An Improved Algorithm for Adversarial Linear Contextual Bandits via Reduction

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present an efficient algorithm for linear contextual bandits with adversarial losses and stochastic action sets. Our approach reduces this setting to misspecification-robust adversarial linear bandits with fixed action sets. Without knowledge of the context distribution or access to a context simulator, the algorithm achieves $\widetilde{\mathcal{O}}(d^2\sqrt{T})$ regret and runs in poly(d, C, T) time, where d is the feature dimension, C is the number of linear constraints defining the action set in each round, and T is number of rounds. This resolves the open question by Liu et al. (2023) on whether one can obtain poly$(d)\sqrt{T}$ regret in polynomial time independent of the number of actions. For the important class of combinatorial bandits with adversarial losses and stochastic action sets, our algorithm is the first to achieve poly$(d)\sqrt{T}$ regret in polynomial time, while no prior algorithm achieves even $o(T)$ regret in polynomial time to our knowledge. When a simulator is available, the regret bound can be improved to $\widetilde{\mathcal{O}}(d\sqrt{L^\star})$, where $L^\star$ is the cumulative loss of the best policy.

1 Introduction

We consider the following linear contextual bandit problem: At each round $t = 1, \ldots, T$, the environment generates a hidden loss vector $\theta_t \in \mathbb{R}^d$ and an action set $\mathcal{A}_t \subset \mathbb{R}^d$. The learner observes $\mathcal{A}_t$, selects an action $a_t \in \mathcal{A}_t$, and incurs loss $\ell_t(a_t) = a_t^\top \theta_t$. The goal is to compete with the best fixed policy—defined as a mapping from an action set to an element in it. This setting generalizes the classical linear bandit model by allowing the action sets $\mathcal{A}_t$ to vary stochastically over time. Crucially, each $\mathcal{A}_t$ encodes the *context* based on which the learner makes decisions. In this work, $\mathcal{A}_t$ is called *context* or *action set* interchangeably.

This framework is applicable in settings such as healthcare and recommendation systems, where decisions must be made conditional on context. Prior work on contextual bandits has studied a variety of assumptions on how losses and contexts are generated. While much of the literature assumes i.i.d. losses and arbitrarily chosen action sets (for which a well-known algorithm is LinUCB Li et al. (2010)), we focus on the complementary regime: the action sets are drawn i.i.d. from a fixed distribution $\mathcal{D}$, while the losses may be chosen adversarially.

A first computationally efficient algorithm for this setting was proposed by Neu and Olkhovskaya (2020) under the assumption that the context (i.e., action set) distribution is *known*. Since an action set is a subset of $\mathbb{R}^d$ (i.e., it lies in the space $2^{\mathbb{R}^d}$), the distribution over action sets is in the space $\Delta(2^{\mathbb{R}^d})$, which is generally intractable to represent efficiently. This assumption was later removed by subsequent efficient algorithms (Luo et al., 2021; Sherman et al., 2023; Dai et al., 2023; Liu et al., 2023). In the setting where the learner has access to a simulator that can generate free contexts (i.e., the learner is able to sample contexts from $\mathcal{D}$ as many times as they want without incurring cost), Dai

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Table 1: Comparison with state-of-the-art results in adversarial linear contextual bandits. d is the feature dimension, K is an upper bound on the number of actions, and C is an upper bound on the number of linear constraints to describe each action set. It holds that $C \leq K + 1$ and in many combinatorial problems we have $C = \text{poly}(d)$ and $K = 2^{\Omega(d)}$. The run time of the linear optimization oracle of Neu and Valko (2014) is bounded by $\text{poly}(d, C)$ but could also be smaller.

Algorithm	Regret	Computation	Simulator	Feedback
Neu and Valko (2014)	$(dT)^{2/3}$	$\text{poly}(d,T)$ plus T oracle calls	no	semi-bandit
Dai et al. (2023)	$\min\{d\sqrt{T}, \sqrt{dT \log K}\}$	$\text{poly}(d,K,T)$	yes	bandit
Liu et al. (2023)	$d\sqrt{T}$	$K \cdot T^{\Omega(d)}$	no	bandit
Liu et al. (2023)	$d^2\sqrt{T}$	$\text{poly}(d,K,T)$	no	bandit
Ours	$d^2\sqrt{T}$	$\text{poly}(d,C,T)$	no	bandit
Ours	$d\sqrt{L^\star}$	$\text{poly}(d,C,T)$	yes	bandit

et al. (2023) shows that a near-optimal regret bound of $\widetilde{\mathcal{O}}(d\sqrt{T})$ is achievable. When the learner has neither knowledge of the context distribution nor simulator access to random context samples, the best known results are by Liu et al. (2023): they provide an algorithm with near-optimal regret $\widetilde{\mathcal{O}}(d\sqrt{T})$ with run time $T^{\Omega(d)}$ and another algorithm with regret $\widetilde{\mathcal{O}}(d^2\sqrt{T})$ with run time $\text{poly}(d, K, T)$ where $K = \max_t |\mathcal{A}_t|$. Notably, while the regret bound of this last algorithm is independent of the number of actions K, its computational complexity scales polynomially in K. In fact, this is the case for all previous algorithms as well (Luo et al., 2021; Sherman et al., 2023; Dai et al., 2023; Liu et al., 2023). This makes them unsuitable for many important combinatorial problems (e.g., m-set, shortest path, flow, bipartite matching), where K is usually exponentially large in the dimension d.

Our work gives the first algorithm whose computational complexity does not explicitly scale with the number of actions, making adversarial linear contextual bandits applicable to a much wider range of problems. Without simulator access, our method achieves regret $\widetilde{\mathcal{O}}(d^2\sqrt{T})$ and with simulator access this can be improved to $\widetilde{\mathcal{O}}(d\sqrt{L^\star})$, where $L^\star = O(T)$ is the cumulative loss of the best policy. Our algorithm runs in $O(d, C, T)$ time, where C is the number of linear constraints to describe the convex hull of each action set. Notice that $C \leq K + 1$ in general, as the convex hull of an action set of size K can be written as a linear program with at most $K + 1$ constraints. On the other hand, in many combinatorial problems, $C = \text{poly}(d)$ while $K = 2^{\Omega(d)}$. For example, in a shortest path problem with d edges, the set of all paths can be described by a linear program with $O(d)$ constraints, while the number of paths could be of order $2^{\Omega(d)}$. For combinatorial problems with stochastic action sets and adversarial losses, we are only aware of Neu and Valko (2014) who studied the case where the learner has *semi-bandit* feedback. Their algorithm achieves $\widetilde{\mathcal{O}}\big((dT)^{2/3}\big)$ regret with one call to the linear optimization oracle of the action set per round. Compared to their work, our work weakens the assumption on the feedback (from semi-bandits to full-bandits) and improves the regret bound, but our method is computationally heavier—there are action sets where linear optimization can be solved in polynomial time while having an exponentially large number of constraints, such as spanning trees. How to further improve our computational complexity to match theirs is left as future work.

Our result is achieved by establishing a novel and computationally efficient reduction from *adversarial linear contextual bandits* to *adversarial linear bandits with misspecification*. Linear bandits can be viewed as linear contextual bandits with a fixed context and is therefore less challenging than linear contextual bandits. This reduction scheme is related to the work of Hanna et al. (2023) from a high level, which reduces stochastic linear contextual bandits (with both stochastic contexts and stochastic losses) to misspecified stochastic linear bandits. However, our problem is strictly more challenging: The algorithm of Hanna et al. (2023) heavily relies on the learner estimates the underlying loss vector $\theta_\star$, while this is impossible in our case where the loss vector changes over time.

Finally, regarding results in terms of $L^\star$, we remark that our bound of $\widetilde{\mathcal{O}}(d\sqrt{L^\star})$ can be specialized to the setting of Olkhovskaya et al. (2023) (a slightly different formulation of adversarial contextual bandits). We then obtain a slightly worse rate but, importantly, remove their unrealistic assumption that the context distribution is log-concave.

2 Problem Description and Main Results

Notation Suppose $\mathcal{A} \subset \mathbb{R}^d$ is a set of vectors. Then the convex hull of $\mathcal{A}$ is denoted as $\mathrm{conv}(\mathcal{A}) = \{x = \sum_{i=1}^k \lambda_i a_i : k \in \mathbb{N}, \sum_{i=1}^k \lambda_i = 1, \lambda_i \geq 0, a_i \in \mathcal{A}\}$. If $\mathcal{A}$ is a polytope, we denote its set of vertices as $\mathrm{vert}(\mathcal{A})$, and denote the normal cone of $\mathcal{A}$ at a vertex $v \in \mathrm{vert}(\mathcal{A})$ as $\mathcal{N}(\mathcal{A}, v) = \{y \in \mathbb{R}^d : \max_{x \in \mathcal{A}} \langle y, x - v \rangle \leq 0\}$.

2.1 Linear Contextual Bandits

For simplicity, we consider an oblivious adversary.[1] Before any interaction with the learner, the adversary secretly chooses T loss vectors $(\theta_t)_{t \in [T]}$ where $\theta_t \in \mathbb{R}^d$ for all t. For each round $t = 1, 2, \ldots, T$, an action set $\mathcal{A}_t \subset \mathbb{R}^d$ is drawn according to $\mathcal{A}_t \overset{\text{i.i.d.}}{\sim} \mathcal{D}$ and revealed to the learner. The learner then chooses an action $a_t \in \mathcal{A}_t$ and observes loss $\ell_t(a_t) \in [-1, 1]$ with $\mathbb{E}[\ell_t(a_t)] = a_t^\top \theta_t$. The regret against a fixed policy $\pi \in \Pi$ is defined as

$$\mathrm{Reg}_T(\pi) = \sum_{t=1}^T (\ell_t(a_t) - \ell_t(\pi(\mathcal{A}_t))) = \sum_{t=1}^T \langle a_t - \pi(\mathcal{A}_t), \theta_t \rangle$$

where policy π maps an action set $\mathcal{A} \subset \mathbb{R}^d$ to a point $\pi(\mathcal{A}) \in \mathrm{conv}(\mathcal{A})$.[2] Note that the policy that minimizes the total expected loss is

$$\operatorname*{arg\,min}_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^T \langle \pi(\mathcal{A}_t), \theta_t \rangle\right] = \operatorname*{arg\,min}_{\pi \in \Pi} \mathbb{E}_{\mathcal{A} \sim \mathcal{D}}\left[\left\langle \pi(\mathcal{A}), \sum_{t=1}^T \theta_t \right\rangle\right],$$

which is attained by the policy $\pi(\mathcal{A}) = \operatorname{arg\,min}_{a \in \mathcal{A}} \left\langle a, \sum_{t=1}^T \theta_t \right\rangle$. Thus, to minimize the expected regret, it suffices to compare the learner to the class of linear classifier policies

$$\Pi_{\text{lin}} = \left\{\pi_\phi \mid \phi \in \mathbb{R}^d\right\} \text{ where } \pi_\phi(\mathcal{A}) = \operatorname*{arg\,min}_{a \in \mathcal{A}} \langle a, \phi \rangle . \tag{1}$$

2.2 Linear Bandits and ϵ-Misspecified Linear Bandits

The adversarial linear bandit problem with oblivious adversary is the case in which the learner decides $(\theta_t)_{t \in [T]}$ before any interaction with the learner. The learner is given a fixed action set $\widehat{\Omega} \subset \mathbb{R}^d$. At every round $t \in [T]$, the learner chooses an action $y_t \in \widehat{\Omega}$ and receives $\ell_t(y_t) \in [-1, 1]$ as feedback with $\mathbb{E}[\ell_t(y_t)] = \langle y_t, \theta_t \rangle$. The regret with respective to a fixed action $y \in \widehat{\Omega}$ is defined as

$$\mathrm{Reg}_T(y) = \sum_{t=1}^T \langle y_t - y, \theta_t \rangle .$$

A *misspecified* linear bandit problem is the case where the learner, instead of receiving an unbiased sample of $\langle y_t, \theta_t \rangle$ as feedback, receives $\ell_t(y_t)$ with $|\,\mathbb{E}[\ell_t(y_t)] - \langle y_t, \theta_t \rangle\,| \leq \epsilon$ for some ϵ known to the algorithm.

2.3 Results Overview

In this section, we present a general framework that can reduce the adversarial contextual bandit problem to a misspecification-robust linear bandit algorithm.

Definition 1 (α-misspecification-robust adversarial linear bandit algorithm)**.** *A α-misspecification-robust linear bandit algorithm over action set $\widehat{\Omega} \subset \mathbb{R}^d$ has the following property: with a given $\epsilon > 0$*

[1]There are two kinds of results we can get: 1) regret compared to the best policy in the full policy set Π under oblivious adversary, and 2) regret compared to the best policy in the linear policy set Π_{lin} under adaptive adversary. To achieve 2), it suffices to use misspecification-robust linear bandit algorithms with high-probability bounds in our reduction. This can be achieved by standard techniques of getting high-probability bounds (Lee et al., 2020; Zimmert and Lattimore, 2022). To simplify the exposition, we only focus on the first case.

[2]Defining $\pi(\mathcal{A}) \in \mathrm{conv}(\mathcal{A})$ instead of $\pi(\mathcal{A}) \in \mathcal{A}$ only makes the guarantee more general and simplifies the notation. Equivalently, it defines a randomized policy: to execute $\pi(\mathcal{A})$, sample $a \in \mathcal{A}$ such that $\mathbb{E}[a] = \pi(\mathcal{A})$.

and the guarantee that every time the learner chooses $y_t \in \widehat{\Omega}$, the loss received $\ell_t(y_t) \in [-1,1]$ satisfies

$$|\,\mathbb{E}\left[\ell_t(y_t)\right] - \langle y_t, \theta_t\rangle\,| \le \epsilon,$$

the algorithm ensures

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle y_t, \theta_t\rangle\right] \le \min_{y \in \widehat{\Omega}} \sum_{t=1}^{T} \langle y, \theta_t\rangle + \widetilde{O}\left(d\sqrt{T} + \alpha\sqrt{d}\epsilon T\right). \tag{2}$$

Notice that there is an α parameter in Definition 1 that specifies the dependence of the regret on the misspecification level ϵ. It is known that $\alpha = 1$ is the statistically optimal dependence. However, for specific algorithms, we might have $\alpha > 1$.

This work establishes the following reductions:

Theorem 1. *Given access to an α-misspecification-robust adversarial linear bandit algorithm, we can achieve $\min_{\pi\in\Pi} \mathrm{Reg}_T(\pi) \le \widetilde{O}(d\sqrt{T} + \alpha d^{1.5}\sqrt{T})$ in adversarial linear contextual bandits without access to a simulator.*

We remark that the $\alpha d^{1.5}\sqrt{T}$ term in Theorem 1 comes from the misspecification term $\alpha\sqrt{d}\epsilon T$ in (2). When the learner has access to a simulator that generates free contexts, the ϵ can be made arbitrarily small, allowing us to achieve the optimal d dependence. This will be discussed in Section 5.

3 Reduction from Linear Contextual Bandits to Linear Bandits

Let π denote a policy, which maps any given action set $\mathcal{A}$ to a randomized action in $\mathrm{conv}(\mathcal{A})$. Let Π denote the set of all possible policies. We define the following map

$$\Psi(\pi) = \mathbb{E}_{\mathcal{A}\sim\mathcal{D}}[\pi(\mathcal{A})],$$

which is the mean action of π. Applying Ψ to all $\pi \in \Pi$, the set $\Omega = \{\Psi(\pi) \mid \pi \in \Pi\}$ is induced. Note that, under this map, the expected loss under actions drawn such that $\mathbb{E}[a_t] = \pi(\mathcal{A}_t)$ may be written as

$$\mathbb{E}\left[\langle a_t, \theta_t\rangle\right] = \mathbb{E}_{\mathcal{A}_t\sim\mathcal{D}}\left[\langle \pi(\mathcal{A}_t), \theta_t\rangle\right] = \langle \Psi(\pi), \theta_t\rangle\,.$$

Accordingly, if the learned draws y_t from policy π_t in round t, the expected regret may be written as

$$\mathbb{E}\left[\mathrm{Reg}_T(\pi)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\langle a_t - \pi(\mathcal{A}_t), \theta_t\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T}\langle \pi_t(\mathcal{A}_t) - \pi(\mathcal{A}_t), \theta_t\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T}\langle \Psi(\pi_t) - \Psi(\pi), \theta_t\rangle\right].$$

3.1 Approximating Ω

Since Ω cannot be accessed directly without full knowledge of the context distribution $\mathcal{D}$, we cannot work with Ω directly. Instead, we will therefore approximate Ω by its empirical counterpart $\widehat{\Omega}$ based on N i.i.d. samples $\mathcal{A}_1, \ldots, \mathcal{A}_N$ from $\mathcal{D}$:

$$\widehat{\Omega} = \left\{\hat{\Psi}(\pi)\middle|\pi \in \Pi\right\} = \left\{x \in \mathbb{R}^d : x = \frac{1}{N}\sum_{i=1}^{N} a_i \;\middle|\; a_i \in \mathrm{conv}(\mathcal{A}_i), a_i = a_j \text{ when } \mathcal{A}_i = \mathcal{A}_j\right\}, \tag{3}$$

where $\hat{\Psi}(\pi) = \frac{1}{N}\sum_{i=1}^{N}\pi(\mathcal{A}_i)$.

We proceed to show that the empirical cumulative loss of any linear classifier policy on the sample $\mathcal{A}_1, \ldots, \mathcal{A}_N$ is close to its expected cumulative loss, as long as N is large enough:

Lemma 1 (Uniform Convergence)**.** *Consider any loss vector $\theta \in \mathbb{R}^d$, and suppose that $|\mathcal{A}| \le K$ and $\max_{a\in\mathcal{A}} |\langle a, \theta\rangle| \le b$ almost surely. Then, for any $\delta \in (0, 1]$, uniformly over all linear classifier policies π_ϕ, the difference in performance of π_ϕ on the sample $\mathcal{A}_1, \ldots, \mathcal{A}_N$ and its expected performance is at most*

$$\sup_{\phi\in\mathbb{R}^d}\left|\left\langle \Psi(\pi_\phi), \theta\right\rangle - \left\langle \hat{\Psi}(\pi_\phi), \theta\right\rangle\right| \le 2b\sqrt{\frac{2d\ln(NK^2)}{N}} + b\sqrt{\frac{2\ln(4/\delta)}{N}}$$

with probability at least $1 - \delta$.

The proof is provided in Appendix B. Its key idea is to rephrase the result as an equivalent statement about uniform convergence for linear multiclass classifiers with K classes in the batch setting, with an unusual loss function. We can then obtain a concentration inequality that holds uniformly over all linear classifiers using standard tools. Specifically, we go via Rademacher complexity and a bound on the growth function of the class of multiclass linear classifiers in terms of its Natarajan dimension, which is known to be at most d.

3.2 Connection between Linear Contextual Bandits and Linear Bandits

---

Algorithm 1: Adversarial Linear Contextual Bandits

Input: An adversarial linear bandit algorithm ALG and a set $\mathcal{S}_N$ of N action sets drawn from $\mathcal{D}$.

Initiate an instance of ALG over action set $\widehat{\Omega}$ constructed from $\mathcal{S}_N$.

for $t = 1, \dots, T$ **do**

- Obtain y_t from ALG.
- Find distribution $\alpha_t \in \Delta(\mathrm{vert}(\widehat{\Omega}))$ such that $\mathbb{E}_{\psi \sim \alpha_t}[\psi] = y_t$.
- Sample $\psi_t \sim \alpha_t$ and let ϕ_t be an arbitrary element in the interior of $-\mathcal{N}(\widehat{\Omega}, \psi_t)$.
- Receive action set $\mathcal{A}_t$, choose action $a_t = \arg\min_{a \in \mathcal{A}_t} \langle a, \phi_t \rangle$, and receive loss ℓ_t.
- Send ℓ_t to ALG.

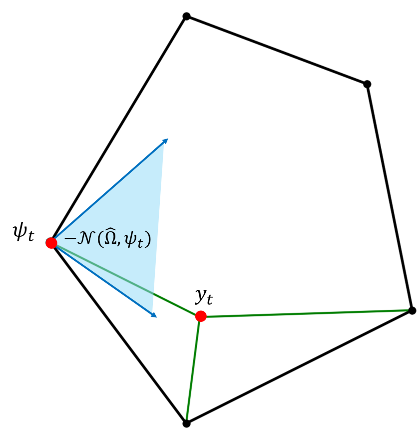

Figure 1: Illustration of $\hat{\Omega}$, y_t, ψ_t and the normal cone $-\mathcal{N}(\hat{\Omega}, \psi_t)$.

The procedure that reduces linear contextual bandits to linear bandits is outlined in Algorithm 1. We also provide a visual illustration of $\hat{\Omega}$ and the main variables in the algorithm in Figure 1. In the following discussion, let us assume for simplicity that the learner has perfect knowledge of Ω, i.e., assume for now that $\widehat{\Omega} = \Omega$. Then, according to Section 3, we can view the problem as linear bandits over $\Omega \subset \mathbb{R}^d$. But when the linear bandit algorithm tells us to sample a point $y_t \in \Omega$, what corresponding policy π_t should we use in order to guarantee that $\Psi(\pi_t) = \mathbb{E}_{\mathcal{A} \sim \mathcal{D}}[\pi_t(\mathcal{A})] = y_t$? This might be theoretically achievable if we knew $\mathcal{D}$ exactly, but even then the support of $\mathcal{D}$ might be too large to make this computationally tractable. To resolve this, we make the following key observation:

Lemma 2. *Fix any action-set distribution $\mathcal{D}$ and its corresponding linear bandit feasible set $\Omega = \{\mathbb{E}_{\mathcal{A} \sim \mathcal{D}}[\pi(\mathcal{A})] :$ all possible policies $\pi\}$. For any vertex ψ of Ω, we can find a linear classifier policy π_ϕ such that $\mathbb{E}_{\mathcal{A} \sim \mathcal{D}}[\pi_\phi(\mathcal{A})] = \psi$. In fact, this holds for any interior point $\phi \in -\mathcal{N}(\Omega, \psi)$.*

Proof of Lemma 2. Assume ψ is a vertex of Ω and ϕ is an interior point of $-\mathcal{N}(\psi, \Omega)$. This implies that for any $\psi' \in \Omega$, $\psi' \neq \psi$, we have $\langle \psi' - \psi, \phi \rangle > 0$.

Let $\tilde{\psi} = \mathbb{E}_{\mathcal{A}\sim\mathcal{D}}\left[\pi_\phi(\mathcal{A})\right]$, and let π be a policy that ψ corresponds to under distribution $\mathcal{D}$ (i.e., $\mathbb{E}_{\mathcal{A}\sim\mathcal{D}}[\pi(\mathcal{A})] = \psi$). Then

$$\begin{aligned}\left\langle \tilde{\psi}, \phi \right\rangle &= \mathbb{E}_{\mathcal{A}\sim\mathcal{D}}\left[\langle \pi_\phi(\mathcal{A}), \phi\rangle\right] \\ &= \mathbb{E}_{\mathcal{A}\sim\mathcal{D}}\left[\min_{a\in\mathcal{A}}\left(a^\top \phi\right)\right] \\ &\le \mathbb{E}_{\mathcal{A}\sim\mathcal{D}}\left[\pi(\mathcal{A})^\top \phi\right] = \langle \psi, \phi\rangle .\end{aligned}$$

Thus, it must be that $\tilde{\psi} = \psi$. This finishes the proof. □

Based on Lemma 2, we have the following strategy to *execute* y_t. First, decompose y_t as a convex combination of the vertices of Ω, i.e., decompose y_t as $\sum_i \alpha_i \psi_i$ where $\psi_1, \psi_2, \ldots$ are vertices of Ω and $(\alpha_1, \alpha_2 \ldots)$ is a distribution over them. Then find $\phi_i \in -\mathcal{N}(\Omega, \psi_i)$ for all i. Finally, let π_t be the policy that mix $\pi_{\phi_1}, \pi_{\phi_2}, \ldots$ with weights $(\alpha_i, \alpha_2, \cdots)$. This way, we have by Lemma 2,

$$\mathbb{E}_{\mathcal{A}\sim\mathcal{D}}\left[\pi_t(\mathcal{A})\right] = \mathbb{E}_{\mathcal{A}\sim\mathcal{D}}\left[\sum_i \alpha_i \pi_{\phi_i}(\mathcal{A})\right] = \sum_i \alpha_i \mathbb{E}_{\mathcal{A}\sim\mathcal{D}}\left[\pi_{\phi_i}(\mathcal{A})\right] = \sum_i \alpha_i \psi_i = y_t.$$

This allows us to execute a policy π_t such that $\mathbb{E}_{\mathcal{A}\sim\mathcal{D}}[\pi_t(\mathcal{A})] = y_t$ without explicit knowledge of $\mathcal{D} \in \Delta(2^{\mathbb{R}^d})$ but only knowledge of $\Omega \subset \mathbb{R}^d$. Notice that we do not have perfect knowledge of Ω, but only the estimated feasible set $\widehat{\Omega}$. However, exactly the same approach can be applied to $\widehat{\Omega}$, resulting in the design of Algorithm 1.

3.3 Bounding the error due to the discrepancy between $\widehat{\Omega}$ and Ω

From Section 3.2, we know how to *execute* the linear contextual bandit algorithm by leveraging a linear bandit procedure in $\widehat{\Omega}$. However, there are errors due to the discrepancy between $\widehat{\Omega}$ and Ω which contribute to the final regret.

First, the loss estimator we constructed will be *biased*. Suppose that $\widetilde{\mathcal{D}}$ is the empirical distribution based on N action sets that are drawn independently from $\mathcal{D}$, and $\widehat{\Omega}$ is constructed from them based on (3). When we sample a point $y_t \in \widehat{\Omega}$ and execute the corresponding policy π_t such that $y_t = \mathbb{E}_{\mathcal{A}\sim\widetilde{\mathcal{D}}}[\pi_t(\mathcal{A})]$, the expected loss feedback is actually $\mathbb{E}_{\mathcal{A}\sim\mathcal{D}}[\langle \pi_t(\mathcal{A}), \theta_t\rangle] \neq \langle y_t, \theta_t\rangle$. This means that from the viewpoint of the linear bandit problem on $\widehat{\Omega}$, the feedback is *misspecified*. This motivates us to develop a misspecification robust linear bandit algorithm (Algorithm 2) which allows the feedback to not fully follow the standard linear bandit protocol. We elaborate more about this in Section 3.4.

The second source of error comes from difference between the action sets $\widehat{\Omega}$ and Ω. With the misspecification-robust linear bandit algorithm on $\widehat{\Omega}$, the learner has good regret bound on $\widehat{\Omega}$. However, the real regret that we care about is on Ω. This requires us to bound the difference between $\mathbb{E}_{\mathcal{A}\sim\widetilde{\mathcal{D}}}\left[\langle \pi_t(\mathcal{A}), \theta_t\rangle\right]$ and $\mathbb{E}_{\mathcal{A}\sim\mathcal{D}}\left[\langle \pi_t(\mathcal{A}), \theta_t\rangle\right]$.

Both errors discussed above are in fact related to the difference between $\mathbb{E}_{\mathcal{A}\sim\widetilde{\mathcal{D}}}\left[\langle \pi(\mathcal{A}), \theta_t\rangle\right]$ and $\mathbb{E}_{\mathcal{A}\sim\mathcal{D}}\left[\langle \pi(\mathcal{A}), \theta_t\rangle\right]$. This can be bounded using Lemma 1, where we establish uniform convergence over the set of all linear policies. According to Lemma 1, we have with probability at least $1-\delta$, for all linear policies π,

$$\left|\mathbb{E}_{\mathcal{A}\sim\widetilde{\mathcal{D}}}\left[\langle \pi(\mathcal{A}), \theta_t\rangle\right] - \mathbb{E}_{\mathcal{A}\sim\mathcal{D}}\left[\langle \pi(\mathcal{A}), \theta_t\rangle\right]\right| \lesssim \sqrt{\frac{d\log(NK/\delta)}{N}}. \tag{4}$$

This allows us to bound the two sources of errors mentioned above.

3.4 Robust Linear Contextual Bandits

As discussed in Section 3.3, we would like to develop a linear bandit algorithm that tolerates misspecification. Although there has been a rich literature about this, most of them are for the stochastic linear bandit problem and do not apply here. For adversarial linear bandits that is misspecification robust,

Algorithm 2: Misspecification-Robust Continuous Exponential Weights

Parameters: $\gamma = 10\log(10dT)$, $\beta = T^{-4}$.

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

Define

$$q_t(y) = \frac{\exp\left(-\eta \sum_{\tau<t} \left\langle y, \widehat{\theta}_\tau - b_\tau \right\rangle\right)}{\int_{\mathcal{A}} \exp\left(-\eta \sum_{\tau<t} \left\langle z, \widehat{\theta}_\tau - b_\tau \right\rangle\right) \mathrm{d}z}, \quad x_t = \mathbb{E}_{y\sim q_t}[y], \quad \Sigma_t = \mathbb{E}_{y\sim q_t}[(y - x_t)(y - x_t)^\top].$$

Define

$$\widehat{q}_t(y) = \frac{q_t(y)\mathbb{I}\left\{\|y - x_t\|^2_{\Sigma_t^{-1}} \leq d\gamma^2\right\}}{\int_{\mathcal{A}} q_t(z)\mathbb{I}\left\{\|z - x_t\|^2_{\Sigma_t^{-1}} \leq d\gamma^2\right\} \mathrm{d}z}, \quad \widehat{\Sigma}_t = \mathbb{E}_{y\sim \widehat{q}_t}[(y - x_t)(y - x_t)^\top].$$

Sample $y_t \sim \widehat{q}_t$ and receive loss $\ell_t \in [0, 1]$.

Define $\widehat{\theta}_t = (\beta I + \widehat{\Sigma}_t)^{-1}(y_t - x_t)\ell_t$ and

$$b_t = 8\eta\left(\epsilon + \frac{1}{T^2}\right)\sum_{\tau<t}(\widehat{\theta}_\tau - b_\tau).$$

we are only aware of the work by Neu and Olkhovskaya (2020) and Liu et al. (2024). However, the bound in Neu and Olkhovskaya (2020) has a worse $T^{2/3}$ regret, while the algorithms of Liu et al. (2024) are either computationally inefficient or highly sub-optimal. Fortunately, our problem is slightly easier than that studied by Liu et al. (2024), as our learner has knowledge of the amount of misspecification ϵ. This allows us to design the computationally efficient Algorithm 2 with an improved dependence on the amount of misspecification. The guarantee of Algorithm 2 is given in the following theorem.

Theorem 2. *Algorithm 2 is a $\sqrt{d}$-misspecification-robust linear bandit algorithm defined in Definition 1.*

We remark that while there exist algorithms that are 1-misspecification-robust (Liu et al., 2024), their run time scales at least with the number of actions. Algorithm 2 achieves α-robustness with the smallest α we are aware of among algorithms that meet the computational requirement.

In fact, Algorithm 2 achieves an even more favorable small-loss regret bound, which can be leveraged to obtain a small-loss bound for linear contextual bandits when the simulator is available. We discuss this in Section 5.

3.5 Combining Everything and Using the Doubling Trick

Combining everything above, we are able to establish the regret bound for the linear contextual bandit problem. The proof of the following theorem is in Appendix C.2.

Theorem 3. *Algorithm 1 with* ALG *instantiated as a α-misspecification-robust linear bandit algorithm ensures*

$$\mathbb{E}\left[\sum_{t=1}^T \ell_t(a_t)\right] \leq \min_{\pi\in\Pi} \mathbb{E}\left[\sum_{t=1}^T \ell_t(\pi(\mathcal{A}_t))\right] + \tilde{O}\left(d\sqrt{T} + \alpha T d\sqrt{\frac{\log(NKT)}{N}}\right).$$

Corollary 1 (Restatement of Theorem 1)**.** *Given access to a α-misspecification-robust adversarial linear bandit algorithm* ALG*, Algorithm 1 with doubling trick achieves $\min_{\pi\in\Pi} \mathrm{Reg}_T(\pi) \leq \widetilde{O}(d\sqrt{T} + \alpha d^{1.5}\sqrt{T})$ in adversarial linear contextual bandits without access to simulators.*

Proof. We will use the doubling trick and restart Algorithm 1 at times $2, 4, 8, 16, \ldots$, each time using the contexts received so far to estimate $\widehat{\Omega}$. Thus, in the k-th epoch, $\widehat{\Omega}$ is constructed by $N = \Theta(2^k)$

contexts, allowing us to bound the regret in epoch k as

$$\tilde{O}\left(d\sqrt{2^k} + \alpha 2^k d\sqrt{\frac{\log(NKT)}{2^k}}\right) = \tilde{O}\left(d\sqrt{2^k} + \alpha d\sqrt{2^k \log(NKT)}\right)$$

using Theorem 3. Summing the regret over all epochs allows us to bound

$$\min_{\pi\in\Pi} \mathrm{Reg}_T(\pi) \leq \tilde{O}\left(\sum_{k=1}^{\log_2 T} d\sqrt{2^k} + \sum_{k=1}^{\log_2 T} \alpha d\sqrt{2^k \log(NKT)}\right) = \tilde{O}\left(d\sqrt{T} + \alpha d^{1.5}\sqrt{T}\right)$$

where in the last inequality we assume $K \leq T^d$ without loss of generality. □

By instantiating ALG as Algorithm 2 (which is an $\sqrt{d}$-misspecification-robust algorithm by Theorem 2) and invoking Corollary 1, we get the final regret bound as $\tilde{O}(d^2\sqrt{T})$.

4 Computational Complexity

Algorithm 1 contains two steps, lines 5 and 6, for which it is not obvious whether they can be implemented efficiently. We will now describe an approach to implement both steps in poly(d, C, N) time, provided that $\mathrm{conv}(\mathcal{A})$ is a polytope that can be described by at most C linear constraints for $\mathcal{D}$-almost all $\mathcal{A}$.

We start with the computational complexity of a separation oracle for $\widehat{\Omega}$. We assume, without loss of generality, that no two sets $\mathcal{A}_i$ and $\mathcal{A}_j$ in the construction of $\widehat{\Omega}$ are equal; otherwise we can replace them by a single set $2\mathcal{A}_i$, which will only decrease N. Thus, $\widehat{\Omega}$ is the Minkowski sum of N convex polytopes that, by assumption, can all be described by at most C constraints. For any $x \in \mathbb{R}^d$, let

$$g(\phi) = \phi^\top x - \max_{x'\in\hat{\Omega}} \phi^\top x' = \phi^\top x - \frac{1}{N}\sum_{i=1}^{N} \max_{x_i\in\mathcal{A}_i} \phi^\top x_i.$$

If $g(\phi) > 0$ for some ϕ, then ϕ gives a hyperplane that separates x from $\hat{\Omega}$; and if $g(\phi) \leq 0$ for all ϕ, then $x \in \hat{\Omega}$. Since g is concave, and we can solve every subproblem $\max_{x_i\in\mathcal{A}_i} \phi^\top x_i$ in poly(d, C) time, we can maximize g in poly(d, N, C) time to obtain our separation oracle.

We then proceed with line 5 of Algorithm 1. To implement it, we need to be able to take any point $y_t \in \widehat{\Omega}$ and find a distribution α_t supported on the vertices of $\widehat{\Omega}$ that we can sample from efficiently. By Carathéodory's theorem, any $y_t \in \widehat{\Omega}$ can be represented as a convex combination of at most $d+1$ vertices of $\widehat{\Omega}$: $y_t = \sum_{l=1}^{k} \lambda_l v_l$ for $k \leq d+1$ where all v_l are vertices of $\widehat{\Omega}$, and $\lambda_1, \ldots, \lambda_k \geq 0$ with $\sum_l \lambda_l = 1$ can be interpreted as the probabilities of selecting the vertices. This is a categorical distribution on at most $d+1$ points, from which we can sample in $O(d)$ time (assuming we have access to an oracle that provides samples from the uniform distribution on $[0, 1]$). Thus the main challenge is to compute the vertices v_l and the probabilities λ_l. By Corollary 14.1g) of Schrijver (1986) (restated in Lemma 3), there exists an algorithm that can do both in time poly(d, h) for any set $\widehat{\Omega}$ given access to a separation oracle that runs in time $O(h)$. As discussed above, we have h of order poly(d, C, N), from which it follows that we can also find α_t and sample from it in poly(d, C, N) time.

After the sampling procedure has chosen a particular vertex $\psi_t \in \{v_1, \ldots, v_k\}$, it remains to find an interior point of the corresponding normal cone $-\mathcal{N}(\widehat{\Omega}, \psi_t)$ to implement line 6. That is, we need to find any ϕ such that $\langle \phi, \psi_t - x\rangle < 0$ for all $x \in \widehat{\Omega}$. Strengthen this requirement slightly to $\langle \phi, \psi_t - x\rangle \leq -\epsilon$ for a small $\epsilon > 0$. By Corollary 14.1a) of Schrijver (1986) (restated in Lemma 4), since we have an efficient separation oracle for $\widehat{\Omega}$, we can compute $x^\star \in \arg\max_{x\in\widehat{\Omega}} \langle \phi, \psi_t - x\rangle$ for a given candidate ϕ in polynomial time, and thus we have a witness $x^\star$ to the constraint violation by ϕ. Furthermore, we have a sub-gradient $\psi_t - x^\star$ of the worst-case constraint at ϕ, which generates a separation oracle for the feasible region via the inner product with any other candidate ϕ'. Incorporating this information into e.g. the ellipsoid method (see e.g. Grötschel et al. (1988), Chapter 3) provides an implementation for line 6 that runs in poly(d, C, N) time.

5 Small-Loss Bound with Access to Simulator

The sub-optimal rate $d^2\sqrt{T}$ we obtained in Section 3 comes from the misspecification term $\alpha\sqrt{d}\epsilon T$ in the regret bound of robust linear bandits (Definition 1). While it is unclear how to further improve α or ϵ, we demonstrate the power of our reduction by further assuming access to simulator: it not only allows us to recover the minimax optimal regret $d\sqrt{T}$ but also allows us to obtain a first-order bound $d\sqrt{L^\star}$, where $L^\star$ is the cumulative loss of the best policy.

By the black-box nature of our reduction, what we additionally need is just a misspecification-robust linear bandit algorithm with *small-loss* regret bound guarantee, formally defined as follows:

Definition 2 (α-misspecification-robust adversarial linear bandit algorithm with small-loss bounds)**.** *A misspecification-robust linear bandit algorithm with small-loss bounds over action set $\widehat{\Omega} \subset \mathbb{R}^d$ has the following property: with a given $\epsilon > 0$ and the guarantee that every time the learner chooses $y_t \in \widehat{\Omega}$, the loss received $\ell_t(y_t) \in [-1, 1]$ satisfies*

$$|\,\mathbb{E}\left[\ell_t(y_t)\right] - \langle y_t, \theta_t\rangle\,| \le \epsilon,$$

the algorithm ensures

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle y_t, \theta_t\rangle\right] \le \min_{y\in\widehat{\Omega}}\sum_{t=1}^{T}\langle y, \theta_t\rangle + \widetilde{O}\left(d\sqrt{\sum_{t=1}^{T}\langle y, \theta_t\rangle} + \alpha\sqrt{d}\epsilon T\right). \tag{5}$$

The next theorem shows that Algorithm 2 satisfies Definition 2 with $\alpha = \sqrt{d}$.

Theorem 4. *Algorithm 2 is a $\sqrt{d}$-misspecification-robust linear bandit algorithm with small-loss bound defined in Definition 2.*

With access to a misspecification-robust linear bandit algorithm with small-loss bounds, we have

Theorem 5. *Given access to simulator that can generate free contexts, and access to an α-misspecification-robust adversarial linear bandit algorithm with small-loss regret bound guarantee, we can achieve $\min_{\pi\in\Pi}\mathrm{Reg}_T(\pi) \le \widetilde{O}(d\sqrt{L^\star})$ in adversarial linear contextual bandits, where*

$$L^\star = \min_{\pi\in\Pi}\mathbb{E}\left[\sum_{t=1}^{T}\langle \pi(\mathcal{A}_t), \theta_t\rangle\right]$$

is the expected total loss of the best policy. This is achieved with $O(\alpha^2 d^2 T^2)$ calls to the simulator.

The proof of Theorem 5 is very similar to Theorem 3, except that now, with access to the simulator, we are able to make N in Theorem 3 large enough that the second term in Theorem 3 is negligible. We provide the omitted proofs in Appendix D.

6 Conclusion and Open Questions

We have provided a general framework that reduces adversarial linear contextual bandits to misspecification-robust linear bandits in a black-box manner. It achieves $\widetilde{O}(d^2\sqrt{T})$ regret without a simulator, and is the first algorithm we know of that handles combinatorial bandits with stochastic action sets and adversarial losses efficiently. The requirement of misspecification robustness stems from our need to use an approximate feasible set $\widehat{\Omega}$ because we do not have direct access to the exact feasible set Ω, which depends on the action set distribution $\mathcal{D}$.

Two open questions are left by our work: First, can the computation cost be improved further to match that of Neu and Valko (2014), who only require a linear optimization oracle for each action set, and do not require a polynomial number of constraints? Second, can the regret be further improved to the near-optimal $\widetilde{O}(d\sqrt{T})$ bound with a polynomial-time algorithm?

References

Bubeck, S. and Eldan, R. (2015). The entropic barrier: a simple and optimal universal self-concordant barrier. In *Conference on Learning Theory*, pages 279–279. PMLR.

Dai, Y., Luo, H., Wei, C.-Y., and Zimmert, J. (2023). Refined regret for adversarial MDPs with linear function approximation. In *International Conference on Machine Learning*, pages 6726–6759. PMLR.

Grötschel, M., Lovász, L., and Schrijver, A. (1988). *Geometric Algorithms and Combinatorial Optimization*, volume 2 of *Algorithms and Combinatorics*. Springer.

Hanna, O. A., Yang, L., and Fragouli, C. (2023). Contexts can be cheap: Solving stochastic contextual bandits with linear bandit algorithms. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1791–1821. PMLR.

Ito, S., Hirahara, S., Soma, T., and Yoshida, Y. (2020). Tight first-and second-order regret bounds for adversarial linear bandits. *Advances in Neural Information Processing Systems*, 33:2028–2038.

Lee, C.-W., Luo, H., Wei, C.-Y., and Zhang, M. (2020). Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. *Advances in neural information processing systems*, 33:15522–15533.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670.

Liu, H., Tajdini, A., Wagenmaker, A., and Wei, C.-Y. (2024). Corruption-robust linear bandits: Minimax optimality and gap-dependent misspecification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Liu, H., Wei, C.-Y., and Zimmert, J. (2023). Bypassing the simulator: Near-optimal adversarial linear contextual bandits. *Advances in Neural Information Processing Systems*, 36.

Luo, H., Wei, C.-Y., and Lee, C.-W. (2021). Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34:22931–22942.

Natarajan, B. K. (1989). On learning sets and functions. *Machine Learning*, 4:67–97.

Neu, G. and Olkhovskaya, J. (2020). Efficient and robust algorithms for adversarial linear contextual bandits. In Abernethy, J. and Agarwal, S., editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3049–3068. PMLR.

Neu, G. and Valko, M. (2014). Online combinatorial optimization with stochastic decision sets and adversarial losses. *Advances in Neural Information Processing Systems*, 27.

Olkhovskaya, J., Mayo, J., van Erven, T., Neu, G., and Wei, C.-Y. (2023). First-and second-order bounds for adversarial linear contextual bandits. *Advances in Neural Information Processing Systems*, 36.

Schrijver, A. (1986). *Theory of Linear and Integer programming*. Wiley-Interscience.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Sherman, U., Koren, T., and Mansour, Y. (2023). Improved regret for efficient online reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 31117–31150. PMLR.

Zimmert, J. and Lattimore, T. (2022). Return of the bias: Almost minimax optimal high probability bounds for adversarial linear bandits. In *Conference on Learning Theory*, pages 3285–3312. PMLR.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims made in the abstract and introduction either have formal proofs.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations and trade-offs for our results compared to the state-of-the-art are provided in Table 1 and the introduction. Computational efficiency is discussed in Section 4.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: all theorems are backed up by formal proofs, either in the main paper or in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: This is a theory paper without experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The paper does not include experiments requiring data or code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This is theoretical work, which did not require human subjects or participants, and involved no data. It contributes to the fundamental understanding of contextual bandits, without any direct foreseeable negative societal impact. Mitigation measures were therefore not applicable.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This is theoretical work, which contributes to the fundamental understanding of contextual bandits. The long term societal impact is too unpredictable to provide a meaningful discussion.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The method development in this research does not involve LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Lemmas for Linear Programming

Lemma 3 (Corollaries 14.1f and 14.1g of Schrijver (1986))**.** *Suppose $P \subset \mathbb{R}^d$ is a bounded polytope defined by rational inequalities of size at most α, and for which there exists a separation oracle* SEP*. There exists an algorithm that solves the following problem: for any rational $x \in P$, find vertices $x_0, x_1, \ldots, x_d \in P$ and $\lambda_0, \lambda_1, \ldots, \lambda_d \geq 0$ such that $x = \sum_{i=0}^d \lambda_i x_i$ and $\sum_{i=0}^d \lambda_i = 1$, in time polynomially bounded by d, α, the running time of* SEP*, and the size of the input x.*

Lemma 4 (Corollary 14.1a of Schrijver (1986))**.** *Suppose $P \subset \mathbb{R}^d$ is a polytope defined by rational inequalities of size at most α, and for which there exists a separation oracle* SEP*. Then there exists an algorithm that solves the linear optimization problem arg $\max_{x \in P} \beta^\top x$ for any rational vector β in time polynomially bounded by d, α, the running time of* SEP*, and the size of the input β.*

B Proof of Lemma 1

Lemma 1 (Uniform Convergence)**.** *Consider any loss vector $\theta \in \mathbb{R}^d$, and suppose that $|\mathcal{A}| \leq K$ and $\max_{a \in \mathcal{A}} |\langle a, \theta \rangle| \leq b$ almost surely. Then, for any $\delta \in (0, 1]$, uniformly over all linear classifier policies π_ϕ, the difference in performance of π_ϕ on the sample $\mathcal{A}_1, \ldots, \mathcal{A}_N$ and its expected performance is at most*

$$\sup_{\phi \in \mathbb{R}^d} \left| \left\langle \Psi(\pi_\phi), \theta \right\rangle - \left\langle \hat{\Psi}(\pi_\phi), \theta \right\rangle \right| \leq 2b \sqrt{\frac{2d \ln(NK^2)}{N}} + b \sqrt{\frac{2 \ln(4/\delta)}{N}}$$

with probability at least $1 - \delta$.

Proof. We will first reduce the result to a statement about uniform convergence for linear multiclass classifiers with K classes and an unusual loss function. To this end, note first that we can assume without loss of generality that $\mathcal{A} = \{V_1, \ldots, V_K\}$, where the $V_1, \ldots, V_K$ are random vectors in $\mathbb{R}^d$. (If $\mathcal{A}$ has fewer than K elements, then consider it as a multiset and add repetitions of one of its elements.) We can assume all randomness in $\mathcal{A}$ is determined by an underlying random variable Z and that $V_y = -g(Z, y)$ for each 'class' $y \in [K]$, where g is a class-sensitive feature map in the sense of Shalev-Shwartz and Ben-David (2014, Section 17.2). Defining the linear multiclass classifier

$$c_\phi(Z) = \operatorname*{arg\,max}_{y \in [K]} \langle g(Z, y), \phi \rangle,$$

we then obtain the correspondence

$$\pi_\phi(\mathcal{A}) = V_{c_\phi(Z)}.$$

For any multiclass classifier c, let $f(Z, c) = \langle V_{c(Z)}, \theta \rangle$ be our 'loss function'. Then

$$\left\langle \Psi(\pi_\phi), \theta \right\rangle = \mathop{\mathbb{E}}_Z [f(Z, c_\phi)], \qquad \left\langle \hat{\Psi}(\pi_\phi), \theta \right\rangle = \frac{1}{N} \sum_{i=1}^N f(Z_i, c_\phi).$$

We therefore need to show the following uniform convergence result, with probability at least $1 - \delta$,

$$\sup_{c \in \mathcal{C}} \left| \mathbb{E}[f(Z, c)] - \frac{1}{N} \sum_{i=1}^N f(Z_i, c) \right| \leq 2b \sqrt{\frac{2d \ln(NK^2)}{N}} + b \sqrt{\frac{2 \ln(4/\delta)}{N}} \tag{6}$$

for the class of linear multiclass classifiers

$$\mathcal{C} = \left\{ c_\phi \mid \phi \in \mathbb{R}^d \right\},$$

with loss function f. In order to establish (6), let $S = (Z_1, \ldots, Z_N)$, and consider the empirical Rademacher complexity

$$\mathrm{Rad}(\mathcal{C}, S) = \frac{1}{N} \mathop{\mathbb{E}}_{\sigma_1, \ldots, \sigma_N} \left[\sup_{c \in \mathcal{C}} \sum_{i=1}^N \sigma_i f(Z_i, c) \right],$$

where the σ_i are independent Rademacher random variables with $\Pr(\sigma_i = -1) = \Pr(\sigma_i = +1) = 1/2$. Since $|f(Z, c)| \leq b$ for $c \in \mathcal{C}$ by assumption, standard concentration bounds in terms of Rademacher complexity imply that

$$\sup_{c \in \mathcal{C}} \Big| \mathop{\mathbb{E}}_{Z}[f(Z, c)] - \frac{1}{N} \sum_{i=1}^{N} f(Z_i, c) \Big| \leq 2 \mathop{\mathbb{E}}_{S'}[\mathrm{Rad}(\mathcal{C}, S')] + b\sqrt{\frac{2\ln(4/\delta)}{N}} \tag{7}$$

with probability at least $1 - \delta$. (This follows, for instance, by observing that

$$\begin{aligned} &\sup_{c \in \mathcal{C}} | \mathop{\mathbb{E}}_{Z}[f(Z, c)] - \frac{1}{N} \sum_{i=1}^{N} f(Z_i, c)| \\ &\qquad = \max \Big\{ \sup_{c \in \mathcal{C}} \mathop{\mathbb{E}}_{Z}[f(Z, c)] - \frac{1}{N} \sum_{i=1}^{N} f(Z_i, c), \sup_{c \in \mathcal{C}} \mathop{\mathbb{E}}_{Z}[-f(Z, c)] - \frac{1}{N} \sum_{i=1}^{N} (-f(Z_i, c)) \Big\} \end{aligned}$$

and applying part 1 of Theorem 26.5 of Shalev-Shwartz and Ben-David (2014) separately to f and $-f$ to control both parts in the maximum separately using a union bound; then noting that f and $-f$ have the same Rademacher complexity.)

We proceed to bound the Rademacher complexity on the right-hand side of (7). First, let $\mathcal{C}_S = \{(c(Z_1), \ldots, c(Z_N)) \mid c \in \mathcal{C}\}$ denote the restriction of $\mathcal{C}$ to the sample S. Then

$$\mathrm{Rad}(\mathcal{C}, S) = \mathrm{Rad}(\mathcal{C}_S, S) \leq b\sqrt{\frac{2\ln|\mathcal{C}_S|}{N}}$$

by Massart's lemma (Shalev-Shwartz and Ben-David, 2014). As discussed by Shalev-Shwartz and Ben-David (2014, Chapter 29), one possible generalization of the Vapnik-Chervonenkis dimension to multiclass classification is Natarajan's dimension $\mathrm{NatDim}(\mathcal{C})$. Natarajan's lemma (Natarajan, 1989, p. 93), (Shalev-Shwartz and Ben-David, 2014, Lemma 29.4) shows that

$$|\mathcal{C}_S| \leq N^{\mathrm{NatDim}(\mathcal{C})} K^{2\,\mathrm{NatDim}(\mathcal{C})},$$

and, for linear multiclass classifiers, it is also known (Shalev-Shwartz and Ben-David, 2014, Theorem 29.7) that

$$\mathrm{NatDim}(\mathcal{C}) \leq d.$$

Putting all inequalities together, it follows that

$$\sup_{c \in \mathcal{C}} \Big| \mathop{\mathbb{E}}_{Z}[f(Z, c)] - \frac{1}{N} \sum_{i=1}^{N} f(Z_i, c) \Big| \leq 2b\sqrt{\frac{2d\ln(NK^2)}{N}} + b\sqrt{\frac{2\ln(4/\delta)}{N}}$$

with probability at least $1 - \delta$, as required. □

C Omitted Details in Section 3

C.1 Robust Linear Bandits

Lemma 5 (Lemma 14 of Zimmert and Lattimore (2022)). *Let F be a ν-self-concordant barrier for $\mathcal{A} \subset \mathbb{R}^d$ for some $\nu \geq 1$. Then for any $x, u \in \mathcal{A}$,*

$$\|x - u\|_{\nabla^2 F(x)} \leq \gamma' \langle x - u, \nabla F(x) \rangle + 6\gamma' \nu$$

where $\gamma' = \frac{8}{3\sqrt{3}} + \frac{7^{\frac{3}{2}}}{6\sqrt{3}\nu}$ ($\gamma' \in [1, 4]$ for $\nu \geq 1$).

It is known that the continuous exponential weight algorithm is equivalent to FTRL with entropic barrier as the regularizer together with a particular sampling scheme (Bubeck and Eldan, 2015; Zimmert and Lattimore, 2022). We summarize the equivalence in the following lemma, the details of which can be found in Zimmert and Lattimore (2022).

Lemma 6 (Facts from Bubeck and Eldan (2015), Zimmert and Lattimore (2022))**.** *Consider Algorithm 2. Let $x_t = \mathbb{E}_{y\sim q_t}[y]$ and let $F : \mathcal{A} \to \mathbb{R}$ be the entropic barrier on $\mathcal{A}$. Then we have*

$$x_t = \underset{y\in\widehat{\Omega}}{\mathit{argmin}} \left\langle y, \sum_{\tau<t}\left(\widehat{\theta}_\tau - b_\tau\right)\right\rangle + \frac{F(y)}{\eta}.$$

Furthermore,

$$\nabla F(x_t) = -\eta\sum_{\tau<t}(\widehat{\theta}_\tau - b_\tau) \quad \mathit{and} \quad \nabla^2 F(x_t) = \left(\mathbb{E}_{y\sim q_t}[(y-x_t)(y-x_t)^\top]\right)^{-1}.$$

Lemma 7 (Lemma 4, Ito et al. (2020))**.** *Let $q_t, \widehat{q}_t, \Sigma_t, \widehat{\Sigma}_t$ be as defined in Algorithm 2. Then for any $f(y) : \mathcal{A} \to [-1, 1]$,*

$$|\mathbb{E}_{y\sim q_t}[f(y)] - \mathbb{E}_{y\sim\widehat{q}_t}[f(y)]| \le 10d\exp(-\gamma) \le \frac{1}{d^5T^{10}}.$$

Furthermore,

$$\frac{3}{4}\Sigma_t \preceq \widehat{\Sigma}_t \preceq \frac{4}{3}\Sigma_t.$$

Proof of Theorem 2. First, we decompose the regret as the following:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \langle y_t - u, \theta_t\rangle\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^T \langle \mathbb{E}_{y\sim q_t}[y] - u, \theta_t\rangle\right] + O(1) \\
&= \underbrace{\mathbb{E}\left[\sum_{t=1}^T \left\langle x_t - u, \widehat{\theta}_t - b_t\right\rangle\right]}_{\textbf{FTRL}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T \left\langle x_t - u, \theta_t - \widehat{\theta}_t\right\rangle\right]}_{\textbf{Bias}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T \langle x_t - u, b_t\rangle\right]}_{\textbf{Bonus}} + O(1).
\end{aligned}$$

Bounding FTRL term The analysis for the **FTRL** term follows that in Ito et al. (2020). Specifically, directly following their Lemma 5, Lemma 6, and the analysis below Lemma 6, we have the following: as long as $\eta\|\widehat{\theta}_t - b_t\|_{\Sigma_t} \le 1$, we have

$$\begin{aligned}
\textbf{FTRL} &\le \frac{d\log T}{\eta} + \eta\mathbb{E}\left[\sum_{t=1}^T \|\widehat{\theta}_t - b_t\|^2_{\Sigma_t}\right] + O(1) \\
&\le \frac{d\log T}{\eta} + 2\eta\mathbb{E}\left[\sum_{t=1}^T \|\widehat{\theta}_t\|^2_{\Sigma_t} + \|b_t\|^2_{\nabla^{-2}F(x_t)}\right] + O(1). && \text{(by Lemma 6)}
\end{aligned}$$

For the two middle terms, we have

$$\begin{aligned}
\|\widehat{\theta}_t\|^2_{\Sigma_t} &\le (y_t - x_t)^\top \left(\beta I + \widehat{\Sigma}_t\right)^{-1} \Sigma_t \left(\beta I + \widehat{\Sigma}_t\right)^{-1} (y_t - x_t)\ell_t^2 \\
&\le 2(y_t - x_t)^\top\Sigma_t^{-1}(y_t - x_t)\ell_t^2 && \text{(by Lemma 7)} \\
&\le 2d\gamma^2\ell_t^2, && \text{(by the truncation in the algorithm)}
\end{aligned}$$

and

$$\|b_t\|^2_{\nabla F(x_t)} = \left(2\sqrt{d}\gamma\epsilon + \frac{1}{T^2}\right)^2 \|\nabla F(x_t)\|^2_{\nabla^{-2}F(x_t)} \le O\left(d^2\gamma^2\epsilon^2 + \frac{1}{T^2}\right),$$

where we used the fact that $F(\cdot)$ is a $O(d)$ self-concordant barrier and thus $\|\nabla F(x_t)\|^2_{\nabla^{-2}F(x_t)} \le O(d)$. Thus, since $\eta \le \frac{1}{d\gamma^2}$, we have

$$\textbf{FTRL} \le \frac{d\log T}{\eta} + O\left(\eta d\gamma^2\sum_{t=1}^T \ell_t^2 + \eta T d^2\gamma^2\epsilon^2\right).$$

Bounding Bias term Let $\mathbb{E}[\ell_t] = y_t^\top \theta_t + \epsilon_t(y_t)$, where $\epsilon_t(y)$ is the amount of misspecification when choosing y. By assumption, we have $|\epsilon_t(y)| \le \epsilon$ for any y.

$$\begin{aligned}
\mathbb{E}_t[\widehat{\theta}_t] &= \mathbb{E}_t\left[(\beta I + \widehat{\Sigma}_t)^{-1}(y_t - x_t)\left(y_t^\top \theta_t + \epsilon_t(y_t)\right)\right] \\
&= \mathbb{E}_t\left[(\beta I + \widehat{\Sigma}_t)^{-1}(y_t - x_t)\left((y_t - x_t)^\top \theta_t + \epsilon_t(y_t)\right)\right] + \mathbb{E}_t\left[(\beta I + \widehat{\Sigma}_t)^{-1}(y_t - x_t)x_t^\top \theta_t\right] \\
&= \theta_t - \beta(\beta I + \widehat{\Sigma}_t)^{-1}\theta_t + \mathbb{E}_t\left[(\beta I + \widehat{\Sigma}_t)^{-1}(y_t - x_t)\epsilon_t(y_t)\right] + (\beta I + \widehat{\Sigma}_t)^{-1}(\widehat{x}_t - x_t)x_t^\top \theta_t.
\end{aligned}$$

Using this we get

$$\begin{aligned}
&\left|\left\langle x_t - u, \theta_t - \mathbb{E}_t[\widehat{\theta}_t]\right\rangle\right| \\
&\le \beta\left|(x_t - u)^\top(\beta I + \widehat{\Sigma}_t)^{-1}\theta_t\right| + \underbrace{\left|(x_t - u)^\top \mathbb{E}_t\left[(\beta I + \widehat{\Sigma}_t)^{-1}(y_t - x_t)\epsilon_t(y_t)\right]\right|}_{(\star)} \\
&\quad + \left|(x_t - u)^\top(\beta I + \widehat{\Sigma}_t)^{-1}(\widehat{x}_t - x_t)x_t^\top \theta_t\right|. \qquad (8)
\end{aligned}$$

We handle $(\star)$ as follows:

$$\begin{aligned}
(\star) &= \mathbb{E}_t\left[\sqrt{(x_t - u)^\top(\beta I + \widehat{\Sigma}_t)^{-1}(y_t - x_t)(y_t - x_t)^\top(\beta I + \widehat{\Sigma}_t)^{-1}(x_t - u)\epsilon_t(y_t)^2}\right] \\
&\le \sqrt{(x_t - u)^\top(\beta I + \widehat{\Sigma}_t)^{-1}\mathbb{E}_t\left[\epsilon_t(y_t)^2(y_t - x_t)(y_t - x_t)^\top\right](\beta I + \widehat{\Sigma}_t)^{-1}(x_t - u)} \\
&\le \epsilon\sqrt{(x_t - u)^\top(\beta I + \widehat{\Sigma}_t)^{-1}\widehat{\Sigma}_t(\beta I + \widehat{\Sigma}_t)^{-1}(x_t - u)} \\
&\le \epsilon\|x_t - u\|_{(\beta I + \widehat{\Sigma}_t)^{-1}}.
\end{aligned}$$

Continuing from (8), we get

$$\begin{aligned}
&\left|\left\langle x_t - u, \theta_t - \mathbb{E}_t[\widehat{\theta}_t]\right\rangle\right| \\
&\le \beta\|x_t - u\|_{(\beta I + \widehat{\Sigma}_t)^{-1}}\|\theta_t\|_{(\beta I + \widehat{\Sigma}_t)^{-1}} + \epsilon\|x_t - u\|_{(\beta I + \widehat{\Sigma}_t)^{-1}} + \|x_t - u\|_{(\beta I + \widehat{\Sigma}_t)^{-1}}\|\widehat{x}_t - x_t\|_{(\beta I + \widehat{\Sigma}_t)^{-1}} \\
&\le \sqrt{d\beta}\|x_t - u\|_{(\beta I + \widehat{\Sigma}_t)^{-1}} + \epsilon\|x_t - u\|_{(\beta I + \widehat{\Sigma}_t)^{-1}} + \sqrt{\frac{1}{\beta}}\|x_t - u\|_{(\beta I + \widehat{\Sigma}_t)^{-1}}\|\widehat{x}_t - x_t\|_2,
\end{aligned}$$

where in the last inequality we use that $\|y_t - x_t\|_{(\beta I + \widehat{\Sigma}_t)^{-1}} \le 2\|y_t - x_t\|_{\Sigma_t^{-1}}$ by Lemma 7 and the assumption that $\|\theta_t\|_2 \le \sqrt{d}$.

By Lemma 7 we have $\|x_t - \widehat{x}_t\| = \|\mathbb{E}_{y \sim q_t}[y] - \mathbb{E}_{y \sim \widehat{q}_t}[y]\| \le \sqrt{d}d^{-5}T^{-10}$. By the choice of $\beta = d^{-2}T^{-4}$, we can bound the expectation of the sum of the last expression as

$$\textbf{Bias} \le \mathbb{E}\left[\sum_{t=1}^{T}\|x_t - u\|_{(\beta I + \widehat{\Sigma}_t)^{-1}}\left(\epsilon + \frac{1}{T^2}\right)\right].$$

Bounding Bonus term Notice that with the equivalence established in Lemma 6, our bonus term b_t can also be written as

$$b_t = -8\left(\epsilon + \frac{1}{T^2}\right)\nabla F(x_t),$$

where $F(\cdot)$ is the entropic barrier on $\mathcal{A}$. Using Lemma 5 and the fact that $F(\cdot)$ is an $O(d)$-self-concordant barrier, we can bound

$$\begin{aligned}
\langle x_t - u, b_t\rangle &\le -\frac{(8\epsilon + 8/T^2)}{4}\|x_t - u\|_{\nabla^2 F(x_t)} + 6\epsilon\nu \\
&= -\left(2\epsilon + \frac{2}{T^2}\right)\|x_t - u\|_{\nabla^2 F(x_t)} + O(d\epsilon) \\
&= -\left(2\epsilon + \frac{2}{T^2}\right)\|x_t - u\|_{\Sigma_t^{-1}} + O(d\epsilon) \\
&\le -\left(\epsilon + \frac{1}{T^2}\right)\|x_t - u\|_{\widehat{\Sigma}_t^{-1}} + O(d\epsilon). \qquad \text{(by Lemma 7)}
\end{aligned}$$

Thus,

$$\textbf{Bonus} \leq \mathbb{E}\left[-\sum_{t=1}^{T}\left(\epsilon+\frac{1}{T^2}\right)\|x_t-u\|_{\widehat{\Sigma}_t^{-1}}+O\left(dT\epsilon\right)\right].$$

Adding up all terms Adding up the three terms, we get

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle y_t-u,\theta_t\rangle\right] \leq \frac{d\log T}{\eta}+O\left(\eta d\gamma^2\mathbb{E}\left[\sum_{t=1}^{T}\ell_t^2\right]+dT\epsilon\right).$$

By the assumption $\ell_t \in [0,1]$ and the assumption $\left|\mathbb{E}\left[\ell_t\right]-y_t^\top\theta_t\right| \leq \epsilon$, we can further bound the right-hand side by

$$O\left(\frac{d\log T}{\eta}+\eta d\gamma^2\mathbb{E}\left[\sum_{t=1}^{T}\ell_t\right]+dT\epsilon\right) \leq O\left(\frac{d\log T}{\eta}+\eta d\gamma^2\mathbb{E}\left[\sum_{t=1}^{T}\langle y_t,\theta_t\rangle\right]+(d+\eta d\gamma^2)T\epsilon\right).$$

Then, by rearranging, we find that

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle y_t-u,\theta_t\rangle\right] \leq O\left(\frac{d\log T}{\eta}+\eta d\gamma^2\sum_{t=1}^{T}\langle u,\theta_t\rangle+dT\epsilon\right).$$

Choosing the optimal η, we further bound

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle y_t-u,\theta_t\rangle\right] = O\left(d\gamma\sqrt{(\log T)\sum_{t=1}^{T}\langle u,\theta_t\rangle}+dT\epsilon\right) = \tilde{O}\left(d\sqrt{T}+dT\epsilon\right). \tag{9}$$

By Definition 1, this is a $\sqrt{d}$-misspecification-robust algorithm.

□

C.2 Regret bound of LCB

Proof of Theorem 3. Let $\widehat{\Pi}$ be the set of linear policies created from the vertices of $\widehat{\Omega}$, and let Π be set of all linear policies.

The regret bound guaranteed by the α-misspecification-robust linear bandit problem is

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle y_t,\theta_t\rangle\right] \leq \min_{y\in\widehat{\Omega}}\sum_{t=1}^{T}y^\top\theta_t+\tilde{O}\left(d\sqrt{T}+\alpha\sqrt{d}\epsilon T\right) \tag{10}$$

for some $\alpha \geq 1$. By Lemma 2, we have

$$y_t = \mathbb{E}_{\mathcal{A}\sim\widetilde{\mathcal{D}}}\left[\pi_t(\mathcal{A})\right].$$

We further define

$$z_t = \mathbb{E}_{\mathcal{A}\sim\mathcal{D}}\left[\pi_t(\mathcal{A})\right], \qquad z^\star = \mathbb{E}_{\mathcal{A}\sim\mathcal{D}}\left[\pi^\star(\mathcal{A})\right], \qquad y^\star = \mathbb{E}_{\mathcal{A}\sim\widetilde{\mathcal{D}}}\left[\pi^\star(\mathcal{A})\right],$$

where $\pi^\star \in \Pi$ is the final regret comparator. Define

$$\epsilon = 4\sqrt{\frac{d\log(NKT/\delta)}{N}},$$

where N is the number of contexts used to construct $\widehat{\Omega}$. Then by Lemma 1 we have $|\langle y_t,\theta_t\rangle-\langle z_t,\theta_t\rangle| \leq \epsilon$ and $|\langle y^\star,\theta_t\rangle-\langle z^\star,\theta_t\rangle| \leq \epsilon$ with probability at least $1-\delta$ for all t. Choosing $\delta = \frac{1}{T^2}$, we obtain

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\langle z_t,\theta_t\rangle\right] &\le \mathbb{E}\left[\sum_{t=1}^{T}\langle y_t,\theta_t\rangle\right] + T\epsilon \\
&\le \sum_{t=1}^{T}\langle y^\star,\theta_t\rangle + \tilde{O}\left(d\sqrt{T} + \alpha\sqrt{d}T\epsilon\right) && \text{(by (10))} \\
&\le \sum_{t=1}^{T}\langle z^\star,\theta_t\rangle + \tilde{O}\left(d\sqrt{T} + \alpha\sqrt{d}T\epsilon\right) \\
&= \sum_{t=1}^{T}\langle z^\star,\theta_t\rangle + \tilde{O}\left(d\sqrt{T} + \alpha dT\sqrt{\frac{\log(NKT)}{N}}\right).
\end{aligned}
$$

This proves the theorem. □

D Omitted Details in Section 5

Proof of Theorem 4. This is by the same proof as Theorem 2, just noticing that it actually achieves a small-loss bound in (9).

□

Proof of Theorem 5. This is similar to the proof of Theorem 3, except that we replace (10) by

$$
\mathbb{E}\left[\sum_{t=1}^{T}\langle y_t,\theta_t\rangle\right] \le \min_{y\in\widehat{\Omega}}\sum_{t=1}^{T} y^\top\theta_t + \tilde{O}\left(d\sqrt{\sum_{t=1}^{T}\langle y,\theta_t\rangle} + \alpha\sqrt{d}\epsilon T\right). \tag{11}
$$

Following the same steps, we get

$$
\mathbb{E}\left[\sum_{t=1}^{T}\langle z_t,\theta_t\rangle\right] \le \sum_{t=1}^{T}\langle z^\star,\theta_t\rangle + \tilde{O}\left(d\sqrt{\sum_{t=1}^{T}\langle z^\star,\theta_t\rangle} + \alpha dT\sqrt{\frac{\log(NKT)}{N}}\right), \tag{12}
$$

where $z_t = \mathbb{E}_{\mathcal{A}\sim\mathcal{D}}\left[\pi_t(\mathcal{A})\right]$ and $z^\star = \mathbb{E}_{\mathcal{A}\sim\mathcal{D}}\left[\pi^\star(\mathcal{A})\right]$.

Since the learner is given simulator access, she can draw $N = \tilde{\Omega}(\alpha^2 d^2 T^2)$ samples and make the last term in (12) be a constant. This will give a final regret bound of $\tilde{O}\left(d\sqrt{\sum_{t=1}^{T}\langle z^\star,\theta_t\rangle}\right) = \tilde{O}(d\sqrt{L^\star})$. □