Bad Values but Good Behavior: Learning Highly Misspecified Bandits with Function Approximation

Debangshu Banerjee
HP AI Research Labs
Bengaluru, India

Aditya Gopalan
Indian Institute of Science
Bengaluru, India

Abstract

Function approximation with parametric, feature-based reward models is widely used to enable decision-making in bandits with large action spaces. While bandit learning is well understood in the case of little or no misspecification in the reward approximation, real-world applications can often involve significantly high model misspecification. We study whether optimal learning is still possible under arbitrary misspecification. We identify structural, instance-dependent conditions, determined jointly by the problem instance and model class, under which standard algorithms like ε-greedy and LinUCB achieve sublinear regret, despite an arbitrarily large misspecification error in the traditional sense. These results contrast sharply with worst-case analyses that predict linear regret, and show that a broad class of instances remains robust to model error. Our findings offer a theoretical explanation for the empirical success of approximate value-based methods in complex environments.

1 INDTRODUCTION

Sequential decision-making algorithms, such as those used for multi-armed bandits and Markov decision processes (MDPs), often rely on learning a parametric reward model over time. Common examples include linear and generalized linear bandits [Rusmevichientong and Tsitsiklis, 2010, Filippi et al., 2010], linear contextual bandits [Chu et al., 2011], and value function approximation methods in MDPs [Sutton and Barto, 2018]. These algorithms are typically designed and analyzed assuming *realizability*, i.e., the true reward function is perfectly representable by the model class [Abbasi-Yadkori et al., 2011, Filippi et al., 2010, Jin et al., 2020]. Yet in practice, models are often misspecified and only approximate the true rewards [Box, 1976, Foster et al., 2018]. Even so, optimal actions may still be identified correctly if their (possibly inaccurate) estimates remain dominant. For example, let us consider the standard YAHOO! LEARNING TO RANK dataset [Chapelle and Chang, 2011]. We randomly sample K-armed bandit instances along with a random d-subset of associated features and run the ε-greedy algorithm on them. The results, summarized in Table 1, show that, despite significant misspecification error, the percentage of instances exhibiting sublinear regret is non-trivial.

K	d	O($\sqrt{n}$) %	O(n) %	Avg. Misspec.
500	100	64.6	35.4	17.63
	200	73.2	26.8	16.19
1000	200	79.2	20.8	24.50
	500	80.0	20.0	22.06
5000	200	84.2	15.8	57.69
	500	89.6	10.4	55.44

Table 1: We randomly sample 5000 K-armed bandit instances along with a random d-subset of features from the YAHOO! LEARNING TO RANK dataset and run the ε-greedy algorithm. The table reports the percentage of instances that exhibit sublinear regret ($O(\sqrt{n})$), linear regret ($O(n)$), and the average misspecification error. Notably, even under considerable model misspecification, a substantial fraction of instances achieve sublinear regret, highlighting the prevalence of robustness in real-world settings.

This raises a fundamental question: Under what structural conditions can decision-making remain robust to misspecification? In this work, we address this question by characterizing how the structural interplay between the true rewards and the model class determines whether learning optimal actions remains possible, even when reward estimates can be highly inaccurate.

Contribution and Organization of the Paper. We identify a broad class of *robust instances* for misspecified bandits where the greedy action remains invariant under ordinary least squares projection. In Section 3, we show

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

that standard ε-greedy with $\varepsilon_t = 1/\sqrt{t}$ achieves $O(\sqrt{T})$ regret asymptotically on robust instances. We also show that standard LinUCB achieves poly-logarithmic instance-dependent regret on robust instances.[1] In Section 4, we prove the existence of robust regions and demonstrate that they occupy a non-trivial portion of all bandit instances. We extend our framework to the ε-best arm identification in misspecified bandits in Section 5.1 and show that uniform sampling is (ε, δ) PAC for robust instances. In Section 5.2, we show how to extend our framework for non-linear function approximation and design non-linear features that support arbitrary misspecification. In Section 6, we empirically demonstrate the existence of robust instances in synthetic as well as real-world data. We generalize our results to contextual bandits in Appendix A.

Related Work. Existing work on misspecified bandits can be broadly categorized into two directions. The *first* concerns negative, worst-case results: for linear bandits, there exist instances where (i) the l_∞ misspecification error from the model subspace Range($\mathbf{\Phi}$) is at most $\rho > 0$, yet (ii) the LinUCB algorithm incurs $\Omega(\rho T)$ regret over T rounds. Such worst-case results are well documented [Ghosh et al., 2017, Lattimore et al., 2020, Zanette et al., 2020]. In particular, [Lattimore et al., 2020] show that even modified versions of LinUCB cannot achieve sublinear regret under model misspecification, while [Zanette et al., 2020] extend this analysis to learning in Markov Decision Processes. The *second* direction seeks positive results: identifying conditions under which misspecified bandits admit sublinear regret. [Liu et al., 2023] analyze LinUCB under the assumption that sub-optimality gaps dominate the misspecification, showing sublinear regret when the misspecification is sufficiently small. [Zhang et al., 2023] extend this to the contextual setting, proposing a phased elimination algorithm akin to SupLinUCB [Chu et al., 2011], though requiring knowledge of sub-optimality gaps. Our work continues along this positive line, providing finer-grained structural conditions—beyond coarse measures like misspecification error—under which ε-greedy and LinUCB achieve robust learning without modifications. We present a more detailed literature survey in Appendix G.

Illustrative Examples. We construct the following two synthetic examples that highlight key observations that form the foundation for the rest of the paper.

Example 1: A misspecified 2-armed (non-contextual) linear bandit. Consider a 2-armed bandit with mean rewards $\boldsymbol{\mu} = (20, 3)$ in $\mathbb{R}^2$ (marked by $\times$ in Fig. 1). Suppose we model this with a 1-dimensional linear approximation, with features for the arms being $\varphi_1 = 3$ and $\varphi_2 = 1$ respectively. Then any modeled instance lies in the range space of $\mathbf{\Phi} = \left[\begin{smallmatrix} 3 \\ 1 \end{smallmatrix}\right]$. However, by the choice of our configuration, $\boldsymbol{\mu}$

[1]We define model error to be the distance between the true rewards and the feature subspace.

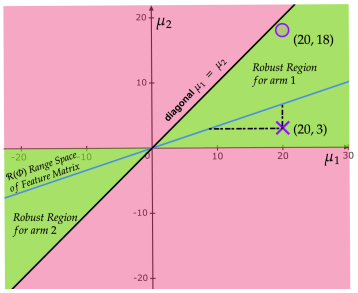

Figure 1: Example 1: A 2-armed, noiseless bandit with 1-dimensional linear approximation. Each point represents the true rewards of the arms. The blue line shows the model's range. Green and red regions mark robust and non-robust areas. Instances $(20, 3)$ and $(20, 18)$ yield no regret under greedy selection.

is a misspecified instance for this feature representation.[2]. Assuming noiseless observations of rewards μ_1 and μ_2, the ordinary least squares estimate after n_1 and n_2 arm pulls are arms 1 and 2 is

$$\widehat{\theta}_t = \frac{n_1\varphi_1\mu_1 + n_2\varphi_2\mu_2}{n_1\varphi_1^2 + n_2\varphi_2^2},$$

which is always a convex combination of $\mu_1/\varphi_1 = 6.7$ and $\mu_2/\varphi_2 = 3$. Consequently any $\mathbf{\Phi}\widehat{\theta}_t$ belongs to the set $\left\{\left(\begin{smallmatrix} 3\theta \\ \theta \end{smallmatrix}\right) : 3 \leqslant \theta \leqslant 6.7\right\}$. This interval is the hypotenuse of the triangle formed by the bases of the two canonical projections from $\boldsymbol{\mu}$ in Fig. 1. Since $\mathbf{\Phi}\widehat{\theta}_t$ remains below the $\mu_1 = \mu_2$ diagonal (black line), a greedy policy ($A_{t+1} = \operatorname{argmax}_i \varphi_i^\top \widehat{\theta}_t$) always selects arm 1, the optimal arm, yielding no regret. Note that, for the misspecified instance $\boldsymbol{\mu}$, the misspecification error (2.75) is much smaller than the reward gap (17). In contrast, for an instance like $\tilde{\boldsymbol{\mu}} = (20, 18)$ (marked by $\circ$), the misspecification error (8.5) exceeds the gap (2), yet a greedy policy still achieves no regret. Such robust instances, colored green in Fig. 1, fall outside the scope of prior misspecified bandit analyses (e.g., [Zhang et al., 2023]) and are addressed in this work.

Example 2: A 'highly misspecified' nonlinear bandit. An extreme form of robustness to misspecification is depicted in Fig. 2, where the model class is the tube

$$\{(x, y) \in \mathbb{R}^2 : |y - x| = \varepsilon\}$$

of effective dimension 1. Any instance within the green region is robust. For example, for $\boldsymbol{\mu} = (-10, 10)$, the model estimate for any sampling distribution α (fraction of arm 1) is

$$\begin{pmatrix} -10\alpha + (1-\alpha)(10-\varepsilon) \\ (-10+\varepsilon)\alpha + (1-\alpha)10 \end{pmatrix},$$

[2]The l_∞ misspecification error is 2.75..

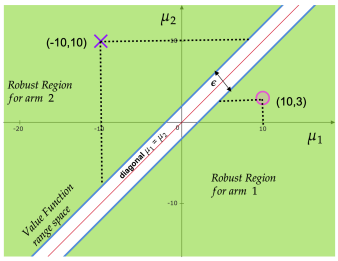

Figure 2: Example 2: A tube around the diagonal illustrating robustness under nonlinear misspecification. The image of the function, denoting the set of all realizable instances, is defined by $\{(x, y) \in \mathbb{R}^2 : |y - x| = \varepsilon\}$. Except for the diagonal itself (measure zero set), almost all instances are robust

a convex combination of the extreme points $(-10,\, -10 + \varepsilon)$ and $(10 - \varepsilon,\, 10)$. Greedy selection thus favors arm 2, ensuring no regret under ε-greedy sampling.

For a parameterized nonlinear example, consider the set defined by

$$\boldsymbol{x}(t) = \begin{cases} \boldsymbol{x}_0 + 400t\,\vec{u}, & t \in [0, \frac{1}{4}], \\ \boldsymbol{x}_0 + 100\vec{u} - (2t-1)\,2\varepsilon\,\vec{v}, & t \in [\frac{1}{4}, \frac{1}{2}], \\ \boldsymbol{x}_0 + 100\boldsymbol{u} - 2\varepsilon\vec{v} - 100(4t-2)\vec{u}, & t \in [\frac{1}{2}, \frac{3}{4}], \\ \boldsymbol{x}_0 - 2\varepsilon\vec{v} + 2\varepsilon(4t-3)\vec{v}, & t \in [\frac{3}{4}, 1]. \end{cases}$$

where $\vec{u} = (1, 1)$, $\vec{v} = (-1, 1)$, and $\boldsymbol{x}_0 = \binom{-50-\varepsilon}{-50+\varepsilon}$. This defines a thin rectangle aligned along $y = x$ (see Fig. 2). For $\boldsymbol{\mu} = (-10, 10)$, if arm 1 and 2 are sampled fractions α_1 and α_2, the model estimate is

$$\begin{pmatrix} -10\alpha_1 + (10 - 2\varepsilon)\alpha_2 \\ (-10 + 2\varepsilon)\alpha_1 + 10\alpha_2 \end{pmatrix},$$

which again lies within the convex hull of $(-10,\, -10+2\varepsilon)$ and $(10 - 2\varepsilon,\, 10)$, ensuring correct greedy decisions.

The remainder of the paper formalizes these ideas for bandits (contextual and non-contextual), under exploration strategies such as ε-greedy and optimism-based methods, and characterizes all reward instances admitting no-regret learning.

2 ROBUST INSTANCES FOR MULTI ARMED BANDITS

Problem Setting and Notation. We consider a K-armed bandit problem where each arm $i \in [K]$ has an associated feature vector $\varphi_i \in \mathbb{R}^d$, and an unknown true mean reward μ_i. Let $\boldsymbol{\mu} \in \mathbb{R}^K$ denote the vector of true rewards and $\boldsymbol{\Phi} \in \mathbb{R}^{K\times d}$ the feature matrix whose i-th row is $\varphi_i^\top$. We assume the number of arms exceeds the feature dimension ($K > d$), and that the feature vectors $\{\varphi_i\}_{i=1}^K$ span $\mathbb{R}^d$. The reward model is approximated using a parametric function class $\{f_\theta : \mathbb{R}^d \to \mathbb{R}\}_{\theta\in\boldsymbol{\Theta}}$, where the goal is to use features to predict rewards. At each round, we observe the empirical mean reward vector $\widehat{\boldsymbol{\mu}} \in \mathbb{R}^K$ based on past samples, and the sampling frequencies are captured by a distribution $\{\lambda_i\}_{i=1}^K \in \Delta_K$, the $(K-1)$-dimensional probability simplex. We write $\boldsymbol{\Lambda} = \mathrm{diag}(\{\lambda_i\}) \in \mathbb{R}^{K\times K}$ for the corresponding diagonal sampling matrix. For any vector x and positive semidefinite matrix A, we define the weighted norm as $\|x\|_A^2 = x^\top A x$. We denote the index of the optimal arm under any reward vector $\boldsymbol{\mu}$ by $\mathrm{OPT}(\boldsymbol{\mu}) \triangleq \mathrm{argmax}_{i\in[K]}\, \mu_i$, and define the maximum sub-optimality gap as $\Delta_{\max} \triangleq \max_{i\in[K]} \left(\mu_{\mathrm{OPT}(\boldsymbol{\mu})} - \mu_i\right)$.

Robust Instances. We aim to identify *sufficient* conditions under which standard bandit algorithms, such as ε-greedy and LinUCB, can achieve sublinear regret—even when the reward model is arbitrarily misspecified. That is, we allow for $\boldsymbol{\mu} \notin \{f_\theta(\varphi_i)\}_{i\in[K]}$ for any $\theta \in \mathbb{R}^d$. Our goal is to characterize the *structural conditions* which would allow misspecified instances to give rise to optimal policies. To this end, we introduce the notion of *robust instances*, which are instances for which greedy selection based on an *oracle* estimate consistently identifies the optimal arm. A crucial step in formalizing robustness is to understand the geometry of *greedy regions*—subsets of the reward space where a unique optimal arm exists. This arises from the observation that any K-armed bandit problem with means $\boldsymbol{\mu}$ can be embedded in $\mathbb{R}^K$. For each arm $k \in [K]$, we define the greedy region: $\mathcal{G}_k \triangleq \left\{\boldsymbol{\mu} \in \mathbb{R}^K : \mu_k > \mu_i \;\forall\, i \neq k\right\}$. The collection $\{\mathcal{G}_k\}_{k=1}^K$ forms a partition of $\mathbb{R}^K$, except for a measure-zero set of ties in the optimal arm. Each region identifies the set of reward vectors where arm k is uniquely optimal under greedy selection.

We assume a linear model class, where the true reward vector $\boldsymbol{\mu}$ is approximated via a feature matrix $\boldsymbol{\Phi} \in \mathbb{R}^{K\times d}$. Given a sampling distribution represented by a diagonal matrix $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_i)$, the least-squares estimate of the parameter vector is $\widehat{\theta} = (\boldsymbol{\Phi}^\top\boldsymbol{\Lambda}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Lambda}\widehat{\boldsymbol{\mu}}$, where $\widehat{\boldsymbol{\mu}}$ is the empirical mean reward vector. We assume that $\boldsymbol{\Phi}^\top\boldsymbol{\Lambda}\boldsymbol{\Phi}$ is invertible, which is guaranteed if the sampling distribution $\{\lambda_i\}$ assigns positive mass to a set of d linearly independent arms. A regularized alternative is discussed in Appendix E. This estimator equivalently solves the weighted least squares objective $\mathrm{argmin}_\theta \|\boldsymbol{\Phi}\theta - \widehat{\boldsymbol{\mu}}\|_{\boldsymbol{\Lambda}^{1/2}}$. We formalize the notion of this biased estimate as a function of the sampling distribution $\boldsymbol{\Lambda}$ and the observations $\boldsymbol{\mu}$ by defining the *oracle* estimate under a given sampling distribution as follows.

Definition 2.1 (*Oracle Estimate*)**.** Given the true reward mean vector $\boldsymbol{\mu} \in \mathbb{R}^K$ and sampling matrix $\boldsymbol{\Lambda}$, define the *oracle* estimate as $\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) \triangleq \mathrm{argmin}_\theta \|\boldsymbol{\Phi}\theta - \boldsymbol{\mu}\|_{\boldsymbol{\Lambda}^{1/2}}$.

Note that $\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ is the least squares solution with the true rewards mean vector $\boldsymbol{\mu}$; a reason behind using the terminology *oracle*. Using the above notation, the vanilla least squares estimate $\widehat{\theta}_t$ would be written as $\mathbf{P}^{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\mu}}_t)$ as it con-

tains the *noisy* estimates, $\boldsymbol{\mu}_t$. This notation explicitly highlights the dependence of the projection on both the observation vector and the sampling distribution.

If $\boldsymbol{\mu} \in \mathcal{G}_k$ and the projection $\boldsymbol{\Phi}\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ also lies in $\mathcal{G}_k$, then greedy selection based on $\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ recovers the optimal arm. This motivates the following definition of robustness.

Definition 2.2 (*Robust Parameter Region*)**.** The k-th robust parameter region is denoted by $\boldsymbol{\Theta}_k$ and defined as $\left\{\theta \in \mathbb{R}^d : \boldsymbol{\Phi}\theta \in \mathcal{G}_k\right\}$, i.e., the set of parameters that predict arm k to be optimal under the model.

Note that even if $\boldsymbol{\Theta}_k = \emptyset$ for some k, the collection $\{\boldsymbol{\Theta}_k\}_{k=1}^K$ partitions $\mathbb{R}^d$ (except for a measure zero set, namely the preimages of the sets where optimal arms are ties) and hence there must exist a $\boldsymbol{\Theta}_k$ which is non empty. The image $\boldsymbol{\Phi}\boldsymbol{\Theta}_k$ includes all realizable reward vectors for which arm k appears optimal under the model.

Definition 2.3 (*Robust Region*)**.** The k-th robust region is defined as

$$\mathcal{R}_k = \left\{\boldsymbol{\mu} \in \mathcal{G}_k : \mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) \in \boldsymbol{\Theta}_k \quad \forall\, \boldsymbol{\Lambda} \in \mathcal{P}(\Delta_K)\right\},$$

where

$$\mathcal{P}(\Delta_K) = \left\{\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_i) : \begin{array}{l}\{\lambda_i\} \in \Delta_K, \\ \boldsymbol{\Phi}^\top \boldsymbol{\Lambda}\boldsymbol{\Phi} \text{ is invertible}\end{array}\right\}.$$

Together, the collection $\{\mathcal{R}_k\}_{k=1}^K$ defines the robust regions with respect to the feature matrix $\boldsymbol{\Phi}$. Any instance $\boldsymbol{\mu} \in \mathcal{R}_k$ is said to be a *robust instance*, since the greedy policy $\operatorname{argmax}_{i\in[K]} \varphi_i^\top \mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ selects the optimal arm k for any valid sampling distribution $\boldsymbol{\Lambda} \in \mathcal{P}(\Delta_K)$. Importantly, robustness does *not* require realizability: $\boldsymbol{\mu}$ need not lie in the image of $\boldsymbol{\Phi}$, allowing robustness to hold even under model misspecification.

3 INSTANCE DEPENDENT ZERO-REGRET ALGORITHMS

Robust instances are defined in an idealized setting where the true mean reward vector $\boldsymbol{\mu}$ is observed without noise. In such cases, the oracle estimate $\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ and its projection $\boldsymbol{\Phi}\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ determine whether a greedy policy selects the optimal arm. However, in practical scenarios, bandit algorithms only observe noisy samples of the reward, and the empirical mean $\widehat{\boldsymbol{\mu}}_t$ may deviate from $\boldsymbol{\mu}$ and fall outside the robust region, even if $\boldsymbol{\mu} \in \mathcal{R}_k$. This raises the question of whether algorithms like ε-greedy and LinUCB continue to achieve sublinear regret under model misspecification. Fortunately, under sub-Gaussian noise, the empirical means $\widehat{\boldsymbol{\mu}}_t$ concentrate around the true means $\boldsymbol{\mu}$ over time. We show that if $\boldsymbol{\mu}$ lies in a robust region, then both ε-greedy and LinUCB attain sublinear regret, even in the presence of arbitrarily large model misspecification. The key technical condition that enables this result is that $\boldsymbol{\mu}$ is an interior point of $\mathcal{R}_k$as defined by the following.

Definition 3.1 (*Interior Robust Instance*)**.** A mean reward vector $\boldsymbol{\mu}$ is said to be an interior point of the robust region $\mathcal{R}_k$ if there exists $\delta > 0$ such that the open rectangle $\mathbf{1}_\delta(\boldsymbol{\mu})$, a box of side length δ centered at $\boldsymbol{\mu}$, is entirely contained in $\mathcal{R}_k$.[3] As illustrated in Figure 1, larger sub-optimality gaps Δ_i allow for larger values of δ, thereby making the instance more robust to noise. The parameter δ thus quantifies the degree of robustness.

ε-Greedy Algorithm. We show that a misspecified bandit instance achieves sublinear regret asymptotically under ε-greedy, provided that the mean reward vector lies in the robust region determined by the feature matrix. We shall assume that the observations are noisy with a sub-Gaussian distribution, that is,

Assumption 3.2 (*Sub-Gaussian Noise*)**.** The observation from playing any arm a with mean μ_a is $1/2$-sub-Gaussian[4].

Theorem 3.3. *Let $\boldsymbol{\Phi}$ be a fixed feature matrix, and let $\boldsymbol{\mu}$ be the mean reward of a bandit instance which is an interior robust instance (Definition 3.1) of $\boldsymbol{\Phi}$ with sub-Gaussian noisy observations (Assumptions 3.2). Then the ε-greedy algorithm with $\varepsilon_t = 1/\sqrt{t}$ satisfies*

$$\lim_{T\to\infty} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbb{E}\left[\mu_{\mathrm{OPT}(\mu)} - \mu_{A_t}\right] \leqslant \Delta_{\max}\,.$$

Moreover, the probability of selecting a suboptimal arm $i \neq \mathrm{OPT}(\boldsymbol{\mu})$ is bounded as:

$$\mathbb{P}(A_t = i) \leq \frac{1}{K\sqrt{t}} + (\sqrt{t+1} - 1)\, e^{-\frac{\sqrt{t+1}-1}{5K}} + \frac{K}{\delta^2}\, e^{-\frac{2\delta^2(\sqrt{t+1}-1)}{K}} \quad \forall\ t \geqslant (K+1)^2 - 1.$$

The proof adapts the argument of [Auer et al., 2002]. Under sub-Gaussian noise, the empirical mean $\widehat{\boldsymbol{\mu}}_t$ concentrates around the true mean $\boldsymbol{\mu}$. If $\boldsymbol{\mu} \in \operatorname{Int}(\mathcal{R}_k)$, then for large enough t, we have $\widehat{\boldsymbol{\mu}}_t \in \mathcal{R}_k$, so greedy selection via $\mathbf{P}^{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\mu}}_t)$ correctly identifies the optimal arm. Notably, the regret bound in Theorem 3.3 depends only on the sub-optimality gap and is agnostic to the l_∞ misspecification error.

LinUCB Algorithm. The LinUCB algorithm enjoys sublinear regret on misspecified instances, provided that these are robust. In the realizable setting, instance-dependent bounds depend upon $\Delta_{\min}$, the minimum suboptimality gap between rewards. We introduce an analogous term,

[3] We use open rectangles rather than open balls for technical convenience.

[4] The constant $1/2$ is chosen purely for convenience; any fixed constant would suffice.

which plays the role of $\Delta_{\min}$ in the feature space. We denote this quantity by Δ_{rob} which is defined as the minimum suboptimality gap between arms in the feature space. We first show the following property, $\Delta_{\text{rob}}(\boldsymbol{\mu})$ is necessarilty positive for any robust $\boldsymbol{\mu}$ and that that the the suboptimal gap estimate calculated under any sampling distribution must always be lower bounded by $\Delta_{\text{rob}}(\boldsymbol{\mu})$.

Lemma 3.4. *Let $\boldsymbol{\mu}$ be an interior robust instance for a given feature matrix $\boldsymbol{\Phi}$. Then there exists a constant $\Delta_{\text{rob}}(\boldsymbol{\mu}) > 0$, such that the oracle estimate $\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ calculated under any sampling distribution $\boldsymbol{\Lambda}$, enjoys*

$$\left\langle \varphi_{\text{OPT}(\boldsymbol{\mu})} - \varphi_i,\, \mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) \right\rangle \geqslant \Delta_{\text{rob}}(\boldsymbol{\mu}),$$

for all suboptimal arm features $i \neq \text{OPT}(\boldsymbol{\mu})$.

For a fixed feature $\boldsymbol{\Phi}$, the larger the robust neighborhood $\mathbf{1}_\delta(\boldsymbol{\mu})$ for a $\boldsymbol{\mu}$, the larger the corresponding $\Delta_{\text{rob}}(\boldsymbol{\mu})$. For the regret result of LinUCB, we adopt the standard assumptions from [Abbasi-Yadkori et al., 2011]:

Assumption 3.5 (Bounded Features)**.** Each feature vector φ_i satisfies $\|\varphi_i\|_2 \leqslant 1$ for all $i \in [K]$.

Assumption 3.6 (Minimum Eigen Value)**.** The matrix V, used in the computation of $V_t = V + \sum_{i=1}^{t} \varphi_i \varphi_i^\top$, is initialized during a *forced exploration* phase as $V = \sum_{i=1}^{2d} \varphi_i \varphi_i^\top$, and satisfies $\lambda_{\min}(V) \geqslant 1$.[5]

Assumption 3.7 (Conditionally Sub-Gaussian Noise)**.** At round t, the observed reward satisfies $y_t = \mu_{A_t} + \eta_t$, where η_t is conditionally R-sub-Gaussian: $\mathbb{E}\left[e^{\lambda \eta_t} \mid A_{1:t}, \eta_{1:t-1}\right] \leqslant \exp\left(\frac{\lambda^2 R^2}{2}\right)$ for all $\lambda \in \mathbb{R}$.

We have the following regret bound for robust instances:

Theorem 3.8. *Let the feature matrix $\boldsymbol{\Phi}$ satisfy Assumption 3.5, and let the initial matrix V used in $V_t = V + \sum_{i=1}^{t} \varphi_i \varphi_i^\top$ satisfy Assumption 3.6. Suppose $\boldsymbol{\mu}$ is an interior robust instance of $\boldsymbol{\Phi}$ under sub-Gaussian noise (Assumption 3.7). Then, setting $\beta_t(\delta) = 2R^2 \log\left(\frac{(1+t/d)^{d/2}}{\delta}\right)$ for any $\delta > 0$, we have with probability at least $1-\delta$, for all $T \geqslant 2d$:*

$$\sum_{s=1}^{T} \mu_{\text{OPT}(\boldsymbol{\mu})} - \mu_{A_s} \leq \frac{2\sqrt{2}\, R\, \Delta_{\max} d\sqrt{T}}{\Delta_{\text{rob}}} \times \sqrt{\log\left(\frac{(1+T/d)}{\delta}\right)} \sqrt{\log(1+T/d)}.$$

This shows that robust instances, despite misspecification, enjoy *anytime* $\widetilde{O}(\frac{d}{\Delta_{\text{rob}(\boldsymbol{\mu})}}\sqrt{T})$ regret. The regret as measured in the feature space, is defined by $\sum_{t=1}^{T} \langle \varphi_{\text{OPT}(\boldsymbol{\mu})} - \varphi_{A_t}, \widehat{\theta}_t \rangle$ where $\widehat{\theta}_t$ is the parameter estimate computed by LinUCB at time t, is bounded by

$$2\sqrt{2} R d \sqrt{T} \sqrt{\log\left(\frac{(1+T/d)}{\delta}\right)} \sqrt{\log(1+T/d)};$$

which is order-wise $\tilde{O}(d\sqrt{T})$, the same as for the perfectly realizable setting. The effect of misspecification is evident in the term Δ_{rob}. Prior works which have focused on positive results in the presence of misspecification, [Liu et al., 2023] introduce a ρ-gap-adjusted misspecification, where it is assumed that the actual regret is related to the feature space regret by a constant factor, that is, $\mu_{\text{OPT}(\boldsymbol{\mu})} - \mu_{A_t} \leqslant \frac{1}{1-\rho}\left(\varphi_{\text{OPT}(\boldsymbol{\mu})}^\top \widehat{\theta}_t - \varphi_{A_t}^\top \widehat{\theta}_t\right)$, where $\rho \ll 1$. This assumption leads to a regret bound (cf. [Liu et al., 2023, Theorem 9]) of $\tilde{O}(\frac{d}{1-\rho}\sqrt{T})$. Thus, it seems at this point that it is not possible to remove the Δ_{rob} dependency. The $\Delta_{\max}$ term plays the role of the maximum possible regret, analogous to the rewards being scaled between -1 and 1 in the analysis of LinUCB in [Abbasi-Yadkori et al., 2011].

As in [Abbasi-Yadkori et al., 2011, Thm 5.] we can arrive at a polylogarithmic regret bound.

Corollary 3.9. *Under the assumptions of Theorem 3.8 and with $\beta_t(\delta) = 2R^2 \log\left(\frac{(1+t/d)^{d/2}}{\delta}\right)$, for any $\delta \in (0,1)$, with probability at least $1-\delta$, for all $T \geq 2d$,*

$$\begin{aligned}
&\sum_{s=1}^{T} \mu_{\text{OPT}(\boldsymbol{\mu})} - \mu_{A_s} \\
&\leq \frac{16R^2 \Delta_{\max}}{\Delta_{\text{rob}}^2}\Big(\log T + (d-1)\log\Big(\tfrac{16R^2}{\Delta_{\text{rob}}^2}\Big) \\
&+ 2(d-1)\log\Big(2\log\tfrac{1}{\delta} + d\log\big(1+\tfrac{T}{d}\big)\Big) + 2\log\tfrac{1}{\delta}\Big)^2,
\end{aligned}$$

The regret bound scales as $O\big(\frac{1}{\Delta_{\text{rob}}^2}(\log^2 T + d\log n + d^2 \log\log n)\big)$. Thus, except for an extra Δ_{rob} term, regret bound scales similarly as for the realizable case which scales as $O\big(\frac{1}{\Delta_{\min}}(\log^2 T + d\log n + d^2\log\log n)\big)$ [Abbasi-Yadkori et al., 2011]. The number of times a feature corresponding to a suboptimal arm is played is $\frac{16R^2}{\Delta_{\text{rob}^2}}\big(2\log\frac{1}{\delta} + d\log(1+\frac{T}{d})\big)^2$, which is orderwise same as for the realizable case. Thus, we see that for the realizable case, the bounds depend upon $\Delta_{\min}$; in the misspecified setting, the bounds depend upon Δ_{rob}.

4 EXISTENCE OF ROBUST REGIONS

In this section we compute the size of a robust region for an arbitrary feature matrix. To this end we begin with the following result.

Theorem 4.1. *A mean vector $\boldsymbol{\mu}$ belongs to the set of robust instances $\mathcal{R}_{\text{OPT}(\boldsymbol{\mu})}$ if and only if for every full-rank*

[5] Note that this can be achieved by forced sampling of d linearly independent features for a fixed number of times. This is done to ensure the invertibility of the design matrix V_t. In Appendix E, we remove the forced exploration phase by adding a regularizer to the design matrix.

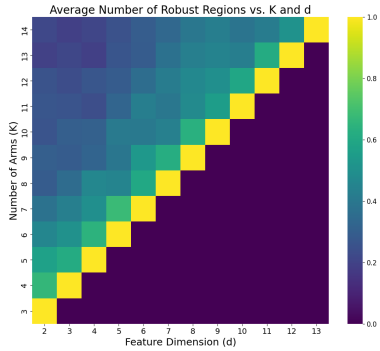

Figure 3: Average number of non-empty robust regions, expressed as a fraction of the total number of arms K, across varying numbers of arms K and feature dimensions d. For each (K, d) pair, we sample 50 standard Gaussian feature matrices and compute how many robust regions $\mathcal{R}_k$ are non-empty. Even low-dimensional features often yield a non-trivial number of robust regions, while higher dimensions tend to support near-complete coverage.

$d \times d$ submatrix Φ_d of $\mathbf{\Phi}$, the d dimensional vector $\boldsymbol{\mu}_d$ constructed by taking the corrsponding d rows of $\boldsymbol{\mu}$ satisfies $\Phi_d^{-1}\boldsymbol{\mu}_d \in \mathbf{\Theta}_{\mathrm{OPT}(\boldsymbol{\mu})}$.

Theorem 4.1 states a sufficient condition for an instance to be robust: namely, its basic solutions should all belong to the same greedy region. This theorem also shows that robust regions are convex cones. However, we shall show something stronger: that the robust regions are K-dimensional subsets of positive Lebesgue measure. The next theorem illustrates that any arbitrary feature matrix must support at least one robust region.

Theorem 4.2. *Given any arbitrary feature matrix $\mathbf{\Phi}$, there exists at least one robust region.*

We now provide the following result, which shows that robust regions in $\mathbb{R}^K$ are nontrivially large in that they have positive (Lebesgue) measure. Without loss of generality, assume that $\mathcal{R}_1$ is non-empty.

Theorem 4.3. *For any robust region $\mathcal{R}_1 \subseteq \mathcal{G}_1$ and any realizable $\boldsymbol{\mu} = \mathbf{\Phi}\theta \in \mathcal{R}_1$, there exists a positive constant α such that the open ball centred at μ with radius $\alpha\|\boldsymbol{\mu}\|$ denoted by $\mathcal{B}(\boldsymbol{\mu}, \alpha\|\boldsymbol{\mu}\|)$ lies entirely in $\mathcal{R}_1$.*

Theorem 4.3 can be spanned over all $\theta \in \mathbf{\Theta}_1$, to show that the robust region contains a convex conical subset of $\mathbb{R}^K$ with the range space being interior points. Section 5.2 is devoted to the construction of a feature representation that allows for almost all instances to be robust.

Empirical Study. The previous paragraph illustrates that there must exist at least one robust region. Here, we conduct an empirical study to determine, on average, how many robust regions an arbitrary feature matrix can support. To this end, we develop a linear program that leverages upon Therem 4.1. For every subset $S = \{i_1, \ldots, i_d\} \subseteq [K]$ such that $\mathbf{\Phi}_S$ has full rank, we define $A^{(S)} := \mathbf{\Phi}\mathbf{\Phi}_S^{-1}$ and denote $\boldsymbol{\mu}_S = (\mu_{i_1}, \ldots, \mu_{i_d})^\top$. To ensure that arm k is optimal in this subset, we require that $(A_k^{(S)} - A_i^{(S)})\boldsymbol{\mu}_S > 0$ for all $i \in S \setminus \{k\}$. This can be equivalently written as a system of linear inequalities:

$$A_{\text{ineq}}^{(S)}\boldsymbol{\mu}_S \leqslant -\varepsilon \cdot \mathbf{1},$$

for a small margin $\varepsilon > 0$. We then expand these local constraints back to the full $\boldsymbol{\mu}$ space by constructing an augmented matrix $A^{(S)} \in \mathbb{R}^{(|S|-1)\times K}$ where

$$A_{m,j}^{(S)} = \begin{cases} (A_{\text{ineq}}^{(S)})_{m,n} & \text{if } j = i_n, \\ 0 & \text{otherwise.} \end{cases}$$

By stacking all such matrices across all valid subsets S, we construct a global constraint matrix $A_{\text{global}} \in \mathbb{R}^{M\times K}$ with right-hand side $b_{\text{global}} = -\varepsilon \cdot \mathbf{1}_M$. We then solve the linear program:

$$\begin{aligned} \text{maximize}_{\boldsymbol{\mu}\in\mathbb{R}^K} \quad & \mathbf{1}^\top\boldsymbol{\mu} \\ \text{subject to} \quad & A_{\text{global}}\boldsymbol{\mu} \leqslant b_{\text{global}}, \\ & l \leqslant \mu_i \leqslant u \quad \forall i \in [K], \end{aligned}$$

where l and u are user-defined lower and upper bounds on rewards. Feasibility of this program certifies that $\mathcal{R}_k \neq \emptyset$.

To empirically assess the prevalence of robust regions, we evaluate various (K, d) configurations. We sample random feature matrices $\mathbf{\Phi} \in \mathbb{R}^{K\times d}$ with i.i.d. standard Gaussian entries. For each arm $k \in [K]$, we apply the LP procedure described above to test whether $\mathcal{R}_k$ is non-empty. Figure 3 reports the average fraction of non-empty robust regions (normalized by K) across sampled matrices. We observe two key trends: *(i)* even low-dimensional features ($d \ll K$) support several robust regions; and *(ii)* as d/K ratio increases, more arms admit robust instances.

5 EXTENDING ROBUSTNESS

Our notion of robustness extends naturally to the PAC settings and general-function classes. This further reinforces the usefulness of our geometric intuition.

5.1 ε-Optimal Arm Selection

In this subsection, we aim to provide PAC guarantees for *ε-optimal arm* selection under misspecification under uniform sampling. We do this by first expanding our definition of robust region to include ε-robust regions, where ε-optimal arms are robust as well. We assume that the optimal arm is k throughout the discussion.

Definition 5.1 (*ε-Optimal Set*)**.** Given a bandit instance $\boldsymbol{\mu}$, the ε-optimal set is defined as

$$A_\varepsilon(\boldsymbol{\mu}) = \{i \in [K] : \mu_i > \mu_k - \varepsilon\}.$$

By construction, $A_\varepsilon(\boldsymbol{\mu})$ is nonempty (containing k) and expands as ε increases. It includes more than one arm if ε exceeds the minimum suboptimality gap $\Delta_{\min}$.

We now generalize the robust region to an *ε-robust region*, which guarantees that greedy policies based on the *oracle* estimate will select an ε-optimal arm regardless of the sampling distribution.

Definition 5.2 (*ε-Robust Region*)**.** The ε-robust region for arm k is defined as

$$\mathcal{R}_k^\varepsilon = \Big\{\boldsymbol{\mu} \in \mathcal{G}_k \;\Big|\; \boldsymbol{\Phi}\,\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) \in \bigsqcup_{i \in A_\varepsilon(\boldsymbol{\mu})} \mathcal{G}_i,\ \forall\, \boldsymbol{\Lambda} \in \mathcal{P}(\Delta_K)\Big\}.$$

In words, if $\boldsymbol{\mu} \in \mathcal{R}_k^\varepsilon$, then for *any* sampling distribution, the oracle estimate guarantees selection of an ε-optimal arm under the greedy policy. Note that $\mathcal{R}_k \subseteq \mathcal{R}_k^\varepsilon$ trivially. As in Theorem 4.1, we can characterize ε-robust regions through linear algebraic conditions.

Theorem 5.3. *A vector $\boldsymbol{\mu}$ belongs to $\mathcal{R}_k^\varepsilon$ if and only if, for every full-rank $d \times d$ submatrix Φ_d of $\boldsymbol{\Phi}$ and the corresponding reduced vector $\boldsymbol{\mu}_d$, we have*

$$\boldsymbol{\Phi}\Phi_d^{-1}\boldsymbol{\mu}_d \in \mathcal{G}_i \quad \textit{for some} \quad i \in A_\varepsilon(\boldsymbol{\mu}).$$

Remark 5.4. Since the union $\bigsqcup_{i \in A_\varepsilon(\boldsymbol{\mu})} \mathcal{G}_i$ is non-convex, we require all basic solutions to lie within at least one of $\bigsqcup_{i \in A_\varepsilon(\boldsymbol{\mu})} \mathcal{G}_i$.

PAC Identification of ε-Optimal Arms. The ε-robust region enables us to derive PAC-style guarantees under misspecification: if the bandit instance is ε-robust, a uniform sampling strategy suffices to reliably identify an ε-optimal arm with high probability.

Theorem 5.5. *Let $\boldsymbol{\mu}$ be a mean reward vector that lies in the interior of the ε-robust region $\mathcal{R}^\varepsilon_{\mathrm{OPT}(\boldsymbol{\mu})}$. Let $\mathbf{1}_\delta(\boldsymbol{\mu})$ denote the largest rectangle centered at $\boldsymbol{\mu}$ and fully contained in this region, with maximum width δ. Suppose the bandit observations are perturbed by sub-Gaussian noise. For any $\alpha > 0$ and $\varepsilon > 0$, if each arm is sampled uniformly $\frac{1}{2\delta^2}\log\left(\frac{K}{\alpha}\right)$ times, then the greedy policy selects an ε-optimal arm with probability at least $1-\alpha$.*

Note that, under uniform sampling, with no misspecification, the sample complexity for identifying an ε optimal arm is $O(\frac{K}{\varepsilon^2}\log K)$ [Even-Dar et al., 2006]. Thus, we see that *uniform sampling* continues to achieve PAC bounds comparable to the realizable setting, highlighting that robustness provides a powerful tool for understanding near-optimal behavior in the presence of model misspecification. We see that the effect of ε has been consumed by δ, which inherently ensures that arms are ε-optimal.

5.2 Robustness in Nonlinear Function Approximation

We extend the notion of robust regions to nonlinear function classes and show that sublinear regret guarantees continue to hold for ε-greedy algorithms even under extreme misspecification, particularly when feature representations are carefully designed.

General Framework. Let $\{f_\theta : [K] \to \mathbb{R}\}_{\theta \in \mathbb{R}^d}$ denote a parametric function class, with $\boldsymbol{f}_\theta \in \mathbb{R}^K$ collecting $\{f_\theta(i)\}_{i \in [K]}$. Given a sampling matrix $\boldsymbol{\Lambda}$ defined as $\mathrm{diag}(\lambda_1, \ldots, \lambda_K)$, the oracle estimate is

$$\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) = \operatorname*{argmin}_\theta \|\boldsymbol{f}_\theta - \boldsymbol{\mu}\|_{\boldsymbol{\Lambda}^{1/2}}.$$

We define the robust parameter region and robust observation region for arm k as

$$\boldsymbol{\Theta}_k := \{\theta : f_\theta(k) > f_\theta(i) \ \ \forall i \neq k\},$$
$$\mathcal{R}_k := \{\boldsymbol{\mu} \in \mathcal{G}_k : \mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) \in \boldsymbol{\Theta}_k \ \ \forall\, \boldsymbol{\Lambda} \in \mathcal{P}(\Delta_K)\},$$

where $\mathcal{P}(\Delta_K)$ denotes the set of sampling matrices with full rank. If $\boldsymbol{\mu} \in \mathcal{R}_k$, then a greedy strategy based on $\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ always selects arm k, regardless of the sampling distribution. The following result shows that robustness extends to the nonlinear setting:

Theorem 5.6. *Suppose the model class $\boldsymbol{f}_\theta$ admits a robust interior instance $\boldsymbol{\mu}$ and the noise is sub-Gaussian (Assumption 3.2). Then ε-greedy with $\varepsilon_t = 1/\sqrt{t}$ satisfies*

$$\lim_{T\to\infty} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbb{E}[\mu_{\mathrm{OPT}(\boldsymbol{\mu})} - \mu_{A_t}] \leqslant \Delta_{\max}.$$

Designing Robust Feature Maps. Beyond arbitrary function classes, we can construct feature representations to ensure large robust regions. Fix $\varepsilon > 0$ and define

$$\mathcal{M}_i(\varepsilon) := \{\boldsymbol{\mu} \in \mathbb{R}^K : \mu_i > \mu_j + \varepsilon \ \ \forall j \neq i\},$$

which identifies reward vectors where arm i is at least ε better than all others. These sets are disjoint and lie within their respective greedy regions. Let

$$\mathcal{F}(\varepsilon) := \bigsqcup_{i=1}^{K} \partial\mathcal{M}_i(\varepsilon),$$

where $\partial\mathcal{M}_i(\varepsilon)$ denotes the boundary of $\mathcal{M}_i(\varepsilon)$. The following result shows that this feature design guarantees robustness:

Theorem 5.7. *For the representation $\mathcal{F}(\varepsilon)$, the union $\bigsqcup_{i=1}^K \mathcal{M}_i(\varepsilon)$ is a robust observation region. Moreover, as $\varepsilon \to 0$, the region expands to cover almost all of $\mathbb{R}^K$.*

To see this, take any $\boldsymbol{\mu} \in \mathcal{M}_k(\varepsilon)$. Any projection of $\boldsymbol{\mu}$ onto $\mathcal{F}(\varepsilon)$ lands on $\partial\mathcal{M}_k(\varepsilon) \subseteq \mathcal{G}_k$, ensuring the greedy policy selects the optimal arm k. Thus, this construction guarantees robustness for a large portion of the space. By tuning ε, one can trade off conservatism for broader coverage.

6 EXPERIMENTS WITH BANDITS

We empirically study the emergence of robustness in misspecified bandit. Our experiments are organized into two parts: (i) synthetic multi-armed bandits, and (ii) large-scale bandits based on the YAHOO! LEARNING TO RANK dataset.

Synthetic Data. We consider a 3-armed bandit with the following feature matrix:

$$\mathbf{\Phi} = \begin{bmatrix} 2 & 3 \\ 4 & 5 \\ 2 & 1 \end{bmatrix}.$$

Using Theorem 4.1, we explicitly compute the robust observation regions $\{\mathcal{R}_k\}_{k=1}^3$. We sample 10 instances from the robust region $\mathcal{R}_2$ and run ε-greedy with $\varepsilon_t = 1/\sqrt{t}$. Figure 4 shows mean cumulative regret with 3σ confidence intervals, annotated with each instance's misspecification error ρ and suboptimality gaps $\Delta_{\max}$, $\Delta_{\min}$. All instances display sublinear regret, supporting our robustness claim. Notably, $\Delta_{\max}$ is a stronger predictor of regret than ρ: an instance with $\rho = 9.02$, $\Delta_{\max} = 12.07$ performs better than one with $\rho = 0.23$, $\Delta_{\max} = 50.69$. Instances with similar $\Delta_{\max}$ (e.g., 12.07) yield comparable regret despite varying ρ (e.g., 1.03 vs. 9.02). Finally, a case with $\Delta_{\min} = 4.83 < \rho = 9.02$ demonstrates that our robust region tolerates higher misspecification than prior thresholds.

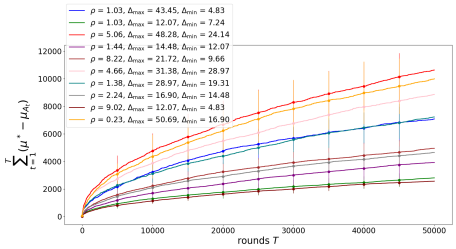

Figure 4: Cumulative regret for 10 misspecified bandit instances sampled from the robust region $\mathcal{R}_2$, evaluated under the ε-greedy algorithm with $\varepsilon_t = 1/\sqrt{t}$. The plot shows the mean over 10 trials. The Y-axis denotes $\sum_{t=1}^{T}(\mu^* - \mu_{A_t})$, and the X-axis denotes rounds T. All instances exhibit sublinear regret growth. Each curve is annotated with the l_∞ misspecification error (ρ), maximum suboptimality gap ($\Delta_{\max}$), and minimum suboptimality gap ($\Delta_{\min}$).

YAHOO! LEARNING TO RANK. Although it is computationally infeasible to verify robustness using our LP formulation, we attempt to provide empirical evidence of the existence of robust instances in natural datasets. We use the ε-greedy algorithm on the YAHOO! LEARNING TO RANK dataset [Chapelle and Chang, 2011]. For each (K, d) configuration, we randomly sample 5000 instances and run ε-greedy with $\varepsilon_t = 1/\sqrt{t}$ for $T = 20000$ rounds. Robustness is determined by checking whether the greedy action of the pathwise oracle estimate (calculated by the trajectory sample distribution induced by the algorithm) matches the true optimal arm at *every* step (See Algorithm 1). Note that this corresponds to an algorithm-dependent robust region being computed for all sampling distributions induced by an algorithm. Our notion of robust instances is one of the strictest, and in fact, the actual set of robust instances can be much larger based on the particular algorithm. For example, it is unlikely that algorithms like LinUCB and ε-greedy will induce severe adversarial sampling distributions; therefore, one can disregard these sampling distributions and define an algorithm-dependent robust region, still enjoying the regret guarantees. Table 2 summarizes the results. Under this empirical estimate of a robust instance, we observe a non-trivial proportion of robust instances, and these *all* enjoy sublinear regret. We also observe a non-trivial portion of non-robust instances that exhibit sublinear regret, thereby reiterating that robustness is a sufficient but not a necessary condition.

Algorithm 1 Testing Robustness with ε-Greedy Algorithm (Bandits)

Input:
Feature matrix $\mathbf{\Phi} = [\varphi_1^\top; \ldots; \varphi_K^\top] \in \mathbb{R}^{K \times d}$
Reward vector $\boldsymbol{\mu} \in \mathbb{R}^K$, exploration schedule ε_t, horizon T
Output: `True` if robust, `False` otherwise
Initialize $\widehat{\theta}_1 \leftarrow \mathbf{0}$, history $\mathcal{H}_0 = \emptyset$
for $t = 1$ to T **do**
  $V_t \leftarrow \sum_{a \in \mathcal{H}_{t-1}} \varphi_a \varphi_a^\top + \lambda I$
  $\widehat{\theta}_t \leftarrow V_t^{-1} \sum_{(a,\mu_a) \in \mathcal{H}_{t-1}} \varphi_a \mu_a$
  With probability ε_t, choose $A_t \leftarrow \mathrm{Uniform}([K])$
  Otherwise, choose $A_t \leftarrow \arg\max_{i \in [K]} \varphi_i^\top \widehat{\theta}_t$
  if $\arg\max_{i \in [K]} \varphi_i^\top \widehat{\theta}_t \neq \arg\max_{i \in [K]} \mu_i$ **then**
    return `False`
  end if
  Update $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{(A_t, \mu_{A_t})\}$
end for
return `True`

7 DISCUSSION AND LIMITATIONS

We presented a systematic study of model misspecification in bandit settings under practical learning algorithms. While existing theory suggests that such algorithms may fail without strong assumptions, such as realizability, they often perform well in practice. To understand this apparent contradiction, we identify structural conditions in the form of robust regions, under which effective learning is still possible despite model error. Our findings suggest that naturally occurring benign instances may play a key role in enabling this robustness.

At the same time, our analysis raises several important open questions. First, while robust regions provide a sufficient condition for sublinear regret, it remains unclear

K	d	**Robustness**	**SubLinear%**	**Linear%**	**Misspec.**
500	100	Robust	51.5	0.0	17.63
		Not Robust	12.2	36.3	
	200	Robust	52.7	0.0	16.23
		Not Robust	20.1	27.2	
1000	100	Robust	48.5	0.0	24.78
		Not Robust	19.4	32.1	
	200	Robust	50.1	0.0	23.50
		Not Robust	28.6	21.3	
	500	Robust	56.7	0.0	21.09
		Not Robust	20.6	19.7	
5000	100	Robust	43.2	0.0	60.38
		Not Robust	33.0	23.8	
	200	Robust	46.2	0.0	58.72
		Not Robust	37.4	16.4	
	500	Robust	47.3	0.0	55.32
		Not Robust	41.2	11.5	

Table 2: Summary of regret behavior, percentage of robust instances, and misspecification error under the ε-greedy algorithm on the YAHOO! LEARNING TO RANK dataset. For fixed K, increasing d reduces the average misspecification error and raises the proportion of robust instances. The observed gap between the percentage of robust instances and the percentage achieving sublinear regret empirically validates that robustness is a *sufficient* but not *necessary* condition for sublinear regret.

whether they are in any sense necessary, or whether weaker, distribution-dependent conditions could suffice. In particular, a more refined analysis might relax robustness to hold only along the trajectory distribution induced by a given algorithm (e.g., LinUCB or ε-greedy). However, such an extension is technically challenging, since a distribution over trajectories induces a distribution over weighted projections, making the characterization of these conditions nontrivial.

Second, even within the class of robust instances, the quality of guarantees can vary significantly. For instances that lie close to the boundary of a robust region, the resulting regret bounds may become vacuous, and standard algorithms may still incur large (potentially linear) regret. This highlights an inherent limitation: algorithms operating with restricted reward representations cannot, in general, overcome arbitrary forms of misspecification.

Third, while our results establish the existence of robust regions, identifying or verifying such instances in practice remains challenging. Although our experiments demonstrate that robust instances can arise in realistic settings, it is not yet clear how one can reliably detect or construct such instances for a given feature representation. This raises an important practical question: to what extent can robustness be engineered through feature design or algorithmic modifications?

Finally, our findings suggest a somewhat surprising perspective: in favorable instances that lie well within robust regions, even simple exploration strategies (such as uniform sampling or ε-greedy) may suffice to achieve good performance. While this provides a potential explanation for the empirical success of basic algorithms, it also underscores a limitation of our framework—namely, that the conditions required for robustness may be restrictive, and their prevalence in real-world problems remains to be fully understood.

Overall, our work takes a step toward explaining the empirical robustness of function-approximation-based bandit algorithms under misspecification. We hope it motivates further research on identifying weaker and more practically verifiable conditions, as well as on designing algorithms that can adapt to or exploit such structure when it is present.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24, pages 2312–2320. Curran Associates, Inc., 2011.

Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling Revisited. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 176–184. PMLR, 20–22 Apr 2017. URL `https://proceedings.mlr.press/v54/abeille17a.html`.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 39.1–39.26, Edinburgh, Scotland, 25–27 Jun 2012. PMLR. URL `https://proceedings.mlr.press/v23/agrawal12.html`.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 127–135, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL `https://proceedings.mlr.press/v28/agrawal13.html`.

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

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

George EP Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

Olivier Chapelle and Yi Chang. Yahoo! learning to rank challenge overview. In Olivier Chapelle, Yi Chang, and Tie-Yan Liu, editors, *Proceedings of the Learning to Rank Challenge*, volume 14 of *Proceedings of Machine Learning Research*, pages 1–24, Haifa, Israel, 25 Jun 2011. PMLR. URL https://proceedings.mlr.press/v14/chapelle11a.html.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, 2008.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(39):1079–1105, 2006. URL http://jmlr.org/papers/v7/evendar06a.html.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. *Advances in neural information processing systems*, 23, 2010.

Anders Forsgren. On linear least-squares problems with diagonally dominant weight matrices. *SIAM Journal on Matrix Analysis and Applications*, 17(4):763–788, 1996. doi: 10.1137/S0895479895284014.

Dylan Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3199–3210. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/foster20a.html.

Dylan J. Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert E. Schapire. Practical contextual bandits with regression oracles, 2018. URL https://arxiv.org/abs/1803.01088.

Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11478–11489. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/84c230a5b1bc3495046ef916957c7238-Paper.pdf.

Avishek Ghosh, Sayak Ray Chowdhury, and Aditya Gopalan. Misspecified linear bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Sanath Kumar Krishnamurthy, Vitor Hadad, and Susan Athey. Tractable contextual bandits beyond realizability. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1423–1431. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/kumar-krishnamurthy21a.html.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in RL with a generative model. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5662–5670. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/lattimore20a.html.

Chong Liu, Ming Yin, and Yu-Xiang Wang. No-regret linear bandits beyond realizability. In Robin J. Evans and Ilya Shpitser, editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1294–1303. PMLR, 31 Jul–04 Aug 2023. URL https://proceedings.mlr.press/v216/liu23c.html.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

Weitong Zhang, Jiafan He, Zhiyuan Fan, and Quanquan Gu. On the interplay between misspecification and sub-optimality gap in linear contextual bandits. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors,

Proceedings of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, pages 41111–41132. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/zhang23n.html`.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]. All proofs are in appendix. Assumptions to theorems are presented in main paper.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes] in Appendix
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Not Applicable. All Open Source code. Does not require propretary material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Not Applicable
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Not Applicable

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

A Contextual Bandits

#Ctx	K	d	O($\sqrt{n}$)%	O(n)%	Misspec. Error
10	20	50	50.54	49.46	8.54
		100	67.01	32.99	7.50
		200	78.72	21.28	8.30
	50	50	51.04	48.96	10.95
		100	67.78	32.22	10.06
		200	78.72	21.28	8.30
20	20	50	16.48	83.52	13.10
		100	57.30	42.70	12.03
		200	71.26	28.74	14.01
	50	50	10.00	90.00	16.70
		100	56.18	43.82	15.22
		200	71.26	28.74	14.01
50	20	50	1.35	98.65	22.36
		100	11.84	88.16	21.01
		200	47.37	52.63	19.50
	50	50	1.41	98.59	27.90
		100	16.90	83.10	26.32
		200	57.97	42.03	24.46

Table 3: We randomly sample 5000 contextual bandit instances from the YAHOO! LEARNING TO RANK dataset by selecting a random set of query IDs (contexts) and a d-dimensional subset of features. For each context, only the associated arms (documents) are available, and the ε-greedy algorithm is run to interact with this structured environment. The table reports the percentage of instances exhibiting sublinear regret ($O(\sqrt{n})$), linear regret ($O(n)$), the average model misspecification error, and the percentage of robust instances (where greedy selection recovers the optimal arm across all rounds). Remarkably, even under substantial model misspecification, a significant portion of contextual instances achieve sublinear regret, underscoring the practical robustness of linear policies in structured real-world settings.

In this section we shall extend our notion of robustness to contextual bandits. We begin with a brief motivation to capture that despite misspecification, ε-greedy is often able to attain sub-linear regret in contextual bandit settings.

Motivation We begin by examining the YAHOO! LEARNING TO RANK dataset [Chapelle and Chang, 2011], where each query ID defines a distinct context with an associated set of candidate arms. We randomly sample 5000 contextual bandit instances by selecting a subset of contexts and a random d-dimensional feature representation, and evaluate the ε-greedy algorithm on each. The results, summarized in Table 3, reveal a striking trend: despite substantial model misspecification, a large fraction of instances consistently achieve sublinear regret. These observations motivate the need for a theoretical framework that explains when and why learning remains effective—even under arbitrarily large model errors.

Problem Setting We consider the more general setting of a contextual bandit in which contexts are drawn from a finite set $\mathcal{X}$, and each context has finite arms in $\mathcal{A}$ giving rewards with means $\{\mu_{x,a}\}_{(x,a)\in\mathcal{X}\times\mathcal{A}}$. We assume that each context-arm pair (x, a) is associated with a known feature $\varphi(x, a)$ in $\mathbb{R}^d$. We also have an available parametric class of functions that serves to (approximately) model the mean reward of each context-arm pair; each function in this class is of the form $f_\theta : \mathbb{R}^d \to \mathbb{R}$ for a parameter $\theta \in \mathbf{\Theta} \subset \mathbb{R}^d$, and applying it to arm a at context x yields the expressed reward $f_\theta(\varphi(x, a)) \in \mathbb{R}$. In linear contextual bandits [Chu et al., 2011], it is assumed that the mean rewards for a context-action pair $\mu_{x,a}$ is a linear function of the features $\varphi(x, a)$, that is there exists a θ^* such that $\boldsymbol{\mu} = \mathbf{\Phi}\theta^*$. Note that we make no such assumption.

Notation We denote the sizes of the context and action spaces by $|\mathcal{X}|$ and $|\mathcal{A}|$, respectively. We assume that the number of arms exceeds the parameter dimension, i.e., $|\mathcal{A}| > d$, and that the set of features $\{\varphi(x, a)\}_{x\in\mathcal{X},\, a\in\mathcal{A}}$ spans $\mathbb{R}^d$.

The true mean rewards are represented by the vector:

$$\boldsymbol{\mu} = \left[\mu_{x,a}\right]_{x\in\mathcal{X},\, a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|},$$

and the feature matrix is given by:

$$\boldsymbol{\Phi} = \left[\cdots \varphi(x,a)^\top \cdots\right]_{x\in\mathcal{X},\, a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|\times d}.$$

For a fixed context $x \in \mathcal{X}$, we denote by $\boldsymbol{\Phi}_x \in \mathbb{R}^{|\mathcal{A}|\times d}$ the context-specific feature matrix, whose rows correspond to the feature vectors $\{\varphi(x,a)\}_{a\in\mathcal{A}}$. Similarly, we denote by $\boldsymbol{\mu}_x \in \mathbb{R}^{|\mathcal{A}|}$ the context-specific reward vector, with entries $\{\mu_{x,a}\}_{a\in\mathcal{A}}$. Finally, we let $\mathrm{OPT}(x)$ denote the optimal arm for context x. We define by $\Delta_{\max} = \max_{x\in\mathcal{X}} \max_{a\in\mathcal{A}} \left(\mu_{x,\mathrm{OPT}(x)} - \mu_{x,a}\right)$ as the largest sub-optimality gap across all context-arm pairs.

A.1 Robust Instances

We aim to identify *sufficient* conditions under which standard contextual bandit algorithms, such as ε-greedy and LinUCB, can achieve sublinear regret—even in the presence of arbitrary model misspecification.

In contrast to prior work, we do not assume *realizability* of the reward function, i.e., the true reward vector $\boldsymbol{\mu}_x$ for context x need not lie in the function class. Formally, we allow $\boldsymbol{\mu}_x \notin \{f_\theta(x,a)\}_{a\in\mathcal{A}}$ for any $\theta \in \mathbb{R}^d$. Our goal is to characterize conditions under which such a misspecified contextual instance still permits identification of the optimal arm with sublinear regret.

To extend our notion of *robust instances* to the contextual bandit setting, we observe that for any fixed context $x \in \mathcal{X}$, the learning problem reduces to a multi-armed bandit instance over the arm set $\mathcal{A}$. The key idea is to define robustness in terms of whether the model-based greedy selection under misspecification is able to consistently identify the optimal arm for each context.

Definition A.1 (*Greedy Region $\mathcal{R}_a^x$ for context x*)**.** For any context $x \in \mathcal{X}$ and arm $a \in \mathcal{A}$, the greedy region $\mathcal{R}_a^x$ is defined as:

$$\mathcal{R}_a^x \triangleq \left\{\boldsymbol{\mu}_x \in \mathbb{R}^{|\mathcal{A}|} : \mu_{x,a} > \mu_{x,b} \quad \forall\, b \neq a\right\},$$

i.e., the set of reward vectors under context x for which arm a is uniquely optimal.

The collection $\{\mathcal{R}_a^x\}_{a\in\mathcal{A}}$ partitions $\mathbb{R}^{|\mathcal{A}|}$ for each $x \in \mathcal{X}$, forming a disjoint cover (modulo boundaries) of the contextual reward space. This structure allows us to define a contextual instance as *robust* if, for every $x \in \mathcal{X}$, the projection of the true reward vector $\boldsymbol{\mu}_x$ under the model class still lies in the same greedy region $\mathcal{R}_{\mathrm{OPT}(x)}^x$ that contains $\boldsymbol{\mu}_x$ itself. In other words, despite model misspecification, the model-estimated best arm for each context agrees with the true best arm.

For clarity, we fix the model class to be linear, allowing for a closed-form least squares estimate. At time t, this estimate is given by:

$$\widehat{\theta}_{t+1} = \left(\boldsymbol{\Phi}^\top \boldsymbol{\Lambda} \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{\Lambda} \widehat{\boldsymbol{\mu}}_t, \tag{1}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix in $\mathbb{R}^{|\mathcal{X}||\mathcal{A}|\times|\mathcal{X}||\mathcal{A}|}$ containing normalized sample frequencies $\{\alpha(x,a)\}_{(x,a)\in\mathcal{X}\times\mathcal{A}}$, which lie in the simplex $\Delta_{\mathcal{X}\mathcal{A}}$. The vector $\widehat{\boldsymbol{\mu}}_t \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$ contains empirical mean estimates $\widehat{\mu}_{x,a,t}$ for each context-arm pair (x,a) up to time t. [6] From the closed-form in (1), the sample frequencies $\alpha(x,a)$ can be interpreted as:

$$\lambda_{x,a} = \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}\{X_s = x, A_s = a\},$$

which ensures that $\{\lambda_{x,a}\}_{(x,a)}$ lies in the simplex $\Delta_{\mathcal{X}\mathcal{A}}$, i.e., $\sum_{(x,a)} \lambda_{x,a} = 1$. The least squares solution can then equivalently be written as a projection under a weighted norm:

$$\widehat{\theta}_{t+1} = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \left\|\boldsymbol{\Phi}\theta - \widehat{\boldsymbol{\mu}}_t\right\|_{\boldsymbol{\Lambda}^{1/2}}.$$

This motivates the following general definition of an *oracle* estimate under a sampling distribution.

[6] As in the bandit setting, we assume $\boldsymbol{\Phi}^\top \Lambda \boldsymbol{\Phi}$ is invertible. This can be ensured by forced sampling of d linearly independent features.

Definition A.2 (*Oracle Estimate*)**.** Let $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$ be the true reward mean of a contextual bandit instance, and let $\boldsymbol{\Lambda} = \mathrm{diag}(\{\lambda_{x,a}\}_{(x,a)\in\mathcal{X}\times\mathcal{A}})$ denote a sampling distribution with $\{\lambda_{x,a}\} \in \Delta_{\mathcal{X}\mathcal{A}}$. Then the *oracle* estimate of $\boldsymbol{\mu}$ is given by:

$$\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) \triangleq \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \|\boldsymbol{\Phi}\theta - \boldsymbol{\mu}\|_{\boldsymbol{\Lambda}^{1/2}} .$$

Remark A.3. Each parameter estimate $\widehat{\theta}_t$ corresponds to an estimate of the empirical reward vector $\widehat{\boldsymbol{\mu}}_t$ under some sampling distribution $\boldsymbol{\Lambda} \in \Delta_{\mathcal{X}\times\mathcal{A}}$, i.e., $\widehat{\theta}_t = \mathbf{P}^{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\mu}}_t)$. The quantity $\boldsymbol{\Phi}\mathbf{P}^{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\mu}}_t)$ represents the projection of $\widehat{\boldsymbol{\mu}}_t$ onto the subspace spanned by the feature matrix $\boldsymbol{\Phi}$, weighted by $\boldsymbol{\Lambda}$. This distinction motivates the study of the image of the projection $\boldsymbol{\Phi}\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ as $\boldsymbol{\mu}$ varies over $\mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$ and $\boldsymbol{\Lambda}$ varies over the simplex $\Delta_{\mathcal{X}\times\mathcal{A}}$.

Let $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$ be a contextual bandit instance, and fix a context $x \in \mathcal{X}$. Let $\boldsymbol{\mu}_x \in \mathbb{R}^{|\mathcal{A}|}$ denote the sub-vector of mean rewards corresponding to context x. By definition, $\boldsymbol{\mu}_x$ lies in the greedy region $\mathcal{R}^x_{\mathrm{OPT}(x)}$, where $\mathrm{OPT}(x)$ denotes the optimal arm at x. Suppose the oracle estimate, $\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ satisfies the property that, for every $x \in \mathcal{X}$, the projected reward vector $\boldsymbol{\Phi}_x\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ lies in the same greedy region $\mathcal{R}^x_{\mathrm{OPT}(x)}$. Then, a greedy policy based on this oracle estimate will always select the optimal arm for each context, thus ensuring optimal play.

Definition A.4 (*Context dependent Robust Parameters*)**.** For any context $x \in \mathcal{X}$ and arm $a \in \mathcal{A}$, define the ath *robust parameter region* $\boldsymbol{\Theta}^x_a$ as the set of parameters $\theta \in \mathbb{R}^d$ for which the predicted reward vector $\boldsymbol{\Phi}_x\theta$ lies entirely within the greedy region $\mathcal{R}^x_a$. That is,

$$\boldsymbol{\Theta}^x_a \triangleq \left\{\theta \in \mathbb{R}^d : \boldsymbol{\Phi}_x\theta \in \mathcal{R}^x_a\right\} .$$

Remark A.5. Note that for any given context x, the collection $\{\boldsymbol{\Theta}^x_a\}_{a\in\mathcal{A}}$ form a disjoint partition of $\mathbb{R}^d$.

Definition A.6 (*Context dependent Robust Region*)**.** For any context $x \in \mathcal{X}$ and arm $a \in \mathcal{A}$, define the ath *robust region* $\mathcal{R}^x_a$ as the set of contextual bandit instances $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$ such that:

(i) Arm a is the optimal arm at context x, i.e., $\boldsymbol{\mu}_x \in \mathcal{G}^x_a$.
(ii) The oracle estimate, $\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ lies in the robust parameter region $\boldsymbol{\Theta}^x_a$ for all sampling distributions $\boldsymbol{\Lambda} \in \mathcal{P}(\Delta_{\mathcal{X}\mathcal{A}})$.

Formally,

$$\mathcal{R}^x_a = \left\{\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|} \ : \ \boldsymbol{\mu}_x \in \mathcal{G}^x_a \text{ and } \mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) \in \boldsymbol{\Theta}^x_a \quad \forall\, \boldsymbol{\Lambda} \in \mathcal{P}(\Delta_{\mathcal{X}\mathcal{A}})\right\},$$

where $\Delta_{\mathcal{X}\mathcal{A}}$ is the $(|\mathcal{X}|\,|\mathcal{A}| - 1)$-dimensional probability simplex, and the admissible sampling distributions are defined as:

$$\mathcal{P}(\Delta_{\mathcal{X}\mathcal{A}}) = \left\{\boldsymbol{\Lambda} = \mathrm{diag}(\{\lambda_{x,a}\}_{(x,a)\in\mathcal{X}\times\mathcal{A}}) : \{\lambda_{x,a}\} \in \Delta_{\mathcal{X}\mathcal{A}},\ \boldsymbol{\Phi}^\top\boldsymbol{\Lambda}\boldsymbol{\Phi} \text{ is invertible}\right\} .$$

We define the *global robust region* as

$$\mathcal{R}^{\mathcal{X}} \triangleq \bigcap_{x\in\mathcal{X}} \bigsqcup_{a\in\mathcal{A}} \mathcal{R}^x_a,$$

which depends on the feature matrix $\boldsymbol{\Phi}$. Any instance $\boldsymbol{\mu} \in \mathcal{R}^{\mathcal{X}}$ is a *robust instance*, meaning it admits optimal greedy play across all contexts. This is formalized in the following theorem.

Theorem A.7 (*Sufficient Condition*)**.** *If $\boldsymbol{\mu} \in \mathcal{R}^{\mathcal{X}}$, then for every context x and any sampling distribution $\boldsymbol{\Lambda}$, the projected oracle estimate satisfies*

$$\boldsymbol{\Phi}_x\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) \in \mathcal{G}^x_{\mathrm{OPT}(x)},$$

i.e.,

$$\operatorname*{argmax}_{a\in\mathcal{A}} \varphi(x,a)^\top\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) = \mathrm{OPT}(x).$$

Proof. If $\boldsymbol{\mu} \in \mathcal{R}^{\mathcal{X}}$, then for every $x \in \mathcal{X}$, we have $\boldsymbol{\mu} \in \mathcal{R}^x_{\mathrm{OPT}(x)}$. By definition, this implies $\boldsymbol{\mu}_x \in \mathcal{G}^x_{\mathrm{OPT}(x)}$ and $\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) \in \boldsymbol{\Theta}^x_{\mathrm{OPT}(x)}$ for all admissible $\boldsymbol{\Lambda}$. Hence, by the definition of $\boldsymbol{\Theta}^x_{\mathrm{OPT}(x)}$, we have $\boldsymbol{\Phi}_x\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) \in \mathcal{G}^x_{\mathrm{OPT}(x)}$, which implies the greedy strategy selects the optimal arm in every context. □

This theorem establishes that membership in $\mathcal{R}^{\mathcal{X}}$ is a sufficient condition for the greedy policy to act optimally in every context, regardless of the sampling distribution, despite possible model misspecification.

A.2 Instance Dependent Zero Regret Algorithms

Robust instances in contextual bandits are defined in an idealized setting where the true mean reward vector $\boldsymbol{\mu}$ is observed without noise. In this setting, the oracle estimate $\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ and its projection onto the feature space, $\boldsymbol{\Phi}\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$, determine whether a greedy strategy can consistently select the optimal arm for each context. However, in practice, algorithms such as ε-greedy and LinUCB operate under noisy observations, and the empirical reward estimate $\widehat{\boldsymbol{\mu}}_t$ may fall outside the robust region, even when $\boldsymbol{\mu} \in \mathcal{R}^{\mathcal{X}}$. As a result, it is not immediately clear that these algorithms achieve low regret under model misspecification.

Nonetheless, under standard sub-Gaussian noise assumptions, the empirical estimate $\widehat{\boldsymbol{\mu}}_t$ concentrates around $\boldsymbol{\mu}$ as t grows. We show that if $\boldsymbol{\mu}$ lies in the interior of the robust region $\mathcal{R}^{\mathcal{X}}$, then both ε-greedy and LinUCB achieve sublinear regret—even when the model class fails to realize $\boldsymbol{\mu}$. This analysis relies on the following technical definition:

Definition A.8 (*Interior Point*)**.** We assume that the mean reward $\boldsymbol{\mu}$ is an interior point of $\mathcal{R}^{\mathcal{X}}$, that is, there exists $\delta > 0$ such that the open rectangle $\mathbf{1}_\delta(\boldsymbol{\mu})$—a box of side length δ centered at $\boldsymbol{\mu}$ in $\mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$—is fully contained within $\mathcal{R}^{\mathcal{X}}$.[7]

A.2.1 ε-Greedy Algorithm

We show that a misspecified contextual bandit instance achieves sublinear regret under the ε-greedy algorithm, provided the true mean reward vector lies in the robust region defined by the feature matrix. The proof is deferred to Appendix D.1.

We assume the following noise and context distribution conditions:

Assumption A.9 (*Sub-Gaussian Noise*)**.** For any context x, the observed reward from playing arm a with mean $\mu_{x,a}$ is $1/2$-sub-Gaussian.

Assumption A.10 (*Context Distribution*)**.** Each context $x \in \mathcal{X}$ is observed with nonzero probability $\mathbf{p}_x > 0$.

Theorem A.11. *Let $\boldsymbol{\Phi}$ be a fixed feature matrix, and let $\boldsymbol{\mu}$ be the mean reward vector of a contextual bandit instance that is an interior robust instance (Definition A.8) of $\boldsymbol{\Phi}$, with noisy observations satisfying Assumption A.9 and context distribution satisfying Assumption A.10. Then, under the ε-greedy algorithm with $\varepsilon_t = 1/\sqrt{t}$, the cumulative regret is asymptotically sublinear:*

$$\lim_{T\to\infty} \frac{\textit{Regret}(T)}{\sqrt{T}} \leqslant \Delta_{\max},$$

where $\Delta_{\max} = \max_{x\in\mathcal{X}} \max_{a\in\mathcal{A}} \left(\mu_{x,\mathrm{OPT}(x)} - \mu_{x,a}\right)$ is the maximum suboptimality gap. Moreover, for all $t \geqslant (1 + A/p_x)^2 - 1$, the probability of selecting a suboptimal arm a at context x is bounded as:

$$\mathbb{P}(A_t = a \mid X_t = x) \leqslant \frac{1}{A\sqrt{t}} + p_x X(\sqrt{t+1} - 1)e^{-\frac{p_x}{5A}(\sqrt{t+1}-1)} + \frac{XA}{\delta^2} e^{-\frac{2\delta^2 p_x}{A}(\sqrt{t+1}-1)}.$$

The proof follows the same structure as in the bandit setting. Notably, the regret bound in Theorem A.11 does not depend on the misspecification error, but only on the suboptimality gap. Experimental results in Appendix B.1 support this claim.

A.2.2 LinUCB Algorithm

We analyze the LinUCB algorithm in the contextual setting and show that misspecified instances can still achieve sublinear regret, provided they are robust. As in the bandit case, the key idea is that for any robust instance, the minimum suboptimality gap in the value-function space remains strictly positive under any sampling distribution.

Lemma A.12. *Let $\boldsymbol{\mu} \in \mathrm{Int}(\mathcal{R}^{\mathcal{X}})$ be an interior robust point for a given feature matrix $\boldsymbol{\Phi}$. Then there exists a constant $\Delta_{\mathrm{rob}} > 0$ such that for any context $x \in \mathcal{X}$ and any sampling distribution $\boldsymbol{\Lambda} \in \mathcal{P}(\Delta_{\mathcal{X}\mathcal{A}})$, we have:*

$$\left\langle \varphi_{x,\mathrm{OPT}(x)} - \varphi_{x,a},\ \mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) \right\rangle \geqslant \Delta_{\mathrm{rob}},$$

for any sub-optimal arm $a \neq \mathrm{OPT}(x)$.

The quantity Δ_{rob} is instance-dependent and determined by both the true reward vector $\boldsymbol{\mu}$ and the feature matrix $\boldsymbol{\Phi}$. With this lower bound in place, we now derive a regret guarantee for LinUCB. The following assumptions mirror those in the bandit setting:

[7] We use open rectangles instead of open balls for technical convenience.

Assumption A.13 (Bounded Features)**.** All feature vectors are uniformly bounded in ℓ_2 norm: $\|\varphi_{x,a}\|_2 \leqslant 1$ for all $x \in \mathcal{X}$, $a \in \mathcal{A}$.

Assumption A.14 (Minimum Eigenvalue Condition)**.** The matrix V is initialized during a *forced exploration* phase as $V = \sum_{i=1}^{2d} \varphi_{x_i,a_i}\varphi_{x_i,a_i}^\top$, and the regularized Gram matrix $V_t = V + \sum_{i=1}^{t} \varphi_{x_i,a_i}\varphi_{x_i,a_i}^\top$ satisfies $\lambda_{\min}(V) \geqslant 1$.

Assumption A.15 (Conditionally Sub-Gaussian Noise)**.** The reward at time t is $y_t = \mu_{X_t,A_t} + \eta_t$, where η_t is conditionally R-sub-Gaussian: $\mathbb{E}\left[e^{\lambda\eta_t} \mid A_{1:t}, \eta_{1:t-1}\right] \leqslant \exp\left(\frac{\lambda^2 R^2}{2}\right)$ for all $\lambda \in \mathbb{R}$.

Under these assumptions, we obtain the following regret bound:

Theorem A.16. *Let the features in $\mathbf{\Phi}$ satisfy Assumption A.13, and let the initialization matrix V satisfy Assumption A.14. Suppose $\boldsymbol{\mu}$ is an interior robust instance (Definition A.8) and observations satisfy Assumption A.15. Then, the LinUCB algorithm with confidence parameter*

$$\beta_t(\delta) = 2R^2 \log\left(\frac{(1+t/d)^{d/2}}{\delta}\right)$$

achieves, with probability at least $1-\delta$ for all $t \geqslant 2d$,

$$\sum_{s=1}^{t} \mu_{X_s,\mathrm{OPT}(X_s)} - \mu_{X_s,A_s} \leqslant \frac{4\sqrt{t}R\Delta_{\max}}{\Delta_{\mathrm{rob}}}\sqrt{\log\left(\frac{(1+t/d)^{d/2}}{\delta}\right)}\sqrt{\log(1+t/d)^d},$$

where Δ_{rob} is as defined in Lemma A.12 and $\Delta_{\max}$ is the maximum suboptimality gap over all context-arm pairs.

We can also derive a poly-logarithmic instance dependent bound

Corollary A.17. *Let the features in $\mathbf{\Phi}$ satisfy Assumption A.13, and let the initialization matrix V satisfy Assumption A.14. Suppose $\boldsymbol{\mu}$ is an interior robust instance (Definition A.8) and the observations satisfy Assumption A.15. Run* LinUCB *with*

$$\beta_t(\delta) \;=\; 2R^2 \log\left(\frac{(1+t/d)^{d/2}}{\delta}\right).$$

Then, with probability at least $1-\delta$, for all $t \geq 2d$,

$$b_t \;:=\; \sum_{s=1}^{t} \mathbf{1}\{A_s \neq \mathrm{OPT}(X_s)\} \;\leq\; \frac{16R^2}{\Delta_{\mathrm{rob}}^2}\Big(2\log\tfrac{1}{\delta} + d\log(1+\tfrac{t}{d})\Big)^2. \tag{2}$$

Consequently, letting $\Delta_{\max}$ denote the maximum suboptimality gap over all context–arm pairs,

$$\sum_{s=1}^{t}\big(\mu_{X_s,\mathrm{OPT}(X_s)} - \mu_{X_s,A_s}\big) \;\leq\; \Delta_{\max}\, b_t \;\leq\; \frac{16R^2\,\Delta_{\max}}{\Delta_{\mathrm{rob}}^2}\Big(2\log\tfrac{1}{\delta} + d\log t\Big)^2. \tag{3}$$

The proofs are present in Appendix D.2. The regret bound in Theorem A.16 implies that the cumulative regret of LinUCB grows as $\tilde{O}(\frac{d}{\Delta_{\mathrm{rob}}}\sqrt{T})$. Thus, even under misspecification, robust instances exhibit performance guarantees comparable to realizable settings. The regret bound in Corollary A.17 is of the order $O\big(\frac{1}{\Delta_{\mathrm{rob}}^2}(d^2\log^2 T + d\log T)\big)$.

Comparison with the results of Zhang et al. [2023] Zhang et al. [2023] study misspecified contextual bandits with the misspecification dominated by the minimum sub-optimality gap. We see this characterization as being more in line with the work of Liu et al. [2023]. Under this robustness criterion, they develop a sophisticated algorithm (DS-OFUL) which they show to be regret-optimal. We believe their characterization of robustness is not directly comparable to our work. For example, our framework allows for the study of problems whose misspecification error can be (much) larger than the sub-optimality gap.

A.3 Characterizing Robust Instances

In the class of linear models, robust instances in the region $\mathcal{R}^{\mathcal{X}}$ admit an analytical characterization, stated in Theorem A.18. This result enables explicit computation of robust regions for any given feature matrix $\mathbf{\Phi}$, as illustrated in Appendix B.1.

Theorem A.18. *Let $\mathbf{\Theta}^{\mathcal{X}} \triangleq \bigcap_{x\in\mathcal{X}} \mathbf{\Theta}^{x}_{\mathrm{OPT}(x)}$ denote the intersection of robust parameter regions across all contexts. A contextual bandit instance $\boldsymbol{\mu}$ belongs to the robust observation region $\mathcal{R}^{\mathcal{X}}$ if and only if, for every $d\times d$ full-rank submatrix $\mathbf{\Phi}_d$ of $\mathbf{\Phi}$ and the corresponding subvector $\boldsymbol{\mu}_d$ of $\boldsymbol{\mu}$, we have*

$$\mathbf{\Phi}_d^{-1}\boldsymbol{\mu}_d \in \mathbf{\Theta}^{\mathcal{X}}.$$

Proof. The proof relies on a result by Forsgren [1996], restated in Lemma F.4, which asserts that for any sampling distribution $\{\lambda_{x,a}\}_{(x,a)\in\mathcal{X}\times\mathcal{A}} \in \Delta_{\mathrm{XA}}$, the oracle estimate $\mathbf{P}^{\mathbf{\Lambda}}(\boldsymbol{\mu})$, where $\mathbf{\Lambda} = \mathrm{diag}(\{\lambda_{x,a}\})$, lies in the convex hull of the basic solutions $\mathbf{\Phi}_d^{-1}\boldsymbol{\mu}_d$.

By definition, $\boldsymbol{\mu}_x \in \mathcal{G}^{x}_{\mathrm{OPT}(x)}$ for every $x \in \mathcal{X}$, so we have:

$$\begin{aligned}
\boldsymbol{\mu} \in \mathcal{R}^{\mathcal{X}} &\iff \mathbf{P}^{\mathbf{\Lambda}}(\boldsymbol{\mu}) \in \mathbf{\Theta}^{\mathcal{X}} \quad \forall\, \mathbf{\Lambda} \in \mathcal{P}(\Delta_{\mathrm{XA}}) \\
&\iff \mathrm{conv}\left\{\mathbf{\Phi}_d^{-1}\boldsymbol{\mu}_d : \mathbf{\Phi}_d \subset \mathbf{\Phi}\right\} \subset \mathbf{\Theta}^{\mathcal{X}} \\
&\iff \mathbf{\Phi}_d^{-1}\boldsymbol{\mu}_d \in \mathbf{\Theta}^{\mathcal{X}} \quad \forall\, \mathbf{\Phi}_d \subset \mathbf{\Phi}.
\end{aligned}$$

The first equivalence follows from Corollary A.19; the second from Lemma F.4; and the third from the convexity of $\mathbf{\Theta}^{\mathcal{X}}$, since a convex hull is contained in a convex set if and only if each of its generating points is. □

Corollary A.19. *A contextual bandit instance $\boldsymbol{\mu}$ belongs to the set $\mathcal{C}^{\mathcal{X}}$ if and only if the oracle estimate $\mathbf{P}^{\mathbf{\Lambda}}(\boldsymbol{\mu})$ lies in $\mathbf{\Theta}^{\mathcal{X}} \triangleq \bigcap_{x\in\mathcal{X}} \mathbf{\Theta}^{x}_{\mathrm{OPT}(x)}$ for every sampling distribution $\mathbf{\Lambda} \in \Delta_{\mathcal{X}\mathcal{A}}$.*

Proof. By definition, $\boldsymbol{\mu} \in \mathcal{C}^{\mathcal{X}}$ if and only if $\boldsymbol{\mu}_x \in \mathcal{C}^{x}_{\mathrm{OPT}(x)}$ for all $x \in \mathcal{X}$. From the definition of $\mathcal{C}^{x}_{\mathrm{OPT}(x)}$, this is equivalent to requiring that for all x,

$$\mathbf{P}^{\mathbf{\Lambda}}(\boldsymbol{\mu}) \in \mathbf{\Theta}^{x}_{\mathrm{OPT}(x)}$$

for every sampling distribution $\mathbf{\Lambda} \in \Delta_{\mathrm{XA}}$. Taking intersection over x gives the result:

$$\boldsymbol{\mu} \in \mathcal{C}^{\mathcal{X}} \iff \mathbf{P}^{\mathbf{\Lambda}}(\boldsymbol{\mu}) \in \bigcap_{x\in\mathcal{X}} \mathbf{\Theta}^{x}_{\mathrm{OPT}(x)} = \mathbf{\Theta}^{\mathcal{X}}$$

for all $\mathbf{\Lambda} \in \Delta_{\mathcal{X}\mathcal{A}}$. □

B Experiments with Contextual Bandits

We empirically study the emergence of robustness in misspecified contextual bandits. Our experiments are organized into two parts: (i) synthetic contextual bandits with controlled feature matrices, and (ii) large-scale contextual bandits based on the same dataset.

B.1 Misspecified Contextual Bandits (Synthetic)

Setup. We study a contextual bandit with two contexts x_1 and x_2, each associated with three arms. The corresponding feature matrices are given by:

$$\mathbf{\Phi}_{x_1} = \begin{bmatrix} 2 & 3 \\ 4 & 5 \\ 2 & 1 \end{bmatrix}, \quad \mathbf{\Phi}_{x_2} = \begin{bmatrix} 2 & 3 \\ 4 & 5 \\ 6 & 7 \end{bmatrix}.$$

Using Theorem A.18, we construct the robust instance region $\mathcal{R}_3^{x_1} \cap \mathcal{R}_1^{x_2}$, where arm 3 is optimal under context x_1, and arm 1 is optimal under x_2.

Results. We sample 10 reward vectors from this intersection and run the ε-greedy algorithm with $\varepsilon_t = 1/\sqrt{t}$ for $T = 5000$ rounds, under varying distributions over the contexts. As shown in Figure 5, the cumulative regret consistently grows sublinearly, validating our robustness condition. We also observe that regret scales more with the suboptimality gap $\Delta_{\max}$ than with the misspecification error ρ.

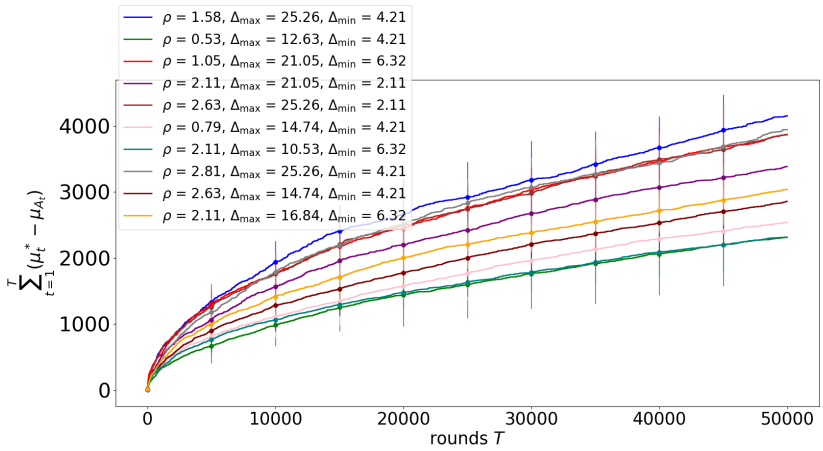

Figure 5: Cumulative regret for 10 misspecified contextual bandit instances sampled from $\mathcal{R}_3^{x_1} \cap \mathcal{R}_1^{x_2}$ under the ε-greedy algorithm with $\varepsilon_t = 1/\sqrt{t}$, for a uniform context distributions $\mathbf{p}_{x_1}, \mathbf{p}_{x_2}$. The plot shows the average regret over 10 runs. Instance-level values of misspecification error ρ, maximum gap $\Delta_{\max}$, and minimum gap $\Delta_{\min}$ are reported. Regret increases more sharply with $\Delta_{\max}$ than with ρ.

B.2 Misspecified Contextual Bandits (YAHOO! LEARNING TO RANK)

We evaluate ε-greedy on contextual bandit instances derived from the YAHOO! LEARNING TO RANK dataset [Chapelle and Chang, 2011]. Each query ID defines a distinct context, and associated documents serve as arms. In each trial, we randomly sample a subset of query IDs and restrict each context to at most K arms. For each context-arm pair, we randomly choose d out of 700 available features. The algorithm runs for $T = 5000$ rounds with $\varepsilon_t = 1/\sqrt{t}$. For each configuration $(\#\text{Contexts}, K, d)$, we conduct 5000 independent trials. We classify regret trajectories using a log-log slope test into $O(\sqrt{n})$ or $O(n)$. We also compute the fraction of robust instances—defined as those where the greedy selection from the model projection of the *oracle* matches the true optimal arm—and report the average misspecification error. Table 4 summarizes the results.

No. of Contexts #Contexts	K	d	**O($\sqrt{n}$)%**	**O(n)%**	**Robust %**	**Misspec. Error**	**feature dimension to number of features ratio** $\frac{d}{\#\text{Contexts}\,K}$
10	20	50	50.54	49.46	29.89	8.54	0.25
		100	67.01	32.99	38.55	7.50	0.5
		200	78.72	21.28	40.11	8.30	1
	50	50	51.04	48.96	23.64	10.95	0.1
		100	67.78	32.22	33.97	10.06	0.2
		200	78.72	21.28	38.55	8.30	0.4
20	20	50	16.48	83.52	9.35	13.10	0.125
		100	57.30	42.70	15.82	12.03	0.25
		200	71.26	28.74	20.05	14.01	0.5
	50	50	10.00	90.00	8.17	16.70	0.05
		100	56.18	43.82	11.13	15.22	0.1
		200	71.26	28.74	15.45	14.01	0.2
50	20	50	1.41	98.65	1.35	22.36	0.05
		100	11.84	88.16	2.48	21.01	0.1
		200	47.37	52.63	2.53	19.50	0.2
	50	50	1.41	98.59	0.86	27.90	0.02
		100	16.90	83.10	1.00	26.32	0.04
		200	57.97	42.03	1.67	24.46	0.08

Table 4: Summary of contextual bandit performance on the YAHOO! LEARNING TO RANK dataset. For each configuration of number of contexts, arms per context K, and feature dimension d, we report the percentage of trials exhibiting $O(\sqrt{n})$ and $O(n)$ regret, the robustness rate, and average misspecification error. Larger d correlates with improved learning behavior, even under significant misspecification.

Observations. We observe a consistent trend that as the model capacity ratio $\frac{d}{\#\text{Contexts}\cdot K}$ increases, the fraction of robust instances also increases. This ratio captures the relative expressiveness of the linear model compared to the number of unique context-arm pairs encountered during learning. For low values of $\frac{d}{|\mathcal{X}|\cdot K}$ (e.g., 0.02 or 0.05), robustness rates are negligible—often below 2%. However, as this ratio increases beyond 0.2, the robustness percentage rises sharply, with several configurations exceeding 30%–40%. This suggests that robustness is not merely a property of the absolute dimensionality d, but depends critically on how d scales relative to the problem complexity (number of arms and contexts). These empirical findings underscore the intuition that higher model capacity relative to the diversity of training signals can mitigate the adverse effects of misspecification, enabling consistent identification of the optimal arm.

C Description of Algorithms and Proofs of Theorems in Bandits

C.1 ε-Greedy Algorithm: Proof

We give a detailed description and proof of the result presented in Theorem 3.3.

Theorem 3.3. *Let $\mathbf{\Phi}$ be a fixed feature matrix, and let $\boldsymbol{\mu}$ be the mean reward of a bandit instance which is an interior robust instance (Definition 3.1) of $\mathbf{\Phi}$ with sub-Gaussian noisy observations (Assumptions 3.2). Then the ε-greedy algorithm with $\varepsilon_t = 1/\sqrt{t}$ satisfies*

$$\lim_{T\to\infty} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbb{E}\left[\mu_{\mathrm{OPT}(\mu)} - \mu_{A_t}\right] \leqslant \Delta_{\max} .$$

Moreover, the probability of selecting a suboptimal arm $i \neq \mathrm{OPT}(\boldsymbol{\mu})$ is bounded as:

$$\mathbb{P}(A_t = i) \leq \frac{1}{K\sqrt{t}} + (\sqrt{t+1} - 1)\, e^{-\frac{\sqrt{t+1}-1}{5K}} + \frac{K}{\delta^2}\, e^{-\frac{2\delta^2(\sqrt{t+1}-1)}{K}} \quad \forall\ t \geqslant (K+1)^2 - 1.$$

The proof follows the approach of Auer et al. [2002]. The key insight is that the least squares estimate $\widehat{\theta}_t$—equivalently, $\mathbf{P}^{\mathbf{\Lambda}}(\widehat{\boldsymbol{\mu}}_t)$ in our notation—guarantees optimal play under a greedy strategy if the sample mean estimate $\widehat{\boldsymbol{\mu}}_t$ lies within the robust region $\mathcal{R}_k$. By the concentration properties of sub-Gaussian random variables, $\widehat{\boldsymbol{\mu}}_t$ converges to the true reward vector $\boldsymbol{\mu}$ and, with high probability, lies within any fixed δ-neighbourhood of it. Since $\boldsymbol{\mu}$ is assumed to be in the interior of $\mathcal{R}_k$, there exists such a δ-neighbourhood fully contained in $\mathcal{R}_k$, ensuring the desired optimality condition eventually holds.

Proof. Assume $\boldsymbol{\mu} \in \mathrm{Int}(\mathcal{R}_k)$, so there exists a δ-open rectangle $\mathbf{1}_\delta(\boldsymbol{\mu}) \subset \mathcal{R}_k$. At round t, the probability of selecting a suboptimal arm $i \neq k$ is:

$$\mathbb{P}(A_t = i) \leqslant \frac{\varepsilon_t}{K} + \left(1 - \frac{\varepsilon_t}{K}\right) \mathbb{P}\left(\widehat{\boldsymbol{\mu}}_t \notin \mathbf{1}_\delta(\boldsymbol{\mu})\right).$$

By union bound:

$$\mathbb{P}(\widehat{\boldsymbol{\mu}}_t \notin \mathbf{1}_\delta(\boldsymbol{\mu})) \leqslant \sum_{i=1}^{K} \mathbb{P}(|\widehat{\mu}_i^{n_{i,t}} - \mu_i| \geqslant \delta).$$

Using sub-Gaussian concentration and standard peeling (cf. Auer et al., 2002), each term is bounded by:

$$\mathbb{P}(|\widehat{\mu}_i^{n_{i,t}} - \mu_i| \geqslant \delta) \leqslant t_0 \mathbb{P}(n_{i,t}^R \leqslant t_0) + \frac{1}{\delta^2} e^{-2t_0\delta^2},$$

where $n_{i,t}^R$ is the number of exploration pulls of arm i up to t, and $t_0 := \frac{1}{2K}\sum_{s=1}^{t} \varepsilon_s$. Since $\mathbb{E}[n_{i,t}^R] = \frac{1}{K}\sum_{s=1}^{t} \varepsilon_s$, Bernstein's inequality yields:

$$\mathbb{P}(n_{i,t}^R \leqslant t_0) \leqslant e^{-t_0/5}.$$

Thus,

$$\mathbb{P}(A_t = i) \leqslant \frac{\varepsilon_t}{K} + K t_0 e^{-t_0/5} + \frac{K}{\delta^2} e^{-2t_0\delta^2}.$$

For $\varepsilon_t = 1/\sqrt{t}$, we have $\sum_{s=1}^{t} \varepsilon_s \geqslant 2(\sqrt{t+1} - 1)$, hence $t_0 \geqslant \frac{1}{K}(\sqrt{t+1} - 1)$, so for $t \geqslant (1+K)^2 - 1$:

$$\mathbb{P}(A_t = i) \leqslant \frac{1}{K\sqrt{t}} + (\sqrt{t+1} - 1) e^{-\frac{\sqrt{t+1}-1}{5K}} + \frac{K}{\delta^2} e^{-\frac{2\delta^2(\sqrt{t+1}-1)}{K}}.$$

Thus, per-round regret is bounded by:

$$\mathrm{Regret}_t \leqslant \Delta_{\max} \cdot \mathbb{P}(A_t \neq k) \leqslant \frac{\Delta_{\max}}{\sqrt{t}} + o(1/t^{\alpha}),$$

for any $\alpha > 1$, and the cumulative regret satisfies:

$$\lim_{T\to\infty} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathrm{Regret}_t \leqslant \Delta_{\max}.$$

□

C.2 Characterization of Robust Regions

Theorem 4.1. *A mean vector $\boldsymbol{\mu}$ belongs to the set of robust instances $\mathcal{R}_{\mathrm{OPT}(\boldsymbol{\mu})}$ if and only if for every full-rank $d \times d$ submatrix Φ_d of $\boldsymbol{\Phi}$, the d dimensional vector $\boldsymbol{\mu}_d$ constructed by taking the corrsponding d rows of $\boldsymbol{\mu}$ satisfies $\Phi_d^{-1}\boldsymbol{\mu}_d \in \boldsymbol{\Theta}_{\mathrm{OPT}(\boldsymbol{\mu})}$.*

Proof. Without loss of generality, assume $\boldsymbol{\mu} \in \mathcal{G}_k$. The proof uses a result from Forsgren [1996] (Lemma F.4), stating that $\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ lies in the convex hull of the basic solutions $\Phi_d^{-1}\boldsymbol{\mu}_d$. Thus, for $\boldsymbol{\mu} \in \mathcal{G}_k$,

$$\begin{aligned}
\boldsymbol{\mu} \in \mathcal{R}_k &\iff \mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) \in \boldsymbol{\Theta}_k \quad \forall\, \boldsymbol{\Lambda} \in \mathcal{P}(\Delta_K) \\
&\iff \mathrm{conv}\left\{\Phi_d^{-1}\boldsymbol{\mu}_d\right\} \subset \boldsymbol{\Theta}_k \\
&\iff \Phi_d^{-1}\boldsymbol{\mu}_d \in \boldsymbol{\Theta}_k \quad \forall\, \Phi_d \subset \boldsymbol{\Phi},
\end{aligned}$$

where conv denotes the convex hull. The last equivalence follows since $\boldsymbol{\Theta}_k$ is convex. □

Theorem 4.2. *Given any arbitrary feature matrix $\boldsymbol{\Phi}$, there exists at least one robust region.*

Proof. The range of $\boldsymbol{\Phi}$ must intersect in at least one greedy region $\mathcal{G}_k$. Any point in the intersection that is sufficiently close to the range space admits a small neighborhood that lies entirely within $\mathcal{G}_k$, for which all basic solutions also lie in $\mathcal{G}_k$. In fact, there exists an ε-neighborhood, with ε sufficiently small, such that any point in the neighborhood is a robust instance. □

Theorem 4.3. *For any robust region $\mathcal{R}_1 \subseteq \mathcal{G}_1$ and any realizable $\boldsymbol{\mu} = \boldsymbol{\Phi}\theta \in \mathcal{R}_1$, there exists a positive constant α such that the open ball centred at μ with radius $\alpha\|\boldsymbol{\mu}\|$ denoted by $\mathcal{B}(\boldsymbol{\mu}, \alpha\|\boldsymbol{\mu}\|)$ lies entirely in $\mathcal{R}_1$.*

Proof. Let $\Phi \in \mathbb{R}^{K\times d}$ be a feature matrix with nonempty robust region $\mathcal{R}_1 \subset \mathbb{R}^K$ and greedy region $\mathcal{G}_1 \subset \mathbb{R}^K$. Consider a realizable bandit instance $\mu^\star = \Phi\theta^\star$. Assume $\mu^\star \in \mathrm{int}(\mathcal{G}_1)$. Then there exists $\delta > 0$ such that the open ball

$$B_\delta(\mu^\star) \;=\; \{\mu \in \mathbb{R}^K :\; \|\mu - \mu^\star\| \leq \delta\}$$

is contained in $\mathcal{G}_1$. Consider the restriction of this neighborhood to the range (image) of Φ, i.e.,

$$U_\delta \;=\; B_\delta(\mu^\star) \cap \mathrm{Range}(\Phi) \;=\; \{\Phi\theta :\; \|\Phi\theta - \mu^\star\| \leq \delta\}.$$

Define $\Theta_1 = \{\theta :\; \Phi\theta \in \mathcal{G}_1\}$ so that $U_\delta = \Phi\Theta_1 \cap B_\delta(\mu^\star)$.

Now perturb the instance to $\mu^\star + \varepsilon$, where $\varepsilon \in \mathbb{R}^K$. Fix any basic index set $S \subseteq \{1, \ldots, K\}$ with $|S| = d$ such that the square submatrix $\Phi_S \in \mathbb{R}^{d\times d}$ is invertible. Consider the basic solution θ'_S defined by matching the S-coordinates of the perturbed instance:

$$\theta' \;=\; \theta^\star + \Phi_S^{-1}\varepsilon_S, \qquad \mu' \;=\; \Phi\theta' \;=\; \Phi\theta^\star + \Phi\Phi_S^{-1}\varepsilon_S \;=\; \mu^\star + \underbrace{\Phi\Phi_S^{-1}\varepsilon_S}_{\text{projected perturbation}}.$$

Hence

$$\|\mu' - \mu^\star\| \;=\; \|\Phi\Phi_S^{-1}\varepsilon_S\| \;\leq\; \|\Phi\|\,\|\Phi_S^{-1}\|\,\|\varepsilon_S\| \;\leq\; \|\Phi\|\,\|\Phi_S^{-1}\|\,\|\varepsilon\|.$$

Therefore, a sufficient condition for $\mu' \in B_\delta(\mu^\star) \subset \mathcal{G}_1$ is

$$\|\varepsilon\| \;\leq\; \frac{\delta}{\|\Phi\|\,\|\Phi_S^{-1}\|}.$$

To make this uniform over all basic choices S with invertible Φ_S, take

$$\|\varepsilon\| \;\leq\; \min_{S:\ \Phi_S \text{ invertible}} \frac{\delta}{\|\Phi\|\,\|\Phi_S^{-1}\|}.$$

Under this bound every corresponding basic solution remains in $\mathcal{G}_1$, so the perturbed instance is robust.

Finally, since $\mathcal{G}_1$ is a cone with apex at 0 (i.e., $x \in \mathcal{G}_1 \Rightarrow \lambda x \in \mathcal{G}_1$ for all $\lambda > 0$), the margin of any $\mu \in \mathcal{G}_1$ to the boundary $\partial\mathcal{G}_1$ scales linearly with $\|\mu\|$ under any norm. Precisely, write $u = \mu/\|\mu\|$ and let

$$c(u) \;=\; \operatorname{dist}\big(u,\ \partial\mathcal{G}_1 \cap \mathbb{S}\big),$$

where $\mathbb{S} = \{x : \|x\| = 1\}$ is the angular margin between a unit vector u to boundary of $\partial\mathcal{G}_1$ restricted to a unit sphere. Then $\operatorname{dist}(\mu, \partial\mathcal{G}_1) = \|\mu\| \cdot c(u)$. Hence for $\mu^\star \in \mathcal{G}_1$ we can choose $\delta = \|\mu^\star\| \cdot c(\mu^\star/\|\mu^\star\|)$. Consequently, there exists $\alpha(\theta) > 0$ such that any perturbation ε obeying $\|\varepsilon\| \leq \alpha(\theta)\,\|\Phi\theta\|$ keeps $\Phi\theta + \varepsilon$ inside $\mathcal{G}_1$. Therefore the set

$$\Phi\Theta_1^\varepsilon \;\triangleq\; \Big\{\mu = \Phi\theta + \varepsilon :\ \theta \in \Theta_1,\ \|\varepsilon\| \leq \alpha(\theta)\,\|\Phi\theta\|\Big\}$$

is contained in the robust region $\mathcal{R}_1$; i.e., it is a norm-proportional blow-up of $\Phi\Theta_1$ that remains inside $\mathcal{R}_1$.

□

C.3 LinUCB Algorithm: Proof and Discussions

LinUCB remains a canonical regret-optimal bandit algorithm. In this section, we show that under standard assumptions, we retain the optimal regret of LinUCB even under misspecified settings, given that the bandit instance belongs to the robust region.

We add a forced exploration of d linearly-independent features to ensure the invertibility of the design matrix. In Appendix E, we remove the forced exploration phase by adding a regularizer to the design matrix. The rest of the assumptions are standard in the analysis of LinUCB [Abbasi-Yadkori et al., 2011].

Assumption 3.5 (Bounded Features)**.** Each feature vector φ_i satisfies $\|\varphi_i\|_2 \leqslant 1$ for all $i \in [K]$.

Assumption 3.6 (Minimum Eigen Value)**.** The matrix V, used in the computation of $V_t = V + \sum_{i=1}^t \varphi_i\varphi_i^\top$, is initialized during a *forced exploration* phase as $V = \sum_{i=1}^{2d} \varphi_i\varphi_i^\top$, and satisfies $\lambda_{\min}(V) \geqslant 1$.[8]

Assumption 3.7 (Conditionally Sub-Gaussian Noise)**.** At round t, the observed reward satisfies $y_t = \mu_{A_t} + \eta_t$, where η_t is conditionally R-sub-Gaussian: $\mathbb{E}\left[e^{\lambda\eta_t} \mid A_{1:t}, \eta_{1:t-1}\right] \leqslant \exp\left(\frac{\lambda^2 R^2}{2}\right)$ for all $\lambda \in \mathbb{R}$.

Remark C.1. Note that, since we have the forced exploration phase to be $2d$ times, we have, $\operatorname{trace}(V) \leqslant 2d\|\varphi\|_2^2 \leqslant 2d$. We also have $\operatorname{trace}(V) \geqslant d\lambda_{\min}(V)$. This ensures that $\lambda_{\min}(V) \leqslant 2$, thus resolving any contradictions with respect to Assumption 3.6.

Lemma 3.4. *Let $\boldsymbol{\mu}$ be an interior robust instance for a given feature matrix $\boldsymbol{\Phi}$. Then there exists a constant $\Delta_{\mathrm{rob}}(\boldsymbol{\mu}) > 0$, such that the oracle estimate $\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ calculated under any sampling distribution $\boldsymbol{\Lambda}$, enjoys*

$$\big\langle \varphi_{\mathrm{OPT}(\boldsymbol{\mu})} - \varphi_i,\ \mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})\big\rangle \geqslant \Delta_{\mathrm{rob}}(\boldsymbol{\mu}),$$

for all suboptimal arm features $i \neq \mathrm{OPT}(\boldsymbol{\mu})$.

Proof. We observe that if $\boldsymbol{\mu} \in \operatorname{Int}\big(\mathcal{R}_{\mathrm{OPT}(\boldsymbol{\mu})}\big)$, then for all $\boldsymbol{\Lambda}$, the oracle estimate $\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ lies in the robust parameter region $\boldsymbol{\Theta}_{\mathrm{OPT}(\boldsymbol{\mu})}$. Since $\boldsymbol{\mu}$ is an interior point, we have

$$\begin{aligned}\big(\varphi_{\mathrm{OPT}(\boldsymbol{\mu})} - \varphi_i\big)^\top \mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) &\geqslant \min_{\theta \in \boldsymbol{\Theta}_{\mathrm{OPT}(\boldsymbol{\mu})}} \big(\varphi_{\mathrm{OPT}(\boldsymbol{\mu})} - \varphi_i\big)^\top \theta \\ &\triangleq \Delta_i > 0 \quad \text{since } \varphi_{\mathrm{OPT}(\boldsymbol{\mu})}^\top\theta > \varphi_i^\top\theta \text{ for all } \theta \in \boldsymbol{\Theta}_{\mathrm{OPT}(\boldsymbol{\mu})}\,.\end{aligned}$$

[8] Note that this can be achieved by forced sampling of d linearly independent features for a fixed number of times. This is done to ensure the invertibility of the design matrix V_t. In Appendix E, we remove the forced exploration phase by adding a regularizer to the design matrix.

Finally, we define the minimum sub-optimality gap in the model space as:

$$\Delta_{\text{rob}} = \min_{i \in [K] \setminus \{\text{OPT}(\boldsymbol{\mu})\}} \Delta_i.$$

□

The following result shows that the number of times LinUCB can play suboptimally is at most $\tilde{O}(d\sqrt{t})$.

Lemma C.2 (Upper Bound of Sub-Optimal Plays)**.** *Let the features of the feature matrix $\boldsymbol{\Phi}$ satisfy Assumptions 3.5. Assume that the initial matrix V used in the computation of $V_t = V + \sum_{i=1}^{t} \varphi_i \varphi_i^\top$ satisfies Assumption 3.6. Let $\boldsymbol{\mu}$ be an interior robust instance of the feature matrix $\boldsymbol{\Phi}$ with sub-Gaussian noise (Assumption 3.7). Then, the LinUCB algorithm, plays sub-optimally at most $\tilde{O}(d\sqrt{t})$. Specifically, setting $\beta_t(\delta)$ set as $2R^2 \log\left(\frac{(1+t/d)^{d/2}}{\delta}\right)$ for any $\delta > 0$, we have with probability at least $1-\delta$, for any $t \geqslant 2d$*

$$\sum_{s=1}^{t} \mathbb{1}\{A_s \neq \text{OPT}(\boldsymbol{\mu})\} \leqslant \frac{4\sqrt{t}R}{\Delta_{\text{rob}}} \sqrt{\log\left(\frac{(1+t/d)^{d/2}}{\delta}\right)} \sqrt{\log(1+t/d)^d},$$

where Δ_{rob} is as defined in Lemma 3.4.

Proof. Fix $t \geqslant 2d$. Define the least squares estimate and the oracle estimate:

$$\widehat{\theta}_t = \left(\sum_{s=1}^{t} \varphi_{A_s} \varphi_{A_s}^\top\right)^{-1} \sum_{s=1}^{t} \varphi_{A_s}(\mu_{A_s} + \eta_s), \quad \theta_t^\star = \left(\sum_{s=1}^{t} \varphi_{A_s} \varphi_{A_s}^\top\right)^{-1} \sum_{s=1}^{t} \varphi_{A_s} \mu_{A_s}.$$

Let $V_t = \sum_{s=1}^{t} \varphi_{A_s} \varphi_{A_s}^\top$. Then for any $x \in \mathbb{R}^d$:

$$x^\top(\widehat{\theta}_t - \theta_t^\star) = x^\top V_t^{-1} \sum_{s=1}^{t} \varphi_{A_s} \eta_s \leqslant \|x\|_{V_t^{-1}} \left\| \sum_{s=1}^{t} \varphi_{A_s} \eta_s \right\|_{V_t^{-1}}.$$

Applying Lemma F.1, with probability at least $1-\delta$:

$$\|\widehat{\theta}_t - \theta_t^\star\|_{V_t} \leqslant R\sqrt{2\log\left(\frac{\det(V_t)^{1/2}}{\delta}\right)} \triangleq \sqrt{\beta_t(\delta)}.$$

At each round t, the algorithm chooses $A_t = \text{argmax}_a \varphi_a^\top \tilde{\theta}_t$ for $\tilde{\theta}_t \in \mathcal{C}_t$, the confidence ellipsoid centered at $\widehat{\theta}_t$. Define function-space regret:

$$r_t = \varphi_{\text{OPT}(\boldsymbol{\mu})}^\top \theta_t^\star - \varphi_{A_t}^\top \theta_t^\star.$$

Since $\tilde{\theta}_t \in \mathcal{C}_t$, we have:

$$\begin{aligned} r_t &\leqslant \varphi_{A_t}^\top \tilde{\theta}_t - \varphi_{A_t}^\top \theta_t^\star \\ &= \varphi_{A_t}^\top (\tilde{\theta}_t - \widehat{\theta}_t) + \varphi_{A_t}^\top (\widehat{\theta}_t - \theta_t^\star) \\ &\leqslant \|\varphi_{A_t}\|_{V_t^{-1}} \cdot \|\tilde{\theta}_t - \widehat{\theta}_t\|_{V_t} + \|\varphi_{A_t}\|_{V_t^{-1}} \cdot \|\widehat{\theta}_t - \theta_t^\star\|_{V_t} \leqslant 2\|\varphi_{A_t}\|_{V_t^{-1}} \sqrt{\beta_t(\delta)}. \end{aligned}$$

Summing over t and applying Cauchy-Schwarz:

$$\begin{aligned} \sum_{s=1}^{t} r_s &\leqslant 2\sum_{s=1}^{t} \|\varphi_{A_s}\|_{V_s^{-1}} \sqrt{\beta_s(\delta)} \leqslant 2\sqrt{\beta_t(\delta)} \sum_{s=1}^{t} \|\varphi_{A_s}\|_{V_s^{-1}} \\ &\leqslant 2\sqrt{\beta_t(\delta)} \sqrt{t \sum_{s=1}^{t} \|\varphi_{A_s}\|_{V_s^{-1}}^2} \leqslant 2\sqrt{\beta_t(\delta)} \sqrt{t \cdot d\log(1+t/d)} \quad \text{(by Lemma F.3).} \end{aligned}$$

Substituting $\beta_t(\delta) = 2R^2 \log\left(\frac{(1+t/d)^{d/2}}{\delta}\right)$ yields:

$$\sum_{s=1}^{t} r_s \leqslant 4R\sqrt{t \log\left(\frac{(1+t/d)^{d/2}}{\delta}\right) \cdot d\log(1+t/d)}.$$

Finally, invoking Lemma 3.4, which implies $r_s \geqslant \Delta_{\text{rob}}$ when $A_s \neq \text{OPT}(\boldsymbol{\mu})$, we obtain:

$$\sum_{s=1}^{t} \mathbb{1}\{A_s \neq \text{OPT}(\boldsymbol{\mu})\} \leqslant \frac{4R\sqrt{t}}{\Delta_{\text{rob}}}\sqrt{\log\left(\frac{(1+t/d)^{d/2}}{\delta}\right) \cdot \log(1+t/d)^d}.$$

□

With the above two lemmas, the regret bound is straightforward. Formally,

Theorem 3.8. *Let the feature matrix $\boldsymbol{\Phi}$ satisfy Assumption 3.5, and let the initial matrix V used in $V_t = V + \sum_{i=1}^{t} \varphi_i \varphi_i^\top$ satisfy Assumption 3.6. Suppose $\boldsymbol{\mu}$ is an interior robust instance of $\boldsymbol{\Phi}$ under sub-Gaussian noise (Assumption 3.7). Then, setting $\beta_t(\delta) = 2R^2 \log\left(\frac{(1+t/d)^{d/2}}{\delta}\right)$ for any $\delta > 0$, we have with probability at least $1-\delta$, for all $T \geqslant 2d$:*

$$\sum_{s=1}^{T} \mu_{\text{OPT}(\boldsymbol{\mu})} - \mu_{A_s} \leq \frac{2\sqrt{2}\, R\, \Delta_{\max} d\sqrt{T}}{\Delta_{\text{rob}}} \times \sqrt{\log\left(\frac{(1+T/d)}{\delta}\right)}\sqrt{\log(1+T/d)}.$$

Proof. The cumulative regret, defined as,

$$\begin{aligned}\sum_{s=1}^{t} \mu_{\text{OPT}(\boldsymbol{\mu})} - \mu_{A_s} &= \sum_{s=1}^{t} \Big(\mu_{\text{OPT}(\boldsymbol{\mu})} - \mu_{A_s}\Big)\mathbb{1}\{A_s \neq \text{OPT}(\boldsymbol{\mu})\} \\ &\leqslant \Delta_{\max} \sum_{s=1}^{t} \mathbb{1}\{A_s \neq \text{OPT}(\boldsymbol{\mu})\}.\end{aligned}$$

The proof is completed using Lemma C.2. □

Corollary 3.9. *Under the assumptions of Theorem 3.8 and with $\beta_t(\delta) = 2R^2 \log\big(\frac{(1+t/d)^{d/2}}{\delta}\big)$, for any $\delta \in (0,1)$, with probability at least $1-\delta$, for all $T \geq 2d$,*

$$\begin{aligned}&\sum_{s=1}^{T} \mu_{\text{OPT}(\boldsymbol{\mu})} - \mu_{A_s} \\ &\leq \frac{16R^2\Delta_{\max}}{\Delta_{\text{rob}}^2}\Big(\log T + (d-1)\log\Big(\tfrac{16R^2}{\Delta_{\text{rob}}^2}\Big) \\ &+ 2(d-1)\log\Big(2\log\tfrac{1}{\delta} + d\log\big(1+\tfrac{T}{d}\big)\Big) + 2\log\tfrac{1}{\delta}\Big)^2,\end{aligned}$$

Proof. Since the regret in the function space r_t can be either 0 or $r_t \geqslant \Delta_{\text{rob}}$, we have from the proof of Lemma C.2,

$$\begin{aligned}\sum_{t=1}^{T} r_t &\leq \frac{1}{\Delta_{\text{rob}}}\sum_{t=1}^{T} r_t^2 \leq \frac{4}{\Delta_{\text{rob}}}\sum_{t=1}^{T} \beta_t(\delta)\, \|\varphi_t\|_{V_t^{-1}}^2 \\ &\leq \frac{4\beta_T(\delta)}{\Delta_{\text{rob}}}\sum_{t=1}^{T} \|\varphi_t\|_{V_t^{-1}}^2 \leq \frac{8\beta_T(\delta)}{\Delta_{\text{rob}}} \operatorname{logdet} V_T. \qquad (4)\end{aligned}$$

Using $\beta_T(\delta)\,\mathrm{logdet}\,V_T \le 2R^2\big(\mathrm{logdet}\,V_T + 2\log\frac{1}{\delta}\big)^2 \le 2R^2\Big(2\log\frac{1}{\delta} + d\log(1+\frac{T}{d})\Big)^2$, we get

$$b_T \;\le\; \frac{1}{\Delta_{\mathrm{rob}}}\sum_{t=1}^{T} r_t \;\le\; \frac{16R^2}{\Delta_{\mathrm{rob}}^2}\Big(2\log\tfrac{1}{\delta} + d\log\big(1+\tfrac{T}{d}\big)\Big)^2, \tag{5}$$

where $b_T = \sum_{t=1}^{T}\mathbf{1}\{A_t \neq \mathrm{OPT}(\boldsymbol{\mu})\}$.

Define $E_t = \sum_{s=1}^{t}\mathbf{1}\{A_s \neq \mathrm{OPT}(\boldsymbol{\mu})\}\varphi_{A_s}\varphi_{A_s}^{\top}$, and let $A_t = V_t - E_t$ which gives $A_t = (t-b_t)\,\varphi_{\mathrm{OPT}(\boldsymbol{\mu})}\varphi_{\mathrm{OPT}(\boldsymbol{\mu})}^{\top}$. Thus A_t is rank-1 with spectrum $\{(t-b_t)\|\varphi_{\mathrm{OPT}(\boldsymbol{\mu})}\|^2, 0, \ldots, 0\}$. If $\lambda_1 \ge \cdots \ge \lambda_d$ are eigenvalues of V_t and $e_1 \ge \cdots \ge e_d$ of E_t, Weyl's inequalities give

$$\lambda_1 \in \big[(t-b_t)\|\varphi_{\mathrm{OPT}(\boldsymbol{\mu})}\|^2 + e_d,\ (t-b_t)\|\varphi_{\mathrm{OPT}(\boldsymbol{\mu})}\|^2 + e_1\big], \qquad \lambda_i \in [e_d, e_1]\ (i>1).$$

Since $\|\varphi_a\| \le 1$ for all $a \in \mathcal{A}$, we have $\mathrm{tr}(E_t) \le b_t$ and hence $e_1 \le b_t$, we obtain

$$\mathrm{logdet}\,V_t \le \log\big((t-b_t)\|\varphi_{\mathrm{OPT}(\boldsymbol{\mu})}\|^2 + e_1\big) + (d-1)\log e_1 \le \log\big((t-b_t)+b_t\big) + (d-1)\log b_t.$$

Thus substituting the relations, we get

$$\begin{aligned}
\sum_{t=1}^{T} r_t &\le \frac{8\,\beta_T(\delta)}{\Delta_{\mathrm{rob}}}\ \mathrm{logdet}\,V_T \;\le\; \frac{16R^2}{\Delta_{\mathrm{rob}}}\Big(\mathrm{logdet}\,V_T + 2\log\tfrac{1}{\delta}\Big)^2 \\
&\leqslant \frac{16R^2}{\Delta_{\mathrm{rob}}}\Big(\log\big((T-b_t)+b_T\big) + (d-1)\log b_T + 2\log\tfrac{1}{\delta}\Big)^2 \\
&\le \frac{16R^2}{\Delta_{\mathrm{rob}}}\left(\log T + 2\log\frac{1}{\delta} + (d-1)\log\Big(\frac{16R^2}{\Delta_{\mathrm{rob}}^2}\Big) + 2(d-1)\log\big(2\log\tfrac{1}{\delta} + d\log(1+\tfrac{T}{d})\big)\right)^2.
\end{aligned} \tag{6}$$

Therefore we have

$$\sum_{t=1}^{T}\mathbf{1}\{A_t \neq \mathrm{OPT}(\boldsymbol{\mu})\} \leqslant \frac{16R^2}{\Delta_{\mathrm{rob}}^2}\left(\log T + 2\log\frac{1}{\delta} + (d-1)\log\Big(\frac{16R^2}{\Delta_{\mathrm{rob}}^2}\Big) + 2(d-1)\log\big(2\log\tfrac{1}{\delta} + d\log(1+\tfrac{T}{d})\big)\right)^2;$$

which allows us to bound the actual regret as

$$\sum_{s=1}^{t}\mu_{\mathrm{OPT}(\boldsymbol{\mu})} - \mu_{A_s} \leqslant \frac{16R^2\Delta_{\max}}{\Delta_{\mathrm{rob}}^2}\left(\log T + 2\log\frac{1}{\delta} + (d-1)\log\Big(\frac{16R^2}{\Delta_{\mathrm{rob}}^2}\Big) + 2(d-1)\log\big(2\log\tfrac{1}{\delta} + d\log(1+\tfrac{T}{d})\big)\right)^2.$$

□

C.4 PAC Identification of ε-Optimal Arms.

Theorem 5.5. *Let $\boldsymbol{\mu}$ be a mean reward vector that lies in the interior of the ε-robust region $\mathcal{R}^{\varepsilon}_{\mathrm{OPT}(\boldsymbol{\mu})}$. Let $\mathbf{1}_{\delta}(\boldsymbol{\mu})$ denote the largest rectangle centered at $\boldsymbol{\mu}$ and fully contained in this region, with maximum width δ. Suppose the bandit observations are perturbed by sub-Gaussian noise. For any $\alpha > 0$ and $\varepsilon > 0$, if each arm is sampled uniformly $\frac{1}{2\delta^2}\log\big(\frac{K}{\alpha}\big)$ times, then the greedy policy selects an ε-optimal arm with probability at least $1-\alpha$.*

Proof. Under sub-Gaussian concentration (Assumption 3.2), after T_i samples per arm, the empirical means $\widehat{\boldsymbol{\mu}}$ lie within a δ-neighborhood of $\boldsymbol{\mu}$ with probability at least $1-\alpha$. Since $\boldsymbol{\mu}$ is an interior point of $\mathcal{R}^{\varepsilon}_{k}$, this neighborhood lies entirely inside $\mathcal{R}^{\varepsilon}_{k}$, ensuring ε-optimal selection by the greedy policy. □

D Description of Algorithms and Proofs of Theorems in Contextual Bandits

D.1 ε-Greedy Algorithm: Proof

We now detail the proof of Theorem A.11.

Theorem A.11. *Let $\mathbf{\Phi}$ be a fixed feature matrix, and let $\boldsymbol{\mu}$ be the mean reward vector of a contextual bandit instance that is an interior robust instance (Definition A.8) of $\mathbf{\Phi}$, with noisy observations satisfying Assumption A.9 and context distribution satisfying Assumption A.10. Then, under the ε-greedy algorithm with $\varepsilon_t = 1/\sqrt{t}$, the cumulative regret is asymptotically sublinear:*

$$\lim_{T\to\infty} \frac{\textit{Regret}(T)}{\sqrt{T}} \leqslant \Delta_{\max},$$

where $\Delta_{\max} = \max_{x\in\mathcal{X}} \max_{a\in\mathcal{A}} \left(\mu_{x,\mathrm{OPT}(x)} - \mu_{x,a}\right)$ is the maximum suboptimality gap. Moreover, for all $t \geqslant (1 + A/p_x)^2 - 1$, the probability of selecting a suboptimal arm a at context x is bounded as:

$$\mathbb{P}(A_t = a \mid X_t = x) \leqslant \frac{1}{A\sqrt{t}} + p_x X(\sqrt{t+1} - 1)e^{-\frac{p_x}{5A}(\sqrt{t+1}-1)} + \frac{XA}{\delta^2} e^{-\frac{2\delta^2 p_x}{A}(\sqrt{t+1}-1)}.$$

Remark D.1. The argument follows the same structure as the bandit case. The key insight is that the greedy policy selects the optimal arm whenever the empirical mean $\widehat{\boldsymbol{\mu}}_{\boldsymbol{t}}$ lies within the robust region $\mathcal{R}^{\mathcal{X}}$. Since $\boldsymbol{\mu}$ is an interior point, a small enough δ-neighborhood around it is also contained in $\mathcal{R}^{\mathcal{X}}$. Sub-Gaussian concentration ensures that $\widehat{\boldsymbol{\mu}}_{\boldsymbol{t}}$ enters this neighborhood with high probability.

Proof. By Assumption A.8, there exists a δ-neighborhood $\mathbf{1}_\delta(\boldsymbol{\mu}) \subset \mathcal{R}^{\mathcal{X}}$. At round t, the probability of selecting a suboptimal arm a in context x satisfies:

$$\mathbb{P}(A_t = a \mid X_t = x) \leqslant \frac{\varepsilon_t}{A} + \mathbb{P}(\widehat{\boldsymbol{\mu}}_t \notin \mathbf{1}_\delta(\boldsymbol{\mu})) \leqslant \frac{\varepsilon_t}{A} + \sum_{x,a} \mathbb{P}(|\widehat{\mu}_{x,a}^{n_{x,a,t}} - \mu_{x,a}| \geqslant \delta).$$

Each term can be bounded via:

$$\sum_{s=1}^{t} \mathbb{P}(n_{x,a,t} = s) \cdot 2e^{-2s\delta^2} \leqslant \sum_{s=1}^{t_0} \mathbb{P}(n_{x,a,t} = s) + \sum_{s>t_0} 2e^{-2s\delta^2} \leqslant t_0 \cdot \mathbb{P}(n_{x,a,t}^R \leqslant t_0) + \frac{1}{\delta^2} e^{-2t_0\delta^2},$$

where $n_{x,a,t}^R$ counts the number of random plays of (x, a) up to time t. From Assumption A.10, we have:

$$\mathbb{E}[n_{x,a,t}^R] = \frac{p_x}{A} \sum_{s=1}^{t} \varepsilon_s, \quad \mathrm{Var}[n_{x,a,t}^R] \leqslant \mathbb{E}[n_{x,a,t}^R].$$

Using Bernstein's inequality with $t_0 = \frac{p_x}{2A} \sum_{s=1}^{t} \varepsilon_s$, we obtain:

$$\mathbb{P}(n_{x,a,t}^R \leqslant t_0) \leqslant \exp(-t_0/5).$$

Combining all bounds, we get:

$$\mathbb{P}(A_t = a \mid X_t = x) \leqslant \frac{\varepsilon_t}{A} + XAt_0 e^{-t_0/5} + \frac{XA}{\delta^2} e^{-2t_0\delta^2}.$$

Using $\varepsilon_t = 1/\sqrt{t}$ and the lower bound $\sum_{s=1}^{t} \varepsilon_s \geqslant \sqrt{t+1} - 1$, we have:

$$t_0 \geqslant \frac{p_x}{A}(\sqrt{t+1} - 1),$$

and for $t \geqslant (1 + A/p_x)^2 - 1$,

$$\mathbb{P}(A_t = a \mid X_t = x) \leqslant \frac{1}{A\sqrt{t}} + p_x X(\sqrt{t+1} - 1)e^{-\frac{p_x}{5A}(\sqrt{t+1}-1)} + \frac{XA}{\delta^2} e^{-\frac{2\delta^2 p_x}{A}(\sqrt{t+1}-1)}.$$

Letting $\Delta_x = \max_a \Delta_{x,a}$, we obtain:

$$\mathrm{Regret}_{x,t} \leqslant \Delta_x \cdot \frac{1}{\sqrt{t}} + o(1/t^\alpha),$$

and summing over contexts,

$$\mathrm{Regret}_t = \sum_x p_x \cdot \mathrm{Regret}_{x,t} \leqslant \frac{\Delta_{\max}}{\sqrt{t}} + o(1/t^\alpha), \quad \text{where } \Delta_{\max} = \max_x \Delta_x.$$

Hence,

$$\lim_{T\to\infty} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathrm{Regret}_t \leqslant \Delta_{\max}.$$

□

D.2 LinUCB in Contextual Bandits: Proof and Discussion

We show that LinUCB achieves sublinear regret for any contextual bandit instance lying within the robust observation region. To ensure the invertibility of the design matrix, we include an initial forced exploration phase with d linearly independent features. (In Appendix E, we eliminate this requirement via ridge regularization.) All other assumptions follow standard LinUCB analysis [Abbasi-Yadkori et al., 2011].

Assumption A.13 (Bounded Features)**.** All feature vectors are uniformly bounded in ℓ_2 norm: $\|\varphi_{x,a}\|_2 \leqslant 1$ for all $x \in \mathcal{X}$, $a \in \mathcal{A}$.

Assumption A.14 (Minimum Eigenvalue Condition)**.** The matrix V is initialized during a *forced exploration* phase as $V = \sum_{i=1}^{2d} \varphi_{x_i,a_i}\varphi_{x_i,a_i}^\top$, and the regularized Gram matrix $V_t = V + \sum_{i=1}^{t} \varphi_{x_i,a_i}\varphi_{x_i,a_i}^\top$ satisfies $\lambda_{\min}(V) \geqslant 1$.

Assumption A.15 (Conditionally Sub-Gaussian Noise)**.** The reward at time t is $y_t = \mu_{X_t,A_t} + \eta_t$, where η_t is conditionally R-sub-Gaussian: $\mathbb{E}\left[e^{\lambda\eta_t} \mid A_{1:t}, \eta_{1:t-1}\right] \leqslant \exp\left(\frac{\lambda^2 R^2}{2}\right)$ for all $\lambda \in \mathbb{R}$.

following lemma mirrors Lemma 3.4 adapted to the contextual setting.

Lemma A.12. *Let $\boldsymbol{\mu} \in \mathrm{Int}(\mathcal{R}^{\mathcal{X}})$ be an interior robust point for a given feature matrix $\boldsymbol{\Phi}$. Then there exists a constant $\Delta_{\mathrm{rob}} > 0$ such that for any context $x \in \mathcal{X}$ and any sampling distribution $\boldsymbol{\Lambda} \in \mathcal{P}(\Delta_{\mathcal{X}\mathcal{A}})$, we have:*

$$\left\langle \varphi_{x,\mathrm{OPT}(x)} - \varphi_{x,a},\ \mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) \right\rangle \geqslant \Delta_{\mathrm{rob}},$$

for any sub-optimal arm $a \neq \mathrm{OPT}(x)$.

Proof. Suppose $\boldsymbol{\mu} \in \mathrm{Int}(\mathcal{R}^{\mathcal{X}})$ is an interior point of the contextual robust region. Then, for each context $x \in \mathcal{X}$, we have $\boldsymbol{\mu}_x \in \mathrm{Int}(\mathcal{R}^x_{\mathrm{OPT}(x)})$. By the definition of the contextual robust region, this implies that for all sampling distributions $\boldsymbol{\Lambda}$, the oracle estimate $\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ satisfies:

$$\varphi_{x,\mathrm{OPT}(x)}^\top \mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) > \varphi_{x,i}^\top \mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) \quad \forall i \in \mathcal{A} \setminus \{\mathrm{OPT}(x)\},\ \ \forall x \in \mathcal{X}.$$

Since $\mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ belongs to $\boldsymbol{\Theta}^x_{\mathrm{OPT}(x)}$, we can lower bound this margin uniformly over all $\theta \in \boldsymbol{\Theta}^x_{\mathrm{OPT}(x)}$:

$$\begin{aligned}\left(\varphi_{x,\mathrm{OPT}(x)} - \varphi_{x,i}\right)^\top \mathbf{P}^{\boldsymbol{\Lambda}}(\boldsymbol{\mu}) &\geqslant \min_{\theta \in \boldsymbol{\Theta}^x_{\mathrm{OPT}(x)}} \left(\varphi_{x,\mathrm{OPT}(x)} - \varphi_{x,i}\right)^\top \theta \\ &\triangleq \Delta_{x,i} > 0 \quad \text{since } \theta \in \boldsymbol{\Theta}^x_{\mathrm{OPT}(x)} \Rightarrow \varphi_{x,\mathrm{OPT}(x)}^\top \theta > \varphi_{x,i}^\top \theta.\end{aligned}$$

We define the minimum sub-optimality gap over all contexts and arms as

$$\Delta_{\mathrm{rob}} = \min_{x \in \mathcal{X}} \min_{i \in \mathcal{A} \setminus \{\mathrm{OPT}(x)\}} \Delta_{x,i}.$$

This quantity is strictly positive and depends only on $\boldsymbol{\Phi}$, $\boldsymbol{\mu}$, and the sampling distributions $\boldsymbol{\Lambda}$. □

The following lemma parallels Lemma C.2; its proof is omitted as it follows identically.

Lemma D.2 (Bound on Suboptimal Plays)**.** *Let the features of the feature matrix $\boldsymbol{\Phi}$ satisfy Assumptions A.13. Assume that the initial matrix V used in the computation of $V_t = V + \sum_{i=1}^{t} \varphi_i \varphi_i^\top$ satisfies Assumption A.14. Let $\boldsymbol{\mu}$ be an interior robust contextual instance of the feature matrix $\boldsymbol{\Phi}$ with sub-Gaussian noise (Assumption A.15). Then, the LinUCB algorithm, plays sub-optimally at most $\tilde{O}(d\sqrt{t})$. Specifically, setting $\beta_t(\delta)$ set as $2R^2 \log\left(\frac{(1+t/d)^{d/2}}{\delta}\right)$ for any $\delta > 0$, we have with probability at least $1-\delta$, for any $t \geqslant 2d$*

$$\sum_{s=1}^{t} \mathbb{1}\{A_s \neq \mathrm{OPT}(\boldsymbol{\mu})\} \leqslant \frac{4\sqrt{t}R}{\Delta_{\mathrm{rob}}} \sqrt{\log\left(\frac{(1+t/d)^{d/2}}{\delta}\right)} \sqrt{\log\left(1+t/d\right)^d},$$

where Δ_{rob} is as defined in Lemma A.12.

Using the above two results give us:

Theorem A.16. *Let the features in $\mathbf{\Phi}$ satisfy Assumption A.13, and let the initialization matrix V satisfy Assumption A.14. Suppose $\boldsymbol{\mu}$ is an interior robust instance (Definition A.8) and observations satisfy Assumption A.15. Then, the LinUCB algorithm with confidence parameter*

$$\beta_t(\delta) = 2R^2 \log\left(\frac{(1+t/d)^{d/2}}{\delta}\right)$$

achieves, with probability at least $1-\delta$ for all $t \geqslant 2d$,

$$\sum_{s=1}^{t} \mu_{X_s,\mathrm{OPT}(X_s)} - \mu_{X_s,A_s} \leqslant \frac{4\sqrt{t}R\Delta_{\max}}{\Delta_{\mathrm{rob}}} \sqrt{\log\left(\frac{(1+t/d)^{d/2}}{\delta}\right)} \sqrt{\log(1+t/d)^d},$$

where Δ_{rob} is as defined in Lemma A.12 and $\Delta_{\max}$ is the maximum suboptimality gap over all context-arm pairs.

Proof. The cumulative regret satisfies:

$$\begin{aligned}\sum_{s=1}^{T} \mu_{X_s,\mathrm{OPT}(X_s)} - \mu_{X_s,A_s} &= \sum_{s=1}^{T} \mathbb{1}\{A_s \neq \mathrm{OPT}(X_s)\}\Delta_{X_s,A_s} \\ &\leqslant \Delta_{\max} \sum_{s=1}^{T} \mathbb{1}\{A_s \neq \mathrm{OPT}(X_s)\}.\end{aligned}$$

Applying Lemma D.2 completes the proof. □

Theorem A.16. *Let the features in $\mathbf{\Phi}$ satisfy Assumption A.13, and let the initialization matrix V satisfy Assumption A.14. Suppose $\boldsymbol{\mu}$ is an interior robust instance (Definition A.8) and observations satisfy Assumption A.15. Then, the LinUCB algorithm with confidence parameter*

$$\beta_t(\delta) = 2R^2 \log\left(\frac{(1+t/d)^{d/2}}{\delta}\right)$$

achieves, with probability at least $1-\delta$ for all $t \geqslant 2d$,

$$\sum_{s=1}^{t} \mu_{X_s,\mathrm{OPT}(X_s)} - \mu_{X_s,A_s} \leqslant \frac{4\sqrt{t}R\Delta_{\max}}{\Delta_{\mathrm{rob}}} \sqrt{\log\left(\frac{(1+t/d)^{d/2}}{\delta}\right)} \sqrt{\log(1+t/d)^d},$$

where Δ_{rob} is as defined in Lemma A.12 and $\Delta_{\max}$ is the maximum suboptimality gap over all context-arm pairs.

Proof. Define the per-round regret in function space by $r_s := \mu_{X_s,\mathrm{OPT}(X_s)} - \mu_{X_s,A_s}$. By the interior-robustness assumption (Definition A.8 and Lemma A.12), either $r_s = 0$ or $r_s \geq \Delta_{\mathrm{rob}}$. The usual LINUCB analysis (cf. Lemma C.2 in the contextual variant) gives, on the $1-\delta$ high-probability event,

$$\sum_{s=1}^{t} r_s^2 \;\leq\; 4\sum_{s=1}^{t} \beta_s(\delta)\, \|\varphi_{X_s,A_s}\|^2_{V_s^{-1}} \;\leq\; 4\,\beta_t(\delta) \sum_{s=1}^{t} \|\varphi_{X_s,A_s}\|^2_{V_s^{-1}} \;\leq\; 8\,\beta_t(\delta)\,\operatorname{logdet} V_t,$$

where $V_s = V + \sum_{u=1}^{s} \varphi_{X_u,A_u}\varphi_{X_u,A_u}^{\top}$ and the last step uses the standard self-normalized identity $\sum_{s=1}^{t} \|\varphi_{X_s,A_s}\|^2_{V_s^{-1}} \leq 2\operatorname{logdet} V_t$. Since $r_s \in \{0\} \cup [\Delta_{\mathrm{rob}}, \infty)$, we have

$$\sum_{s=1}^{t} r_s \;\leq\; \frac{1}{\Delta_{\mathrm{rob}}} \sum_{s=1}^{t} r_s^2 \;\leq\; \frac{8\,\beta_t(\delta)}{\Delta_{\mathrm{rob}}}\,\operatorname{logdet} V_t.$$

With $\|\varphi_{x,a}\| \leq 1$ and V full-rank (Assumptions A.13, A.14), $\operatorname{logdet} V_t \leq \operatorname{logdet} V + d\log(1+\frac{t}{d})$, and absorbing $\operatorname{logdet} V$ into constants yields

$$\sum_{s=1}^{t} r_s \;\leq\; \frac{8\,\beta_t(\delta)}{\Delta_{\mathrm{rob}}}\, d\log\!\big(1+\tfrac{t}{d}\big).$$

Now plug in $\beta_t(\delta) = 2R^2 \log\big(\frac{(1+t/d)^{d/2}}{\delta}\big) = R^2\big(d\log(1+\frac{t}{d}) + 2\log\frac{1}{\delta}\big)$, apply the inequality $(x+y) \le 2\max\{x,y\} \le 2(x+y)$ to keep the expression compact, and use $b_t \le \Delta_{\mathrm{rob}}^{-1}\sum_{s=1}^{t} r_s$ to obtain

$$b_t \;\le\; \frac{16R^2}{\Delta_{\mathrm{rob}}^2}\Big(2\log\tfrac{1}{\delta} + d\log(1+\tfrac{t}{d})\Big)^2,$$

where $b_T = \sum_{t=1}^{T} \mathbf{1}\{A_t \neq \mathrm{OPT}(\boldsymbol{\mu})\}$.

Define

$$E_t \;:=\; \sum_{s=1}^{t} \mathbf{1}\{A_s \neq \mathrm{OPT}(X_s)\}\,\varphi_{X_s,A_s}\varphi_{X_s,A_s}^{\top},$$

and

$$A_t \;:=\; V_t - E_t \;=\; V + \sum_{s=1}^{t} \mathbf{1}\{A_s = \mathrm{OPT}(X_s)\}\,\varphi_{X_s,\mathrm{OPT}(X_s)}\varphi_{X_s,\mathrm{OPT}(X_s)}^{\top}.$$

Then $\mathrm{tr}(E_t) = \sum_{s:\,A_s\neq\mathrm{OPT}(X_s)} \|\varphi_{X_s,A_s}\|^2 \le b_t$, so the largest eigenvalue e_1 of E_t satisfies $e_1 \le \mathrm{tr}(E_t) \le b_t$. Write the eigenvalues of V_t and E_t as $\lambda_1 \ge \cdots \ge \lambda_d$ and $e_1 \ge \cdots \ge e_d$, respectively. By Weyl's inequalities,

$$\lambda_1 \;\le\; \lambda_{\max}(A_t) + e_1, \qquad \lambda_i \;\le\; \lambda_{\max}(A_t) + e_1 \quad (i>1).$$

Moreover,

$$\begin{aligned}
\lambda_{\max}(A_t) \;\le\; \mathrm{tr}(A_t) &= \mathrm{tr}\left(V + \sum_{s:\,A_s=\mathrm{OPT}(X_s)} \varphi_{X_s,\mathrm{OPT}(X_s)}\varphi_{X_s,\mathrm{OPT}(X_s)}^{\top}\right)\\
&\le \mathrm{tr}(V) + \sum_{s:\,A_s=\mathrm{OPT}(X_s)} \|\varphi_{X_s,\mathrm{OPT}(X_s)}\|^2\\
&\le \mathrm{tr}(V) + (t-b_t).
\end{aligned}$$

Absorbing the fixed $\mathrm{tr}(V)$ into constants (or assuming $V \succeq I$ and starting the count at $2d$), we use $\lambda_{\max}(A_t) \le (t-b_t)$ for notational simplicity. Hence

$$\det(V_t) \;=\; \prod_{i=1}^{d}\lambda_i \;\le\; \big((t-b_t)+e_1\big)^d \;\le\; \big((t-b_t)+b_t\big)^d \;=\; t^{\,d}.$$

Taking logs,

$$\mathrm{logdet}\,V_t \;\le\; d\log t \tag{7}$$

Plug (7) into the high-probability bound $\sum_{s=1}^{t} r_s \;\le\; \frac{8\,\beta_t(\delta)}{\Delta_{\mathrm{rob}}}\,\mathrm{logdet}\,V_t$ and continue exactly as in your non-contextual derivation:

$$\begin{aligned}
\sum_{s=1}^{t} r_s \le \frac{8\,\beta_t(\delta)}{\Delta_{\mathrm{rob}}}\;\mathrm{logdet}\,V_t \;&\le\; \frac{16R^2}{\Delta_{\mathrm{rob}}}\Big(\mathrm{logdet}\,V_t + 2\log\tfrac{1}{\delta}\Big)^2\\
&\le \frac{16R^2}{\Delta_{\mathrm{rob}}}\Big(d\log t + 2\log\tfrac{1}{\delta}\Big)^2,
\end{aligned}$$

yielding

$$b_t \;=\; \sum_{s=1}^{t}\mathbf{1}\{A_s \neq \mathrm{OPT}(X_s)\} \;\le\; \frac{16R^2}{\Delta_{\mathrm{rob}}^2}\left(d\log t + 2\log\frac{1}{\delta}\right)^2.$$

Finally,

$$\sum_{s=1}^{t}\big(\mu_{X_s,\mathrm{OPT}(X_s)} - \mu_{X_s,A_s}\big) \;\le\; \Delta_{\max}\,b_t,$$

which gives the stated contextual regret bound through the same substitution. □

E Weighted Ridge Regression

E.1 Regularized Robust Instances

In our definition of the *robust region*, we explicitly assumed the invertibility of $\mathbf{\Phi}^\top \mathbf{\Lambda}\mathbf{\Phi}$. This necessitates a forced exploration phase in our algorithms. In practice, however, one typically uses a regularization parameter $\lambda > 0$ to ensure invertibility and numerical stability. We now describe how our theory of robust regions naturally extends to the case where the algorithm uses a weighted, regularized least squares estimate in place of the ordinary least squares estimate.

Consider the regularized *oracle* estimate:

$$\mathbf{P}_\lambda^{\mathbf{\Lambda}}(\boldsymbol{\mu}) \triangleq \left(\mathbf{\Phi}^\top \mathbf{\Lambda}\mathbf{\Phi} + \lambda I\right)^{-1} \mathbf{\Phi}^\top \mathbf{\Lambda}\boldsymbol{\mu}, \quad \text{for some } \lambda > 0.$$

We define the corresponding *regularized robust region* as:

$$\mathcal{R}_k^\lambda = \left\{\boldsymbol{\mu} \in \mathcal{G}_k : \mathbf{P}_\lambda^{\mathbf{\Lambda}}(\boldsymbol{\mu}) \in \mathbf{\Theta}_k \quad \forall\, \mathbf{\Lambda} \in \Delta_K\right\},$$

where

$$\mathbf{\Theta}_k \triangleq \left\{\theta \in \mathbb{R}^d : \ \mathbf{\Phi}\theta \in \mathcal{G}_k\right\} \ .$$

We now provide an explicit characterization of the set $\mathcal{R}_k^\lambda$. Given a feature matrix $\mathbf{\Phi}$, a sampling distribution $\mathbf{\Lambda} \in \mathcal{P}(\Delta_K)$, and a reward vector $\boldsymbol{\mu} \in \mathbb{R}^K$, define the following augmented quantities:

$$\mathbf{\Phi}^* = \begin{bmatrix} \mathbf{\Phi} \\ \sqrt{\lambda} I_{d\times d} \end{bmatrix}, \quad \mathbf{\Lambda}^* = \begin{bmatrix} \mathbf{\Lambda} & 0 \\ 0 & I_{d\times d} \end{bmatrix}, \quad \boldsymbol{\mu}^* = \begin{bmatrix} \boldsymbol{\mu} \\ 0 \end{bmatrix},$$

where $\mathbf{\Phi}^* \in \mathbb{R}^{(K+d)\times d}$, $\mathbf{\Lambda}^* \in \mathbb{R}^{(K+d)\times(K+d)}$, and $\boldsymbol{\mu}^* \in \mathbb{R}^{K+d}$.

Let $\mathcal{J}(\mathbf{\Phi}^*)$ denote the collection of index sets corresponding to non-singular $d \times d$ submatrices of $\mathbf{\Phi}^*$, and define $\mathcal{I} = \{K+1, \ldots, K+d\}$ as the indices of the regularization rows.

Theorem E.1. *Let $\boldsymbol{\mu} \in \mathbb{R}^K$ be any reward vector with optimal arm k. Then,*

$$\boldsymbol{\mu} \in \mathcal{R}_k^\lambda \quad \iff \quad {\Phi_J^*}^{-1} \boldsymbol{\mu}_J^* \in \mathbf{\Theta}_k \quad \textit{for all } J \in \mathcal{J}(\mathbf{\Phi}^*) \setminus \mathcal{I}.$$

Proof. For any sampling distribution $\mathbf{\Lambda}$, the regularized oralce estimate can be written as:

$$\mathbf{P}_\lambda^{\mathbf{\Lambda}}(\boldsymbol{\mu}) = \left({\mathbf{\Phi}^*}^\top \mathbf{\Lambda}^* \mathbf{\Phi}^*\right)^{-1} {\mathbf{\Phi}^*}^\top \mathbf{\Lambda}^* \boldsymbol{\mu}^*.$$

By Lemma F.4, this estimate lies in the convex hull of at most $\binom{K+d}{d}$ basic solutions of the form ${\Phi_J^*}^{-1}\boldsymbol{\mu}_J^*$, for $J \in \mathcal{J}(\mathbf{\Phi}^*)$, and for all $\mathbf{\Lambda} \in \mathcal{P}(\Delta_K)$. That is,

$$\mathbf{P}_\lambda^{\mathbf{\Lambda}}(\boldsymbol{\mu}) = \sum_{J \in \mathcal{J}(\mathbf{\Phi}^*)} \left(\frac{\det(\mathbf{\Lambda}_J^*) \cdot \det(\Phi_J^*)^2}{\sum_{K \in \mathcal{J}(\mathbf{\Phi}^*)} \det(\mathbf{\Lambda}_K^*) \cdot \det(\Phi_K^*)^2} \right) {\Phi_J^*}^{-1} \boldsymbol{\mu}_J^*.$$

Now consider the index set $\mathcal{I} = \{K+1, \ldots, K+d\}$ corresponding to the regularization rows. For $J = \mathcal{I}$, we have ${\Phi_J^*}^{-1}\boldsymbol{\mu}_J^* = 0$. Partitioning $\mathcal{J}(\mathbf{\Phi}^*)$ into $\mathcal{J}(\mathbf{\Phi}^*) \setminus \{\mathcal{I}\}$ and $\mathcal{I}$, the above expression becomes:

$$\begin{aligned}
\mathbf{P}_\lambda^{\mathbf{\Lambda}}(\boldsymbol{\mu}) &= \sum_{J \in \mathcal{J}(\mathbf{\Phi}^*)\setminus\{\mathcal{I}\}} \left(\frac{\det(\mathbf{\Lambda}_J^*) \cdot \det(\Phi_J^*)^2}{\sum_{K \in \mathcal{J}(\mathbf{\Phi}^*)\setminus\{\mathcal{I}\}} \det(\mathbf{\Lambda}_K^*) \cdot \det(\Phi_K^*)^2 + \lambda^d} \right) {\Phi_J^*}^{-1} \boldsymbol{\mu}_J^* \\
&= c^\lambda \sum_{J \in \mathcal{J}(\mathbf{\Phi}^*)\setminus\{\mathcal{I}\}} \left(\frac{\det(\mathbf{\Lambda}_J^*) \cdot \det(\Phi_J^*)^2}{\sum_{K \in \mathcal{J}(\mathbf{\Phi}^*)\setminus\{\mathcal{I}\}} \det(\mathbf{\Lambda}_K^*) \cdot \det(\Phi_K^*)^2} \right) {\Phi_J^*}^{-1} \boldsymbol{\mu}_J^*,
\end{aligned}$$

where the scaling factor

$$c^\lambda = \frac{\sum_{K \in \mathcal{J}(\mathbf{\Phi}^*)\setminus\{\mathcal{I}\}} \det(\mathbf{\Lambda}_K^*) \cdot \det(\Phi_K^*)^2}{\sum_{K \in \mathcal{J}(\mathbf{\Phi}^*)\setminus\{\mathcal{I}\}} \det(\mathbf{\Lambda}_K^*) \cdot \det(\Phi_K^*)^2 + \lambda^d}$$

is strictly positive for any $\mathbf{\Lambda} \in \mathcal{P}(\Delta_K)$.

Hence, $\boldsymbol{\mu} \in \mathcal{R}_k^\lambda$ if and only if $c^\lambda {\Phi_J^*}^{-1}\boldsymbol{\mu}_J^* \in \mathbf{\Theta}_k$ for all $J \in \mathcal{J}(\mathbf{\Phi}^*) \setminus \{\mathcal{I}\}$. Since $\mathbf{\Theta}_k$ is a convex cone, scaling by a positive constant does not affect membership, so it suffices to check that ${\Phi_J^*}^{-1}\boldsymbol{\mu}_J^* \in \mathbf{\Theta}_k$ for all such J. □

E.2 LinUCB Algorithm: Discussions and Results

We now present an analysis of LinUCB [Abbasi-Yadkori et al., 2011] without a forced exploration phase and under the standard assumptions. That is without the Assumption 3.6. This is the result that one would find, say in, Abbasi-Yadkori et al. [2011]. But we present here without the realizability assumption.

Assumption E.2 (Bounded Features)**.** We assume that for any arm i in the arm set $[K]$, the corresponding feature φ_i is bounded in the l^2 norm by L, that is,

$$\|\varphi_i\|_2 \leqslant L \quad \forall\, i \in [K].$$

Assumption E.3 (Regularization)**.** We assume that the regularizer λ, so chosen satisfies

$$\lambda \geqslant \max\{1, L^2\}\ ,$$

where L is as defined in Assumption E.2

Theorem E.4. *Given a feature matrix $\mathbf{\Phi}$ satisfying Assumptions E.2, regularization parameter λ satisfying Assumption E.3, and any bandit parameter $\boldsymbol{\mu}$ which is an interior point of the regularized robust observation region, $\mathcal{R}^{\lambda}_{\mathrm{OPT}(\boldsymbol{\mu})}$ and satisfies Assumption 3.7, the LinUCB algorithm, [Abbasi-Yadkori et al., 2011], achieves regret of at most $\tilde{O}(d\sqrt{t})$.*

That is, with $\beta_t(\delta)$ set as $2R^2 \log\left(\frac{(\lambda + tL^2/d)^{d/2}}{\delta}\right)$ for any $\delta > 0$, we have with probability at least $1-\delta$, for $t \geqslant 1$,

$$\sum_{s=1}^{t} \mu_{\mathrm{OPT}(\boldsymbol{\mu})} - \mu_{A_s} \leqslant \frac{4\sqrt{t}R\Delta_{\max}}{\Delta_{\mathrm{rob}}} \sqrt{\log\left(\frac{(1+tL^2/\lambda d)^{d/2}}{\delta}\right)} \sqrt{\log\left(1+tL^2/\lambda d\right)^{d}},$$

where Δ_{rob} is as defined above and $\Delta_{\max}$ is defined as the worst sub-optimal gap, that is, $\Delta_{\max} = \max_{i\in[K]} \mu_{\mathrm{OPT}(\boldsymbol{\mu})} - \mu_i$.

We provide a sketch of the proof: As in Lemma, 3.4, we have the following result for any bandit instance $\boldsymbol{\mu}$ belonging to the interior of the regularized robust observation region $\mathcal{R}^{\lambda}_{\mathrm{OPT}(\boldsymbol{\mu})}$.

Lemma E.5 (Lower Bound of per instant Regret in the Function Space)**.** *If $\boldsymbol{\mu}$ is an interior point of the regularized robust observation region, then there exists a constant $\Delta_{\mathrm{rob}} > 0$, depending only on the features $\mathbf{\Phi}$, the mean instance $\boldsymbol{\mu}$, and the regularizer λ, such that for any sampling distribution $\mathbf{\Lambda} \in \Delta_K$, we have*

$$\varphi^{\top}_{\mathrm{OPT}(\boldsymbol{\mu})} \mathbf{P}^{\mathbf{\Lambda}}_{\lambda}(\boldsymbol{\mu}) - \varphi_i^{\top} \mathbf{P}^{\mathbf{\Lambda}}_{\lambda}(\boldsymbol{\mu}) \geqslant \Delta_{\mathrm{rob}} \quad \forall\, i \neq \mathrm{OPT}(\boldsymbol{\mu}).$$

Proof. If $\boldsymbol{\mu}$ is an interior point of the regularized robust observation region $\mathcal{R}^{\lambda}_{\mathrm{OPT}(\boldsymbol{\mu})}$, then by definition, for all sampling distributions $\mathbf{\Lambda}$, the regularized oracle estimate $\mathbf{P}^{\mathbf{\Lambda}}_{\lambda}(\boldsymbol{\mu})$ lies in the robust parameter region $\mathbf{\Theta}_{\mathrm{OPT}(\boldsymbol{\mu})}$.

Since $\boldsymbol{\mu}$ is an interior point, there exists a positive margin between the optimal arm and every sub-optimal arm in the value space induced by the oracle estimate. That is, for every $i \neq \mathrm{OPT}(\boldsymbol{\mu})$ and any $\mathbf{\Lambda} \in \Delta_K$,

$$\left(\varphi_{\mathrm{OPT}(\boldsymbol{\mu})} - \varphi_i\right)^{\top} \mathbf{P}^{\mathbf{\Lambda}}_{\lambda}(\boldsymbol{\mu}) \geqslant \min_{\theta \in \mathbf{\Theta}_{\mathrm{OPT}(\boldsymbol{\mu})}} \left(\varphi_{\mathrm{OPT}(\boldsymbol{\mu})} - \varphi_i\right)^{\top} \theta \triangleq \Delta_i > 0.$$

Define the minimum sub-optimality gap as

$$\Delta_{\mathrm{rob}} = \min_{i \in [K] \setminus \{\mathrm{OPT}(\boldsymbol{\mu})\}} \Delta_i.$$

Then for any sampling distribution $\mathbf{\Lambda}$ and any sub-optimal arm i, we have

$$\varphi^{\top}_{\mathrm{OPT}(\boldsymbol{\mu})} \mathbf{P}^{\mathbf{\Lambda}}_{\lambda}(\boldsymbol{\mu}) - \varphi_i^{\top} \mathbf{P}^{\mathbf{\Lambda}}_{\lambda}(\boldsymbol{\mu}) \geqslant \Delta_{\mathrm{rob}} > 0,$$

which completes the proof. □

We, also have the following result, as in Lemma C.2,

Lemma E.6 (Upper Bound of Sub-Optimal Plays)**.** *Given a feature matrix $\mathbf{\Phi}$ satisfying Assumptions E.2, regularization parameter λ satisfying Assumption E.3, and any bandit parameter $\boldsymbol{\mu}$ which is an interior point of the regularized robust observation region, $\mathcal{R}^{\lambda}_{\mathrm{OPT}(\boldsymbol{\mu})}$ and satisfies Assumption 3.7, the LinUCB algorithm, [Abbasi-Yadkori et al., 2011], plays sub-optimally at most $\tilde{O}(d\sqrt{t})$.*

That is, with $\beta_t(\delta)$ set as $2R^2 \log\left(\frac{(1+tL^2/\lambda d)^{d/2}}{\delta}\right)$ for any $\delta > 0$, we have with probability at least $1-\delta$, for $t \geqslant 1$

$$\sum_{s=1}^{t} \mathbb{1}\{A_s \neq \mathrm{OPT}(\boldsymbol{\mu})\} \leqslant \frac{4\sqrt{t}R}{\Delta_{\mathrm{rob}}} \sqrt{\log\left(\frac{(1+tL^2/\lambda d)^{d/2}}{\delta}\right)} \sqrt{\log\left(1+tL^2/\lambda d\right)^{d}},$$

where Δ_{rob} is as defined in the previously.

Proof. By Lemma F.1, with probability at least $1-\delta$, for all $t \geqslant 1$,

$$\left\|\mathbf{P}^{\mathbf{\Lambda_t}}_{\lambda}(\widehat{\boldsymbol{\mu}}_{\boldsymbol{t}}) - \mathbf{P}^{\mathbf{\Lambda_t}}_{\lambda}(\boldsymbol{\mu})\right\|_{V_t} \leqslant R\sqrt{2\log\left(\frac{\det(V_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} \triangleq \sqrt{\beta_t(\delta)}.$$

Thus, the per-round value regret satisfies

$$\varphi^{\top}_{\mathrm{OPT}(\boldsymbol{\mu})}\mathbf{P}^{\mathbf{\Lambda_t}}_{\lambda}(\boldsymbol{\mu}) - \varphi^{\top}_{A_t}\mathbf{P}^{\mathbf{\Lambda_t}}_{\lambda}(\boldsymbol{\mu}) \leqslant 2\|\varphi_{A_t}\|_{V_t^{-1}}\sqrt{\beta_t(\delta)}.$$

Using Lemma F.2, we bound $\beta_t(\delta) \leqslant 2R^2\log\left(\frac{(1+tL^2/\lambda d)^{d/2}}{\delta}\right)$. Then, following Lemma C.2, the cumulative function-space regret is:

$$\begin{aligned}\sum_{s=1}^{t}\left[\varphi^{\top}_{\mathrm{OPT}(\boldsymbol{\mu})}\mathbf{P}^{\mathbf{\Lambda_s}}(\boldsymbol{\mu}) - \varphi^{\top}_{A_s}\mathbf{P}^{\mathbf{\Lambda_s}}(\boldsymbol{\mu})\right] &\leqslant 2\sqrt{2}R\sum_{s=1}^{t}\|\varphi_{A_s}\|_{V_s^{-1}}\sqrt{\log\left(\frac{(1+sL^2/\lambda d)^{d/2}}{\delta}\right)} \\ &\leqslant 4\sqrt{t}R\sqrt{\log\left(\frac{(1+tL^2/\lambda d)^{d/2}}{\delta}\right)}\sqrt{\log(1+tL^2/\lambda d)^{d}},\end{aligned}$$

where the last step uses Lemmas F.3 and F.2. The rest follows as in Lemma C.2. □

The proof of Theorem E.4 is completed analogously.

F Technical Lemmas

Lemma F.1 (Self Normalized Bound for Vector Valued Martingales Abbasi-Yadkori et al. [2011])**.** *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=1}^{\infty}$ be a real valued stochastic process such that η_t is $\mathcal{F}_t$- measurable and η_t is conditionally R-sub-Gaussian for some $R > 0$, i.e.*

$$\forall \lambda \in \mathbb{R} \quad \mathbf{E}[e^{\lambda\eta_t} \mid \mathcal{F}_{t-1}] \leqslant \exp\left(\frac{\lambda^2 R^2}{2}\right).$$

Let $\{\varphi_t\}_{t=1}^{\infty}$ be a $\mathbb{R}^d$-valued stochastic process such that φ_t is $\mathcal{F}_{t-1}$-measurable. Assume V is a $d \times d$ positive definite matrix and for any $t \geqslant 0$ define

$$V_t = V + \sum_{s=1}^{t}\varphi_s\varphi_s^{\top} \qquad S_t = \sum_{s=1}^{t}\eta_s\varphi_s.$$

Then for any $\delta > 0$, with probability at least $1-\delta$, for all $t \geqslant 0$,

$$\|S_t\|^2_{V_t^{-1}} \leqslant 2R^2\log\left(\frac{\det V_t^{1/2}\det V^{-1/2}}{\delta}\right).$$

Lemma F.2 (Determinant Trace Inequality Abbasi-Yadkori et al. [2011])**.** *Suppose $\{\varphi_s\}_{s=1}^{t} \subset \mathbb{R}^d$ be such that $\|\varphi_s\|_2 \leqslant L \;\forall\; s \in [t]$. Let $V_t = \lambda I + \sum_{s=1}^{t}\varphi_s\varphi_s^{\top}$ for some $\lambda > 0$. Then*

$$\det V_t \leqslant (\lambda + tL^2/d)^d$$

Lemma F.3 (Abbasi-Yadkori et al. [2011]). *Let $\{\varphi_t\}_{t=1}^{\infty}$ be a sequence in $\mathbb{R}^d$, V a $d \times d$ positive definite matrix and $V_t = V + \sum_{s=1}^{t} \varphi_s \varphi_s^\top$. Then*

$$\log\Big(\frac{\det V_t}{\det V}\Big) \leqslant \sum_{s=1}^{t} \|\varphi_s\|^2_{V_s^{-1}}.$$

Moreover, if $\|\varphi_s\|_2 \leqslant L$ for all s and if $\lambda_{\min}(V) \geqslant \max\{1, L^2\}$, then

$$\sum_{s=1}^{t} \|\varphi_s\|^2_{V_s^{-1}} \leqslant 2\log\Big(\frac{\det V_t}{\det V}\Big).$$

Lemma F.4 (Forsgren [1996]). *Let Φ be a $K \times d$ full column matrix, let $\boldsymbol{\mu}$ be a K dimensional matrix. Let $\boldsymbol{\Lambda}$ be the set of all positive semi-definite diagonal matrices such that $\Phi^\top \Lambda \Phi$ is invertible for any $\Lambda \in \boldsymbol{\Lambda}$, that is*

$$\boldsymbol{\Lambda} = \{\Lambda \in \mathbb{R}^{K\times K} \ : \ \Lambda \text{ is diagonal and positive semi-definite } \wedge \Phi^\top \Lambda \Phi \text{ is invertible}\}.$$

Then the solution to the weighted least-squares problem lies in the convex hull of the basic solutions, that is,

$$\Big(\Phi^\top \Lambda \Phi\Big)^{-1} \Phi^\top \Lambda \mu = \sum_{J \in \mathcal{J}(\Phi)} \Big(\frac{\det \Lambda_J \det \Phi_J^2}{\sum_{K \in \mathcal{J}(\Phi)} \det \Lambda_K \det \Phi_K^2}\Big) \Phi_J^{-1} \boldsymbol{\mu}_J,$$

where $\mathcal{J}(\Phi)$ is the is the set of column indices associated with non-singular $d \times d$ sub-matrices of Φ.

Lemma F.5 (sub-Gaussian Concentration). *Assume $\{X_i - \mu\}_{i=1}^{n}$ are n independent σ-sub-Gaussian random variables, then*

$$\mathbb{P}\left[|\widehat{\mu}_n - \mu| \geqslant \varepsilon\right] \leqslant 2\exp(-\frac{n\varepsilon^2}{2\sigma^2}) \ ,$$

where $\widehat{\mu}_n = \frac{\sum_{i=1}^{N} X_i}{n}$.

Lemma F.6 (Bernstein inequality). *Let $\{T_i\}_{i=1}^{n}$ be random variables in $[0,1]$, such that*

$$\sum_{i=1}^{n} \mathbb{V}[T_i \mid T_{i-1}, T_{i-2}, \cdots, 1] = \sigma^2 \ ,$$

then

$$\mathbb{P}\left[\sum_{i=1}^{n} T_i \geqslant \mathbb{E}[\sum_{i=1}^{n} T_i] + \varepsilon\right] \leqslant \exp\Big\{-\frac{\varepsilon^2/2}{\sigma^2 + \varepsilon/2}\Big\}.$$

G Extended Related Work

1. **Linear Bandits:** The monograph by Lattimore and Szepesvári [2020] offers a comprehensive treatment of bandit algorithms. Stochastic linear bandits with finitely many arms were first explored in Auer [2002]. Algorithms based on optimism [Dani et al., 2008, Abbasi-Yadkori et al., 2011] and Thompson Sampling [Agrawal and Goyal, 2012, Abeille and Lazaric, 2017] achieve $\widetilde{O}(\sqrt{T})$ regret. Phased elimination with optimal design yields $\widetilde{O}(\sqrt{dT})$ regret [Lattimore and Szepesvári, 2020].
2. **Contextual Linear Bandits:** In contextual bandits, the arm set changes over time. Classical algorithms such as SupLinRel [Auer, 2002] and SupLinUCB [Chu et al., 2011] achieve $\widetilde{O}(\sqrt{dT})$ regret. In practice, LinUCB [Abbasi-Yadkori et al., 2011] and Thompson Sampling [Agrawal and Goyal, 2013] also yield sub-linear regret.
3. **Misspecified Bandits with Linear Regret:** Model misspecification in linear bandits was introduced in Ghosh et al. [2017], where LinUCB was shown to suffer linear regret in the worst case. Their RLB algorithm combines linear and finite-armed bandits to achieve $O(\min\{\sqrt{K}, d\}\sqrt{T})$ regret. Under l_∞ model error ε, Lattimore et al. [2020] and Foster and Rakhlin [2020] show regret bounds of $O(d\sqrt{T} + \varepsilon\sqrt{dT})$ using phased elimination and contextual Thompson Sampling, respectively. Zanette et al. [2020] reports similar findings. Other formulations of misspecification (e.g., expected squared loss [Kumar Krishnamurthy et al., 2021] or empirical deviation [Foster et al., 2020]) also lead to linear regret.
4. **Misspecified Bandits with Sublinear Regret:** A positive line of work focuses on conditions under which misspecified bandits yield sublinear regret. Liu et al. [2023] show that LinUCB achieves sublinear regret when the misspecification is small relative to suboptimality gaps. Zhang et al. [2023] extend this to the contextual setting with a phased elimination algorithm that requires gap knowledge.