\Omega(d/\sqrt{T})$, which causes $\Omega(T \cdot d/\sqrt{T}) = \Omega(d\sqrt{T})$-regret in general. To overcome this issue, we *truncate* (the support of) the distribution p_t instead. More specifically, we truncate the distribution to ensure that the magnitude of chosen actions would be controlled with respect to an appropriately designed norm. We will show here that this approach ensures the stability of $\hat{\ell}_t$ with almost no degradation of the expected performance, with the help of a concentration property of log-concave distributions [42]. We should note that similar techniques of truncating distributions can be found in the literature of bandit optimization, such as combinatorial semi-bandits [43] and bandit convex optimization [14]. This paper, however, employs a different way of truncation and analyses as the problem settings are different.

Another essential element in our algorithms is the technique called *optimistic online optimization* [44; 45]. Rakhlin and Sridharan [44; 45] introduced the framework of online optimization with predicted loss m_t, and proposed algorithms referred to as *optimistic online mirror descent* and *optimistic follow the regularized leader* that exploit the predicted loss m_t to improve the regret. Our first algorithm employs their techniques to achieve the regret in Theorem 1. The second appropriately chooses m_t on the basis of an online optimization method, and it achieves regret bounds noted in Theorems 2 and 3. A similar technique for computing m_t can be found in the study by Cutkosky [23], though this existing work uses a different loss as it is for the full-information setting.

2 Related Work

The linear bandit problem generalizes, as a special case, the well-studied multi-armed bandit (MAB) problem [11], in which the action set $\mathcal{A} = [K] := \{1, 2, \ldots, K\}$ is just a finite set of actions. Other important special cases are *combinatorial bandits* [13; 18], in which the action set $\mathcal{A} \subseteq 2^{[K]}$ is a family of subsets of a finite set, and the loss incurred by choosing an action $a \in \mathcal{A}$ is $\sum_{i \in a} \ell_{ti}$. For example, given a directed graph $G = (V, A)$ and nodes $s, t \in V$, by setting $\mathcal{A}$ to be the family of edge sets representing st-paths, one can model the *bandit shortest path* or *adaptive routing* problem [12] as a combinatorial bandit. Further, online recommendation problems have been formulated as linear bandits [41]. To solve linear bandits, many algorithms have been proposed for stochastic settings [1; 10; 21], as well as for adversarial settings [2; 3; 12; 13; 18; 32]. It was shown that we can achieve $\tilde{O}(d\sqrt{T})$ regret [13; 20], which nearly matches the lower bound of $\Omega(d\sqrt{T})$ shown in [25]. The truncation technique in this paper has recently been applied to the delayed-feedback setting as well [37].

A seminal work by Freund and Schapire [26] provided a first-order regret bound for the *expert problem*, a full-information counterpart of MAB. For MAB, Allenberg et al. [6] proposed an algorithm with a first-order bound. This result has been extended in two directions. First one is the *contextual bandit problem* [11; 39], in which the player is given a *context* $x_t \in X$ before choosing the action and the regret is measured by means of a *hypothesis set* $\Pi \subseteq \{\pi : X \to [K]\}$ as $R_T = \sum_{t=1}^{T} \ell_{ta_t} - \min_{\pi^* \in \Pi} \sum_{t=1}^{T} \ell_{t\pi^*(x_t)}$. Offering a first-order regret bound for the contextual bandit was posed as an open problem [4], and Allen-Zhu et al. [5] solved it affirmatively. Another direction is combinatorial semi bandits [9; 28; 47], a variant of combinatorial bandits with more informative feedback, in which the player who chose $a_t \in \mathcal{A} \subseteq 2^{[d]}$ can observe loss ℓ_{ti} for each $i \in a_t$. For combinatorial semi bandits, Neu [43] proposed an algorithm with a first-order regret bound. This algorithm, however, does not apply directly to the full-bandit setting in which only $\sum_{i \in a_t} \ell_{ti}$ is observable.

The notion of second-order regret bound was introduced by Cesa-Bianchi et al. [19] for the (full-information) expert problem, in which a regret bound of $O(\sqrt{\log K \cdot Q^*})$ where $Q^* \leq \max_{a \in [K]} \sum_{t=1}^{T} \ell_{ta}^2$ is known beforehand. Hazan and Kale [29] improved this result by replacing Q^* with the *variation* $V^* \leq \max_{a \in [K]} \sum_{t=1}^{T} \left(\ell_{ta} - \bar{\ell}_{ta}\right)^2$ of the loss sequence.[4] For MAB, Hazan and Kale [30] proposed an algorithm achieving $\tilde{O}(K^2\sqrt{V})$-regret with $V = \sum_{t=1}^{T} \|\ell_t - \bar{\ell}\|_2^2$, and they conjectured that there exists an efficient algorithm with an $\tilde{O}(\sqrt{V})$-regret bound [31]. Bubeck et al. [15] proved this conjecture by providing such a regret upper bound, which almost matches a lower bound of $\Omega(\sqrt{V})$ provided by Gerchinovitz and Lattimore [27]. Wei and Luo [48] provided an MAB algorithm with $\tilde{O}(\sqrt{KS})$ regret, where we denote $S = \sum_{t=1}^{T} \left(\ell_{ta^*} - \bar{\ell}_{ta^*}\right)^2$ for $a^* \in \operatorname{argmin}_{a \in [K]} \sum_{t=1}^{T} \ell_{ta}$, which is incomparable to $\tilde{O}(\sqrt{V})$ in general. It is worth noting that MAB algorithms mentioned here require a priori knowledge of parameters V and S, in contrast to our algorithms. Our algorithms can be applied to MAB to achieve the regret bound (Theorem 2) of $\tilde{O}\left(K\sqrt{\sum_{t=1}^{T} \|\ell_t - \bar{\ell}_t\|_\infty^2}\right)$ for this special case, which is inferior to previous results of $\tilde{O}(\sqrt{V})$ and $\tilde{O}(\sqrt{KS})$ achieved by MAB-specialized algorithms. Algorithms based on continuous MWU, such as the one by Hazan and Karnin [32] and ours, may achieve worst-case optimal regret for general linear bandits, but, for MAB, they seem not competitive with MAB-specialized algorithms as they do not exploit specific structures of the action set.

[4] In literature [19; 29], values Q^* and V^* were originally defined as $Q^* = \max_{\tau \in [T]} \sum_{t=1}^{\tau} \ell_{ta_\tau^*}^2$ and $V^* = \max_{\tau \in [T]} \sum_{t=1}^{\tau} \left(\ell_{ta_\tau^*} - \bar{\ell}_{\tau a_\tau^*}\right)^2$ with $a_\tau^* \in \operatorname{argmin}_{a \in [K]} \sum_{t=1}^{\tau} \ell_{ta}$ and $\bar{\ell}_\tau = \frac{1}{\tau} \sum_{t=1}^{\tau} \ell_t$.

3 Preliminaries

Given a vector $x \in \mathbb{R}^d$ and a positive-semidefinite matrix $M \in \mathbb{R}^{d\times d}$, let $\|x\|_M = \sqrt{x^\top M x}$. For symmetric matrices A, B, we denote $A \succeq B$ if $A - B$ is positive-semidefinite. We denote the convex hull of $\mathcal{A}$ by $\mathcal{A}'$. Given a distribution p over $\mathcal{A}'$, define a vector $\mu(p) \in \mathbb{R}^d$ and a matrix $S(p) \in \mathbb{R}^{d\times d}$ by

$$\mu(p) = \mathop{\mathbf{E}}_{x\sim p}[x], \quad S(p) = \mathop{\mathbf{E}}_{x\sim p}[xx^\top]. \tag{2}$$

For ease of exposition, we also use p to denote its density function simultaneously. If the density function $p : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ of a probability distribution has a convex support and $\log(p(x))$ is a concave function (on the support), then we call the distribution *log-concave*. We will make use of the following concentration property of log-concave distributions:

Lemma 1. *If x follows a log-concave distribution p over $\mathbb{R}^d$ and $S(p) \preceq I$, we have*

$$\Pr[\|x\|_2^2 \geq d\alpha^2] \leq d\exp(1-\alpha) \tag{3}$$

for arbitrary $\alpha \geq 0$.

This lemma follows from, e.g., Lemma 5.7 in [42]. A complete proof can be found in Appendix A.

3.1 Adversarial Linear Bandits

In the adversarial linear bandit problem, a player is given an action set $\mathcal{A} \subseteq \mathbb{R}^d$, which is assumed to be a compact set in $\mathbb{R}^d$, before the game starts. Without loss of generality, we assume that $\mathcal{A}$ is not contained in any proper linear subspace. Note that the action set $\mathcal{A}$ can be discrete. In each round $t \in [T]$, the player chooses an action $a_t \in \mathcal{A}$, and then the environment reveals the loss $\langle \ell_t, a_t\rangle$, where the loss vector $\ell_t \in \mathbb{R}^d$ is in the convex set $\mathcal{L} \subseteq \mathbb{R}^d$ defined as

$$\mathcal{L} = \{\ell \in \mathbb{R}^d \mid -1 \leq \langle \ell, a\rangle \leq 1 \text{ for all } a \in \mathcal{A}\}, \tag{4}$$

and hence $\langle \ell_t, a_t\rangle \in [-1, 1]$ holds.

In Section 4.1, we consider a problem setting with *predicted loss vectors* $m_t \in \mathbb{R}^d$. In this problem setting, the player is given m_t before choosing a_t. The predicted loss vectors (m_t) are arbitrary sequences over $\mathbb{R}^d$, and each m_t may be chosen depending on $\{(a_j, \ell_j)\}_{j<t}$.

When we discuss first-order regret bounds, we assume that the loss $\langle \ell_t, a\rangle$ is non-negative for any $a \in \mathcal{A}$. This assumption is a standard one when discussing first-order bounds (e.g., [6]) and is indispensable for ensuring that $\sum_{t=1}^{T} \langle \ell_t, a^*\rangle \geq 0$. When we discuss second-order regret bounds, we fix a norm $\|\cdot\|$ over $\mathbb{R}^d$ such that $\max_{a\in\mathcal{A}} \|a\| \leq 1$. Let $\|\cdot\|_*$ denote the dual norm of $\|\cdot\|$, i.e., $\|\ell\|_* = \max_{\|x\|\leq 1} \langle \ell, x\rangle$.

When we consider computational complexity, we assume that we can solve linear optimization over $\mathcal{A}$, i.e., there exists an oracle $\mathcal{O}$ with which given $\ell \in \mathbb{R}^d$, we can compute $\mathcal{O}(\ell) \in \mathrm{argmin}_{a\in\mathcal{A}}\{\langle \ell, a\rangle\}$. Such an assumption is standard in the context of online optimization [24; 33; 38] as this is an almost minimum assumption for developing computationally efficient online optimization algorithms with sublinear regret bounds.

4 Algorithms and Regret Upper Bounds

In this section, we explain our algorithms and analyze their regret bounds. In Section 4.1, we provide an algorithm (Algorithm 1) for a scenario in which predicted loss vectors m_t are available, and we analyze its regret bound, which holds for arbitrary m_t. In Section 4.2, we show that Algorithm 1 enjoys a second-order regret bound when we choose m_t in a sophisticated way on the basis of the observed feedback. In Section 4.3, this algorithm with a second-order bound is shown to have a first-order regret bound as well,

4.1 Algorithm with predicted loss vectors

Let us first consider here the case in which predicted loss vectors $m_t \in \mathbb{R}^d$ for ℓ_t are available. In this setting, the player is given m_t before choosing a_t. We assume that $\langle m_t, a\rangle \in [-1, 1]$ holds for any $a \in \mathcal{A}$.

Algorithm 1 An algorithm for adversarial linear bandits with predicted loss

for $t = 1, 2, \ldots, T$ **do**
  repeat
    Pick x_t from the distribution p_t, defined by (5).
  until $\|x_t\|^2_{S(p_t)^{-1}} \leq d\gamma_t^2$
  Choose $a_t \in \mathcal{A}$ so that $\mathbf{E}[a_t] = x_t$, play a_t, and receive a loss $\langle \ell_t, a_t \rangle$ as feedback.
  Compute an unbiased estimator $\hat{\ell}_t$ of ℓ_t as $\hat{\ell}_t = m_t + \langle \ell_t - m_t, a_t \rangle \cdot S(\tilde{p}_t)^{-1} x_t$.
  Update p_t as in (5).
end for

Our algorithm maintains probability distributions over $\mathcal{A}'$, the convex hull of $\mathcal{A}$, following the multiplicative weight update method [7].

$$w_t(x) = \exp\left(-\eta_t \left\langle \sum_{j=1}^{t-1} \hat{\ell}_j + m_t, x \right\rangle\right), \quad p_t(x) = \frac{w_t(x)}{\int_{y \in \mathcal{A}'} w_t(y) \mathrm{d}y} \quad (x \in \mathcal{A}'), \tag{5}$$

where $\eta_j > 0$ are parameters referred to as *learning rates*, which we will determine later, and each $\hat{\ell}_j \in \mathbb{R}^d$ is an unbiased estimator of ℓ_j defined below. First, define the *truncated distribution* $\tilde{p}_t$ of p_t as

$$\tilde{p}_t(x) = \frac{p_t(x)\mathbf{1}\{\|x\|^2_{S(p_t)^{-1}} \leq d\gamma_t^2\}}{\Pr_{y \sim p_t}[\|y\|^2_{S(p_t)^{-1}} \leq d\gamma_t^2]} \propto p_t(x)\mathbf{1}\{\|x\|^2_{S(p_t)^{-1}} \leq d\gamma_t^2\}, \tag{6}$$

where $\gamma_t > 1$ is a parameter we will define later. In each round, the algorithm samples $x_t \in \mathcal{A}'$ according to $\tilde{p}_t$, and then chooses $a_t \in \mathcal{A}$ so that $\mathbf{E}[a_t | x_t] = x_t$. More precisely, we compute $\lambda_1, \ldots, \lambda_{d+1} \geq 0$ and $b_1, \ldots, b_{d+1} \in \mathcal{A}$ such that $\sum_{i=1}^{d+1} \lambda_i = 1$ and $\sum_{i=1}^{d+1} \lambda_i b_i = x_t$, and then output $a_t = b_i$ with probability λ_i. Such $\{(\lambda_i, b_i)\}_{i=1}^{d+1}$ can be efficiently computed given a linear optimization oracle for $\mathcal{A}$, e.g., via the ellipsoid method as stated in Corollary 14.1g in [46]. After taking the action a_t, the algorithm receives a loss $\langle \ell_t, a_t \rangle$ as feedback, and it constructs an unbiased estimator $\hat{\ell}_t$ of ℓ_t as follows:

$$\hat{\ell}_t = m_t + (\langle \ell_t, a_t \rangle - \langle m_t, a_t \rangle) S(\tilde{p}_t)^{-1} x_t. \tag{7}$$

We note that $S(\tilde{p}_t)$ is invertible. This follows from the assumption that $\mathcal{A}$ is not contained in any proper linear subspace. Indeed, under this assumption, $\mathcal{A}'$ is a full-dimensional convex set with a positive Lebesgue measure. Combining this and Lemma 1, we can see that the domain of $\tilde{p}_t$ is full-dimensional as well. Therefore, the distribution $\tilde{p}_t$ has a density function taking positive values over a full-dimensional convex set, which implies that $S(\tilde{p}_t)$ is positive-definite. A similar argument can be found, e.g., in [36] (between Eq. (4) and (5)).

Lemma 2. *The vector $\hat{\ell}_t$ is an unbiased estimator of ℓ_t, i.e., we have $\mathbf{E}[\hat{\ell}_t | \ell_t] = \ell_t$.*

Proof. The expectation of $\hat{\ell}_t$ is

$$\mathop{\mathbf{E}}_{x_t, a_t}[\hat{\ell}_t | \ell_t] = m_t + S(\tilde{p}_t)^{-1} \mathop{\mathbf{E}}_{x_t, a_t}[x_t a_t^\top](\ell_t - m_t). \tag{8}$$

Since $\mathbf{E}[a_t | x_t] = x_t$, we have $\mathbf{E}_{x_t, a_t}[x_t a_t^\top] = \mathbf{E}_{x_t}[x_t x_t^\top] = S(\tilde{p}_t)$. Combining this and (8), we obtain $\mathbf{E}[\hat{\ell}_t | \ell_t] = \ell_t$. □

Our algorithm can be summarized in Algorithm 1. Algorithm 1 enjoys the following regret bound:

Theorem 1. *Suppose $\gamma_t \geq 4\log(10dt)$ and $\eta_t \leq \frac{1}{\sqrt{800d}\gamma_t}$ for all t. Then, for any $a^* \in \mathcal{A}$, Algorithm 1 satisfies*

$$\mathbf{E}[R_T(a^*)] \leq d \cdot \mathbf{E}\left[4 \sum_{t=1}^{T} \eta_t \gamma_t^2 (\langle \ell_t - m_t, a_t \rangle)^2 + \frac{\log T}{\eta_T}\right] + 3. \tag{9}$$

Consequently, by setting $\gamma_t = 4\log(10dt)$ *and* $\eta_t = \left(800d\gamma_t^2 + 16\sum_{j=1}^{t-1}\gamma_j^2(\langle \ell_j - m_j, a_j\rangle)^2\right)^{-\frac{1}{2}}$, *we obtain*

$$\mathbf{E}[R_T(a^*)] \le 32d\log T \cdot \log(10dT) \cdot \mathbf{E}\left[\sqrt{\sum_{t=1}^{T}(\langle \ell_t - m_t, a_t\rangle)^2 + 50d}\right]. \tag{10}$$

This regret bound can be shown via analyses for the optimistic follow-the-regularized-leader algorithm [45; 44] and the continuous multiplicative weight update method [7; 32], combined with Lemmas 1 and 2. A complete proof is given in Section B in the appendix.

4.2 Second-order regret bound

In this section, we show that we can obtain a second-order bound by appropriately choosing m_t in Algorithm 1, by means of an online learning technique. Consider m_t defined by

$$m_t \in \operatorname*{argmin}_{m\in\mathcal{L}}\left\{\|m\|_S^2 + \sum_{j=1}^{t-1}(\langle \ell_j - m, a_j\rangle)^2\right\}, \tag{11}$$

where $\mathcal{L} \subseteq \mathbb{R}^d$ is defined as in (4) and $S \in \mathbb{R}^{d\times d}$ is an arbitrary positive-definite matrix.

Lemma 3. *If* m_t *is given by* (11)*, we have, for any* $m^* \in \mathcal{L}$*,*

$$\sum_{t=1}^{T}(\langle \ell_t - m_t, a_t\rangle)^2 \le \sum_{t=1}^{T}(\langle \ell_t - m^*, a_t\rangle)^2 + \|m^*\|_S^2 + 16d\log\left(1 + \frac{T}{d}\max_{a\in\mathcal{A}}\|a\|_{S^{-1}}^2\right). \tag{12}$$

This lemma can be shown by following the analysis of *online ridge regression*, e.g., see the proof of Theorem 11.7 in [17]. To further bound the right-hand side of (12), we provide a specific example of S. We first note that there exists a matrix $S \in \mathbb{R}^{d\times d}$ such that

$$\|m\|_S^2 \le d \text{ for any } m \in \mathcal{L}, \quad \|a\|_{S^{-1}}^2 \le 4d \text{ for any } a \in \mathcal{A}, \tag{13}$$

and given a linear optimization oracle over $\mathcal{A}$, one can compute such an S efficiently, via a *barycentric spanner* [12] for $\mathcal{A}$. More precise method for constructing S is described in Section C in the appendix.

Combining Theorem 1, Lemma 3 and (13), we obtain the following regret bound:

Theorem 2. *Suppose that* γ_t*,* η_t*, and* m_t *are given by* (32)*,* (11) *(with* $\mathcal{L}$ *as in* (4)*), and* (34)*, respectively. Then the actions* a_t *of Algorithm 1 satisfy*

$$\begin{aligned}\mathbf{E}[R_T(a^*)] &\le 32d\log T \cdot \log(10dT) \cdot \mathbf{E}\left[\sqrt{\sum_{t=1}^{T}(\langle \ell_t - m^*, a_t\rangle)^2 + 51d + 16d\log(1+4T)}\right]\\ &\le 32d\log T\log(10dT)\cdot \mathbf{E}\left[\sqrt{\sum_{t=1}^{T}\|\ell_t - m^*\|_*^2 + 51d + 16d\log(1+4T)}\right].\end{aligned}$$

for arbitrary $a^* \in \mathcal{A}$ *and* $m^* \in \mathcal{L}$*.*

4.3 First-order regret bound

In this section, we show that the bound in the Theorem 2 can be used to obtain a first-order regret bound assuming that $0 \le \langle \ell_t, a_t\rangle \le 1$.

Theorem 3. *Suppose that the assumptions in Theorem 2 hold and that the observed losses are non-negative. Then we have, for* $\xi = O(\log d \cdot \log^2 T)$*,*

$$\mathbf{E}[R_T(a^*)] = O\left(\xi d\sqrt{\mathbf{E}\left[\sum_{t=1}^{T}\langle \ell_t, a^*\rangle\right] + \xi^2 d^2}\right). \tag{14}$$

Proof. For notational simplicity, let $C_1 = 32d \log T \cdot \log(10dT)$ and $C_2 = 51d + 16d \log(1 + 4T)$. From the inequality in Theorem 2 with $m^* = 0$, we have

$$\mathbf{E}[R_T(a^*)] \le C_1 \mathbf{E}\left[\sqrt{\sum_{t=1}^{T} (\langle \ell_t, a_t \rangle)^2 + C_2}\right] \le C_1 \mathbf{E}\left[\sqrt{\sum_{t=1}^{T} \langle \ell_t, a_t \rangle + C_2}\right]$$
$$= C_1 \mathbf{E}\left[\sqrt{\sum_{t=1}^{T} \langle \ell_t, a^* \rangle + R_T(a^*) + C_2}\right] \le C_1 \sqrt{\mathbf{E}\left[\sum_{t=1}^{T} \langle \ell_t, a^* \rangle\right] + \mathbf{E}[R_T(a^*)] + C_2},$$

where the second inequality follows from the assumption of $0 \le \langle \ell_t, a_t \rangle \le 1$, the equality follows from the definition (1) of R_T, and the last inequality follows from Jensen's inequality. By solving the quadratic inequality with respect to $\mathbf{E}[R_T(a^*)]$, we obtain $\mathbf{E}[R_T(a^*)] \le C_1 \left(\frac{C_1}{2} + \frac{1}{2}\sqrt{C_1^2 + 4\left(C_2 + \mathbf{E}\left[\sum_{t=1}^{T} \langle \ell_t, a^* \rangle\right]\right)}\right) \le C_1 \sqrt{C_1^2 + 2\left(C_2 + \mathbf{E}\left[\sum_{t=1}^{T} \langle \ell_t, a^* \rangle\right]\right)}$, where the last inequality follows from $\frac{\sqrt{x}+\sqrt{y}}{2} \le \sqrt{\frac{x+y}{2}}$ for $x, y \ge 0$. □

Remark 1. The regret bound in Theorem 3 holds even when m_t is chosen to be $m_t = 0$ for all t, as can be seen in the proof. However, we would like to stress here that, by setting m_t as in Theorem 2, a *single* algorithm *simultaneously* enjoys two different regret bounds as in Theorems 2 and 3.

4.4 Computationally efficient implementation

Algorithm 1 can be implemented in a computationally efficient way, assuming that linear optimization over $\mathcal{A}$ can be efficiently solved. As shown in [42], we can get a sample from a log-concave distribution p_t in a polynomial time in d, if we can compute $w_t(x) \propto p_t(x)$ and can access a membership oracle for $\mathrm{supp}(p_t) = \mathcal{A}'$, i.e., we can decide whether a given vector $x \in \mathbb{R}^d$ belongs to $\mathcal{A}'$. The membership problem can be reduced to linear optimization problems by ellipsoid methods, as shown, e.g., in Corollary 14.1b in [46]. Consequently, we can get a sample from p_t in polynomial time. The matrix $S(\tilde{p}_t)$ (7) can be efficiently computed as well. Indeed, since $\tilde{p}_t$ is log-concave, for any $\varepsilon > 0$, we can get an ε-approximation of $S(\tilde{p}_t)$ w.h.p. by generating $(d/\varepsilon)^{O(1)}$ samples from $\tilde{p}_t$, from Corollary 2.7 of [42]. Samples from $\tilde{p}_t$ can be generated with their polynomial-time sampling algorithm as mentioned in Section 4.4 of our manuscript. A similar discussion can be found in Lemma 5.17 of [14].

The vector m_t defined in (11) can be computed efficiently as well. In fact, a linear optimization oracle for $\mathcal{A}$ immediately leads to a separation oracle for $\mathcal{L}$ defined by (4). Hence, we can solve a convex optimization over $\mathcal{L}$ such as (11), e.g., by using ellipsoid methods.

5 Lower Bound

In this section, we provide instance-dependent regret lower bounds. In what follows, we assume $d \ge 2$ and $\mathcal{A} \subseteq \{-1, 1\}^d$. Note that we then have $\|a\|_\infty \le 1$ for any $a \in \mathcal{A}$.

Theorem 4. *Let $\mathcal{A} = \{-1,1\}^{d-1} \times \{1\}$. For any algorithm and for any L with $d^2 \le L \le T$, there exists a sequence $(\ell_t)_{t=1}^T$ of d-dimensional loss vectors such that the following hold: (i) $0 \le \langle \ell_t, a \rangle \le 1$ for any $a \in \mathcal{A}$, (ii) $\min_{a^* \in \mathcal{A}} \sum_{t=1}^{T} \langle \ell_t, a^* \rangle \le L$, (iii) $\sum_{t=1}^{T} \|\ell_t\|_1^2 \le L$, and (iv) any algorithm satisfies $\max_{a^* \in \mathcal{A}} \mathbf{E}[R_T(a^*)] = \Omega(d\sqrt{L})$.*

This theorem complements Theorems 1, 2 and 3 by providing almost matching lower bounds. Indeed, for any problem instances satisfying (iii) with $L \ge d^2$, both Theorem 1 (with $m_t = 0$) and Theorem 2 imply $\max_{a^* \in \mathcal{A}} \mathbf{E}[R_T(a^*)] = \tilde{O}(d\sqrt{L})$, which matches the lower bound of (iv) in Theorem 4. Similarly, for any problem instances satisfying (i) and (ii) with $L \ge d^2$, Theorem 3 implies $\max_{a^* \in \mathcal{A}} \mathbf{E}[R_T(a^*)] = \tilde{O}(d\sqrt{L})$.

Theorem 4 can be shown by adopting the hard instances used to show the worst-case lower bound of $\Omega(d\sqrt{T})$, e.g., by Dani et al. [25]. The proof of Theorem 4 is given in Appendix D.

6 Conclusion

In this paper, we provided algorithms with nearly tight first- and second-order regret bounds for adversarial linear bandit problems, with the aid of techniques such as optimistic online optimization and properties of log-concave distributions. A future research direction is to obtain improved path-length regret bounds, as discussed, e.g., in [16; 48]. Another direction would be to improve practical computational efficiency. Our proposed algorithms require large computational time due to the complexity of continuous multiplicative weight update, though it is of polynomial in dimensions. Hence, algorithms by Hazan and Kale [30]; Rakhlin and Sridharan [44] have smaller runtime if the action set admits a self-concordant barrier that can be computed efficiently.

Broader Impact

This is a theoretical work and does not present any foreseeable societal consequences.

Acknowledgments and Disclosure of Funding

SI was supported by JST, ACT-I, Grant Number JPMJPR18U5, Japan. SH was supported by JST, ACT-I, Grant Number JPMJPR17UM, Japan. TS was supported by JST, ERATO, Grant Number JPMJER1903, Japan. YY was supported by JSPS KAKENHI Grant Number 18H05291, Japan.

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *NeurIPS*, pages 2312–2320, 2011.

[2] N. Abe and P. M. Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pages 3–11, 1999.

[3] J. D. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *COLT*, 2008.

[4] A. Agarwal, A. Krishnamurthy, J. Langford, H. Luo, et al. Open problem: First-order regret bounds for contextual bandits. In *COLT*, pages 4–7, 2017.

[5] Z. Allen-Zhu, S. Bubeck, and Y. Li. Make the minority great again: First-order regret bound for contextual bandits. In *ICML*, pages 186–194, 2018.

[6] C. Allenberg, P. Auer, L. Györfi, and G. Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *ALT*, pages 229–243. Springer, 2006.

[7] S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

[8] J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

[9] J.-Y. Audibert, S. Bubeck, and G. Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2013.

[10] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

[11] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[12] B. Awerbuch and R. D. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *STOC*, pages 45–53, 2004.

[13] S. Bubeck, N. Cesa-Bianchi, and S. Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *COLT*, volume 23, pages 41.1–41.14, 2012.

[14] S. Bubeck, Y. T. Lee, and R. Eldan. Kernel-based methods for bandit convex optimization. In *STOC*, pages 72–85, 2017.

[15] S. Bubeck, M. Cohen, and Y. Li. Sparsity, variance and curvature in multi-armed bandits. In *ALT*, pages 111–127, 2018.

[16] S. Bubeck, Y. Li, H. Luo, and C.-Y. Wei. Improved path-length regret bounds for bandits. In *COLT*, pages 508–528, 2019.

[17] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge university press, 2006.

[18] N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

[19] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2-3):321–352, 2007.

[20] N. Cesa-Bianchi, C. Gentile, and Y. Mansour. Nonstochastic bandits with composite anonymous feedback. In *COLT*, pages 750–773, 2018.

[21] W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *AISTATS*, pages 208–214, 2011.

[22] R. Combes, M. S. T. M. Shahi, A. Proutiere, and M. Lelarge. Combinatorial bandits revisited. In *NeurIPS*, pages 2116–2124, 2015.

[23] A. Cutkosky. Combining online learning guarantees. In *COLT*, pages 895–913, 2019.

[24] V. Dani and T. P. Hayes. Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary. In *SODA*, pages 937–943, 2006.

[25] V. Dani, S. M. Kakade, and T. P. Hayes. The price of bandit information for online optimization. In *NeurIPS*, pages 345–352, 2008.

[26] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[27] S. Gerchinovitz and T. Lattimore. Refined lower bounds for adversarial bandits. In *NeurIPS*, pages 1198–1206, 2016.

[28] A. György, T. Linder, G. Lugosi, and G. Ottucsák. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8(Oct):2369–2403, 2007.

[29] E. Hazan and S. Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine Learning*, 80(2-3):165–188, 2010.

[30] E. Hazan and S. Kale. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12:1287–1311, 2011.

[31] E. Hazan and S. Kale. A simple multi-armed bandit algorithm with optimal variation-bounded regret. In *COLT*, pages 817–820, 2011.

[32] E. Hazan and Z. Karnin. Volumetric spanners: an efficient exploration basis for learning. *The Journal of Machine Learning Research*, 17(1):4062–4095, 2016.

[33] E. Hazan and T. Koren. The computational power of optimization in online learning. In *STOC*, pages 128–141, 2016.

[34] D. Hoeven, T. Erven, and W. Kotłowski. The many faces of exponential weights in online learning. In *COLT*, pages 2067–2092, 2018.

[35] S. Ito. Submodular function minimization with noisy evaluation oracle. In *NeurIPS*, pages 12103–12113, 2019.

[36] S. Ito, D. Hatano, H. Sumita, K. Takemura, T. Fukunaga, N. Kakimura, and K.-I. Kawarabayashi. Oracle-efficient algorithms for online linear optimization with bandit feedback. In *NeurIPS*, pages 10590–10599, 2019.

[37] S. Ito, D. Hatano, H. Sumita, K. Takemura, T. Fukunaga, N. Kakimura, and K.-I. Kawarabayashi. Delay and cooperation in nonstochastic linear bandits. In *NeurIPS*, 2020, to appear.

[38] A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

[39] J. Langford and T. Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *NeurIPS*, pages 817–824, 2007.

[40] T. Lattimore and C. Szepesvári. Bandit algorithms. *preprint, Revision: 1699*, 2019.

[41] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *WWW*, pages 661–670, 2010.

[42] L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

[43] G. Neu. First-order regret bounds for combinatorial semi-bandits. In *COLT*, pages 1360–1375, 2015.

[44] A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *COLT*, pages 993–1019, 2013.

[45] S. Rakhlin and K. Sridharan. Optimization, learning, and games with predictable sequences. In *NeurIPS*, pages 3066–3074, 2013.

[46] A. Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, 1998.

[47] T. Uchiya, A. Nakamura, and M. Kudo. Algorithms for adversarial bandit problems with multiple plays. In *ALT*, pages 375–389, 2010.

[48] C.-Y. Wei and H. Luo. More adaptive algorithms for adversarial bandits. In *COLT*, pages 1263–1291, 2018.

A Proof of Lemma 1

Proof. Since a linear transformation of a log-concave random variable follows a log-concave distribution as well (Theorem 5.1 in [42]), each x_i follows a log-concave distribution and we have $\mathbf{E}[x_i^2] \leq 1$ from the assumption of $S(p) \preceq I$. Then, we have

$$\Pr[\|x\|_2^2 \geq d\alpha^2] \leq \Pr[\exists i \in [d], x_i^2 \geq \alpha^2] \leq \sum_{i=1}^{d} \Pr[|x_i| \geq \alpha] \leq d \exp(1 - \alpha), \tag{15}$$

where the last inequality follows from Lemma 5.7 in [42]. □

B Proof of Theorem 1

Here, we provide a proof of this Theorem 1, and hereafter, we assume that γ_t and ℓ_t satisfy the assumptions in Theorem 1.

Since we have $\mathbf{E}[a_t | \tilde{p}_t] = \mathbf{E}[x_t | \tilde{p}_t] = \mu(\tilde{p}_t)$, the expected regret can be expressed as

$$\begin{aligned} \mathbf{E}[R_T(a^*)] &= \mathbf{E}\left[\sum_{t=1}^{T} \langle \ell_t, a_t - a^* \rangle\right] = \mathbf{E}\left[\sum_{t=1}^{T} \langle \ell_t, \mu(\tilde{p}_t) - a^* \rangle\right] \\ &= \mathbf{E}\left[\sum_{t=1}^{T} \langle \ell_t, \mu(\tilde{p}_t) - \mu(p_t) \rangle\right] + \mathbf{E}\left[\sum_{t=1}^{T} \left\langle \hat{\ell}_t, \mu(p_t) - a^* \right\rangle\right] \end{aligned} \tag{16}$$

where the last equality follows from Lemma 2. The first term in (16) can be bounded by using Lemma 1. Indeed, we can show that $\tilde{p}_t$ is *close* to p_t using the following lemma:

Lemma 4. *For any function $f : \mathcal{A}' \to [-1, 1]$ we have*

$$\left| \mathop{\mathbf{E}}_{x \sim p_t}[f(x)] - \mathop{\mathbf{E}}_{x \sim \tilde{p}_t}[f(x)] \right| \leq 10 d \exp(-\gamma_t) \leq \frac{1}{2t^2}. \tag{17}$$

Further, we have

$$\frac{3}{4} S(p_t) \preceq S(\tilde{p}_t) \preceq \frac{4}{3} S(p_t) \tag{18}$$

Proof. From the definition (6) of $\tilde{p}_t$, we have

$$\begin{aligned} \mathop{\mathbf{E}}_{x \sim \tilde{p}_t}[f(x)] &= \frac{1}{\Pr_{x \sim p_t}[\|x\|_{S(p_t)^{-1}}^2 \leq d\gamma_t^2]} \int_{x \in \mathcal{A}'} f(x) \mathbf{1}\{\|x\|_{S(p_t)^{-1}}^2 \leq d\gamma_t^2\} p_t(x) \mathrm{d}x \\ &= \frac{1}{1 - \delta} \int_{x \in \mathcal{A}'} f(x) \mathbf{1}\{\|x\|_{S(p_t)^{-1}}^2 \leq d\gamma_t^2\} p_t(x) \mathrm{d}x \\ &= \frac{1}{1 - \delta} \left(\mathop{\mathbf{E}}_{x \sim p_t}[f(x)] - \int_{x \in \mathcal{A}'} f(x) \mathbf{1}\{\|x\|_{S(p_t)^{-1}}^2 > d\gamma_t^2\} p_t(x) \mathrm{d}x \right), \end{aligned}$$

where we denote $\delta = \Pr_{x \sim p_t}[\|x\|_{S(p_t)^{-1}}^2 > d\gamma_t^2]$. From this expression, we have

$$\begin{aligned} \left| \mathop{\mathbf{E}}_{x \sim p_t}[f(x)] - \mathop{\mathbf{E}}_{x \sim \tilde{p}_t}[f(x)] \right| &= \frac{1}{1 - \delta} \left| \delta \mathop{\mathbf{E}}_{x \sim p_t}[f(x)] + \int_{x \in \mathcal{A}'} f(x) \mathbf{1}\{\|x\|_{S(p_t)^{-1}}^2 > d\gamma_t^2\} p_t(x) \mathrm{d}x \right| \\ &\leq \frac{1}{1 - \delta} \left(\delta \mathop{\mathbf{E}}_{x \sim p_t}[1] + \int_{x \in \mathcal{A}'} \mathbf{1}\{\|x\|_{S(p_t)^{-1}}^2 > d\gamma_t^2\} p_t(x) \mathrm{d}x \right) = \frac{2\delta}{1 - \delta}, \end{aligned} \tag{19}$$

where the inequality follows from the assumption that $f(x) \in [-1, 1]$. Since p_t is a log-concave distribution, we can apply Lemma 1 to $x = S(p_t)^{-1/2} y$ with $y \sim p_t$. In fact, assumptions in Lemma 1 hold since we have $\mathbf{E}[x x^\top] = S(p_t)^{-1/2} \mathbf{E}[y y^\top] S(p_t)^{-1/2} = S(p_t)^{-1/2} S(p_t) S(p_t)^{-1/2} = I$ and since log-concavity is preserved under any liner transformation. Using Lemma 1 for $x = S(p_t)^{-1/2}y$, we have

$$\delta = \Pr_{x\sim p_t}[\|x\|^2_{S(p_t)^{-1}} > d\gamma_t^2] \leq d\exp(1-\gamma_t) \leq 3d\exp(-\gamma_t) \leq \frac{1}{6t^2}, \tag{20}$$

where the last inequality follows from $\gamma_t \geq 4\log(10dt)$. Combining (19) and (20), we obtain (17). We next show (18). For any $y \in \mathbb{R}^d$, we have

$$\begin{aligned} y^\top S(\tilde{p}_t)y &= \mathop{\mathbf{E}}_{x\sim\tilde{p}_t}\left[(y^\top x)^2\right] = \frac{1}{1-\delta}\mathop{\mathbf{E}}_{x\sim p_t}\left[(y^\top x)^2\mathbf{1}\{\|x\|^2_{S(p_t)^{-1}} \leq d\gamma_t^2\}\right] \\ &\leq \frac{1}{1-\delta}\mathop{\mathbf{E}}_{x\sim p_t}\left[(y^\top x)^2\right] = \frac{1}{1-\delta}y^\top S(p)y. \end{aligned}$$

Since this holds for all $y \in \mathbb{R}^d$ and $\frac{1}{1-\delta} \leq 4/3$, the second inequality in (18) holds. Furthermore, we have

$$\begin{aligned} y^\top S(p_t)y - y^\top S(\tilde{p}_t)y &= \mathop{\mathbf{E}}_{x\sim p_t}\left[(y^\top x)^2\right] - \frac{1}{1-\delta}\mathop{\mathbf{E}}_{x\sim p_t}\left[(y^\top x)^2\mathbf{1}\{\|x\|^2_{S(p_t)^{-1}} \leq d\gamma_t^2\}\right] \\ &\leq \mathop{\mathbf{E}}_{x\sim p_t}\left[(y^\top x)^2\mathbf{1}\{\|x\|^2_{S(p_t)^{-1}} > d\gamma_t^2\}\right] \\ &\leq y^\top S(p_t)y\mathop{\mathbf{E}}_{x\sim p_t}\left[\|x\|^2_{S(p_t)^{-1}}\mathbf{1}\{\|x\|^2_{S(p_t)^{-1}} > d\gamma_t^2\}\right], \end{aligned} \tag{21}$$

where the last inequality follows from the Cauchy–Schwarz inequality:

$$(y^\top x)^2 = \left(\left\langle S(p_t)^{\frac{1}{2}}y, S(p_t)^{-\frac{1}{2}}x\right\rangle\right)^2 \leq \|S(p_t)^{\frac{1}{2}}y\|_2^2 \cdot \|S(p_t)^{-\frac{1}{2}}x\|_2^2 = y^\top S(p_t)y \cdot \|x\|^2_{S(p_t)^{-1}}.$$

The right-hand side of (21) can be bounded by using Lemma 1 as follows:

$$\begin{aligned} &\mathop{\mathbf{E}}_{x\sim p_t}\left[\|x\|^2_{S(p_t)^{-1}}\mathbf{1}\{\|x\|^2_{S(p_t)^{-1}} > d\gamma_t^2\}\right] \\ &\leq \sum_{n=1}^{\infty}(n+1)^2 d\gamma_t^2 \Pr_{x\sim p_t}\left[n^2 d\gamma_t^2 \leq \|x\|^2_{S(p)^{-1}} \leq (n+1)^2 d\gamma_t^2\right] \\ &\leq \sum_{n=1}^{\infty}(n+1)^2 d\gamma_t^2 \cdot d\exp(1-n\gamma_t) \\ &\leq d^2\gamma_t^2\sum_{n=1}^{\infty}\exp(2+n-n\gamma_t) = d^2\gamma_t^2\frac{\exp(3-\gamma_t)}{1-\exp(1-\gamma_t)} \leq \frac{1}{4} \end{aligned} \tag{22}$$

where the second inequality follows from Lemma 1, the second inequality comes from $y^2 \leq \exp(y)$ for $y \leq 0$, and the last inequality follows from the assumption of $\gamma_t \geq 4\log(10dt)$. Combining (21), (22) and the assumption of $\gamma_t \geq 4\log(10dt)$, we obtain the first inequality of (18). □

Since $\langle \ell_t, x\rangle \in [-1,1]$ for all $x \in \mathcal{A}'$, (17) implies $|\langle \ell_t, \mu(\tilde{p}_t) - \mu(p_t)\rangle| \leq 1/(2t^2)$.

The second term in (16) can be bounded by following the analysis of optimistic mirror descent [45]:

Lemma 5. *For any $a^* \in \mathcal{A}$, we have*

$$\sum_{t=1}^{T}\left\langle \hat{\ell}_t, \mu(p_t) - a^*\right\rangle \leq \sum_{t=1}^{T}\left(\frac{1}{\eta_t}\mathop{\mathbf{E}}_{x\sim p_t}\left[\psi\left(-\eta_t\left\langle\hat{\ell}_t - m_t, x\right\rangle\right)\right]\right) + \frac{d\log T}{\eta_T} + \frac{1}{T}\sum_{t=1}^{T}\left\langle\hat{\ell}_t, \bar{a} - a^*\right\rangle,$$

where

$$\psi(y) = \exp(y) - y - 1, \quad \bar{a} = \mu(p_0). \tag{23}$$

Proof. Define $v_t : \mathcal{A}' \to \mathbb{R}$ and $u_t : \mathcal{A}' \to \mathbb{R}$ by

$$u_t(x) = \exp\left(-\eta_t\left\langle\sum_{j=1}^{t}\hat{\ell}_j, x\right\rangle\right), \quad v_t(x) = \exp\left(-\eta_{t+1}\left\langle\sum_{j=1}^{t}\hat{\ell}_j, x\right\rangle\right), \tag{24}$$

and define

$$U_t = \int_{x\in\mathcal{A}'} u_t(x)\mathrm{d}x, \quad V_t = \int_{x\in\mathcal{A}'} v_t(x)\mathrm{d}x, \quad W_t = \int_{x\in\mathcal{A}'} w_t(x)\mathrm{d}x. \tag{25}$$

Since u_t can be expressed as $u_t(x) = w_t(x)\exp\left(-\eta_t\left\langle \hat{\ell}_t - m_t, x\right\rangle\right)$ from the definitions (5) and (24) of w_t and u_t, respectively, we have

$$\begin{aligned} U_t &= \int_{x\in\mathcal{A}'} w_t(x)\exp\left(-\eta_t\left\langle \hat{\ell}_t - m_t, x\right\rangle\right)\mathrm{d}x = W_t \cdot \mathop{\mathbf{E}}_{x\sim p_t}\left[\exp\left(-\eta_t\left\langle \hat{\ell}_t - m_t, x\right\rangle\right)\right] \\ &= W_t \cdot \left(1 - \eta_t\left\langle \hat{\ell}_t - m_t, \mu(p_t)\right\rangle + \mathop{\mathbf{E}}_{x\sim p_t}\left[\psi\left(-\eta_t\left\langle \hat{\ell}_t - m_t, x\right\rangle\right)\right]\right). \end{aligned}$$

By taking the logarithms of both sides, we obtain

$$\begin{aligned} \log U_t &= \log W_t + \log\left(1 - \eta_t\left\langle \hat{\ell}_t - m_t, \mu(p_t)\right\rangle + \mathop{\mathbf{E}}_{x\sim p_t}\left[\psi\left(-\eta_t\left\langle \hat{\ell}_t - m_t, x\right\rangle\right)\right]\right) \\ &\le \log W_t - \eta_t\left\langle \hat{\ell}_t - m_t, \mu(p_t)\right\rangle + \mathop{\mathbf{E}}_{x\sim p_t}\left[\psi\left(-\eta_t\left\langle \hat{\ell}_t - m_t, x\right\rangle\right)\right], \end{aligned}$$

where we used the inequality $\log(1+x) \le x$ for $x > -1$. The condition $x > -1$ indeed holds since x here corresponds to $x = -1 + \mathbf{E}[\exp(-\eta_t\langle\hat{\ell}_t - m_t, x\rangle)]$. Hence, we have

$$\left\langle \hat{\ell}_t - m_t, \mu(p_t)\right\rangle \le \frac{1}{\eta_t}\left(\log\frac{W_t}{U_t} + \mathop{\mathbf{E}}_{x\sim p_t}\left[\psi\left(-\eta_t\left\langle \hat{\ell}_t - m_t, x\right\rangle\right)\right]\right). \tag{26}$$

Similarly, since we have

$$V_{t-1} = \int_{x\in\mathcal{A}'} w_t(x)\exp\left(\eta_t\left\langle m_t, x\right\rangle\right)\mathrm{d}x = W_t \cdot \mathop{\mathbf{E}}_{x\sim p_t}\left[\exp\left(\eta_t\left\langle m_t, x\right\rangle\right)\right] \ge W_t \cdot \exp(\eta_t\left\langle m_t, \mu(p_t)\right\rangle),$$

where we applied Jensen's inequality, it holds that

$$\langle m_t, \mu(p_t)\rangle \le \frac{1}{\eta_t}\log\frac{V_{t-1}}{W_t}. \tag{27}$$

Combining (26) and (27) and taking the sum over $t = 1, \ldots, T$, we obtain

$$\sum_{t=1}^T \left\langle \hat{\ell}_t, \mu(p_t)\right\rangle \le \sum_{t=1}^T \frac{1}{\eta_t}\left(\log\frac{V_{t-1}}{U_t} + \mathop{\mathbf{E}}_{x\sim p_t}\left[\psi\left(-\eta_t\left\langle \hat{\ell}_t - m_t, x\right\rangle\right)\right]\right). \tag{28}$$

Furthermore, noting that $V_0 = U_0 = \mathrm{vol}(\mathcal{A}')$, we have

$$\begin{aligned} \sum_{t=1}^T \frac{1}{\eta_t}\log\frac{V_{t-1}}{U_t} &= \sum_{t=1}^T \frac{1}{\eta_t}\left(\log\frac{V_{t-1}}{V_0} - \log\frac{U_t}{U_0}\right) \\ &= \sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}}\log\frac{V_t}{V_0} - \frac{1}{\eta_t}\log\frac{U_t}{U_0}\right) - \frac{1}{\eta_T}\log\frac{U_T}{U_0} \le -\frac{1}{\eta_T}\log\frac{U_T}{U_0}, \end{aligned} \tag{29}$$

where the inequality follows from the assumption of $\eta_{t+1} \le \eta_t$ and Jensen's inequality, as follows:

$$\begin{aligned} \frac{1}{\eta_{t+1}}\log\frac{V_t}{V_0} &= \frac{1}{\eta_{t+1}}\log\mathop{\mathbf{E}}_{x\sim p_0}\left[\exp\left(-\eta_{t+1}\left\langle\sum_{j=1}^t \hat{\ell}_j, x\right\rangle\right)\right] \\ &= \frac{1}{\eta_{t+1}}\log\mathop{\mathbf{E}}_{x\sim p_0}\left[\exp\left(-\eta_t\left\langle\sum_{j=1}^t \hat{\ell}_j, x\right\rangle\right)^{\frac{\eta_{t+1}}{\eta_t}}\right] \\ &\le \frac{1}{\eta_{t+1}}\log\mathop{\mathbf{E}}_{x\sim p_0}\left[\exp\left(-\eta_t\left\langle\sum_{j=1}^t \hat{\ell}_j, x\right\rangle\right)\right]^{\frac{\eta_{t+1}}{\eta_t}} \\ &= \frac{1}{\eta_t}\log\mathop{\mathbf{E}}_{x\sim p_0}\left[\exp\left(-\eta_t\left\langle\sum_{j=1}^t \hat{\ell}_j, x\right\rangle\right)\right] = \frac{1}{\eta_t}\log\frac{U_t}{U_0}, \end{aligned}$$

where the first and the last equalities follow from the definitions (24) and (25) of U_t and V_t, and the inequality holds since the function $x \mapsto x^{\frac{\eta_{t+1}}{\eta_t}}$ $(x > 0)$ is a concave function. Set $\mathcal{A}_{a^*} := \{(1 - \frac{1}{T})a^* + \frac{1}{T}y \mid y \in \mathcal{A}\} \subseteq \mathcal{A}'$ and let p^* denote a uniform distribution over $\mathcal{A}_{a^*}$. We then have

$$\begin{aligned} U_T &\geq \int_{x \in \mathcal{A}_{a^*}} \exp\left(-\eta_T \left\langle \sum_{t=1}^T \hat{\ell}_t, x \right\rangle\right) \mathrm{d}x \\ &= T^{-d} \int_{y \in \mathcal{A}'} \exp\left(-\eta_T \left\langle \sum_{t=1}^T \hat{\ell}_t, \left(1 - \frac{1}{T}\right) a^* + \frac{1}{T} y \right\rangle\right) \mathrm{d}y \\ &\geq T^{-d} U_0 \exp\left(-\eta_T \left\langle \sum_{t=1}^T \hat{\ell}_t, \left(1 - \frac{1}{T}\right) a^* + \frac{1}{T} \bar{a} \right\rangle\right), \end{aligned}$$

where $\bar{a} = \mu(p_0)$ and the last inequality follows from Jensen's inequality. Taking the logarithms of both sides, we have

$$-\frac{1}{\eta_T} \log \frac{U_T}{U_0} \leq \left\langle \sum_{t=1}^T \hat{\ell}_t, \left(1 - \frac{1}{T}\right) a^* + \frac{1}{T} \bar{a} \right\rangle + \frac{d \log T}{\eta_T} = \left\langle \sum_{t=1}^T \hat{\ell}_t, a^* + \frac{1}{T}(\bar{a} - a^*) \right\rangle + \frac{d \log T}{\eta_T}.$$

Combining this, (28), and (29), we obtain the desired inequality in the statement of Lemma 5. □

We next evaluate the term $\mathbf{E}_{x \sim p_t}\left[\psi\left(-\eta_t \left\langle \hat{\ell}_t - m_t, x \right\rangle\right)\right]$ in Lemma 5. From the definition of ψ, $\psi(y) \leq y^2$ for $|y| \leq 1$, and hence $\mathbf{E}[\psi(y)]$ can be well approximated with $\mathbf{E}[y^2]$ if y follows a log-concave distribution and $\mathbf{E}[y^2]$ is small enough:

Lemma 6. *If y follows a log-concave distribution over $\mathbb{R}$ and if $\mathbf{E}[y^2] \leq 1/100$, we have*

$$\mathbf{E}[\psi(y)] \leq \mathbf{E}[y^2] + 30 \exp\left(-\frac{1}{\sqrt{\mathbf{E}[y^2]}}\right) \leq 2\,\mathbf{E}[y^2] \quad \textit{where} \quad \psi(x) = \exp(x) - x - 1. \tag{30}$$

Proof. Let $s = \sqrt{\mathbf{E}[x^2]}$. We have $s \leq 1/10$ from the assumption. We can bound $\mathbf{E}[\psi(x)]$ as follows:

$$\begin{aligned} \mathbf{E}[\psi(x)] &= \mathbf{E}[\psi(x)\mathbf{1}\{x \leq 1\}] + \mathbf{E}[\psi(x)\mathbf{1}\{x > 1\}] \leq \mathbf{E}[x^2 \mathbf{1}\{x \leq 1\}] + \mathbf{E}[\exp(x)\mathbf{1}\{x > 1\}] \\ &\leq s^2 + \sum_{n=1}^\infty \exp(n+1) \Pr[n < x \leq n+1] \leq s^2 + \sum_{n=1}^\infty \exp(n+1) \Pr\left[\frac{n}{s} < \frac{x}{s}\right] \\ &\leq s^2 + \sum_{n=1}^\infty \exp(n+1) \exp\left(1 - \frac{n}{s}\right) = s^2 + \frac{\exp(3 - s^{-1})}{1 - \exp(1 - s^{-1})} \leq s^2 + 30 \exp(-s^{-1}), \end{aligned}$$

where the first inequality follows from $\psi(x) \leq x^2$ for $x \leq 1$ and $\psi(x) \leq \exp(x)$ for $x > 1$, the forth inequality follows from Lemma 1, and the last inequality follows from $s \leq 1/10$. By combining this and the fact that $30 \exp(-s^{-1}) \leq s^2$ for $0 < s \leq 1/10$, we obtain Lemma 6. □

We give a bound for $\mathbf{E}_{x \sim p_t}\left[\psi\left(-\eta_t \left\langle \hat{\ell}_t - m_t, x \right\rangle\right)\right]$ by applying Lemma 6 with $y = -\eta_t \left\langle \hat{\ell}_t - m_t, x \right\rangle$. We can confirm that y satisfies the assumption of Lemma 6 thanks to the truncated distribution that ensures $\|x_t\|^2_{S(p_t)^{-1}} \leq d\gamma_t^2$. In fact, we have

$$\begin{aligned} &\mathop{\mathbf{E}}_{x \sim p_t}\left[\left(-\eta_t \left\langle \hat{\ell}_t - m_t, x \right\rangle\right)^2\right] = \eta_t^2 (\hat{\ell}_t - m_t)^\top S(p_t)(\hat{\ell}_t - m_t) \\ &= \eta_t^2 (\langle \ell_t - m_t, a_t \rangle)^2 x_t^\top S(\tilde{p}_t)^{-1} S(p_t) S(\tilde{p}_t)^{-1} x_t \leq \frac{4}{3} \eta_t^2 (\langle \ell_t - m_t, a_t \rangle)^2 x_t^\top S(\tilde{p}_t)^{-1} S(\tilde{p}_t) S(\tilde{p}_t)^{-1} x_t \\ &= \frac{4}{3} \eta_t^2 (\langle \ell_t - m_t, a_t \rangle)^2 \|x_t\|^2_{S(\tilde{p}_t)^{-1}} \leq 2\eta_t^2 (\langle \ell_t - m_t, a_t \rangle)^2 \|x_t\|^2_{S(p_t)^{-1}} \leq 2d\gamma_t^2 \eta_t^2 (\langle \ell_t - m_t, a_t \rangle)^2 \leq 1/100, \end{aligned}$$

where the first and second inequalities follow from (18), the third inequality follows from x_t being sampled from $\tilde{p}_t$ and the definition (6) of $\tilde{p}_t$, and the last inequality follows from the assumption of $\eta_t \leq \frac{1}{\sqrt{800 d \gamma_t}}$. Hence, by Lemma 6, we have

$$\mathop{\mathbf{E}}_{x \sim p_t}\left[\psi\left(-\eta_t \left\langle \hat{\ell}_t - m_t, x \right\rangle\right)\right] \leq 4 d \gamma_t^2 \eta_t^2 (\langle \ell_t - m_t, a_t \rangle)^2. \tag{31}$$

Combining this, (16), (17) and Lemma 5, we obtain

$$\mathbf{E}[R_T(a^*)] \leq \sum_{t=1}^{T} \frac{1}{2t^2} + \mathbf{E}\left[\sum_{t=1}^{T} 4 d \gamma_t^2 \eta_t (\langle \ell_t - m_t, a_t \rangle)^2 + \frac{d \log T}{\eta_T} + \frac{1}{T}\sum_{t=1}^{T}\left\langle \hat{\ell}_t, \bar{a} - a^* \right\rangle\right].$$

Since we have $\sum_{t=1}^{T} \frac{1}{t^2} \leq 2$ and since Lemma 2 implies that $\mathbf{E}\left[\left\langle \hat{\ell}_t, \bar{a} - a^* \right\rangle\right] = \langle \ell_t, \bar{a} - a^* \rangle \leq 2$, we obtain (9).

We can now show (10) on the basis of (9). Suppose that γ_t and η_t are defined as

$$\gamma_t = 4 \log(10 d t), \quad \eta_t = \frac{1}{\sqrt{800 d \gamma_t^2 + 16 \sum_{j=1}^{t-1} \gamma_j^2 (\langle \ell_j - m_j, a_j \rangle)^2}}. \tag{32}$$

Denote $\beta_t = 16 \gamma_t^2 (\langle \ell_t - m_t, a_t \rangle)^2$ and $\eta_t' = \frac{1}{\sqrt{800 d \gamma_{t-1}^2 + 16 \sum_{j=1}^{t-1} \gamma_j^2 (\langle \ell_j - m_j, a_j \rangle)^2}}$. We then have

$$\begin{aligned}
\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \geq \frac{1}{\eta_{t+1}'} - \frac{1}{\eta_t} &= \sqrt{800 d \gamma_t^2 + \sum_{j=1}^{t} \beta_j} - \sqrt{800 d \gamma_t^2 + \sum_{j=1}^{t-1} \beta_j} \\
&\geq \frac{\beta_t}{2}\left(800 d \gamma_t^2 + \sum_{j=1}^{t} \beta_j\right)^{-1/2} \geq \frac{\beta_t}{4}\left(800 d \gamma_t^2 + \sum_{j=1}^{t-1} \beta_j\right)^{-1/2} = \frac{1}{4}\beta_t \eta_t,
\end{aligned} \tag{33}$$

where the first inequality follows from $\gamma_{t+1} \geq \gamma_t$, the second inequality follows from $\sqrt{y} - \sqrt{y - x} \geq \frac{x}{2\sqrt{y}}$ for $0 \leq x < y$, and the last inequality follows from $\beta_t \leq 800 d \gamma_t^2$. We now have

$$\begin{aligned}
\mathbf{E}[R_T(a^*)] &\leq d \cdot \mathbf{E}\left[\frac{1}{4}\sum_{t=1}^{T} \eta_t \beta_t + \frac{\log T}{\eta_T}\right] + 3 \\
&\leq d \cdot \mathbf{E}\left[\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) + \frac{1}{\eta_{T+1}'} - \frac{1}{\eta_T} + \frac{\log T}{\eta_T}\right] + 3 \\
&= d \cdot \mathbf{E}\left[\frac{1}{\eta_{T+1}} - \frac{1}{\eta_1} + \frac{\log T}{\eta_T}\right] + 3 \\
&\leq 2 d \log T \cdot \mathbf{E}\left[\frac{1}{\eta_{T+1}'}\right] = 2 d \log T \cdot \mathbf{E}\left[\sqrt{800 d \gamma_T^2 + \sum_{t=1}^{T} \beta_t}\right] \\
&\leq 32 d \log T \cdot \log(10 d T) \cdot \mathbf{E}\left[\sqrt{\sum_{t=1}^{T} (\langle \ell_t - m_t, a_t \rangle)^2 + 50 d}\right],
\end{aligned}$$

where the first inequality follows from (9) and the definition of β_t, the second inequality follows from (33), and the last inequality follows from (32) and the definition of β_t. □

C Construction of a Matrix Satisfying (13)

Here for $C > 1$, a subset $X = \{x_1, \ldots, x_d\} \subseteq \mathcal{A}$ is said to be a *C-barycentric spanner* for $\mathcal{A}$ if every $a \in \mathcal{A}$ can be expressed as a linear combination of elements in X with coefficients in $[-C, C]$.

Theorem 5 (Proposition 2.4 in [12]). *Suppose $\mathcal{A} \subseteq \mathbb{R}^d$ is a full-dimensional compact set. Given $C > 1$ and an algorithm for linear optimization over $\mathcal{A}$, we can compute a C-barycentric spanner for $\mathcal{A}$ in polynomial time, making $O(d^2 \log_C d)$ calls to the linear optimization oracle.*

Given a 2-barycentric spanner $X = \{x_1, \ldots, x_d\}$, define S by

$$M = (x_1 \; x_2 \; \cdots \; x_d), \quad S = MM^\top = \sum_{i=1}^{d} x_i x_i^\top. \tag{34}$$

Recalling that $x_i \in \mathcal{A}$ and $m \in \mathcal{L}$, we have $\|m\|_S^2 = m^\top \left(\sum_{i=1}^d x_i x_i^\top \right) m = \sum_{i=1}^d (\langle x_i, m \rangle)^2 \leq d$. Further, from the definition of a 2-barycentric spanner, for any $a \in \mathcal{A}$, there exists $u \in [-2, 2]^d$ such that $a = Mu$ and hence we have $\|a\|_{S^{-1}}^2 = a^\top (M^{-1})^\top M^{-1} a = u^\top u \leq 4d$. Consequently, the matrix S defined by (34) satisfies (13).

Remark 2. Given a *d-optimal design* for $\mathcal{A}$, one can construct a matrix S such that

$$\|m\|_S^2 \leq d \text{ for all } m \in \mathcal{L}, \quad \|a\|_{S^{-1}}^2 \leq 1 \text{ for all } a \in \mathcal{A}, \tag{35}$$

which helps improve the regret bound by reducing the term $\log(1 + 4T)$ in Theorem 2 into $\log(1 + T/d)$. Computing *d-optimal design* is, however, harder than computing barycentric spanners for many examples of $\mathcal{A}$.

D Proof of Theorem 4

To show Theorem 4, we use the following lower bound:

Theorem 6. *Theorem 4.3. in [25], Lemmas 3. and 4. in [35] Suppose that ℓ_t is generated as follows: Let $\varepsilon = \min\{\frac{1}{6}, \frac{d}{\sqrt{8T}}\}$ be a fixed parameter. Pick $a^* \in \{-1, 1\}^d$ from a uniform distribution over $\{-1, 1\}^d$. For $t = 1, \ldots, T$, $i(t)$ is chosen from a uniform distribution over $[d]$ independently. Then $\ell_{ti(t)}$ follows a Bernoulli distribution, where $\Pr[\ell_{ti(t)} = 1] = \frac{1 - \varepsilon a^*_{i(t)}}{2}$ and $\Pr[\ell_{i(t)} = -1] = \frac{1 + \varepsilon a^*_{i(t)}}{2}$. For $i \neq i(t)$, $\ell_{ti} = 0$. Then any algorithm for $\mathcal{A} = \{-1, 1\}^d$ suffers regret of*

$$\mathbf{E}\left[\max_{a^* \in \mathcal{A}} R_T(a^*)\right] \geq \frac{\varepsilon T}{2} = \min\left\{\frac{d\sqrt{T}}{\sqrt{32}}, \frac{T}{12}\right\}, \tag{36}$$

where the expectation is taken over the choice of ℓ_t and the randomness of the algorithm.

Let us prove Theorem 4. Note that $\mathcal{A}$ is given by $\mathcal{A} = \{-1, 1\}^{d-1} \times \{1\}$. Let ℓ_t be a random vector defined as follows: for $t \leq L$, $(\ell_{t1}, \ldots, \ell_{t,d-1})$ are generated in a similar way as in Theorem 6 with T and d replaced by $\lfloor L \rfloor$ and $d - 1$, respectively, multiplied by $1/2$, and we set $\ell_{td} = 1/2$. For $t > L$ we set $\ell_t = 0$. We then have $\|\ell_t\|_1 \leq 1$ and $0 \leq \langle \ell_t, a \rangle \leq 1$ for any $a \in \mathcal{A}$ and for $t \leq L$. Combining this and $\ell_t = 0$ for $t > L$, we can confirm that (i), (ii), and (iii) in Theorem 4 hold. From Theorem 6, we have

$$\mathbf{E}\left[\max_{a^* \in \mathcal{A}} R_T(a^*)\right] = \Omega\left(\min\{(d-1)\sqrt{\lfloor L \rfloor}, \lfloor L \rfloor\}\right) = \Omega(d\sqrt{L}), \tag{37}$$

where the last equality follows from the assumptions of $L = \Omega(d^2)$ and $d \geq 1$, and the expectation is taken over the choice of $(\ell_t)_{t=1}^T$ and the randomness of the algorithm. Since (37) holds for the expectation, there is a (fixed) realization of $(\ell_t)_t^T$ for which (37) holds. If we fix $(\ell_t)_t^T$ to them, $\mathrm{argmax}_{a^* \in \mathcal{A}} R_T(a^*)$ has no randomness, and hence we can exchange $\max_{a^* \in \mathcal{A}}$ and $\mathbf{E}$, i.e., we have $\mathbf{E}[\max_{a^* \in \mathcal{A}} R_T(a^*)] = \max_{a^* \in \mathcal{A}} \mathbf{E}[R_T(a^*)]$, which implies that (iv) in Theorem 4 holds.

E Proof of Lemma 3

For $t = 0, 1, \ldots, T$, define convex functions $f_t : \mathcal{L} \to \mathbb{R}$ and $F_t : \mathcal{L} \to \mathbb{R}$ as follows:

$$\begin{aligned}
f_0(m) &= \frac{1}{2}\|m\|_S^2, \\
f_t(m) &= \frac{1}{2}(\langle \ell_t - m, a_t \rangle)^2 && (t \in [T]), \\
F_t(m) &= \sum_{j=0}^{t} f_j(m) && (t \in \{0, 1, \ldots, T\}).
\end{aligned}$$

Then, the definition (11) of m_t can be rewritten as

$$m_t \in \operatorname*{argmin}_{m \in \mathcal{L}} F_{t-1}(m). \tag{38}$$

Using this fact recursively, we have

$$\begin{aligned}F_T(m^*) &\geq F_T(m_{T+1}) = F_{T-1}(m_{T+1}) + f_T(m_{T+1}) \geq F_{T-1}(m_T) + f_T(m_{T+1}) \\ &= f_{T-2}(m_T) + f_{T-1}(m_T) + f_T(m_{T+1}) \geq \cdots \geq f_0(m_1) + \sum_{t=1}^{T} f_t(m_{t+1}) \\ &\geq \sum_{t=1}^{T} f_t(m_{t+1})\end{aligned}$$

for arbitrary $m^* \in \mathcal{L}$. From this, we have

$$\begin{aligned}&\sum_{t=1}^{T} (\langle \ell_t - m_t, a_t \rangle)^2 - \sum_{t=1}^{T} (\langle \ell_t - m^*, a_t \rangle)^2 = 2\sum_{t=1}^{T} f_t(m_t) - 2\sum_{t=1}^{T} f_t(m^*) \\ &= 2\sum_{t=1}^{T} f_t(m_t) - 2(F_T(m^*) - f_0(m^*)) \leq 2f_0(m^*) + 2\sum_{t=1}^{T}(f_t(m_t) - f_t(m_{t+1})) \\ &= \|m^*\|_S^2 + 2\sum_{t=1}^{T}(f_t(m_t) - f_t(m_{t+1})). \end{aligned} \tag{39}$$

We next show

$$f_t(m_t) - f_t(m_{t+1}) \leq 4\|a_t\|^2_{A_t^{-1}} \tag{40}$$

where we define positive semi-definite matrices $A_t \in \mathbb{R}^{d \times d}$ by

$$A_t = S + \sum_{j=1}^{t} a_j a_j^\top \tag{41}$$

for $t = 0, 1, \ldots, T$. To show (40), we use the fact that F_t is A_t-strongly convex, i.e., it holds for any $m, m' \in \mathcal{L}$ that

$$F_t(m') \geq F_t(m) + \langle \nabla F_t(m), m' - m \rangle + \|m' - m\|^2_{A_t}. \tag{42}$$

Further, (38) implies that

$$\langle \nabla F_{t-1}(m_t), m - m_t \rangle \geq 0 \tag{43}$$

for any $m \in \mathcal{L}$ and $t \in [T]$. From (42) and (43), we can show (40) as follows:

$$\begin{aligned}&f_t(m_t) - f_t(m_{t+1}) = F_t(m_t) - F_t(m_{t+1}) - F_{t-1}(m_t) + F_{t-1}(m_{t+1}) \\ &\leq \langle \nabla F_t(m_t), m_t - m_{t+1} \rangle - \|m_t - m_{t+1}\|^2_{A_t} + \langle \nabla F_{t-1}(m_{t+1}), m_{t+1} - m_t \rangle \\ &\leq \langle \nabla F_t(m_t) - \nabla F_{t-1}(m_t), m_t - m_{t+1} \rangle + \langle \nabla F_{t-1}(m_{t+1}) - \nabla F_t(m_{t+1}), m_{t+1} - m_t \rangle \\ &\quad - \|m_t - m_{t+1}\|^2_{A_t} \\ &= \langle \nabla f_t(m_t), m_t - m_{t+1} \rangle - \|m_t - m_{t+1}\|^2_{A_t} - \langle \nabla f_t(m_{t+1}), m_{t+1} - m_t \rangle \\ &= \langle \nabla f_t(m_t) + \nabla f_t(m_{t+1}), m_t - m_{t+1} \rangle - \|m_t - m_{t+1}\|^2_{A_t} \\ &\leq \|\nabla f_t(m_t) + \nabla f_t(m_{t+1})\|_{A_t^{-1}} \|m_t - m_{t+1}\|_{A_t} - \|m_t - m_{t+1}\|^2_{A_t} \\ &\leq \frac{1}{4}\|\nabla f_t(m_t) + \nabla f_t(m_{t+1})\|^2_{A_t^{-1}} = \frac{1}{4}\|(\langle m_t - \ell_t, a_t \rangle + \langle m_{t+1} - \ell_t, a_t \rangle) a_t\|^2_{A_t^{-1}} \leq 4\|a_t\|^2_{A_t^{-1}},\end{aligned}$$

where the first and second inequalities follow from (42) and (43) respectively, the third inequality follows from Cauchy–Schwarz inequatlity, the forth inequality follows from the fact that $a^2 - ab + b^2/4 = (a - b/2)^2 \geq 0$ for $a, b \in \mathbb{R}$. Combining (39) and (40), we obtain

$$\sum_{t=1}^{T} (\langle \ell_t - m_t, a_t \rangle)^2 - \sum_{t=1}^{T} (\langle \ell_t - m^*, a_t \rangle)^2 \leq \|m^*\|^2_S + 8\sum_{t=1}^{T} \|a_t\|^2_{A_t^{-1}}. \tag{44}$$

We next show

$$\sum_{t=1}^{T} \|a_t\|_{A_t^{-1}}^2 \leq d \log\left(1 + \frac{T}{d} \max_{a \in \mathcal{A}} \|a\|_{S^{-1}}^2\right). \tag{45}$$

Each $\|a_t\|_{A_t^{-1}}^2$ can be bounded by $\log\det A_t - \log\det A_{t-1}$. In fact, we have

$$\begin{aligned}
&\log\det A_t - \log\det A_{t-1} = -(\log\det(A_t - a_t a_t^\top) - \log\det A_t) \\
&= -\log\det(A_t^{-\frac{1}{2}}(A_t - a_t a_t^\top)A_t^{-\frac{1}{2}}) = -\log\det(I - A_t^{-\frac{1}{2}} a_t a_t^\top A_t^{-\frac{1}{2}}) \\
&= -\log(1 - \|A_t^{-\frac{1}{2}} a_t\|_2^2) \geq \|A_t^{-\frac{1}{2}} a_t\|_2^2 = \|a_t\|_{A_t^{-1}}^2,
\end{aligned}$$

where the forth equality holds since the matrix $(I - A_t^{-\frac{1}{2}} a_t a_t^\top A_t^{-\frac{1}{2}})$ has eigenvalues $\lambda_1' = 1 - \|A_t^{-\frac{1}{2}} a_t\|_2^2$ and $\lambda_2' = \lambda_3' = \cdots = \lambda_d' = 1$, and the inequality follows from $\log(1+y) \leq y$ for $y > -1$. From this, we have

$$\sum_{t=1}^{T} \|a_t\|_{A_t^{-1}}^2 \leq \log\det A_T - \log\det A_0 = \log\det\left(I + \sum_{t=1}^{T} S^{-\frac{1}{2}} a_t a_t^\top S^{-\frac{1}{2}}\right) = \sum_{i=1}^{d} \log(1+\lambda_i), \tag{46}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_d \geq 0$ are eigenvalues of $\sum_{t=1}^{T} S^{-\frac{1}{2}} a_t a_t^\top S^{-\frac{1}{2}}$. Since we have

$$\sum_{i=1}^{d} \lambda_i = \mathrm{tr}\left(\sum_{t=1}^{T} S^{-\frac{1}{2}} a_t a_t^\top S^{-\frac{1}{2}}\right) = \sum_{t=1}^{T} \|a_t\|_{S^{-1}}^2 \leq T \max_{a \in \mathcal{A}} \|a\|_{S^{-1}}^2,$$

the right-hand side of (46) can be bounded as

$$\sum_{i=1}^{d} \log(1+\lambda_i) \leq d \log\left(1 + \frac{T}{d} \max_{a \in \mathcal{A}} \|a\|_{S^{-1}}^2\right) \tag{47}$$

which implies that (45) holds. Combining (44) and (45), we obtain the inequality in Lemma 3.