Delay and Cooperation in Nonstochastic Linear Bandits

Shinji Ito
NEC Corporation
i-shinji@nec.com

Daisuke Hatano
RIKEN AIP
daisuke.hatano@riken.jp

Hanna Sumita
Tokyo Institute of Technology
sumita@c.titech.ac.jp

Kei Takemura
NEC Corporation
kei_takemura@nec.com

Takuro Fukunaga
Chuo University, RIKEN AIP, JST PRESTO
fukunaga.07s@g.chuo-u.ac.jp

Naonori Kakimura
Keio University
kakimura@math.keio.ac.jp

Ken-ichi Kawarabayashi
National Institute of Informatics
k-keniti@nii.ac.jp

Abstract

This paper offers a nearly optimal algorithm for online linear optimization with delayed bandit feedback. Online linear optimization with bandit feedback, or nonstochastic linear bandits, provides a generic framework for sequential decision-making problems with limited information. This framework, however, assumes that feedback can be observed just after choosing the action, and, hence, does not apply directly to many practical applications, in which the feedback can often only be obtained after a while. To cope with such situations, we consider problem settings in which the feedback can be observed d rounds after the choice of an action, and propose an algorithm for which the expected regret is $\tilde{O}(\sqrt{m(m+d)T})$, ignoring logarithmic factors in m and T, where m and T denote the dimensionality of the action set and the number of rounds, respectively. This algorithm achieves nearly optimal performance, as we are able to show that arbitrary algorithms suffer the regret of $\Omega(\sqrt{m(m+d)T})$ in the worst case. To develop the algorithm, we introduce a technique we refer to as *distribution truncation*, which plays an essential role in bounding the regret. We also apply our approach to cooperative bandits, as studied by Cesa-Bianchi et al. [18] and Bar-On and Mansour [12], and extend their results to the linear bandits setting.

1 Introduction

Bandit linear optimization (nonstochastic linear bandits) models various sequential decision-making problems under partial-information conditions and has a wide range of applications including combinatorial bandits [17] and adaptive routing problems [11]. In this model, a player is given a set $\mathcal{A} \subseteq \mathbb{R}^m$ of actions, each of which corresponds to an m-dimensional feature vector, and chooses an action $a_t \in \mathcal{A}$ in each round $t \in [T]$. Just after choosing action a_t, the player gets feedback $\ell_t^\top a_t$ of the loss for the chosen action, where $\ell_t \in \mathbb{R}^m$ is a *loss vector*. The goal of the player is to minimize cumulative loss $\sum_{t=1}^{T} \ell_t^\top a_t$. A number of studies have proposed algorithms for this model, achieving sublinear regret bounds of $\tilde{O}(m\sqrt{T})$ [17; 22; 31].

However, in many applications, we cannot always get feedback *right after* choosing actions, as noted, e.g., in [5]. For example, in the application of advertisement optimization [3; 6; 33], it takes a certain

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

Table 1: Regret bounds for nonstochastic bandit problems with/without delay

	Multi-armed bandit (K-arms)	Linear bandit (m: dim. of action set)
Without delay	$O(\sqrt{KT\log K})$ [10]	$O(m\sqrt{T\log m})$ [15; 17; 22]
	$O(\sqrt{KT})$ [8]	
	$\Omega(\sqrt{KT})$ [10]	$\Omega(m\sqrt{T})$ [21]
With delay	$O(\sqrt{dKT})$ [35; 36]	$\tilde{O}(m\sqrt{dT})$ [19]
(d-rounds delay)	$O(\sqrt{(d+K)T\log K})$ [18]	([19] applies to more general models)
	$O(\sqrt{(d\log K+K)T})$ [46]	$\tilde{O}(\sqrt{m(d+m)T})$ **[Theorem 1]**
	$\Omega(\sqrt{(d\log K+K)T})$ [18]	$\Omega(\sqrt{m(d+m)T})$ **[Theorem 2]**

amount of time until the advertisements begin to affect consumers' behavior. Previous studies are not directly applicable to such situations. For the multi-armed bandit problems, algorithms that work well even for delayed-feedback settings have been proposed, as shown in Table 1, but they are not available for more general linear bandits settings. Vernade et al. [43] provided algorithms for *stochastic* linear bandits with delayed feedback, and they work well under assumptions of time-invariant generative models for the loss, without regret bounds for nonstochastic settings.

This paper introduces *bandit linear optimization with delayed feedback*, which includes the multi-armed bandit problems [18], and proposes a min-max optimal algorithm for the model. In this model, there is an additional parameter $d > 1$ representing the rounds of feedback delay. In contrast to the standard bandit feedback model, in which a player observes $\ell_t^\top a_t$ at the end of t-th round, in our model, the feedback $\ell_t^\top a_t$ can be observed at the end of the $(t+d)$-th round, i.e., d-rounds later.

Our contribution

The main contribution of this paper is to construct a nearly-optimal algorithm for online linear optimization with delayed bandit feedback. More precisely, for online linear optimization with an m-dimensional action set and d-rounds delay, Algorithm 1, described in Section 4, enjoys the following regret bound:

Theorem 1. *For arbitrary loss sequences $(\ell_t)_{t=1}^T$, the regret for Algorithm 1 is bounded as*

$$\mathbf{E}[R_T] \leq \max\left\{\sqrt{8m(d+\mathrm{e}m)T\log T}+3, Cm(d\sqrt{m}+m)\log^3(dmT)\right\},$$

where $\mathbf{E}[\cdot]$ means the expectation taken w.r.t. the internal randomization of the algorithm and $C > 0$ is a global constant.

We note that this paper considers *oblivious adversarial model*, i.e., ℓ_t is assumed not to depend on the output of the algorithm. Our results can easily be generalized to an adaptive adversarial model, i.e., a model in which ℓ_t may depend on $\{a_j\}_{j<t}$. As shown in Table 1, our regret bound improves upon the results presented by Cesa-Bianchi et al. [19]. We should note that their algorithm works for more general settings with *composite* delayed feedback. Their work will be discussed in Section 2.

Our regret bound can be shown to be min-max optimal up to logarithmic factors. In fact, we provide the following regret lower bound:

Theorem 2. *Suppose that $\mathcal{A} = \{-1,1\}^m$ and $\|\ell_t\|_1 \leq 1$. There is a distribution of $(\ell_t)_{t=1}^T$ for which an arbitrary algorithm suffers regret as*

$$\mathbf{E}[R_T] = \Omega(\min\{\sqrt{m(d+m)T}, T\}), \tag{1}$$

where $\mathbf{E}[\cdot]$ means the expectation w.r.t. the randomness of (ℓ_t) and the algorithm.

We show this lower bound by combining the result for bandit linear optimization without delay [21; 26] and for online linear optimization with delayed full-information [44]. This lower bound implies that there is no room to improve the upper bound in Theorem 1, up to logarithmic factors.

Our proposed algorithm is based on the multiplicative weight update (MWU) method [7] with an unbiased estimator $\hat{\ell}_t$ of the loss vector ℓ_t. MWU methods manage probability distributions p_t over action set $\mathcal{A}$, and choose an action a_t following p_t. Some existing algorithms [17; 22] for bandit linear optimization employ MWU to achieve an optimal regret bound up to logarithmic factors. These algorithms, however, have not been proven to work well for delayed-feedback settings. This issue appears to be due to the behavior of the probability distribution p_t. In existing algorithms, p_t is updated using $\hat{\ell}_t$, and p_t may change drastically per round since $\hat{\ell}_t$ is unbounded, which can worsen the regret, especially in delayed-feedback settings.

To deal with the above issue, we employ two techniques to construct a new, more stable unbiased estimator $\hat{\ell}_t$. The first technique is to manage probability distributions over the convex hull $\mathcal{B} := \mathrm{conv}(\mathcal{A})$ of an action set $\mathcal{A}$, rather than $\mathcal{A}$ itself. If we apply MWU to $\mathcal{B}$, the probability distribution p_t will have a property referred to as *log-concavity* [34], which plays an important role in our analysis. The other, and more essential, technique is to *truncate* the distribution, which ensures that the estimator $\hat{\ell}_t$ is bounded in terms of a specific norm depending on the distribution. Additionally, thanks to log-concavity, we can also show that this truncation does not change the distribution drastically. Similar techniques to this truncation can be found in [16; 28], though much difference can be found as well. For example, in contrast to our truncation technique, the focus region introduced in [16] is updated so that the new one is included in the prior one. This property seems essential for stabilizing their kernel-based estimators, but makes the algorithm and the analysis much complicated.

Our algorithm and analysis can be generalized to a multi-agent cooperative bandit setting [12; 18] in which N agents cooperate to solve a common bandit optimization problem while communicating via a network. In this problem, there is an underlying undirected communication graph $G = (V, E)$, each node of which corresponds to an individual agent. Each agent solves a common bandit optimization problem, and the observation of agent $u \in V$ can be shared with another agent $v \in V$ in $d_G(u, v)$-rounds later, where $d_G(u, v)$ denotes the length of the shortest path in G between u and v. As shown in Table 2, for cooperative nonstochastic multi-armed bandit problems, Cesa-Bianchi et al. [18] provides an algorithm achieving regret of $\tilde{O}(\sqrt{(1 + \frac{K}{|V|}\alpha(G))T})$ averaged over all agents, where $\alpha(G)$ denotes the independence number of G. Bar-On and Mansour [12] achieved $\tilde{O}(\sqrt{(1 + \frac{K}{|N(v)|})T})$-regret for each agent v simultaneously, where $N(v)$ denotes the neighbors of v in G. In this paper, we construct an algorithm for a more general linear optimization setting. By combining the techniques in cooperative multi-armed bandit [12; 18] and linear bandits with delayed feedback as in Theorem 1, we obtain an algorithm such that the expected regret of each agent is bounded by $\tilde{O}(\sqrt{m(1 + \frac{m}{|N(v)|})T})$.

Theorem 3. *For cooperative nonstochastic linear bandits, there is an algorithm for which the regret of each agent v is bounded as*

$$\mathbf{E}[R_T(v)] \leq \max\left\{16\sqrt{m\left(1 + \log m + \frac{m}{|N(v)|}\right)T\log T} + 3, Cm^2\log^3(mT)\right\}, \tag{2}$$

where $\mathbf{E}[\cdot]$ means the expectation taken w.r.t. the internal randomization of each agent's algorithm as well as the randomness of loss vectors, and C is a global constant.

This bound is tight up to logarithmic factors in a special case in which G is a complete graph. Indeed, we provide the following regret lower bound:

Theorem 4. *Let $G = (V, E)$ be a complete graph. There is an environment of a cooperative linear bandit problem over $\mathcal{A} = \{-1, 1\}^m$ with the communication graph G, for which each agent $v \in V$ suffers regret $R_T(v)$ of at least*

$$\mathbf{E}[R_T(v)] = \Omega\left(\min\left\{\sqrt{m\left(1 + \frac{m}{|V|}\right)T}, T\right\}\right) \tag{3}$$

for any arbitrary algorithm.

This lower bound matches the upper bound of Theorem 3 when $|N(v)| = \Omega(|V|)$.

2 Related Work

Bandit linear optimization [17; 21; 22] is a generic model for sequential decision-making with partial information. This model includes a well-studied multi-armed bandit problem [10] as a special

Table 2: Regret bounds for cooperative nonstochastic bandit problems

	Multi-armed bandit (K-arms)	Linear bandit (m: dim. of action set)
Upper bound	$\tilde{O}(\sqrt{(1+\frac{K}{\lvert V\rvert}\alpha(G))T})$ [18] $\tilde{O}(\sqrt{(1+\frac{K}{\lvert N(v)\rvert})T})$ [12]	$\tilde{O}(\sqrt{m(1+\frac{m}{\lvert N(v)\rvert})T})$ **[Theorem 3]**
Lower bound	$\Omega(\sqrt{(\log K+\frac{K}{\lvert V\rvert})T})$ [40]	$\Omega(\sqrt{m(1+\frac{m}{\lvert V\rvert})T})$ **[Theorem 4]**

case, where the action set $\mathcal{A} = [K]$ is just a finite set of actions. Other important special cases are *combinatorial bandits* [15; 17], where the action set $\mathcal{A} \subseteq 2^{[K]}$ is a set of subsets of a fixed finite set, and the loss for choosing action $a \in \mathcal{A}$ is given by $\sum_{i \in a} \ell_{ti}$, the sum of losses for the items in a. For solving bandit linear optimization, many algorithms have been proposed for stochastic settings [1; 9; 20] as well as for nonstochastic settings [2; 4; 11; 15; 17; 22]. There are known to exist algorithms [15; 19] achieving $\tilde{O}(m\sqrt{T})$-regret, which nearly matches the lower bound of $\Omega(m\sqrt{T})$ shown in [21].

In the context of online optimization, delayed feedback has been considered for a wide range of settings [18; 19; 29; 43; 32; 37; 42; 45; 46] due to its practical significance. As has been noted, e.g., in [18; 29; 44], a regret bound of $R_T \leq U(T)$ for the no-delay setting immediately leads to a bound of $R_T \leq d \cdot U(T/d)$ for the setting with d-rounds delay. Our question is, then, how one can achieve a regret bound better than $d \cdot U(T/d)$. For nonstochastic multi-armed bandit settings, some algorithms have been found to achieve better bounds than $d \cdot U(T/d)$ [18; 42; 46], as can be seen in Table 1. For nonstochastic linear bandits, however, such an improvement upon $O(d \cdot U(T/d))$-regret cannot be found in the literature, and this paper offers the first. We should note that Vernade et al. [43] provided an algorithm for linear bandits that works for stochastic settings with delayed feedback. It is also worth noting that Cesa-Bianchi et al. [19] provided a generic framework for reducing bandit problems with delayed and *composite* feedback to those without delay. In their model, the feedback at each step t depends on all chosen actions in the last d steps $t, t-1, \ldots, t-d+1$, i.e., the feedback is expressed as $\sum_{s=0}^{d-1} \ell_{t-s}^{(s)}(a_{t-s})$. This model is applicable to wider problem settings than the delayed feedback model in this paper, as each feedback depends on the chosen action at a single round in the latter model. In their paper, it was shown that any algorithm for a bandit optimization problem without delay can be transformed into one for a counterpart problem with delayed composite feedback.

Multi-agent cooperative bandit online learning has been considered for various settings [11; 12; 14; 19; 25; 30; 38; 40; 41], with applications including, e.g., peer-to-peer recommendation services serving a large number of users connected in a network [13]. While many of these studies are focusing on stochastic models, only a limited number of studies dealing with nonstochastic models can be found. Among them, Cesa-Bianchi et al. [18] have considered a cooperative nonstochastic multi-armed bandit, taking communication delays into account, and proposed an algorithm for which the average (or the sum) of the regret for all agents is bounded. The cooperative bandit model in this paper is based on their model. Bar-On and Mansour [12] provided algorithms for the multi-armed bandit problem in [19], by which the regret for each agent is bounded well. In this paper, we generalize these results to linear bandit problems.

3 Problem Settings

3.1 Online linear optimization with delayed bandit feedback

A player is given the number T of rounds and an action set $\mathcal{A} \subseteq \mathbb{R}^m$ before the game starts. The action set $\mathcal{A}$ is an arbitrary compact set in $\mathbb{R}^m$, which is not contained in any proper linear subspace. We note that the assumption does not affect the generality of the problem. Indeed, if $\mathcal{A}$ is contained in a proper linear subspace, we can find such a subspace using the linear optimization oracle for $\mathcal{A}$ (e.g., from Corollary 14.1 of [39]). Hence, by reducing the entire vector space into this linear subspace, we can transform the problem so that the assumption holds. In each round $t \in [T]$, the player chooses action a_t, and then, if $t > d$, the environment reveals the loss $\ell_{t-d}^\top a_{t-d} \in \mathbb{R}$. Without loss of generality, we assume $d \leq T-1$. The loss vector $\ell_t \in \mathbb{R}^m$ is assumed to satisfy $|\ell_t^\top a| \leq 1$

for all $a \in \mathcal{A}$. In this paper, we assume that the sequence $(\ell_t)_{t=1}^T$ of loss vectors is an arbitrary *non-adaptive* sequence, i.e., we do not assume any generative model for ℓ_t, but each ℓ_t is assumed not to depend on the output of the algorithm. Player performance is measured by means of *regret* R_T, defined as $R_T = \sum_{t=1}^T \ell_t^\top a_t - \min_{a^* \in \mathcal{A}} \sum_{t=1}^T \ell_t^\top a^*$.

3.2 Cooperative nonstochastic linear bandits

The model for cooperative bandits in this paper is based on the problem setting considered in [12; 18]. There is a communication graph $G = (V, E)$, an undirected graph each vertex of which corresponds to an agent that plays a common linear bandit problem. In each round $t = 1, 2, \ldots, T$, each agent $v \in V$ chooses an action $a_t(v) \in \mathcal{A}$ from a common action set $\mathcal{A} \subseteq \mathbb{R}^m$, and then observes loss $\ell_t(v)^\top a_t(v) \in [-1, 1]$ for the chosen action, where we assume that $\mathbf{E}[\ell_t(v)] = \ell_t$ for each agent v.[1] At the end of each round, each agent v sends a message $m_t(v)$ to neighbors $u \in N(v) = \{u \in V \mid \{u, v\} \in E\}$. The message $m_t(v)$ consists of the chosen action, observed loss, and the distribution q_t^v for choosing an action such that $a_t(v) \sim q_t^v$:

$$m_t(v) = \langle t, v, a_t(v), \ell_t(v)^\top a_t(v), q_t^v \rangle. \tag{4}$$

Note that each agent chooses an action independently, i.e., the action $a_t(v)$ independently follows $q_t(v)$ for each agent v. The goal is to construct an algorithm for which the regret $R_T(v) = \sum_{t=1}^T \ell_t^\top a_t(v) - \min_{a^* \in \mathcal{A}} \sum_{t=1}^T \ell_t^\top a^*$ for each agent v is as small as possible.

4 Algorithms and Regret Upper Bounds

4.1 Preliminary

In this subsection, we introduce a technique that we refer to as *distribution truncation*, which plays a central role in bounding the regret. We denote the convex hull of $\mathcal{A}$ by $\mathcal{B}$. Given a distribution p over $\mathcal{B}$, define $\mu(p) \in \mathbb{R}^m$ and $S(p) \in \mathrm{Sym}(m)$ by $\mu(p) = \mathbf{E}_{x \sim p}[x]$ and $S(p) = \mathbf{E}_{x \sim p}[xx^\top]$. For any vector $x \in \mathbb{R}^m$ and positive semidefinite matrix $A \in \mathrm{Sym}(m)$, denote $\|x\|_A = \|A^{\frac{1}{2}} x\|_2 = \sqrt{x^\top A x}$. Given a distribution p over $\mathcal{B}$, define a *truncated distribution* p' by

$$p'(x) = \frac{p(x)\mathbf{1}\{\|x\|_{S(p)^{-1}}^2 \leq m\gamma^2\}}{\mathrm{Prob}_{y \sim p}[\|y\|_{S(p)^{-1}}^2 \leq m\gamma^2]} \propto p(x)\mathbf{1}\{\|x\|_{S(p)^{-1}}^2 \leq m\gamma^2\}, \tag{5}$$

where $\gamma \geq 4 \log(10mT)$ is a parameter that we refer to as the *truncation level*. From the definition of p', any sample b chosen from a truncated distribution p' is bounded in terms the norm $\|\cdot\|_{S(p)^{-1}}$, as $\|b\|_{S(p)^{-1}}^2 \leq m\gamma^2$. This property ensures that the estimated loss vector (defined in (10)), constructed from a sample from a truncated distribution, has a bounded norm as in (11), thanks to which action distributions do not change drastically per round, as will be shown in Lemmas 5 and 6.

Properties of log-concave distributions If a probability distribution has a density function $p : \mathcal{B} \to \mathbb{R}_{\geq 0}$ such that $\log(p(x))$ is a concave function, then we call it a *log-concave* distribution. We use the following concentration property of log-concave distributions:

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

$$\mathrm{Prob}[\|x\|_2^2 \geq m\alpha^2] \leq m \exp(1 - \alpha) \tag{6}$$

for an arbitrary $\alpha \geq 0$.

Missing proofs of lemmas are provided in the appendix. From this lemma 1, we can show that p and p' defined as (5) are close if p is a log-concave distribution, in the following sense:

Lemma 2. *Suppose that p is a log-concave distribution over $\mathcal{B}$. For any function $f : \mathcal{B} \to [-1, 1]$ and $\gamma \geq 4 \log(10mT)$, we have*

$$\left| \mathop{\mathbf{E}}_{x \sim p}[f(x)] - \mathop{\mathbf{E}}_{x \sim p'}[f(x)] \right| \leq \frac{1}{T} \quad \textit{and} \quad \frac{T}{T+1} \cdot S(p) \preceq S(p') \preceq \frac{T+1}{T} \cdot S(p). \tag{7}$$

[1] In previous studies [18; 12], all players share a common loss vector ℓ_t, i.e., the loss vectors $\ell_t(v)$ are equal to ℓ_t for all v. This case is a special case of our problem setting.

Lemma 3. *If y follows a one-dimensional log-concave distribution such that $\mathbf{E}[y^2] \le s^2 \le 1/100$, we have*

$$\mathbf{E}[g(y)] \le s^2 + 30 \exp\left(-\frac{1}{s}\right) \le 2s^2 \quad \textit{where} \quad g(y) = \exp(y) - y - 1. \tag{8}$$

4.2 Algorithm for linear bandits with delayed feedback

In our algorithm, we update distribution p_t over $\mathcal{B} := \mathrm{conv}(\mathcal{A})$, by the multiplicative weight update method (MWU) as follows:

$$w_t(x) := \exp\left(-\eta \sum_{j=1}^{t-d-1} \hat{\ell}_j^\top x\right), \quad p_t(x) = \frac{w_t(x)}{\int_{y \in \mathcal{B}} w_t(y) \mathrm{d}y}, \tag{9}$$

where $\eta > 0$ is a parameter referred to as the *learning rate*, and $\hat{\ell}_t$ is as defined below. In each round, we pick $b_t \in \mathcal{B}$ from the *truncated distribution* p'_t of p_t. We can get a sample from p'_t by iteratively sampling $b \sim p_t$ until $\|b\|^2_{S(p_t)^{-1}} \le m\gamma^2$. There is a computationally efficient way for sampling $b \sim p_t$ under mild assumptions since p_t is a log-concave distribution. In fact, we can use the techniques developed in [34] to get samples from p_t with polynomial-time computation, given a membership oracle for $\mathcal{B}$. A membership oracle for $\mathcal{B}$ can be constructed from a linear optimization oracle for $\mathcal{A}$, as stated e.g., in [39]. After getting b_t, we choose $a_t \in \mathcal{A}$ so that $\mathbf{E}[a_t | b_t] = b_t$. We can compute such an a_t efficiently, given a linear optimization oracle for $\mathcal{A}$. Indeed, as shown in Corollary 14.1g in [39], given $b \in \mathcal{B} = \mathrm{conv}(\mathcal{A})$ we can compute $\lambda_1, \ldots, \lambda_{m+1} \ge 0$ and $c_1, \ldots, c_{m+1} \in \mathcal{A}$ such that $\sum_{i=1}^{m+1} \lambda_i = 1$ and $\sum_{i=1}^{m+1} \lambda_{ti} c_i = b$ via solving linear optimization over $\mathcal{A}$ polynomial times. By setting $a = c_i$ with probability λ_i, we obtain $\mathbf{E}[a|b] = b$. The algorithm then plays a_t at the t-th round, and the feedback $\ell_t^\top a_t$ will be observed at the end of the $(t+d)$-th round. We define $\hat{\ell}_t$ by

$$\hat{\ell}_t = \ell_t^\top a_t S(p'_t)^{-1} b_t. \tag{10}$$

We note that $S(p'_t)$ is invertible, which can be concluded from the assumption that $\mathcal{A}$ is not contained in any proper linear subspace. Under this assumption, indeed, $\mathcal{B} = \mathrm{Conv}(\mathcal{A})$ is a full-dimensional convex set with a positive Lebesgue measure. It follows from this fact and Lemma 1 that the domain of p'_t is full-dimensional as well. Thus, the distribution p'_t has a density function taking positive values over a full-dimensional set, which implies that the matrix $S(p'_t)$ is invertible. A similar argument can be found, e.g., in [27] (between Eq. (4) and (5)), and is implicitly used in [16] as well. Further, we can compute $S(p'_t)$ efficiently. In fact, since p'_t is a log-concave distribution, for any $\varepsilon > 0$, we can calculate an ε-approximation of $S(p'_t)$ w.h.p. using $(d/\varepsilon)^{O(1)}$ samples generated from p'_t, as can be seen from Corollary 2.7 of [34].

The vector $\hat{\ell}_t$ defined as (10) is an unbiased estimator of ℓ_t and bounded in terms of the norm $\|\cdot\|^2_{S(p_t)}$, i.e., we have

$$\mathbf{E}\left[\hat{\ell}_t\right] = \ell_t, \quad \left\|\hat{\ell}_t\right\|^2_{S(p_t)} \le 4m\gamma^2. \tag{11}$$

In fact, we have $\mathbf{E}[\hat{\ell}_t] = \mathbf{E}\left[S(p'_t)^{-1} b_t a_t^\top \ell_t\right] = \mathbf{E}\left[S(p'_t)^{-1} b_t b_t^\top \ell_t\right] = \mathbf{E}\left[S(p'_t)^{-1} S(p'_t) \ell_t\right] = \ell_t$, which means the first part in (11) holds. To show the second part in (11), we use $\|b_t\|^2_{S(p_t)^{-1}} \le m\gamma^2$, which is ensured by the fact that b_t is sampled from the truncated distribution p'_t. It follows that $\|\hat{\ell}_t\|^2_{S(p_t)} = (\ell_t^\top a_t)^2 \|S(p'_t)^{-1} b_t\|^2_{S(p_t)} \le 2\|S(p'_t)^{-1} b_t\|^2_{S(p'_t)} = 2\|b_t\|^2_{S(p'_t)^{-1}} \le 4\|b_t\|^2_{S(p_t)^{-1}} \le 4m\gamma^2$, where we use the second part of (7) in the first and the second inequalities. The inequality in (11) for $\hat{\ell}_t$ will be used in the analysis of MWU. The procedure can be summarized in Algorithm 1.

Let us next show that Algorithm 1 enjoys the regret bound given in Theorem 1. Since we have $\mathbf{E}[a_t | p'_t] = \mathbf{E}[b_t | p'_t] = \mu(p'_t)$, the regret can be bounded as

$$\begin{aligned}
\mathbf{E}[R_T] &= \mathbf{E}\left[\sum_{t=1}^T \ell_t^\top (a_t - a^*)\right] = \mathbf{E}\left[\sum_{t=1}^T \ell_t^\top (\mu(p'_t) - a^*)\right] \\
&= \mathbf{E}\left[\sum_{t=1}^T \ell_t^\top (\mu(p_t) - a^*) + \sum_{t=1}^T \ell_t^\top (\mu(p'_t) - \mu(p_t))\right] \le \mathbf{E}\left[\sum_{t=1}^T \ell_t^\top (\mu(p_t) - a^*)\right] + 1,
\end{aligned} \tag{12}$$

Algorithm 1 An algorithm for online linear optimization with delayed bandit feedback

Require: Action set $\mathcal{A}$, parameters T, $d \le T-1$

Set $\gamma = 4\log(10mT)$ and $\eta = \min\left\{\sqrt{\frac{m\log T}{2(d+\mathrm{e}m)T}}, \frac{1}{100\gamma^2(d\sqrt{m}+m)}\right\}$.

Define $w_1 : \mathcal{B} \to \mathbb{R}_{>0}$ by $w_1(x) = 1$ for all $x \in \mathcal{B}$.

for $t = 1, 2, \ldots, T$ **do**

    Let p_t be a distribution whose density function is proportional to w_t as in (9).

    Pick $b_t \sim p'_t$, e.g., by iteratively sampling $b \sim p_t$ until $\|b\|^2_{S(p_t)^{-1}} \le m\gamma^2$ holds.

    If $t > d$, get feedback of $\ell_{t-d}^\top a_{t-d}$, construct an unbiased estimator of ℓ_{t-d} as $\hat{\ell}_{t-d} = \ell_{t-d}^\top a_{t-d} \cdot S(p'_{t-d})^{-1} b_{t-d}$, and update w_t by $w_{t+1}(x) = w_t(x)\exp(-\eta\hat{\ell}_{t-d}^\top x)$.

    If $t \le d$, let $w_{t+1} = w_t$.

end for

where $a^* \in \mathrm{argmin}_{a\in\mathcal{A}} \sum_{t=1}^T \ell_t^\top a$ and the last inequality follows from the first part of (7) and the assumption that $\ell_t^\top a \in [-1,1]$ for all $a \in A$. From this inequality and a standard analysis for continuous multiplicative weight update methods [7; 22; 23], we obtain the following regret bounds:

Lemma 4. *If p_t is defined by* (9) *with $\hat{\ell}_t$ such that $\mathbf{E}[\hat{\ell}_t] = \ell_t$, the regret for a_t is bounded as*

$$\mathbf{E}[R_T] \le \mathbf{E}\left[\sum_{t=1}^T \left(\ell_t^\top(\mu(p_t) - \mu(p_{t+d})) + \frac{1}{\eta}\mathop{\mathbf{E}}_{x\sim p_{t+d}}\left[g(-\eta\hat{\ell}_t^\top x)\right]\right)\right] + \frac{m\log T}{\eta} + 3, \tag{13}$$

where $g : \mathbb{R} \to \mathbb{R}$ is defined in (8)*.*

In the following, we give bounds for the right-hand side of (13), separately for the terms $\ell_t^\top(\mu(p_t) - \mu(p_{t+d}))$ and $\mathbf{E}_{x\sim p_{t+d}}\left[g(-\eta\hat{\ell}_t^\top x)\right]$. The first can be bounded via the following lemma:

Lemma 5. *Suppose that $\ell \in \mathbb{R}^m$ satisfies $|\ell^\top a| \le 1$ for all $a \in \mathcal{A}$ and $\eta \le 1/(48\gamma^2 m)$. Then, if p_t is defined by* (9) *with $\hat{\ell}_t$ satisfying* (11)*, it holds for all $t \in [T]$ that $|\,\mathbf{E}[\ell^\top(\mu(p_t) - \mu(p_{t+1}))]| \le 2\eta$.*

This lemma can be shown using (11) and Lemma 3. Finally, to bound the term $\mathbf{E}_{x\sim p_{t+d}}\left[g(-\eta\hat{\ell}_t^\top x)\right]$ in Lemma 4, we use the following lemma:

Lemma 6. *Assume $\eta \le \frac{1}{100\gamma^2(d+1)\sqrt{m}}$. If p_t is defined by* (9) *with $\hat{\ell}_t$ satisfying* (11)*, for all t, we have $S(p_{t+1}) \preceq \left(1 + \frac{1}{d+1}\right) S(p_t)$.*

This lemma follows from (11) and Lemma 1, by induction in t. We are now ready to prove Theorem 1.

Proof of Theorem 1 Combining Lemmas 4 and 5, we have

$$\mathbf{E}[R_T] \le 2\eta dT + \mathbf{E}\left[\sum_{t=1}^T \left(\frac{1}{\eta}\mathop{\mathbf{E}}_{x\sim p_{t+d}}\left[g(-\eta\hat{\ell}_t^\top x)\right]\right)\right] + \frac{m\log T}{\eta} + 3. \tag{14}$$

Let us bound $\mathbf{E}_{x\sim p_{t+d}}\left[g(-\eta\hat{\ell}_t^\top x)\right]$ using Lemma 3 and (11). We have

$$\mathop{\mathbf{E}}_{x\sim p_{t+d}}\left[(-\eta\hat{\ell}_t^\top x)^2\right] = \eta^2\|\hat{\ell}_t\|^2_{S(p_{t+d})} \le \eta^2\left(1 + \frac{1}{d+1}\right)^d \|\hat{\ell}_t\|^2_{S(p_t)} \le 4\mathrm{e}\eta^2 m\gamma^2 \le \frac{1}{100}, \tag{15}$$

where the first inequality follows from Lemma 6, the second inequality follows from (11), and the last inequality comes from the assumption of $\eta \le \frac{1}{100\gamma^2(d\sqrt{m}+m)}$. From this inequality and the fact that $\eta\hat{\ell}_t^\top x$ follows a log-concave distribution when $x \sim p_{t+d}$, we have

$$\mathbf{E}\left[\mathop{\mathbf{E}}_{x\sim p_{t+d}}\left[g(-\eta\hat{\ell}_t^\top x)\right]\right] \le 2\,\mathbf{E}\left[\mathop{\mathbf{E}}_{x\sim p_{t+d}}\left[(-\eta\hat{\ell}_t^\top x)^2\right]\right] \le 2\eta^2\left(1 + \frac{1}{d+1}\right)^d \mathbf{E}\left[\left\|\hat{\ell}_t\right\|^2_{S(p_t)}\right],$$

where the first and second inequalities follow from (3) and (15), respectively. From the definition (10) of $\hat{\ell}_t$, we have

$$\mathbf{E}\left[\left\|\hat{\ell}_t\right\|^2_{S(p_t)}\right] = \mathbf{E}\left[(\ell_t^\top a_t)^2 \left\|S(p_t')^{-1}b_t\right\|^2_{S(p_t)}\right] \leq \frac{T+1}{T}\,\mathbf{E}\left[\left\|S(p_t')^{-1}b_t\right\|^2_{S(p_t')}\right]$$
$$\leq \left(1+\frac{1}{d+1}\right) S(p_t') \bullet S(p_t')^{-1}S(p_t')S(p_t')^{-1} = \left(1+\frac{1}{d+1}\right) m,$$

where the first inequality follows from $|\ell_t^\top a_t| \leq 1$ and the second part of (7). Combining the above two inequalities, we obtain $\mathbf{E}\left[\mathbf{E}_{x\sim p_{t+d}}\left[g(-\eta\hat{\ell}_t^\top x)\right]\right] \leq 2\eta^2(1+\frac{1}{d+1})^{d+1}m \leq 2\mathrm{e}\eta^2 m$. From this and (14), we have $\mathbf{E}[R_T] \leq 2\eta(d+\mathrm{e}m)T + \frac{m\log T}{\eta} + 3$. By substituting the parameter setting in Step 1 of Algorithm 1, we obtain the inequality of Theorem 1. $\square$

4.3 Algorithm for cooperative nonstochastic linear bandits

For the cooperative bandit problem, we consider a *center-based algorithm* [12]. A major difference between our algorithm and the one by Bar-On and Mansour [12] is that ours applies to linear bandit settings while theirs focus on multi-armed bandit settings.

Similarly to [12], we first choose *center agents* $C \subseteq V$ and a corresponding *partition* $\{V_c\}_{c\in C}$ of agents with the following properties:

Theorem 5 (Theorem 12. in [12])**.** *Given an undirected graph $G=(V,E)$ and a parameter $m \geq 2$, one can find center agents $C \subseteq V$ and a partition $\{V_c\}_{c\in C}$ of V such that the following hold for all $c \in C$: (i) $c \cup N(c) \subseteq V_c$, (ii) subgraphs of G induced by V_c are connected, and (iii) for all $v \in V_c$, it holds that*

$$\frac{\min\{|N(v)|, m\}}{\min\{|N(c)|, m\}} \leq \exp\left(1 - \frac{d_c(v)}{6}\right), \tag{16}$$

where $d_c(v)$ denotes the distance from c to v in the subgraph of G induced by V_c.

In the center-based algorithm, each center agent $c \in C$ updates its distribution by means of MWU, and the other agents $v \in V_c$ imitate center agent c on the basis of the message m_t in (4). More precisely, for each $c \in C$ we construct breadth-first search tree T_c for the subgraph induced by V_c with root node c, and each agent $v \in V_c \setminus \{c\}$ uses the distribution q_t^v defined by $q_t^v = q_{t-1}^{\mathrm{pa}(v)}$, where $\mathrm{pa}(v)$ denotes the parent node of v in T_c. Consequently, for each $v \in V_c \setminus \{c\}$, the action distribution q_t^v satisfies $q_t^v = q_{t-d_c(v)}^c$ for $t > d_c(v)$. We define the distribution q_t^c of each center agent $c \in C$ as follows: define a distribution p_t^c over $\mathcal{B} = \mathrm{conv}(\mathcal{A})$ by MWU as follows:

$$w_t^c(x) := \exp\left(-\eta(c)\sum_{j=1}^{t-1}\hat{\ell}_j(c)^\top x\right), \quad p_t^c(x) = \frac{w_t^c(x)}{\int_{y\in\mathcal{B}} w_t^c(y)\mathrm{d}y}, \tag{17}$$

where we set the truncation level γ and the learning rate $\eta(c)$ as $\gamma = 4\log(10mT)$ and $\eta(c) = \min\{\frac{1}{4}\sqrt{\frac{m\log T}{T(1+\log m + m/\min\{|N(c)|,m\})}}, \frac{1}{100\gamma^2 m}\}$. Pick $b_t(c)$ from the truncated distribution $p_t^{c\prime}$ defined by (5), and pick $a_t(c)$ so that $\mathbf{E}[a_t(c)|b_t(c)] = b_t(c)$. The distribution q_t^c is defined to be the distribution that $a_t(c)$ follows. The estimated loss $\hat{\ell}_t(c)$ is defined on the basis of the messages m_t (4) from neighbors $N(c)$ of c. Each center agent $c \in C$ computes $\hat{\ell}_t(c)$ using $a_t(v)$ and $\ell_t(v)^\top a_t(v)$ for $v \in N(v)$ as follows:

$$\hat{\ell}_t(c) = \frac{1}{|N(c)|}\sum_{v\in N(c)} \ell_t(v)^\top a_t(v) S({p_{t-1}^c}')^{-1} b_t'(v), \tag{18}$$

where $b_t'(v)$ is chosen from the posterior probability distribution for $b_t(v)$ given $a_t(v)$, i.e., $\mathrm{Prob}[b_t'(v) = b] \propto \mathrm{Prob}[a_t(v)|b_t(v) = b] \cdot {p_{t-1}^c}'(b)$. Combining Lemmas 3, 4, 5, 6 and Theorem 5, we obtain the regret bounds in Theorem 3. The proof of Theorem 3 is given in Section B in the appendix.

5 Lower Bounds

In this section, we briefly describe how we can prove regret lower bounds in Theorems 2 and 4. Inspired by the proof of the lower bound for the multi-armed bandit problem with delayed feedback given in [18], we prove lower bounds by combining existing bounds for different settings such as linear bandits without delay [21; 26] and full-information online optimization with delay [44].

To prove Theorem 2, we combine lower bounds of $\Omega(m\sqrt{T})$ and of $\Omega(\sqrt{mdT})$. We note that the first one comes from a lower bound for linear bandits without delay, shown, e.g., in [21]. The second one is also a lower bound for online linear optimization with delayed feedback. As shown in [44], if an online optimization problem without delay admits a regret lower bound of $\Omega(L(T))$, a counterpart with d-round delayed feedback has an $\Omega(dL(T/d))$-lower bound. Since there is a lower bound of $\Omega(\sqrt{mT})$ for the online linear optimization (see, e.g., Theorem 3.2. in [24]), we have $\Omega(d\sqrt{mT/d}) = \Omega(\sqrt{mdT})$.

So the only question left is that both the regret bounds of $\Omega(m\sqrt{T})$ and of $\Omega(\sqrt{mdT})$ must come from the same instance. To this end, we construct the problem instance over $\mathcal{A} = \{-1, 1\}^m$ such that $\|\ell_t\|_1 \leq 1$. By considering a mixed distribution of loss vectors, we obtain a lower bound of $\Omega(m\sqrt{T} + \sqrt{mdT}) = \Omega(\sqrt{m(d+m)T})$. A complete proof is provided in Section C in the appendix.

Theorem 4 for cooperative linear bandits can be, again, obtained by combining lower bounds of $\Omega(\sqrt{mT})$ and of $\Omega(m\sqrt{T/|V|})$. The first one comes from a lower bound for full-information online linear optimization, as cooperative bandits are at least harder than full-information online optimization. The second can come from an $\Omega(m\sqrt{T})$-lower bound for signle-player linear bandits [21]. Since we can regard cooperative bandits as a harder version of single-player bandits with $(T \cdot |V|)$ rounds, the sum of regrets over all agents is at least $\Omega(m\sqrt{T \cdot |V|})$, which implies that the regret per agent is of $\Omega(m\sqrt{T/|V|})$. A complete proof of Theorem 4 to show how we construct the problem instance that gives both the regret bounds of $\Omega(\sqrt{mT})$ and of $\Omega(m\sqrt{T/|V|})$ is given in Section D in the appendix.

6 Conclusion

In this paper, we considered online linear optimization with d-round-delayed bandit feedback, where d was a given parameter fixed for all rounds. We provided an algorithm that achieves nearly-tight regret bounds, and extended this result to the cooperate bandit setting.

An important future work would be to extend the model to deal with the unknown and round-dependent delay as in [46]. We believe that an adaptive way of tuning parameters, such as α, η and γ in our algorithms, would work well for this general setting. Another future direction is to improve practical computational cost. The proposed algorithms in this paper rely on continuous relaxation and sampling from log-concave distributions, which causes a large computational time in practice, though they run in polynomial time.

Broader Impact

The authors believe that this paper presents neither ethical nor societal issues, as this is a theoretical work.

Acknowledgments and Disclosure of Funding

This work was supported by JST, ERATO, Grant Number JPMJER1201, Japan. SI was supported by JST, ACT-I, Grant Number JPMJPR18U5, Japan. TF was supported by JST, PRESTO, Grant Number JPMJPR1759, Japan. NK and KK were supported by JSPS, KAKENHI, Grant Number JP18H05291, Japan.

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

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

[3] N. Abe, A. W. Biermann, and P. M. Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.

[4] J. D. Abernethy, E. Hazan, and A. Rakhlin. An efficient algorithm for bandit linear optimization. In *Conference on Learning Theory*, 2008.

[5] A. Agarwal, S. Bird, M. Cozowicz, L. Hoang, J. Langford, S. Lee, J. Li, D. Melamed, G. Oshri, O. Ribas, et al. Making contextual decisions with low technical debt. *arXiv preprint arXiv:1606.03966*, 2016.

[6] D. Agarwal, B.-C. Chen, P. Elango, N. Motgi, S.-T. Park, R. Ramakrishnan, S. Roy, and J. Zachariah. Online models for content optimization. In *Advances in Neural Information Processing Systems*, pages 17–24, 2009.

[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 *Conference on Learning Theory*, pages 217–226, 2009.

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

[10] 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.

[11] B. Awerbuch and R. D. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing*, pages 45–53, 2004.

[12] Y. Bar-On and Y. Mansour. Individual regret in cooperative nonstochastic multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3110–3120, 2019.

[13] R. Baraglia, P. Dazzi, M. Mordacchini, and L. Ricci. A peer-to-peer recommender system for self-emerging user communities based on gossip overlays. *Journal of Computer and System Sciences*, 79(2):291–308, 2013.

[14] I. Bistritz and A. Leshem. Distributed multi-player bandits-a game of thrones approach. In *Advances in Neural Information Processing Systems*, pages 7222–7232, 2018.

[15] S. Bubeck, N. Cesa-Bianchi, and S. Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, 2012.

[16] S. Bubeck, Y. T. Lee, and R. Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 72–85. ACM, 2017.

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

[18] N. Cesa-Bianchi, C. Gentile, Y. Mansour, and A. Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pages 605–622, 2016.

[19] N. Cesa-Bianchi, C. Gentile, and Y. Mansour. Nonstochastic bandits with composite anonymous feedback. In *Conference on Learning Theory*, pages 750–773, 2018.

[20] W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

[21] V. Dani, S. M. Kakade, and T. P. Hayes. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, pages 345–352, 2008.

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

[23] E. Hazan, W. Hu, Y. Li, et al. Online improper learning with an approximation oracle. In *Advances in Neural Information Processing Systems*, pages 5652–5660, 2018.

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

[25] E. Hillel, Z. S. Karnin, T. Koren, R. Lempel, and O. Somekh. Distributed exploration in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 854–862, 2013.

[26] S. Ito. Submodular function minimization with noisy evaluation oracle. In *Advances in Neural Information Processing Systems*, pages 12080–12090, 2019.

[27] 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 *Advances in Neural Information Processing Systems*, pages 10590–10599, 2019.

[28] S. Ito, S. Hirahara, T. Soma, and Y. Yoshida. Tight first- and second-order regret bounds for adversarial linear bandits. In *Advances in Neural Information Processing Systems*, 2020, to appear.

[29] P. Joulani, A. Gyorgy, and C. Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461, 2013.

[30] P. Landgren, V. Srivastava, and N. E. Leonard. On distributed cooperative decision-making in multiarmed bandits. In *2016 European Control Conference (ECC)*, pages 243–248. IEEE, 2016.

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

[32] B. Li, T. Chen, and G. B. Giannakis. Bandit online learning with unknown delays. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 993–1002, 2019.

[33] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670, 2010.

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

[35] G. Neu, A. Antos, A. György, and C. Szepesvári. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems*, pages 1804–1812, 2010.

[36] G. Neu, A. Gyorgy, C. Szepesvari, and A. Antos. Online markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 59(3):676–691, 2014.

[37] C. Pike-Burke, S. Agrawal, C. Szepesvari, and S. Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pages 4105–4113, 2018.

[38] A. Sankararaman, A. Ganesh, and S. Shakkottai. Social learning in multi agent multi armed bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3): 1–35, 2019.

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

[40] Y. Seldin, P. Bartlett, K. Crammer, and Y. Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In *International Conference on Machine Learning*, pages 280–287, 2014.

[41] B. Szorenyi, R. Busa-Fekete, I. Hegedüs, R. Ormandi, M. Jelasity, and B. Kégl. Gossip-based distributed stochastic bandit algorithms. In *30th International Conference on Machine Learning*, volume 28, pages 19–27, 2013.

[42] T. S. Thune, N. Cesa-Bianchi, and Y. Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems*, pages 6538–6547, 2019.

[43] C. Vernade, A. Carpentier, T. Lattimore, G. Zappella, B. Ermis, and M. Brueckner. Linear bandits with stochastic delayed feedback. *arXiv preprint arXiv:1807.02089*, 2018.

[44] M. J. Weinberger and E. Ordentlich. On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7):1959–1976, 2002.

[45] Z. Zhou, R. Xu, and J. Blanchet. Learning in generalized linear contextual bandits with stochastic delays. In *Advances in Neural Information Processing Systems*, pages 5198–5209, 2019.

[46] J. Zimmert and Y. Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. *arXiv preprint arXiv:1910.06054*, 2019.

A Proofs of Lemmas

A.1 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 [34]), each x_i follows a log-concave distribution and we have $\mathbf{E}[x_i^2] \leq 1$ from the assumption of $S(p) \preceq I$. Hence, we have

$$\mathrm{Prob}[\|x\|_2^2 \geq m\alpha^2] \leq \mathrm{Prob}[\exists i \in [m], x_i^2 \geq \alpha^2] \leq \sum_{i=1}^{m} \mathrm{Prob}[|x_i| \geq \alpha] \leq m \exp(1-\alpha), \quad (19)$$

where the last inequality follows from Lemma 5.7 in [34]. $\square$

A.2 Proof of Lemma 2

Proof. From the definition 5 of p', we have

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

where we denote $\delta = \mathrm{Pr}_{x\sim p}[\|x\|_{S(p)^{-1}}^2 > m\gamma^2]$. From this expression, we have

$$\begin{aligned}
\left| \mathop{\mathbf{E}}_{x\sim p}[f(x)] - \mathop{\mathbf{E}}_{x\sim p'}[f(x)] \right| &= \frac{1}{1-\delta} \left| -\delta \mathop{\mathbf{E}}_{x\sim p}[f(x)] + \int_{x\in\mathcal{B}} f(x)\mathbf{1}\{\|x\|_{S(p)^{-1}}^2 > m\gamma^2\} p(x) \mathrm{d}x \right| \\
&\leq \frac{1}{1-\delta} \left(\delta \mathop{\mathbf{E}}_{x\sim p}[1] + \int_{x\in\mathcal{B}} \mathbf{1}\{\|x\|_{S(p)^{-1}}^2 > m\gamma^2\} p(x) \mathrm{d}x \right) = \frac{2\delta}{1-\delta},
\end{aligned} \quad (20)$$

where the inequality follows from the assumption that $f(x) \in [-1, 1]$. The value δ can be bounded via Lemma 1. In fact, when x follows p, a log-concave distribution, $S(p)^{-\frac{1}{2}}x$ follows a log-concave distribution as well. In addition, we have $\mathbf{E}[S(p)^{-\frac{1}{2}}xx^\top S(p)^{-\frac{1}{2}}] = S(p)^{-\frac{1}{2}}S(p)S(p)^{-\frac{1}{2}} = I$. Hence, from Lemma 1, we have

$$\delta = \mathop{\mathrm{Pr}}_{x\sim p}[\|x\|_{S(p)^{-1}}^2 > m\gamma^2] = \mathop{\mathrm{Pr}}_{x\sim p}[\|S(p)^{-\frac{1}{2}}x\|_2^2 > m\gamma^2] \leq m\exp(1-\gamma) \leq 3m\exp(-\gamma) \leq \frac{1}{6T}, \quad (21)$$

where the last inequality follows from $\gamma \geq 4\log(10mT)$. Combining (20) and (21), we obtain the first part of (7). We next show the second part of (7). For any $y \in \mathbb{R}^d$, we have

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

Since this holds for all $y \in \mathbb{R}^d$ and $\frac{1}{1-\delta} \leq \frac{T+1}{T}$, the last inequality in (7) holds. Furthermore, we have

$$\begin{aligned}
y^\top S(p)y - y^\top S(p')y &= \mathop{\mathbf{E}}_{x\sim p}\left[(y^\top x)^2 \right] - \frac{1}{1-\delta} \mathop{\mathbf{E}}_{x\sim p}\left[(y^\top x)^2 \mathbf{1}\{\|x\|_{S(p)^{-1}}^2 \leq m\gamma^2\} \right] \\
&\leq \mathop{\mathbf{E}}_{x\sim p}\left[(y^\top x)^2 \mathbf{1}\{\|x\|_{S(p)^{-1}}^2 > m\gamma^2\} \right] \\
&\leq y^\top S(p)y \mathop{\mathbf{E}}_{x\sim p}\left[\|x\|_{S(p)^{-1}}^2 \mathbf{1}\{\|x\|_{S(p)^{-1}}^2 > m\gamma^2\} \right],
\end{aligned} \quad (22)$$

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

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

The right-hand side of (22) can be bounded by using Lemma 1 as follows:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{x\sim p}\left[\|x\|_{S(p)^{-1}}^2 \mathbf{1}\{\|x\|_{S(p)^{-1}}^2 > m\gamma^2\}\right] \\
&\leq \sum_{n=1}^{\infty} (n+1)^2 m\gamma^2 \Pr_{x\sim p}\left[n^2 m\gamma^2 \leq \|x\|_{S(p)^{-1}}^2 \leq (n+1)^2 m\gamma^2\right] \\
&\leq \sum_{n=1}^{\infty} (n+1)^2 m\gamma^2 \cdot m\exp(1-n\gamma) \\
&\leq m^2\gamma^2\exp(2-\gamma)\sum_{n=1}^{\infty}(n+1)^2\exp(-n) \leq 40m^2\gamma^2\exp(-\gamma) \leq \frac{1}{2T},
\end{aligned} \tag{23}$$

where the second inequality follows from Lemma 1, the third inequality comes from $\gamma \geq 1$ and hence $n\gamma \geq n+\gamma-1$, the forth inequality follows from the fact that $\sum_{i=1}^{\infty}(i+1)^2\exp(-i) = \frac{1-2\mathrm{e}+4\mathrm{e}^2}{(\mathrm{e}-1)^3} \leq 5$, and the last inequality follows from the assumption of $\gamma \geq 4\log(10mT)$. Combining (22) and (23), we have

$$y^\top S(p')y \geq \left(1-\frac{1}{2T}\right)y^\top S(p)y \geq \frac{T}{T+1}y^\top S(p)y.$$

Since this holds for all $y \in \mathbb{R}^m$, we have the second inequality of (7). $\square$

A.3 Proof of Lemma 3

Proof. From Lemma 1, and $\mathbf{E}[(y/s)^2] \leq 1$, we have

$$\mathrm{Prob}[y \geq n] = \mathrm{Prob}\left[\frac{y}{s} \geq \frac{n}{s}\right] \leq \exp\left(1-\frac{n}{s}\right). \tag{24}$$

Using this and the fact that $g(y) \leq y^2$ for $y \leq 1$ and that $g(y) \leq \exp(y)$ for $y > 1$, we have

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

where the last inequality follows from the assumption of $0 \leq s \leq 1/10$ and $\frac{\exp(3)}{1-\exp(-9)} \leq 30$. From this and the fact that $30\exp(-x^{-1}) \leq x^2$ holds for $0 < x \leq 1/10$, we obtain $\mathbf{E}[g(y)] \leq 2s^2$ $\square$

A.4 Proof of Lemma 4

Proof. Since we have $\mathbf{E}[a_t|p'_t] = \mu(p'_t)$, the expected regret can be bounded as follows:

$$\begin{aligned}
\mathbf{E}[R_T] &= \mathbf{E}\left[\sum_{t=1}^{T}\ell_t^\top(a_t-a^*)\right] = \mathbf{E}\left[\sum_{t=1}^{T}\ell_t^\top(\mu(p'_t)-a^*)\right] \\
&= \mathbf{E}\left[\sum_{t=1}^{T}\ell_t^\top(\mu(p_t)-a^*)\right] + \mathbf{E}\left[\sum_{t=1}^{T}\ell_t^\top(\mu(p'_t)-\mu(p_t))\right] \\
&\leq \mathbf{E}\left[\sum_{t=1}^{T}\ell_t^\top(\mu(p_t)-a^*)\right] + 1 \\
&= \mathbf{E}\left[\sum_{t=1}^{T}\ell_t^\top(\mu(p_t)-\mu(p_{t+d}))\right] + \mathbf{E}\left[\sum_{t=1}^{T}\ell_t^\top(\mu(p_{t+d})-a^*)\right] + 1,
\end{aligned} \tag{25}$$

where the inequality follows from the first part of (7). Since $\hat{\ell}_t$ is an unbiased estimator of ℓ_t, i.e., from (11), the second term in (25) can be expressed as

$$\mathbf{E}\left[\sum_{t=1}^{T} \ell_t^\top (\mu(p_{t+d}) - a^*)\right] = \mathbf{E}\left[\sum_{t=1}^{T} \hat{\ell}_t^\top (\mu(p_{t+d}) - a^*)\right]. \tag{26}$$

The right-hand side of (26) can be bounded via a standard analysis for continuous MAB as follows. We have

$$\begin{aligned}\int_{x\in\mathcal{B}} \exp\left(-\eta \sum_{j=1}^{t} \hat{\ell}_j^\top x\right) \mathrm{d}x &= \int_{x\in\mathcal{B}} \exp\left(-\eta \sum_{j=1}^{t-1} \hat{\ell}_j^\top x\right) \exp\left(-\eta \hat{\ell}_t^\top x\right) \mathrm{d}x \\ = \int_{x\in\mathcal{B}} \exp\left(-\eta \sum_{j=1}^{t-1} \hat{\ell}_j^\top x\right) \mathrm{d}x \cdot \mathop{\mathbf{E}}_{x\sim p_{t+d}} &\left[\exp(-\eta \hat{\ell}_t^\top x)\right],\end{aligned}$$

where the second equality follows from the definition (9) of p_t. Since this holds for all $t \in [T]$, we have

$$\begin{aligned}\log \int_{x\in\mathcal{B}} \exp\left(-\eta \sum_{t=1}^{T} \hat{\ell}_t^\top x\right) \mathrm{d}x - \log \int_{x\in\mathcal{B}} 1 \mathrm{d}x &= \sum_{t=1}^{T} \log \mathop{\mathbf{E}}_{x\sim p_{t+d}} \left[\exp(-\eta \hat{\ell}_t^\top x)\right] \\ = \sum_{t=1}^{T} \log \mathop{\mathbf{E}}_{x\sim p_{t+d}} \left[1 - \eta \hat{\ell}_t^\top x + g(-\eta \hat{\ell}_t^\top x)\right] &\le \sum_{t=1}^{T} \left(-\eta \hat{\ell}_t^\top \mu(p_{t+d}) + \mathop{\mathbf{E}}_{x\sim p_{t+d}} \left[g(-\eta \hat{\ell}_t^\top x)\right]\right),\end{aligned} \tag{27}$$

where the second equality follows from the definition of $g(y) = \exp(y) - y - 1$, and the inequality holds since we have $\log(1+z) \le z$ for $z > -1$. We note that this condition $z > -1$ indeed holds since z here can be expressed as $z = \mathbf{E}[-\eta \hat{\ell}_t^\top x + g(-\eta \hat{\ell}_t^\top x)] = \mathbf{E}[\exp(-\eta \hat{\ell}_t^\top x)] - 1 > -1$. The left-hand side of (27) can be bounded via an integration over a subset $\mathcal{B}' = \{x = (1 - \frac{1}{T})a^* + y \mid y \in \mathcal{B}\} \subseteq \mathcal{B}$, as follows:

$$\begin{aligned}\int_{x\in\mathcal{B}} \exp\left(-\eta \sum_{t=1}^{T} \hat{\ell}_t^\top x\right) \mathrm{d}x &\ge \int_{x\in\mathcal{B}'} \exp\left(-\eta \sum_{t=1}^{T} \hat{\ell}_t^\top x\right) \mathrm{d}x \\ &= \frac{1}{T^m} \int_{y\in\mathcal{B}} \exp\left(-\eta \sum_{t=1}^{T} \hat{\ell}_t^\top \left(\left(1 - \frac{1}{T}\right) a^* + \frac{1}{T} y\right)\right) \mathrm{d}y \\ &\ge \frac{1}{T^m} \int_{y\in\mathcal{B}} 1 \mathrm{d}y \cdot \exp\left(-\eta \sum_{t=1}^{T} \hat{\ell}_t^\top \left(\left(1 - \frac{1}{T}\right) a^* + \frac{1}{T} \mu(p_0)\right)\right),\end{aligned}$$

where the last inequality follows from the convexity of $\exp(z)$ and Jensen's inequality. Combining this and (27), we obtain

$$\begin{aligned}\sum_{t=1}^{T} \left(-\eta \hat{\ell}_t^\top \mu(p_{t+d}) + \mathop{\mathbf{E}}_{x\sim p_{t+d}} \left[g(-\eta \hat{\ell}_t^\top x)\right]\right) &\ge -m \log T - \eta \sum_{t=1}^{T} \hat{\ell}_t^\top \left(\left(1 - \frac{1}{T}\right) a^* + \frac{1}{T} \mu(p_0)\right) \\ &= -m \log T - \eta \sum_{t=1}^{T} \hat{\ell}_t^\top a^* - \frac{\eta}{T} \sum_{j=1}^{T} \hat{\ell}_t^\top (\mu(p_0) - a^*),\end{aligned}$$

and hence, we have

$$
\begin{aligned}
&\mathbf{E}\left[\sum_{t=1}^{T} \hat{\ell}_t^\top (\mu(p_{t+d}) - a^*)\right] \\
&\le \frac{1}{\eta}\mathbf{E}\left[\sum_{t=1}^{T} \mathop{\mathbf{E}}_{x\sim p_{t+d}}\left[g(-\eta\hat{\ell}_t^\top x)\right]\right] + \frac{m\log T}{\eta} + \frac{1}{T}\sum_{j=1}^{T}\mathbf{E}\left[\hat{\ell}_t^\top(\mu(p_0) - a^*)\right] \\
&= \frac{1}{\eta}\mathbf{E}\left[\sum_{t=1}^{T} \mathop{\mathbf{E}}_{x\sim p_{t+d}}\left[g(-\eta\hat{\ell}_t^\top x)\right]\right] + \frac{m\log T}{\eta} + \frac{1}{T}\sum_{j=1}^{T}\mathbf{E}\left[\ell_t^\top(\mu(p_0) - a^*)\right] \\
&\le \frac{1}{\eta}\mathbf{E}\left[\sum_{t=1}^{T} \mathop{\mathbf{E}}_{x\sim p_{t+d}}\left[g(-\eta\hat{\ell}_t^\top x)\right]\right] + \frac{m\log T}{\eta} + 2,
\end{aligned}
$$

where the equality follows from (11) and the last inequality follows from the assumption of $|\ell_t^\top a| \le 1$ for all $a \in \mathcal{A}$. Combining this, (25) and (26), we obtain the desired inequality in Lemma 4. $\square$

A.5 Proof of Lemma 5

Proof. Lemma 5 holds for $t \le d$ since $p_t = p_{t+1}$ follows from the definition (9) for this case. We consider the case of $t > d$ in the following. We start by introducing some notations. Define $\alpha > 1$ and $\beta \in \mathbb{R}$ by

$$
\alpha = \mathop{\mathbf{E}}_{x\sim p_t}\left[\exp\left(-\eta\hat{\ell}_{t-d}^\top(x - \mu(p_t))\right)\right], \quad \beta = \mathop{\mathbf{E}}_{x\sim p_t}\left[\ell^\top x \cdot g\left(-\eta\hat{\ell}_{t-d}^\top(x - \mu(p_t))\right)\right]. \tag{28}
$$

We can confirm that $\alpha \ge 1$ by using Jensen's inequality:

$$
\alpha = \mathop{\mathbf{E}}_{x\sim p_t}\left[\exp\left(-\eta\hat{\ell}_{t-d}^\top(x - \mu(p_t))\right)\right] \ge \exp\left(\mathop{\mathbf{E}}_{x\sim p_t}\left[-\eta\hat{\ell}_{t-d}^\top(x - \mu(p_t))\right]\right) = \exp(0) = 1.
$$

Since $p_{t+1}(x) \propto p_t(x)\exp(-\eta\hat{\ell}_{t-d}^\top x) \propto p_t(x)\exp(-\eta\hat{\ell}_{t-d}^\top(x - \mu(p_t)))$ from the definition (9) of p_t, we can express p_{t+1} as

$$
p_{t+1}(x) = \frac{1}{\alpha}p_t(x)\exp\left(-\eta\hat{\ell}_{t-d}^\top(x - \mu(p_t))\right). \tag{29}
$$

Hence, we have

$$
\begin{aligned}
\ell^\top\mu(p_{t+1}) &= \frac{1}{\alpha}\int \ell^\top x \cdot p_t(x)\exp\left(-\eta\hat{\ell}_{t-d}^\top(x - \mu(p_t))\right)\mathrm{d}x \\
&= \frac{1}{\alpha}\int \ell^\top x \cdot p_t(x)\left(1 - \eta\hat{\ell}_{t-d}^\top(x - \mu(p_t)) + g\left(-\eta\hat{\ell}_{t-d}^\top(x - \mu(p_t))\right)\right)\mathrm{d}x \\
&= \frac{1}{\alpha}\left(\ell^\top\mu(p_t) - \eta\ell^\top\mathrm{Cov}(p_t)\hat{\ell}_{t-d} + \beta\right)
\end{aligned} \tag{30}
$$

Hence, using (11), we have

$$
\begin{aligned}
\left|\mathbf{E}\left[\ell^\top(\mu(p_t) - \mu(p_{t+1}))\right]\right| &= \left|\mathbf{E}\left[\frac{\eta}{\alpha}\ell^\top\mathrm{Cov}(p_t)\hat{\ell}_{t-d} + \left(1 - \frac{1}{\alpha}\right)\ell^\top\mu(p_t) - \frac{\beta}{\alpha}\right]\right| \\
&\le \left|\mathbf{E}\left[\frac{\eta}{\alpha}\ell^\top\mathrm{Cov}(p_t)\ell_{t-d}\right]\right| + \left(1 - \frac{1}{\alpha}\right) + \frac{|\beta|}{\alpha} \le \eta + \alpha - 1 + |\beta|,
\end{aligned} \tag{31}
$$

where the first inequality follows from (11) and $|\ell^\top\mu(p_t)| \le 1$, and the last inequality follows from $\alpha \ge 1$ and

$$
\ell^\top\mathrm{Cov}(p_t)\ell_{t-d} \le \sqrt{\|\ell\|_{\mathrm{Cov}(p_t)}^2\|\ell_{t-d}\|_{\mathrm{Cov}(p_t)}^2} \le \sqrt{\|\ell\|_{S(p_t)}^2\|\ell_{t-d}\|_{S(p_t)}^2} \le 1.
$$

Let us bound α and β using Lemma 3. We have

$$
\begin{aligned}
\alpha &= \mathop{\mathbf{E}}_{x\sim p_t}\left[g\left(-\eta\hat{\ell}_{t-d}^\top(x - \mu(p_t))\right) + \left(-\eta\hat{\ell}_{t-d}^\top(x - \mu(p_t))\right) + 1\right] \\
&= \mathop{\mathbf{E}}_{x\sim p_t}\left[g\left(-\eta\hat{\ell}_{t-d}^\top(x - \mu(p_t))\right)\right] + 1.
\end{aligned}
$$

The right-hand side of this can be bounded by means of Lemma 3. Indeed, for any fixed $\hat{\ell}_{t-d}^\top$ and $\mu(p_t)$, when x follows the log-concave distribution p_t then $\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))$ follows a one-dimensional log-concave distribution as well since any marginal of a log-concave distribution is log-concave (see, e.g., Theorem 5.1 in [34]). Further, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{x\sim p_t}\left[\left(-\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))\right)^2\right] &= \eta^2\|\hat{\ell}_{t-d}\|^2_{\mathrm{Cov}(p_t)} \leq \eta^2\|\hat{\ell}_{t-d}\|^2_{S(p_t)} \\ &\leq \mathrm{e}\eta^2\|\hat{\ell}_{t-d}\|^2_{S(p_{t-d})} \leq 4\mathrm{e}\eta^2 m\gamma^2 \leq \frac{1}{100},\end{aligned}$$

where the second inequality follows from Lemma 6 the third inequality follows from (11), and the last inequality comes from the assumption of $\eta \leq \frac{1}{48\gamma^2 m}$.[2] Hence, we can apply Lemma 3 to have

$$\alpha - 1 = \mathop{\mathbf{E}}_{x\sim p_t}\left[g\left(-\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))\right)\right] \leq 8\mathrm{e}\eta^2 m\gamma^2 \leq \frac{\eta}{2}, \tag{32}$$

where the last inequality follows from the assumption of $\eta \leq \frac{1}{48\gamma^2 m}$. Furthermore, since we have $|\ell^\top x| \leq 1$ for $x \in \mathcal{B}$, β defined in (28) can be bounded as

$$|\beta| \leq \mathop{\mathbf{E}}_{x\sim p_t}\left[g\left(-\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))\right)\right] \leq 8\mathrm{e}\eta^2 m\gamma^2 \leq \frac{\eta}{2}.$$

Combining this, (31) and (32), we obtain the desired inequality in Lemma 5. □

A.6 Proof of Lemma 6

Proof. For $t \leq d$, the inequality in Lemma 6 holds since $p_{t+1} = p_t$ follows from the definition (9) of p_t. We show the inequality in 6 for $t > d$ by induction in t. We denote

$$\varepsilon = \frac{1}{2(1+d)}. \tag{33}$$

For arbitrary $y \in \mathbb{R}^m$, we have

$$\begin{aligned}
&y^\top S(p_{t+1})y = \frac{\mathbf{E}_{x\sim p_t}\left[(y^\top x)^2\exp\left(-\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))\right)\right]}{\mathbf{E}_{x\sim p_t}\left[\exp\left(-\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))\right)\right]} \\
&\leq \mathop{\mathbf{E}}_{x\sim p_t}\left[(y^\top x)^2\exp\left(-\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))\right)\right] \\
&= \mathop{\mathbf{E}}_{x\sim p_t}\left[(y^\top x)^2\exp\left(-\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))\right)\mathbf{1}\left\{\exp\left(-\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))\right)\leq 1+\varepsilon\right\}\right] \\
&\quad + \mathop{\mathbf{E}}_{x\sim p_t}\left[(y^\top x)^2\exp\left(-\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))\right)\mathbf{1}\left\{\exp\left(-\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))\right)> 1+\varepsilon\right\}\right] \\
&\leq (1+\varepsilon)y^\top S(p_t)y + \mathop{\mathbf{E}}_{x\sim p_t}\left[(y^\top x)^2\exp\left(-\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))\right)\mathbf{1}\left\{-\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))>\frac{\varepsilon}{2}\right\}\right],
\end{aligned} \tag{34}$$

where the first inequality follows from Jensen's inequality, and the last inequality holds as $\exp(y) > 1+\varepsilon$ implies $y > \varepsilon/2$ for $0 < \varepsilon < 1/2$. Let us evaluate the second term in (34), using Lemma 1. When $x \sim p_t$, we have

$$\mathop{\mathbf{E}}_{x\sim p_t}[(y^\top x)^2] = \|y\|^2_{S(p_t)}.$$

Futhermore, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{x\sim p_t}[(-\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t)))^2] &\leq \eta^2\|\hat{\ell}_{t-d}\|^2_{S(p_t)} \leq \left(1+\frac{1}{d+1}\right)^d\eta^2\|\hat{\ell}_{t-d}\|^2_{S(p_{t-d})} \\ &\leq \mathrm{e}\eta^2\|\hat{\ell}_{t-d}\|^2_{S(p_{t-d})} \leq 4\mathrm{e}\eta^2 m\gamma^2\end{aligned}$$

[2] We should note that Lemma 6 can be applied here in the proof of Lemma 5, as Lemma 5 is *not* used in the proof of Lemma 6.

where the first inequality follows from $\mathbf{E}[(x-\mu(p_t))(x-\mu(p_t))^\top] \preceq S(p_t)$, the second inequality follows from the inductive assumption and the second part of (7), and the forth inequality follows from the second part of (11). From Lemma 1, we have

$$
\begin{aligned}
&\operatorname*{Prob}_{x\sim p_t}\left[\frac{|\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))|}{2\eta\gamma\sqrt{\mathrm{e}m}}+\frac{|y^\top x|}{\|y\|_{S(p_t)}}>\alpha\right] \\
&\leq \operatorname*{Prob}_{x\sim p_t}\left[\frac{|\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))|}{2\eta\gamma\sqrt{\mathrm{e}m}}>\frac{\alpha}{2}\right]+\operatorname*{Prob}_{x\sim p_t}\left[\frac{|y^\top x|}{\|y\|_{S(p_t)}}>\frac{\alpha}{2}\right] \leq 2\exp\left(1-\frac{\alpha}{2}\right)
\end{aligned} \tag{35}
$$

for any $\alpha > 0$. Using this, we have

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{x\sim p_t}\left[(y^\top x)^2\exp\left(-\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))\right)\mathbf{1}\left\{-\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))>\frac{\varepsilon}{2}\right\}\right] \\
&\leq \mathop{\mathbf{E}}_{x\sim p_t}\left[(y^\top x)^2\exp\left(-\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))\right)\mathbf{1}\left\{\frac{|\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))|}{2\eta\gamma\sqrt{\mathrm{e}m}}+\frac{|y^\top x|}{\|y\|_{S(p_t)}}>\frac{\varepsilon}{4\eta\gamma\sqrt{\mathrm{e}m}}\right\}\right] \\
&\leq \frac{\varepsilon^2\|y\|_{S(p_t)}^2}{16\mathrm{e}\eta^2\gamma^2 m}\sum_{n=1}^{\infty}(n+1)^2\exp\left(\frac{(n+1)\varepsilon}{2}\right)\operatorname*{Prob}_{x\sim p_t}\left[\frac{n\varepsilon}{4\eta\gamma\sqrt{\mathrm{e}m}}<\frac{|\eta\hat{\ell}_{t-d}^\top(x-\mu(p_t))|}{2\eta\gamma\sqrt{\mathrm{e}m}}+\frac{|y^\top x|}{\|y\|_{S(p_t)}}\leq\frac{(n+1)\varepsilon}{4\eta\gamma\sqrt{\mathrm{e}m}}\right] \\
&\leq \frac{\varepsilon^2\|y\|_{S(p_t)}^2}{8\mathrm{e}\eta^2\gamma^2 m}\sum_{n=1}^{\infty}(n+1)^2\exp\left(\frac{(n+1)\varepsilon}{2}\right)\exp\left(1-\frac{n\varepsilon}{8\eta\gamma\sqrt{\mathrm{e}m}}\right) \\
&\leq \frac{\|y\|_{S(p_t)}^2}{2\eta^2\gamma^2 m}\sum_{n=1}^{\infty}\exp\left((n+1)\varepsilon\right)\exp\left(-\frac{n\varepsilon}{8\eta\gamma\sqrt{\mathrm{e}m}}\right)=\frac{\exp(\varepsilon)\|y\|_{S(p_t)}^2}{2\eta^2\gamma^2 m}\sum_{n=1}^{\infty}\exp\left(\varepsilon-\frac{\varepsilon}{8\eta\gamma\sqrt{\mathrm{e}m}}\right)^n \\
&\leq \frac{\mathrm{e}\|y\|_{S(p_t)}^2}{2\eta^2\gamma^2 m}\sum_{n=1}^{\infty}\exp\left(-\frac{\varepsilon}{16\eta\gamma\sqrt{\mathrm{e}m}}\right)^n \leq \frac{\mathrm{e}\|y\|_{S(p_t)}^2}{\eta^2\gamma^2 m}\exp\left(-\frac{\varepsilon}{16\eta\gamma\sqrt{\mathrm{e}m}}\right)\leq \|y\|_{S(p_t)^2}\cdot\varepsilon,
\end{aligned}
$$

where the third inequality follows from (35), the forth inequality follows from $(\frac{\varepsilon(n+1)}{2})^2 \leq \exp(\frac{\varepsilon(n+1)}{2})$, and the forth, fifth and the last inequalities follow from (33) and the assumption of $\eta \leq \frac{1}{100(d+1)\gamma^2\sqrt{m}}$ with $\gamma = 4\log(10mT)$. Combining this and (34), we have

$$
y^\top S(p_{t+1})y \leq (1+2\varepsilon)y^\top S(p_{t+1})y = \left(1+\frac{1}{d+1}\right)y^\top S(p_{t+1})y.
$$

Since this holds for any $y \in \mathbb{R}^d$, we have the inequality in Lemma 6. □

B Proof of Theorem 3

We start by showing the following properties of $\hat{\ell}_t(c)$ defined in (18):

Lemma 7. *The vector $\hat{\ell}_t(c)$ defined by (18) satisfies*

$$
\mathbf{E}\left[\hat{\ell}_t(c)\right]=\ell_t, \quad \left\|\hat{\ell}_t(c)\right\|_{S(p_{t-1}^c)}^2 \leq 4m\gamma^2, \quad \mathbf{E}\left[\left\|\hat{\ell}_t(c)\right\|_{S(p_{t-1}^c)}^2\right] \leq \frac{4m}{\min\{|N(c)|,m\}}. \tag{36}
$$

Proof. Since $b_t'(v)$ is chosen from the posterior of $b_t(v)$ given the observation $a_t(v)$, $b_t'(v)$ and $b_t(v)$ are identically distributed conditioned on $a_t(v)$. Hence, we have

$$
\mathbf{E}\left[b_t'(v)a_t(v)^\top|p_{t-1}^c{}'\right]=\mathbf{E}\left[b_t(v)a_t(v)^\top|p_{t-1}^c{}'\right]=\mathbf{E}\left[b_t(v)b_t(v)^\top|p_{t-1}^c{}'\right]=S(p_{t-1}^c{}') \tag{37}
$$

for all $v \in N(c)$. From this, we can show the first equality in (36) as follows:

$$
\begin{aligned}
\mathbf{E}\left[\hat{\ell}_t(c)\right] &= \frac{1}{|N(c)|}\sum_{v\in N(c)}\mathbf{E}\left[S(p_{t-1}^c{}')^{-1}b_t'(v)a_t(v)^\top\ell_t(v)\right] \\
&= \frac{1}{|N(c)|}\sum_{v\in N(c)}\mathbf{E}\left[\ell_t(v)\right]=\frac{1}{|N(c)|}\sum_{v\in N(c)}\ell_t=\ell_t.
\end{aligned} \tag{38}
$$

The second part in (36) can be shown similarly to the second part of (11): From the definition (18) of $\hat{\ell}_t(c)$, we have

$$\left\|\hat{\ell}_t(c)\right\|_{S(p^c_{t-1})} \leq \frac{1}{|N(c)|}\sum_{v\in N(c)}\left\|\ell_t(v)^\top a_t(v) S({p^c_{t-1}}')^{-1} b'_t(v)\right\|_{S(p^c_{t-1})}$$
$$\leq \frac{1}{|N(c)|}\sum_{v\in N(c)}\left\|S({p^c_{t-1}}')^{-1} b'_t(v)\right\|_{S(p^c_{t-1})} \tag{39}$$

where the last inequality follows from the assumption of $|\ell_t(v)^\top a_t(v)| \leq 1$. For all $v \in N(c)$, we have

$$\|S({p^c_{t-1}}')^{-1} b'_t(v)\|^2_{S(p^c_{t-1})} \leq 2\|S({p^c_{t-1}}')^{-1} b'_t(v)\|^2_{S({p^c_{t-1}}')}$$
$$= 2\|b'_t\|_{S({p^c_{t-1}}')^{-1}} \leq 4\|b'_t\|_{S(p^c_{t-1})^{-1}} \leq 4m\gamma^2, \tag{40}$$

where the first and second inequalities follow from the second part of Lemma 2, and the last inequality follows from the fact that b'_t is chosen from the posterior for $b_t \sim {p^c_{t-1}}'$, the truncated distribution of p^c_{t-1}. Combining (39) and (40), we obtain the second part of (36). To show the last part of (36), we evaluate the variance of $\hat{\ell}_t(c)$. As shown in (38), for all $v \in N(c)$ and any fixed $\ell_t(v)$, we have $\mathbf{E}\left[S({p^c_{t-1}}')^{-1} b'_t(v) a_t(v)^\top \ell_t(v)\right] = \ell_t(v)$ and $S({p^c_{t-1}}')^{-1} b'_t(v) a_t(v)^\top \ell_t(v)$ are independent for $v \in N(c)$. Hence, we have

$$\mathbf{E}\left[\left\|\hat{\ell}_t(c)\right\|^2_{S(p^c_{t-1})}\right] = \frac{1}{|N(c)|^2}\mathbf{E}\left[\left\|\sum_{v\in N(c)} S({p^c_{t-1}}')^{-1} b'_t(v) a_t(v)^\top \ell_t(v)\right\|^2_{S(p^c_{t-1})}\right]$$
$$= \frac{1}{|N(c)|^2}\mathbf{E}\left[\left\|\sum_{v\in N(c)}\left(S({p^c_{t-1}}')^{-1} b'_t(v) a_t(v)^\top \ell_t(v) - \ell_t(v)\right)\right\|^2_{S(p^c_{t-1})} + \left\|\sum_{v\in N(c)}\ell_t(v)\right\|^2_{S(p^c_{t-1})}\right]$$
$$= \frac{1}{|N(c)|^2}\left(\sum_{v\in N(c)}\mathbf{E}\left[\left\|S({p^c_{t-1}}')^{-1} b'_t(v) a_t(v)^\top \ell_t(v) - \ell_t(v)\right\|^2_{S(p^c_{t-1})}\right]\right)$$
$$+ \frac{1}{|N(c)|^2}\mathbf{E}\left[\left\|\sum_{v\in N(c)}\ell_t(v)\right\|^2_{S(p^c_{t-1})}\right]. \tag{41}$$

Since we have $\mathbf{E}\left[S({p^c_{t-1}}')^{-1} b'_t(v) a_t(v)^\top \ell_t(v)\right] = \ell_t(v)$ for any fixed $\ell_t(v)$, we have

$$\mathbf{E}\left[\left\|S({p^c_{t-1}}')^{-1} b'_t(v) a_t(v)^\top \ell_t(v) - \ell_t(v)\right\|^2_{S(p^c_{t-1})}\right] \leq \mathbf{E}\left[\left\|S({p^c_{t-1}}')^{-1} b'_t(v) a_t(v)^\top \ell_t(v)\right\|^2_{S(p^c_{t-1})}\right]$$
$$\leq 2\,\mathbf{E}\left[\left\|S({p^c_{t-1}}')^{-1} b'_t(v)\right\|^2_{S({p^c_{t-1}}')}\right] = 2\,\mathbf{E}\left[\|b'_t(v)\|^2_{S({p^c_{t-1}}')^{-1}}\right] = 2S({p^c_{t-1}}') \bullet S({p^c_{t-1}}')^{-1} = 2m, \tag{42}$$

where the last inequality follows from the second part of Lemma 2 and $|a_t(v)^\top \ell_t(v)| \leq 1$ and the second inequality follows from the fact that $b'_t(v) \sim {p^c_{t-1}}'$ after marginalizing $a_t(v)$ out. Further, the second term in the right-hand side of (41) can be bounded as

$$\mathbf{E}\left[\left\|\sum_{v\in N(c)}\ell_t(v)\right\|^2_{S(p^c_{t-1})}\right] = \mathbf{E}\left[\mathop{\mathbf{E}}_{x\sim S(p^c_{t-1})}\left[\left(\left(\sum_{v\in N(c)}\ell_t(v)\right)^\top x\right)^2\right]\right] \leq |N(c)|^2, \tag{43}$$

where the last inequality follows from the assumption of $|\ell_t(v)^\top a| \leq 1$ for all $a \in \mathcal{A}$. Combining (41), (42), and (43), we have

$$\mathbf{E}\left[\left\|\hat{\ell}_t(c)\right\|^2_{S(p^c_{t-1})}\right] \leq \frac{|N(c)|\cdot 2m}{|N(c)|^2} + \frac{|N(c)|^2}{|N(c)|^2} = \frac{\cdot 2m}{|N(c)|} + 1 \leq \frac{4m}{\min\{|N(c)|, m\}}.$$

□

From Lemma 7, we can apply Lemmas 5 and 6 to p^c_{t-1}. In fact, p^c_{t-1} is identical to the distribution p_t defined by (9) with $\hat{\ell}_t = \hat{\ell}_t(c)$ and $d = 1$, and the first and second parts of Lemma 7 implies that the conditions in (11) hold. Hence, from Lemmas 5 and 6, we have

$$\mathbf{E}\left[|\ell^\top(\mu(p^c_t) - \mu(p^c_{t+1}))|\right] \leq 2\eta, \quad S(p^c_{t+1}) \preceq 2S(p^c_t) \tag{44}$$

for any $t \in [T]$, $c \in C$ and $\ell \in \mathbb{R}^m$ such that $|\ell^\top a| \leq 1$ for all $a \in \mathcal{A}$.

Let us evaluate the regret of agent $v \in V_c$. We denote $p^v_t = p^c_{t-d_c(v)}$. From $q^v_t = q^c_{t-d_c(v)}$, we have $\mu(q^v_t) = \mu(p^v_t{}')$. Further, since $p^v_t = p^c_{t-d_c(v)}$ is identical to the distribution p_t defined by (9) with $\hat{\ell}_t = \hat{\ell}_t(c)$ and $d = d_c(v)$, from Lemma 4, the regret of v is bounded as

$$\begin{aligned}
&\mathbf{E}[R_T(v)] \\
&\leq \mathbf{E}\left[\sum_{t=1}^T \left(\ell_t^\top \left(\mu(p^v_t) - \mu(p^v_{t+d_c(v)})\right) + \frac{1}{\eta(c)} \mathop{\mathbf{E}}_{x\sim p^v_{t+d_c(v)}}\left[g\left(-\eta(c)\hat{\ell}_t(c)^\top x\right)\right]\right)\right] + \frac{m\log T}{\eta(c)} + 3 \\
&= \mathbf{E}\left[\sum_{t=1}^T \left(\ell_t^\top \left(\mu(p^c_{t-d_c(v)}) - \mu(p^c_t)\right) + \frac{1}{\eta(c)} \mathop{\mathbf{E}}_{x\sim p^c_t}\left[g\left(-\eta(c)\hat{\ell}_t(c)^\top x\right)\right]\right)\right] + \frac{m\log T}{\eta(c)} + 3.
\end{aligned} \tag{45}$$

From the first part of (44), we have

$$\begin{aligned}
\mathbf{E}\left[\ell_t^\top \left(\mu(p^c_{t-d_c(v)}) - \mu(p^c_t)\right)\right] &= \sum_{i=0}^{d_c(v)-1} \mathbf{E}\left[\ell_t^\top \left(\mu(p^c_{t-d_c(v)+i}) - \mu(p^c_{t-d_c(v)+i+1})\right)\right] \\
&\leq 2\eta(c)d_c(v).
\end{aligned} \tag{46}$$

We can bound the term $\mathbf{E}\left[\mathbf{E}_{x\sim p^c_t}\left[g\left(-\eta(c)\hat{\ell}_t(c)^\top x\right)\right]\right]$ in (45) using Lemma 3 and (44). In fact, we can confirm that the assumption of Lemma 3 holds, as follows:

$$\mathop{\mathbf{E}}_{x\sim p^c_t}\left[\left(-\eta(c)\hat{\ell}_t(c)^\top x\right)^2\right] \leq \eta(c)^2\|\hat{\ell}_t(c)\|^2_{S(p^c_t)} \leq 2\eta(c)^2\|\hat{\ell}_t(c)\|^2_{S(p^c_{t-1})} \leq 8\eta(c)^2 m\gamma^2 \leq 1/100,$$

where the second and third inequalities follows from (44). Hence, by applying Lemma 3 to $y = -\eta(c)\hat{\ell}_t(c)^\top x$, we have

$$\begin{aligned}
\mathbf{E}\left[\mathop{\mathbf{E}}_{x\sim p^c_t}\left[g\left(-\eta(c)\hat{\ell}_t(c)^\top x\right)\right]\right] &\leq 2\eta(c)^2\,\mathbf{E}\left[\|\hat{\ell}_t(c)\|^2_{S(p^c_t)}\right] \\
&\leq 4\eta(c)^2\,\mathbf{E}\left[\|\hat{\ell}_t(c)\|^2_{S(p^c_{t-1})}\right] \leq \frac{16\eta(c)^2 m}{\min\{|N(c)|, m\}},
\end{aligned} \tag{47}$$

where the second inequality follows from the second part of (44), and the last inequality follows from the last part of Lemma 7. Combining (45), (46), and (47), we obtain

$$\mathbf{E}[R_T(v)] \leq \eta(c)\left(2d_c(v) + \frac{16m}{\min\{|N(c)|, m\}}\right)T + \frac{m\log T}{\eta(c)} + 3. \tag{48}$$

From the inequality in Theorem 5, we have

$$d_c(v) \leq 6\left(1 + \log\frac{\min\{|N(v)|, m\}}{\min\{|N(c)|, m\}}\right) \leq 6(1 + \log m), \quad \frac{1}{\min\{|N(c)|, m\}} \leq \frac{\mathrm{e}}{\min\{|N(v)|, m\}}. \tag{49}$$

Combining (48) and (49), we have

$$
\begin{aligned}
\mathbf{E}[R_T(v)] &\leq \eta(c)\left(12(1+\log m) + \frac{16m}{\min\{|N(c)|, m\}}\right)T + \frac{m\log T}{\eta(c)} + 3 \\
&\leq 16\eta(c)\left(1+\log m + \frac{m}{\min\{|N(c)|, m\}}\right)T + \frac{m\log T}{\eta(c)} + 3 \\
&\leq \max\left\{8\sqrt{m\left(1+\log m + \frac{m}{\min\{|N(c)|, m\}}\right)T\log T}, 100m^2\gamma^2\log T\right\} + 3 \\
&\leq \max\left\{8\sqrt{m\left(1+\log m + \frac{\mathrm{e}m}{\min\{|N(v)|, m\}}\right)T\log T}, 100m^2\gamma^2\log T\right\} + 3, \\
&\leq \max\left\{16\sqrt{m\left(1+\log m + \frac{m}{|N(v)|}\right)T\log T}, 100m^2\gamma^2\log T\right\} + 3,
\end{aligned}
$$

where the first inequality follows from (48) and the first part of (49) the third inequality follows from the parameter setting of $\eta(c) = \min\{\frac{1}{4}\sqrt{\frac{m\log T}{T(1+\log m + m/\min\{|N(c)|,m\})}}, \frac{1}{100\gamma^2 m}\}$, and the forth inequality follows from the second part of (49).

C Proof of Theorem 2

Proof of Theorem 2 We first construct a problem instance for which $R_T = \Omega(\sqrt{mdT})$. Let T be a multiple of md, i.e., we denote $T = Smd$ with an integer S. For each $s = 0, 1, \ldots, S-1$ and $i = 1, 2, \ldots, m$, we define ℓ_t for $t \in [smd + d(i-1) + 1, smd + id]$ by $\ell_t = b_{si}\chi_i$, where b_{si} follows a Bernoulli distribution over $\{-1, 1\}$ with parameter $1/2$ for s and i, independently. Since ℓ_t and a_t are independent for all t and $\mathbf{E}[\ell_t] = 0$, we have

$$
\mathbf{E}\left[\sum_{t=1}^{T} \ell_t^\top a_t\right] = 0. \tag{50}
$$

On the other hand, we also have

$$
\mathbf{E}\left[\min_{a\in\mathcal{A}} \sum_{t=1}^{T} \ell_t^\top a\right] = -\sum_{i=1}^{m} \mathbf{E}\left[\left|\sum_{t=1}^{T} \ell_{ti}\right|\right] = -d\sum_{i=1}^{m} \mathbf{E}\left[\left|\sum_{s=1}^{S} b_{si}\right|\right]. \tag{51}
$$

Since $\sum_{s=1}^{S} b_{si}$ follows a binomial distribution $Bi(S, 1/2)$, we have $\mathrm{Prob}[|\sum_{s=1}^{S} b_{si}| \geq \sqrt{S}/10] \geq 1/5$. Hence we have

$$
\mathbf{E}\left[\min_{a\in\mathcal{A}} \sum_{t=1}^{T} \ell_t^\top a\right] \leq -\frac{1}{50} dm\sqrt{S}. \tag{52}
$$

This implies that $\mathbf{E}[R_T] \geq dm\sqrt{S}/50 = \sqrt{dmT}/50$. Even when $T \leq dm$, we can show $\mathbf{E}[R_T] \geq T$ similarly. Hence, we have

$$
\mathbf{E}[R_T] \geq \min\left\{\frac{\sqrt{dmT}}{50}, T\right\}. \tag{53}
$$

We next provide a distribution of ℓ_t for which the regret is $\Omega(m\sqrt{T})$. Let $\varepsilon = \min\{\frac{1}{6}, \frac{m}{\sqrt{8T}}\}$. Consider generating ℓ_t in the following process: First, pick $a^* \in \{-1, 1\}^d$ uniformly at random. Then for $t = 1, \ldots, T$, pick $i_t \in [m]$ uniformly at random and set $\ell_t = s_t\chi_{i_t}$ where $s_t = a^*_{i_t}$ with probability $\frac{1-\varepsilon}{2}$ and $s_t = -a^*_{i_t}$ with probability $\frac{1+\varepsilon}{2}$. Then, the regret is bounded as

$$
\mathbf{E}[R_T] \geq \mathbf{E}\left[\sum_{t=1}^{T} \ell_t^\top a_t - \sum_{t=1}^{T} \ell_t^\top a^*\right] = \sum_{t=1}^{T} \mathbf{E}\left[\ell_t^\top a_t - \ell_t^\top a^*\right]. \tag{54}
$$

As shown in Lemmas 3 and 4 in [26], we have $\mathbf{E}\left[\ell_t^\top a_t - \ell_t^\top a^*\right] \geq \varepsilon/2$ for any algorithm if $\varepsilon \leq \min\{\frac{1}{6}, \frac{m}{\sqrt{8T}}\}$. Hence, the regret is bounded as

$$\mathbf{E}[R_T] \geq \frac{T\varepsilon}{2} = \min\left\{\frac{m\sqrt{T}}{32}, \frac{T}{12}\right\}. \tag{55}$$

If $(\ell_t)_{t=1}^T$ follows the first distribution with probability $1/2$ and the second distribution with probability $1/2$, the regret is bounded as

$$\mathbf{E}[R_T] \geq \min\left\{\frac{m\sqrt{T}}{64} + \frac{\sqrt{dmT}}{100}, \frac{T}{12}\right\} \geq \min\left\{\frac{\sqrt{m(d+m)T}}{100}, \frac{T}{12}\right\}. \tag{56}$$

□

D Proof of Theorem 4

Proof of Theorem 4 We first construct a problem instance for which $R_T(v) = \Omega(\sqrt{mT})$. Let T be a multiple of m, i.e., we denote $T = Sm$ with an integer S. For each $s \in [S]$ and $i \in [m]$, we define ℓ_t by $\ell_t = b_{si}\chi_i$ for $t = (s-1)m + i$, where b_{si} follows a Bernoulli distribution over $\{-1, 1\}$ with parameter $1/2$ independently for $s \in [S]$ and $i \in [m]$. We set $\ell_t(v) = \ell_t$ for all $v \in V$. Since ℓ_t and $a_t(v)$ are independent for all t and $\mathbf{E}[\ell_t] = 0$, we have

$$\mathbf{E}\left[\sum_{t=1}^T \ell_t^\top a_t(v)\right] = 0. \tag{57}$$

On the other hand, we also have

$$\mathbf{E}\left[\min_{a\in\mathcal{A}} \sum_{t=1}^T \ell_t^\top a\right] = -\sum_{i=1}^m \mathbf{E}\left[\left|\sum_{t=1}^T \ell_{ti}\right|\right] = -\sum_{i=1}^m \mathbf{E}\left[\left|\sum_{s=1}^S b_{si}\right|\right]. \tag{58}$$

Since $\sum_{s=1}^S b_{si}$ follows a binomial distribution $Bi(S, 1/2)$, we have $\mathrm{Prob}[|\sum_{s=1}^S b_{si}| \geq \sqrt{S}/10] \geq 1/5$. Hence, we have

$$\mathbf{E}\left[\min_{a\in\mathcal{A}} \sum_{t=1}^T \ell_t^\top a\right] \leq -\frac{1}{50} m\sqrt{S}. \tag{59}$$

This implies that $\mathbf{E}[R_T(v)] \geq m\sqrt{S}/50 = \sqrt{mT}/50$. Even when $T \leq m$, we can show $\mathbf{E}[R_T] \geq T$ similarly. Hence, we have

$$\mathbf{E}[R_T] \geq \min\left\{\frac{\sqrt{dmT}}{50}, T\right\}. \tag{60}$$

We next provide a lower bound of $\Omega(m\sqrt{T/|V|})$. Let $\varepsilon = \min\{\frac{1}{6}, \frac{m}{\sqrt{8T|V|}}\}$. Consider generating $\ell_t(v)$ in the following process: First, pick $a^* \in \{-1, 1\}^d$ uniformly at random. Then for $t = 1, \ldots, T$ and $v \in V$, pick $i_t(v) \in [m]$ uniformly at random and set $\ell_t(v) = s_t(v)\chi_{i_t(v)}$ where $s_t(v) = a^*_{i_t(v)}$ with probability $\frac{1-\varepsilon}{2}$ and $s_t(v) = -a^*_{i_t(v)}$ with probability $\frac{1+\varepsilon}{2}$. Then, as can be shown from Lemmas 3 and 4 in [26], the regret is bounded as

$$\mathbf{E}[R_T(v)] \geq \frac{T\varepsilon}{2} = \min\left\{\frac{m}{32}\sqrt{\frac{T}{|V|}}, \frac{T}{12}\right\}. \tag{61}$$

If $(\ell_t)_{t=1}^T$ follows the first distribution with probability $1/2$ and the second distribution with probability $1/2$, the regret is bounded as

$$\mathbf{E}[R_T] \geq \min\left\{\frac{m}{64}\sqrt{\frac{T}{|V|}} + \frac{\sqrt{mT}}{100}, \frac{T}{12}\right\} \geq \min\left\{\frac{1}{100}\sqrt{m\left(1 + \frac{m}{|V|}\right)T}, \frac{T}{12}\right\}. \tag{62}$$

□