No-Regret Linear Bandits beyond Realizability

Abstract

We study linear bandits when the underlying reward function is *not* linear. Existing work relies on a uniform misspecification parameter ϵ that measures the sup-norm error of the best linear approximation. This results in an unavoidable linear regret whenever $\epsilon > 0$. We describe a more natural model of misspecification which only requires the approximation error at each input x to be proportional to the suboptimality gap at x. It captures the intuition that, for optimization problems, near-optimal regions should matter more and we can tolerate larger approximation errors in suboptimal regions. Quite surprisingly, we show that the classical LinUCB algorithm — designed for the realizable case — is automatically robust against such gap-adjusted misspecification. It achieves a near-optimal $\sqrt{T}$ regret for problems that the best-known regret is almost linear in time horizon T. Technically, our proof relies on a novel self-bounding argument that bounds the part of the regret due to misspecification by the regret itself.

1 INTRODUCTION

Stochastic linear bandit is a classical problem of online learning and decision-making with many influential applications, e.g., A/B testing [Claeys et al., 2021], recommendation systems [Chu et al., 2011], advertisement placements [Wang et al., 2021], clinical trials [Moradipari et al., 2020], hyperparameter tuning [Alieva et al., 2021], and new material discovery [Katz-Samuels et al., 2020].

More formally, stochastic bandits is a sequential game between an agent who chooses a sequence of actions $x_0, ..., x_{T-1} \in \mathcal{X}$ and nature who decides on a sequence of noisy observations (rewards) $y_0, ..., y_{T-1}$ according to $y_t = f_0(x_t) + \textit{noise}$ for some underlying function f_0. The goal of the learner is to minimize the *cumulative regret* the agent experiences relative to an oracle who knows the best action to choose ahead of time, i.e.,

$$R_T(x_0, ..., x_{T-1}) = \sum_{t=0}^{T-1} r_t = \sum_{t=0}^{T-1} \max_{x \in \mathcal{X}} f_0(x) - f_0(x_t),$$

where r_t is called *instantaneous regret*.

Despite being highly successful in the wild, existing theory for stochastic linear bandits (or more generally learning-oracle based bandits problems) relies on a *realizability* assumption, i.e., the learner is given access to a function class $\mathcal{F}$ such that the true expected reward $f_0 : \mathcal{X} \to \mathbb{R}$ satisfies that $f_0 \in \mathcal{F}$. Realizability is considered one of the strongest and most restrictive assumptions in the standard statistical learning setting, but in the bandits theory literature the common sentiment is that it is mild and acceptable, despite the "elephant in the room" that everyone sees but voluntarily ignores — Realizability is never true in practice! Why is this the case? The argument to justify the realizability assumption is legitmate: all known attempts to deviate from the realizability assumption results in a regret that grows linearly with T.

In practical applications, it is often observed that feature-based representation of the actions with function approximations in estimating the reward can result in very strong policies even if the estimated reward functions are far from being correct. Intuitively, it should be sufficient for the estimated reward function to clearly *differentiate* good actions from bad ones, rather than requiring the function to perfectly estimate the rewards numerically.

Contributions. In this paper, we formalize this intuition by defining a new family of misspecified bandits problems based on a condition that adjusts the need for an accurate approximation pointwise at every $x \in \mathcal{X}$ according to the suboptimality gap at x. Unlike the existing misspecified linear bandits problems with a linear regret, our problem admits a nearly optimal $\tilde{O}(\sqrt{T})$ regret despite being heavily misspecified. Specifically:

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

- We define *ρ-gap-adjusted misspecified* (ρ-GAM) function approximations and characterize how they preserve important properties of the true function that are relevant for optimization.
- We show that the classical LinUCB algorithm [Abbasi-yadkori et al., 2011] can be used *as is* (up to some mild hyperparameter) to achieve an $\tilde{O}(\sqrt{T})$ regret under a moderate level of gap-adjusted misspecification ($\rho \leq O(1/\sqrt{\log T})$). In comparison, the regret bound one can obtain under the corresponding uniform-misspecification setting is only $\tilde{O}(T/\sqrt{\log T})$. This represents an exponential improvement in the average regret metric R_T/T.

To the best of our knowledge, the suboptimality-gap-adjusted misspecification problem was not studied before and we are the first to obtain $\sqrt{T}$-style regrets without a realizability assumption.

Technical novelty. Due to misspecification, we have technical challenges that appear in bounding the instantaneous regret and parameter uncertainty region. We tackle the challenge by self bounding trick, i.e., bounding the instantaneous regret by the instantaneous regret itself, which can be of independent interest in more settings, e.g., Gaussian process bandit optimization and reinforcement learning.

2 RELATED WORK

Linear bandits have been well studied for a long time. It was first introduced in Abe and Long [1999]. Then Auer et al. [2002] proposed the upper confidence bound to study linear bandits where the number of actions is finite. Based on it, Dani et al. [2008] proposed an algorithm based on confidence ellipsoids and then Abbasi-yadkori et al. [2011] simplified the proof with a novel self-normalized martingale bound. Later Chu et al. [2011] proposed a simpler and more robust linear bandit algorithm and showed $\tilde{O}(\sqrt{dT})$ regret cannot be improved by proving a lower bound. See Lattimore and Szepesvári [2020] for a detailed survey of linear bandits.

In terms of misspecification, Ghosh et al. [2017] first studied the misspecified linear bandit with a fixed action set. They found that LinUCB [Abbasi-yadkori et al., 2011] is not robust when misspecification is large. They showed that in a favourable case when one can test the linearity of the reward function, their RLB algorithm is able to switch between the linear bandit algorithm and finite-armed bandit algorithm to address misspecification issue and achieve the $\tilde{O}(\min\{\sqrt{K}, d\}\sqrt{T})$ regret where K is number of arms.

The most studied setting of model misspecification is uniform misspecification where the ℓ_∞ distance between the best-in-class function and the true function is always upper bounded by some parameter ϵ. Under this definition, Lattimore et al. [2020] proposed the optimal design-based phased elimination algorithm for misspecified linear bandits and achieved $\tilde{O}(d\sqrt{T}+\epsilon\sqrt{d}T)$ regret when number of actions is infinite. They also found that with modified confidence band in LinUCB, LinUCB is able to achieve the same regret. With the same misspecification model, Foster and Rakhlin [2020] studied contextual bandit with regression oracle, Neu and Olkhovskaya [2020] studied multi-armed linear contextual bandit, and Zanette et al. [2020] studied misspecified contextual linear bandits after reduction of the algorithm. All of their results suffer from linear regrets. Later Bogunovic and Krause [2021] studied misspecified Gaussian process bandit optimization problem and achieved $\tilde{O}(d\sqrt{T}+\epsilon\sqrt{d}T)$ regret when linear kernel is used in Gaussian process. Moreover, their lower bound shows that $\tilde{\Omega}(\epsilon T)$ term is unavoidable in this setting.

Besides uniform misspecification, there are some work considered different definitions of misspecification in contextual bandits. Krishnamurthy et al. [2021] defines misspecification error as expected squared error between true function and best-in-class function where expectation is taken over distribution of context space and action space. Foster et al. [2020] considered average misspecification, which is weaker than uniform misspecification and allows tighter regret bound. However, they also have linear regrets and their results do not directly apply to our problem because our action space is unbounded. Our work is different from all related work mentioned above because we are working under a newly defined misspecifiation condition and show that LinUCB is a no-regret algorithm in this case.

3 PRELIMINARIES

3.1 NOTATIONS

Let $[n]$ denote the integer set $\{1, 2, ..., n\}$. The algorithm runs in T rounds in total. Let f_0 denote the true function, so the maximum function value is defined as $f^* = \max_{x\in\mathcal{X}} f_0(x)$ and the maximum point is defined as $x^* = \operatorname{argmax}_{x\in\mathcal{X}} f_0(x)$. Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$ denote the domain and range of f_0. We use $\mathcal{W}$ to denote the parameter class of a family of linear functions $\mathcal{F} := \{f_w : \mathcal{X} \to \mathcal{Y} | w \in \mathcal{W}\}$ where $f_w(x) = w^\top x$. $\|w\|_2 \leq C_w, \forall w \in \mathcal{W}$ and $\|x\|_2 \leq C_b, \forall x \in \mathcal{X}$. For a vector x, its ℓ_2 norm is denoted by $\|x\|_2 = \sqrt{\sum_{i=1}^d x_i^2}$ and for a matrix A its operator norm is denoted by $\|A\|_{\text{op}}$. For a vector x and a square matrix A, define $\|x\|_A^2 = x^\top Ax$.

3.2 PROBLEM SETUP

We consider the following optimization problem:

$$x_* = \operatorname*{argmax}_{x\in\mathcal{X}} f_0(x),$$

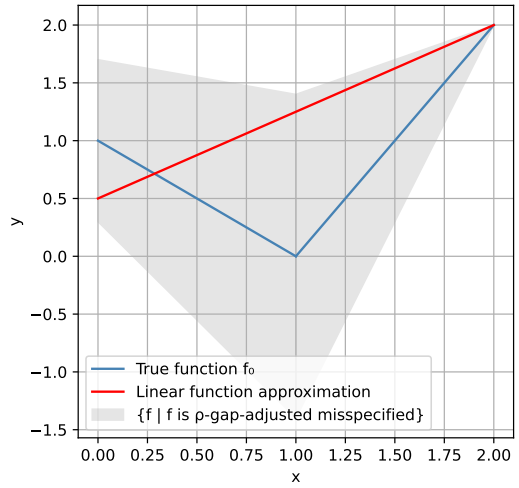

Figure 1: An example of ρ-gap-adjusted misspecification (Definition 1) in 1-dimension where $\rho = 0.7$. The blue line shows a non-linear true function and the gray region shows the gap-adjusted misspecified function class. Note the vertical range of gray region at a certain point x depends on the suboptimal gap. For example, at $x = 1$ suboptimal gap is 2 and the vertical range is $4\rho = 2.8$. The red line shows a feasible linear function that is able to optimize the true function by taking $x_* = 2$.

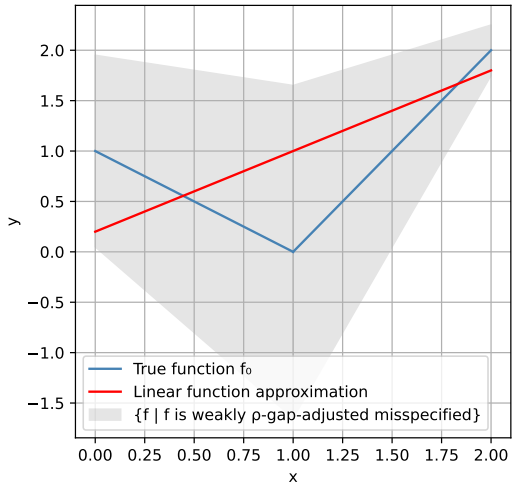

Figure 2: An example of weaker ρ-gap-adjusted misspecification (Definition 3) in 1-dimension where $\rho = 0.7$. The difference to Figure 1 is at optimal point $x_* = 2$ where some tolerence is allowed, but optimizing the linear function ($y = 0.8x + 0.2$, red line) is still able to optimize the true function (blue line) by taking $x_* = 2$.

where f_0 is the true function which might not be linear in $\mathcal{X}$. We want to use a linear function $f_w = w^\top x \in \mathcal{F}$ to approximate f_0 and maximize f_0. At time $0 \le t \le T-1$, after querying a data point x_t, we will receive a noisy feedback:

$$y_t = f_0(x_t) + \eta_t, \tag{1}$$

where η_t is independent, zero-mean, and σ-sub-Gaussian.

The major highlight of our study is that we do not rely on the popular *realizability* assumption (*i.e.* $f_0 \in \mathcal{F}$) that is frequently assumed in the existing function approximation literature. Alternatively, we propose the following gap-adjusted misspecification condition.

Definition 1 (ρ-gap-adjusted misspecification)**.** *We say a function f is a ρ-gap-adjusted misspecified (or ρ-GAM in short) approximation of f_0 if for parameter $0 \le \rho < 1$,*

$$\sup_{x \in \mathcal{X}} \left| \frac{f(x) - f_0(x)}{f^* - f_0(x)} \right| \le \rho.$$

We say function class $\mathcal{F} = \{f_w | w \in \mathcal{W}\}$ satisfies ρ-GAM in short) for f_0, if there exists $w^* \in \mathcal{W}$ such that f_{w_*} is a ρ-GAM approximation of f_0.

Observe that when $\rho = 0$, this recovers the standard realizability assumption, but when $\rho > 0$ it could cover many misspecified function classes.

Figure 1 shows a 1-dimensional example with $f_w(x) = 0.75x + 0.5$ and piece-wise linear function $f_0(x)$ that satisfies local misspecification. With Definition 1, we have the following proposition.

Proposition 2. *Let f be a ρ-GAM approximation of f_0 (Definition 1). Then it holds:*

- *(Preservation of maximizers)*

$$\operatorname*{argmax}_x f(x) = \operatorname*{argmax}_x f_0(x).$$

- *(Preservation of max value)*

$$\max_{x \in \mathcal{X}} f(x) = f^*.$$

- *(Self-bounding property)*

$$|f(x) - f_0(x)| \le \rho(f^* - f_0(x)) = \rho r(x).$$

This tells f_{w_*} and f_0 coincide on the same global maximum points and the same global maxima if Definition 1 is satisfied, while allowing f_{w_*} and f_0 to be different (potentially large) at other locations. Therefore, Definition 1 is a "local" assumption that does not require f_{w_*} to be uniformly close to f_0 (e.g. the "uniform" misspecification assumption $\sup_{x \in \mathcal{X}} |f_{w_*}(x) - f_0(x)| \le \rho$).

In addition, we can modify Definition 1 with a slightly weaker condition that only requires $\operatorname{argmax}_x f_{w_*}(x) = \operatorname{argmax}_x f_0(x)$ but not necessarily $\max_{x \in \mathcal{X}} f_{w_*}(x) = f^*$.

Definition 3 (Weaker ρ-gap-adjusted misspecification)**.** *Denote $f_w^* = \max_{x\in\mathcal{X}} f_w(x)$. There exists $w \in \mathcal{W}$ such that for a parameter $0 \le \rho < 1$,*

$$\sup_{x\in\mathcal{X}} \left| \frac{f_w(x) - f_w^* + f^* - f_0(x)}{f^* - f_0(x)} \right| \le \rho.$$

Remark 4. *See Figure 2 for an example satisfying Definition 3. Both Definition 1 and Definition 3 are defined in the generic way that does not require any assumption on the parametric form of f_w. While in this paper we focus on the linear bandit setting, this notion can be applied to arbitrary parametric function approximation learning problem. In this paper, we stick to Definition 1 and linear function approximation for conciseness and clarity.*

3.3 ASSUMPTIONS

Assumption 5 (Boundedness)**.** *For any $x \in \mathcal{X}$, $\|x\|_2 \le C_b$. For any $w \in \mathcal{W}$, $\|w\|_2 \le C_w$. Moreover, for any $x, \tilde{x} \in \mathcal{X}$, the true expected reward function $|f_0(x) - f_0(\tilde{x})| \le F$.*

These are mild assumptions that we assume for convenience. Relaxations of these are possible but not the focus of this paper. Note that the additional assumption is not required when f_0 is realizable.

Assumption 6. *Suppose $\mathcal{X} \in \mathbb{R}^d$ is a compact set, and all the global maximizers of f_0 live on the $d-1$ dimensional hyperplane. i.e., $\exists a \in \mathbb{R}^d, b \in \mathbb{R}^1$, s.t.*

$$\operatorname*{argmax}_{x\in\mathcal{X}} f_0(x) \subset \{x \in \mathbb{R}^d : x^\top a = b\}.$$

For instance, when $d = 1$, the above reduces to that f_0 has a unique maximizer. This is a compatibility assumption for Definition 1, since any linear function that violates Assumption 6 will not satisfy Definition 1.

In addition, to obtain an $\tilde{O}(\sqrt{T})$ regret, for any finite sample T, we require the following condition.

Assumption 7 (Low misspecification)**.** *The linear function class is a ρ-GAM approximation of f_0 with*

$$\rho < \frac{1}{8d\sqrt{\log\left(1 + \frac{TC_b^2C_w^2}{d\sigma^2}\right)}} = O\left(\frac{1}{d\sqrt{\log T}}\right). \quad (2)$$

The condition is required for technical reasons. Relaxing this condition for LinUCB may require fundamental breakthroughs that knock out logarithmic factors from its regret analysis. This will be further clarified in the proof. In general, however, we conjecture that this condition is not needed and there are algorithms that can achieve $\tilde{O}(\sqrt{T}/(1-\rho))$ regret for any $\rho < 1$, but a new algorithm needs to be designed.

While this assumption may suggest that we still require realizability in a truly asymptotic world, handling a $O(1/\sqrt{\log T})$ level of misspecification is highly nontrivial in finite sample. For instance, if T is a trillion, $1/\sqrt{\log(1e12)} \approx 0.19$. This means that for most practical cases, LinUCB is able to tolerate a constant level of misspecification under the GAM model.

3.4 LINUCB ALGORITHM

We will focus analyzing the classical Linear Upper Confidence Bound (LinUCB) algorithm due to [Dani et al., 2008, Abbasi-yadkori et al., 2011], shown below.

Algorithm 1 LinUCB [Abbasi-yadkori et al., 2011]

Input: Predefined sequence β_t for $t = 1, 2, 3, ...$; Set $\lambda = \sigma^2/C_w^2$ and $\text{Ball}_0 = \mathcal{W}$.

1: **for** $t = 0, 1, 2, ...$ **do**
2: Select $x_t = \operatorname{argmax}_{x\in\mathcal{X}} \max_{w\in\text{Ball}_t} w^\top x$.
3: Observe $y_t = f_0(x_t) + \eta_t$.
4: Update

$$\Sigma_{t+1} = \lambda I + \sum_{i=0}^{t} x_i x_i^\top \text{ where } \Sigma_0 = \lambda I. \quad (3)$$

5: Update

$$\hat{w}_{t+1} = \operatorname*{argmin}_x \lambda\|w\|_2^2 + \sum_{i=0}^{t}(w^\top x_i - y_i)_2^2. \quad (4)$$

6: Update $\text{Ball}_{t+1} = \{w | \|w - \hat{w}_{t+1}\|_{\Sigma_{t+1}}^2 \le \beta_{t+1}\}$.
7: **end for**

4 MAIN RESULTS

In this section, we show that the classical LinUCB algorithm [Abbasi-yadkori et al., 2011] works in ρ-gap-adjusted misspecified linear bandits and achieves cumulative regret at the order of $\tilde{O}(\sqrt{T}/(1-\rho))$. The following theorem shows the cumulative regret bound.

Theorem 8. *Suppose Assumptions 5, 6, and 7 hold. Set*

$$\beta_t = 8\sigma^2\left(1 + d\log\left(1 + \frac{tC_b^2C_w^2}{d\sigma^2}\right) + 2\log\left(\frac{\pi^2t^2}{3\delta}\right)\right). \quad (5)$$

Then w.p. $> 1-\delta$ for simultaneously for all $T = 1, 2, ...$

$$R_T \le F + \sqrt{\frac{8(T-1)\beta_{T-1}d}{(1-\rho)^2}\log\left(1 + \frac{TC_b^2C_w^2}{d\sigma^2}\right)}.$$

Remark 9. *The cumulative regret bound shows that LinUCB achieves $\tilde{O}(\sqrt{T})$ cumulative regret bound and thus it is a no-regret algorithm in ρ-gap-adjusted misspecified linear bandits. In contrast, LinUCB can only achieve $\tilde{O}(\sqrt{T} + \epsilon T)$*

regret in uniformly misspecified linear bandits. Even if $\epsilon = \tilde{O}(1/\sqrt{\log T})$, the resulting regret $\tilde{O}(T/\sqrt{\log T})$ is still exponentially worse than ours.

Proof. By definition of cumulative regret, function range absolute bound F, and Cauchy-Schwarz inequality,

$$\begin{aligned} R_T &= r_0 + \sum_{t=1}^{T-1} r_t \\ &\leq F + \sqrt{\left(\sum_{t=1}^{T-1} 1\right)\left(\sum_{t=1}^{T-1} r_t^2\right)} \\ &= F + \sqrt{(T-1)\sum_{t=1}^{T-1} r_t^2}. \end{aligned}$$

Observe that the choice of β_t is monotonically increasing in t. Also by Lemma 14, we get that with probability $1-\delta$, $w^* \in \text{Ball}_t \forall t = 1, 2, 3, ...$, which verifies the condition to apply Lemma 12 simultaneously for all $T = 1, 2, 3, ...$, thereby completing the proof. □

4.1 REGRET ANALYSIS

The proof follows the LinUCB analysis closely. The main innovation is a self-bounding argument that controls the regret due to misspecification by the regret itself. This appears in Lemma 11 and then again in the proof of Lemma 14.

Before we proceed, let Δ_t denote the deviation term of our linear function from the true function at x_t, formally,

$$\Delta_t = f_0(x_t) - w_*^\top x_t, \tag{6}$$

And our observation model (eq. (1)) becomes

$$y_t = f_0(x_t) + \eta_t = w_*^\top x_t + \Delta_t + \eta_t. \tag{7}$$

Moreover, we have the following lemma showing the property of deviation term Δ_t.

Lemma 10 (Bound of deviation term)**.** $\forall t \in \{0, 1, \ldots, T-1\}$,

$$|\Delta_t| \leq \frac{\rho}{1-\rho} w_*^\top (x_* - x_t).$$

Proof. Recall the definition of deviation term in eq. (6):

$$\Delta_t = f_0(x_t) - w_*^\top x_t.$$

By Definition 1, $\forall t \in \{0, 1, \ldots, T-1\}$,

$$\begin{aligned} -\rho(f^* - f_0(x_t)) \leq \Delta_t \leq \rho(f^* - f_0(x_t)) \\ -\rho(f^* - w_*^\top x_t - \Delta_t) \leq \Delta_t \leq \rho(f^* - w_*^\top x_t - \Delta_t) \\ -\rho(w_*^\top x_* - w_*^\top x_t - \Delta_t) \leq \Delta_t \leq \rho(w_*^\top x_* - w_*^\top x_t - \Delta_t) \\ \frac{-\rho}{1-\rho}(w_*^\top x_* - w_*^\top x_t) \leq \Delta_t \leq \frac{\rho}{1+\rho}(w_*^\top x_* - w_*^\top x_t), \end{aligned}$$

where the third line is by Proposition 2 and the proof completes by taking the absolute value of the lower and upper bounds. □

Next, we prove instantaneous regret bound and its sum of squared regret version in the following two lemmas:

Lemma 11 (Instantaneous regret bound)**.** *Define* $u_t := \|x_t\|_{\Sigma_t^{-1}}$*, assume* $w_* \in \text{Ball}_t$ *then for each* $t \geq 1$

$$r_t \leq \frac{2\sqrt{\beta_t} u_t}{1-\rho}.$$

Proof. By definition of instantaneous regret,

$$\begin{aligned} r_t &= f^* - f_0(x_t) \\ &= w_*^\top x_* - (w_*^\top x_t + \Delta(x_t)) \\ &\leq w_*^\top x_* - w_*^\top x_t + \rho(f^* - f_0(x_t)) \\ &= w_*^\top x_* - w_*^\top x_t + \rho r_t, \end{aligned}$$

where the inequality is by Definition 1. Therefore, by rearranging the inequality we have

$$r_t \leq \frac{1}{1-\rho}(w_*^\top x_* - w_*^\top x_t) \leq \frac{2\sqrt{\beta_t} u_t}{1-\rho},$$

where the last inequality is by Lemma 13. □

Lemma 12. *Assume* β_t *is monotonically nondecreasing and* $w_* \in \text{Ball}_t$ *for all* $t = 1, ..., T-1$*, then*

$$\sum_{t=1}^{T-1} r_t^2 \leq \frac{8\beta_{T-1} d}{(1-\rho)^2} \log\left(1 + \frac{TC_b^2}{d\lambda}\right).$$

Proof. By definition $u_t = \sqrt{x_t^\top \Sigma_t^{-1} x_t}$ and Lemma 11,

$$\begin{aligned} \sum_{t=1}^{T-1} r_t^2 &\leq \sum_{t=1}^{T-1} \frac{4}{(1-\rho)^2} \beta_t u_t^2 \\ &\leq \frac{4\beta_{T-1}}{(1-\rho)^2} \sum_{t=1}^{T-1} u_t^2 \leq \frac{4\beta_{T-1}}{(1-\rho)^2} \sum_{t=0}^{T-1} u_t^2 \\ &\leq \frac{8\beta_{T-1} d}{(1-\rho)^2} \log\left(1 + \frac{TC_b^2}{d\lambda}\right), \end{aligned}$$

where the second inequality is by the monotonic increasing property of β_t and the last inequality uses the elliptical potential lemma (Lemma 16). □

Previous two lemmas hold on the following lemma, bounding the gap between f^* and the linear function value at x_t, shown below.

Lemma 13. *Define* $u_t = \|x_t\|_{\Sigma_t^{-1}}$ *and assume* β_t *is chosen such that* $w_* \in \text{Ball}_t$*. Then*

$$w_*^\top (x_* - x_t) \leq 2\sqrt{\beta_t} u_t.$$

Proof. Let $\tilde{w}$ denote the parameter that achieves $\operatorname{argmax}_{w \in \text{Ball}_t} w^\top x_t$, by the optimality of x_t,

$$\begin{aligned}
& w_*^\top x_* - w_*^\top x_t \\
\leq & \tilde{w}^\top x_t - w_*^\top x_t \\
= & (\tilde{w} - \hat{w}_t + \hat{w}_t - w_*)^\top x_t \\
\leq & \|w_* - \hat{w}_t\|_{\Sigma_t} \|x_t\|_{\Sigma_t^{-1}} + \|\hat{w}_t - w_*\|_{\Sigma_t} \|x_t\|_{\Sigma_t^{-1}} \\
\leq & 2\sqrt{\beta_t} u_t
\end{aligned}$$

where the second inequality applies Holder's inequality; the last line uses the definition of Ball_t (note that both $w_*, \tilde{w} \in \text{Ball}_t$). □

4.2 CONFIDENCE ANALYSIS

All analysis in the previous section requires $w_* \in \text{Ball}_t, \forall t \in [T]$. In this section, we show that our choice of β_t in (5) is valid and w_* is trapped in the uncertainty set Ball_t with high probability.

Lemma 14 (Feasibility of Ball_t)**.** *Suppose Assumptions 5, 6, and 7 hold. Set β_t as in eq. (5). Then, w.p. $> 1 - \delta$,*

$$\|w_* - \hat{w}_t\|_{\Sigma_t}^2 \leq \beta_t, \forall t = 1, 2, ...$$

Proof. By setting the gradient of objective function in eq. (4) to be 0, we obtain the closed form solution of eq. (4):

$$\hat{w}_t = \Sigma_t^{-1} \sum_{i=0}^{t-1} y_i x_i.$$

Therefore,

$$\begin{aligned}
& \hat{w}_t - w_* \\
= & -w_* + \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i y_i \\
= & -w_* + \Sigma_t^{-1} \sum_{i=0}^{t-1} x_i (x_i^\top w_* + \eta_i + \Delta_i) \\
= & -w_* + \Sigma_t^{-1} \left(\sum_{i=0}^{t-1} x_i x_i^\top \right) w_* + \Sigma_t^{-1} \sum_{i=0}^{t-1} \eta_i x_i \\
& + \Sigma_t^{-1} \sum_{i=0}^{t-1} \Delta_i x_i, \qquad (8)
\end{aligned}$$

where the second equation is by eq. 7 and the first two terms of eq. (8) can be further simplified as

$$\begin{aligned}
& - w_* + \Sigma_t^{-1} \left(\sum_{i=0}^{t-1} x_i x_i^\top \right) w_* \\
= & -w_* + \Sigma_t^{-1} \left(\lambda I + \sum_{i=0}^{t-1} x_i x_i^\top - \lambda I \right) w_* \\
= & -w_* + \Sigma_t^{-1} \Sigma_t w_* - \lambda \Sigma_t^{-1} w_* \\
= & -\lambda \Sigma_t^{-1} w_*,
\end{aligned}$$

where the second equation is by definition of Σ_t (eq. (3)). Therefore, eq. (8) can be rewritten as

$$\hat{w}_t - w_* = -\lambda \Sigma_t^{-1} w_* + \Sigma_t^{-1} \sum_{i=0}^{t-1} \eta_i x_i + \Sigma_t^{-1} \sum_{i=0}^{t-1} \Delta_i x_i.$$

Multiply both sides by $\Sigma_t^{\frac{1}{2}}$ and we have

$$\begin{aligned}
& \Sigma_t^{\frac{1}{2}} (\hat{w}_t - w_*) \\
= & -\lambda \Sigma_t^{-\frac{1}{2}} w_* + \Sigma_t^{-\frac{1}{2}} \sum_{i=0}^{t-1} \eta_i x_i + \Sigma_t^{-\frac{1}{2}} \sum_{i=0}^{t-1} \Delta_i x_i.
\end{aligned}$$

Take a square of both sides and apply generalized triangle inequality, we have

$$\begin{aligned}
& \|\hat{w}_t - w_*\|_{\Sigma_t}^2 \\
\leq & 4\lambda^2 \|w_*\|_{\Sigma_t^{-1}}^2 + 4 \left\| \sum_{i=0}^{t-1} \eta_i x_i \right\|_{\Sigma_t^{-1}}^2 + 4 \left\| \sum_{i=0}^{t-1} \Delta_i x_i \right\|_{\Sigma_t^{-1}}^2. \qquad (9)
\end{aligned}$$

The remaining task is to bound these three terms separately. The first term of eq. (9) is bounded as

$$4\lambda^2 \|w_*\|_{\Sigma_t^{-1}}^2 \leq 4\lambda \|w_*\|_2^2 \leq 4\sigma^2,$$

where the first inequality is by definition of Σ_t and $\|\Sigma_t^{-1}\|_{\text{op}} \leq 1/\lambda$ and the second inequality is by choice of $\lambda = \sigma^2 / C_w^2$.

The second term of eq. (9) can be bounded by Lemma 17 and Lemma 20:

$$\begin{aligned}
4 \left\| \sum_{i=0}^{t-1} \eta_i x_i \right\|_{\Sigma_t^{-1}}^2 & \leq 4\sigma^2 \log \left(\frac{\det(\Sigma_t) \det(\Sigma_0)^{-1}}{\delta_t^2} \right) \\
& \leq 4\sigma^2 \left(d \log \left(1 + \frac{t C_b^2}{d\lambda} \right) - \log \delta_t^2 \right),
\end{aligned}$$

where δ_t is chosen as $3\delta / (\pi^2 t^2)$ so that the total failure probabilities over T rounds can always be bounded by $\delta/2$:

$$\sum_{t=1}^{T} \frac{3\delta}{\pi^2 t^2} < \sum_{t=1}^{\infty} \frac{3\delta}{\pi^2 t^2} = \frac{3\delta \pi^2}{6\pi^2} = \frac{\delta}{2}.$$

And the third term of eq. (9) can be bounded as

$$\begin{aligned}
4 \left\| \sum_{i=0}^{t-1} \Delta_i x_i \right\|_{\Sigma_t^{-1}}^2 & = 4 \left(\sum_{i=0}^{t-1} \Delta_i x_i \right)^\top \Sigma_t^{-1} \left(\sum_{j=0}^{t-1} \Delta_j x_j \right) \\
& = 4 \sum_{i=0}^{t-1} \sum_{j=0}^{t-1} \Delta_i \Delta_j x_i \Sigma_t^{-1} x_j \\
& \leq 4 \sum_{i=0}^{t-1} \sum_{j=0}^{t-1} |\Delta_i| |\Delta_j| \|x_i\|_{\Sigma_t^{-1}} \|x_j\|_{\Sigma_t^{-1}},
\end{aligned}$$

where the last line is by taking the absolute value and Cauchy-Schwarz inequality. Continue the proof and we have

$$
\begin{aligned}
&4\sum_{i=0}^{t-1}\sum_{j=0}^{t-1}|\Delta_i||\Delta_j|\|x_i\|_{\Sigma_t^{-1}}\|x_j\|_{\Sigma_t^{-1}} \\
&= 4\left(\sum_{i=0}^{t-1}|\Delta_i|\|x_i\|_{\Sigma_t^{-1}}\right)\left(\sum_{j=0}^{t-1}|\Delta_j|\|x_j\|_{\Sigma_t^{-1}}\right) \\
&= 4\left(\sum_{i=0}^{t-1}|\Delta_i|\|x_i\|_{\Sigma_t^{-1}}\right)^2 \\
&\le 4\left(\sum_{i=0}^{t-1}|\Delta_i|^2\right)\left(\sum_{i=0}^{t-1}\|x_j\|^2_{\Sigma_t^{-1}}\right) \\
&\le 4d\rho^2\sum_{i=0}^{t-1}r_i^2.
\end{aligned}
$$

where the first inequality is due to Cauchy-Schwarz inequality and the second uses the self-bounding properties $|\Delta_i| \le \rho r_i$ from Proposition 2 and Lemma 15.

To put things together, we have shown that w.p. $> 1-\delta$, for any $t \ge 1$,

$$
\begin{aligned}
&\|\hat{w}_t - w_*\|^2_{\Sigma_t^{-1}} \\
&\le 4\sigma^2 + 4\sigma^2\left(d\log\left(1+\frac{tC_b^2}{d\lambda}\right) + 2\log\left(\frac{\pi^2 t^2}{3\delta}\right)\right) \\
&\quad + 4\rho^2 d\sum_{i=0}^{t-1}r_i^2, \qquad (10)
\end{aligned}
$$

where we condition on (10) for the rest of the proof.

Observe that this implies that the feasibility of w_* in Ball_t can be enforced if we choose β_t to be larger than (10). The feasiblity of w_* in turn allows us to apply Lemma 11 to bound the RHS with $\beta_0, ..., \beta_{t-1}$. We will use induction to prove that our choice

$$\beta_t := 2\sigma^2\iota_t \text{ for } t = 1, 2, ...$$

is valid, where short hand

$$\iota_t := 4 + 4\left(d\log\left(1+\frac{tC_b^2}{d\lambda}\right) + 2\log\left(\frac{\pi^2 t^2}{3\delta}\right)\right).$$

For the base case $t = 1$, by eq. (10) and the definition of β_1 we directly have $\|\hat{w}_1 - w_*\|^2_{\Sigma_1^{-1}} \le \beta_1$. Assume our choice of β_i is feasible for $i = 1, ..., t-1$, then we can write

$$
\begin{aligned}
\|\hat{w}_t - w_*\|^2_{\Sigma_t^{-1}} &\le \sigma^2\iota_t + 4\rho^2 d\sum_{i=1}^{t-1}\beta_i u_i^2 \\
&\le \sigma^2\iota_t + 4\rho^2 d\beta_{t-1}\sum_{i=1}^{t-1}u_i^2,
\end{aligned}
$$

where the second line is due to non-decreasing property of β_t. Then by Lemma 16 and Assumption 7, we have

$$
\begin{aligned}
\|\hat{w}_t - w_*\|^2_{\Sigma_t^{-1}} &\le \sigma^2\iota_t + 8\rho^2 d^2\beta_{t-1}\log\left(1+\frac{tC_b^2}{d\lambda}\right) \\
&\le \sigma^2\iota_t + \frac{1}{2}\beta_{t-1} \le 2\sigma^2\iota_t = \beta_t, \qquad (11)
\end{aligned}
$$

The critical difference from the standard LinUCB analysis here is that if β_{t-1} appears on the LHS of the bound and if its coefficient is larger, any valid bound for β_t will have to grow exponentially in t. This is where Assumption 7 helps us. Assumption 7 ensures that the coefficient of β_{t-1} is smaller than $1/2$, so we can take $\beta_{t-1} \le \beta_t$ and move $\beta_t/2$ to the right-hand side. □

Proof of previous lemma needs the following two lemmas.

Lemma 15 (Upper bound of $\sum_{i=0}^{t-1} x_i^\top \Sigma_t^{-1} x_i$)**.**

$$\sum_{i=0}^{t-1} x_i^\top \Sigma_t^{-1} x_i \le d.$$

Proof. Recall that $\Sigma_t = \sum_{i=0}^{t-1} x_i x_i^T + \lambda I_d$.

$$
\begin{aligned}
\sum_{i=0}^{t-1} x_i^\top \Sigma_t^{-1} x_i &= \sum_{i=0}^{t-1} \text{tr}\left[\Sigma_t^{-1} x_i x_i^T\right] \\
&= \text{tr}\left[\Sigma_t^{-1}\sum_{i=0}^{t-1} x_i x_i^T\right] \\
&= \text{tr}\left[\Sigma_t^{-1}(\Sigma_t - \lambda I_d)\right] \\
&= \text{tr}\left[I_d\right] - \text{tr}\left[\lambda\Sigma_t^{-1}\right] \le d.
\end{aligned}
$$

The last line follows from the fact that Σ_t^{-1} is positive semidefinite. □

Lemma 16 (Upper bound of $\sum_{i=0}^{t-1} x_i^\top \Sigma_i^{-1} x_i$ (adapted from Abbasi-yadkori et al. [2011]))**.**

$$\sum_{i=0}^{t-1} x_i^\top \Sigma_i^{-1} x_i \le 2d\log\left(1+\frac{tC_b^2}{d\lambda}\right).$$

Proof. First we prove that $\forall i \in \{0, 1, ..., t-1\}, 0 \le x_i^\top \Sigma_i^{-1} x_i < 1$. Recall the definition of Σ_i and we know Σ_i^{-1} is a positive semidefinite matrix and thus $0 \le x_i^\top \Sigma_i^{-1} x_i$. To prove $x_i^\top \Sigma_i^{-1} x_i < 1$, we need to decompose Σ_i and write

$$
\begin{aligned}
&x_i^\top \Sigma_i^{-1} x_i \\
&= x_i^\top\left(\lambda I + \sum_{j=0}^{i-1} x_j x_j^\top\right)^{-1} x_i \\
&= x_i^\top\left(x_i x_i^\top - x_i x_i^\top + \lambda I + \sum_{j=0}^{i-1} x_j x_j^\top\right)^{-1} x_i.
\end{aligned}
$$

Let $A = -x_i x_i^\top + \lambda I + \sum_{j=0}^{i-1} x_j x_j^\top$ and it becomes

$$x_i^\top \Sigma_i^{-1} x_i = x_i^\top (x_i x_i^\top + A)^{-1} x_i.$$

By Sherman-Morrison lemma (Lemma 18), we have

$$\begin{aligned} x_i^\top \Sigma_i^{-1} x_i &= x_i^\top \left(A^{-1} - \frac{A^{-1} x_i x_i^\top A^{-1}}{1 + x_i^\top A^{-1} x_i} \right) x_i \\ &= x_i^\top A^{-1} x_i - \frac{x_i^\top A^{-1} x_i x_i^\top A^{-1} x_i}{1 + x_i^\top A^{-1} x_i} \\ &= \frac{x_i^\top A^{-1} x_i}{1 + x_i^\top A^{-1} x_i} < 1. \end{aligned}$$

Next we use the fact that $\forall x \in [0, 1), x \leq 2\log(x+1)$ and we have

$$\begin{aligned} \sum_{i=0}^{t-1} x_i^\top \Sigma_i^{-1} x_i &\leq \sum_{i=0}^{t-1} 2\log\left(1 + x_i^\top \Sigma_i^{-1} x_i\right) \\ &\leq 2\log\left(\frac{\det(\Sigma_{t-1})}{\det(\Sigma_0)}\right) \\ &\leq 2d\log\left(1 + \frac{tC_b^2}{d\lambda}\right), \end{aligned}$$

where the last two lines are by Lemma 19 and Lemma 20. □

5 TECHNICAL LEMMAS

Lemma 17 (Self-normalized bound for vector-valued martingales (Lemma A.9 of Agarwal et al. [2021])). *Let $\{\eta_i\}_{i=1}^\infty$ be a real-valued stochastic process with corresponding filtration $\{\mathcal{F}_i\}_{i=1}^\infty$ such that η_i is $\mathcal{F}_i$ measurable, $\mathbb{E}[\eta_i | \mathcal{F}_{i-1}] = 0$, and η_i is conditionally σ-sub-Gaussian with $\sigma \in \mathbb{R}^+$. Let $\{X_i\}_{i=1}^\infty$ be a stochastic process with $X_i \in \mathcal{H}$ (some Hilbert space) and X_i being $\mathcal{F}_t$ measurable. Assume that a linear operator $\Sigma : \mathcal{H} \to \mathcal{H}$ is positive definite, i.e., $x^\top \Sigma x > 0$ for any $x \in \mathcal{H}$. For any t, define the linear operator $\Sigma_t = \Sigma_0 + \sum_{i=1}^t X_i X_i^\top$ (here $xx^\top$ denotes outer-product in $\mathcal{H}$). With probability at least $1-\delta$, we have for all $t \geq 1$:*

$$\left\| \sum_{i=1}^t X_i \eta_i \right\|_{\Sigma_t^{-1}}^2 \leq \sigma^2 \log\left(\frac{\det(\Sigma_t) \det(\Sigma_0)^{-1}}{\delta^2} \right).$$

Lemma 18 (Sherman-Morrison lemma [Sherman and Morrison, 1950]). *Let A denote a matrix and b, c denote two vectors. Then*

$$(A + bc^\top)^{-1} = A^{-1} - \frac{A^{-1} b c^\top A^{-1}}{1 + c^\top A^{-1} b}.$$

Lemma 19 (Lemma 6.10 of Agarwal et al. [2021]). *Define $u_t = \sqrt{x_t^\top \Sigma_t^{-1} x_t}$ and we have*

$$\det \Sigma_T = \det \Sigma_0 \prod_{t=0}^{T-1} (1 + u_t^2).$$

Lemma 20 (Potential function bound (Lemma 6.11 of Agarwal et al. [2021])). *For any sequence $x_0, ..., x_{T-1}$ such that for $t < T$, $\|x_t\|_2 \leq C_b$, we have*

$$\begin{aligned} \log\left(\frac{\det \Sigma_{T-1}}{\det \Sigma_0}\right) &= \log\det\left(I + \frac{1}{\lambda}\sum_{t=0}^{T-1} x_t x_t^\top\right) \\ &\leq d\log\left(1 + \frac{TC_b^2}{d\lambda}\right). \end{aligned}$$

6 CONCLUSION

We study linear bandits with the underlying reward function being non-linear, which falls into the misspecified bandit framework. Existing work on misspecified bandit usually assumes uniform misspecification where the ℓ_∞ distance between the best-in-class function and the true function is upper bounded by the misspecification parameter ϵ. Existing lower bound shows that the $\tilde{\Omega}(\epsilon T)$ term is unavoidable where T is the time horizon, thus the regret bound is always linear. However, in solving optimization problems, one only cares about the approximation error near the global optimal point and approximation error is allowed to be large in highly suboptimal regions. In this paper, we capture this intuition and define a natural model of misspecification, called ρ-gap-adjusted misspecificaiton, which only requires the approximation error at each input x to be proportional to the suboptimality gap at x with ρ being the proportion parameter.

Previous work found that classical LinUCB algorithm is not robust in ϵ-uniform misspecified linear bandit when ϵ is large. However, we show that LinUCB is automatically robust against such gap-adjusted misspecification. Under mild conditions, e.g., $\rho \leq O(1/\sqrt{\log T})$, we prove that it achieves the near-optimal $\tilde{O}(\sqrt{T})$ regret for problems that the best-known regret is almost linear. Also, LinUCB doesn't need the knowledge of ρ to run. However, if the upper bound of ρ is revealed to LinUCB, the β_t term can be carefully chosen according to eq. (11). Our technical novelty lies in a new self-bounding argument that bounds part of the regret due to misspecification by the regret itself, which can be of independent interest in more settings.

Our paper opens a brand new door for research in model misspecification, including misspecified linear bandits, misspecified kernelized bandits, and even reinforcement learning with misspecified function approximation. Moreover, we hope our paper make people rethink about the relationship between function optimization and function approximation. In the future, there is a lot of work can be done. For example, study LinUCB algorithm under the weaker gap-adjusted misspecification, design a new no-regret algorithm that works under gap-adjusted misspecification framework where ρ is a constant, and study ρ-gap-adjusted misspecified Gaussian process bandit optimization.

References

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

Naoki Abe and Philip M Long. Associative reinforcement learning using linear probabilistic concepts. In *International Conference on Machine Learning*, 1999.

Alekh Agarwal, Nan Jiang, Sham M. Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms, 2021.

Ayya Alieva, Ashok Cutkosky, and Abhimanyu Das. Robust pure exploration in linear bandits with limited budget. In *International Conference on Machine Learning*, 2021.

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

Ilija Bogunovic and Andreas Krause. Misspecified gaussian process bandit optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, 2011.

Emmanuelle Claeys, Pierre Gancarski, Myriam Maumy-Bertrand, and Hubert Wassner. Dynamic allocation optimization in a/b-tests using classification-based preprocessing. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):335–349, 2021.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, 2008.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, 2020.

Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

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

Julian Katz-Samuels, Lalit Jain, Kevin G Jamieson, et al. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

Sanath Kumar Krishnamurthy, Vitor Hadad, and Susan Athey. Tractable contextual bandits beyond realizability. In *International Conference on Artificial Intelligence and Statistics*, 2021.

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 *International Conference on Machine Learning*, 2020.

Ahmadreza Moradipari, Christos Thrampoulidis, and Mahnoosh Alizadeh. Stage-wise conservative linear bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

Gergely Neu and Julia Olkhovskaya. Efficient and robust algorithms for adversarial linear contextual bandits. In *Conference on Learning Theory*, 2020.

Jack Sherman and Winifred J Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Annals of Mathematical Statistics*, 21 (1):124–127, 1950.

Shiyao Wang, Qi Liu, Tiezheng Ge, Defu Lian, and Zhiqiang Zhang. A hybrid bandit model with visual priors for creative ranking in display advertising. In *The Web Conference*, 2021.

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