THE POWER OF FEEL-GOOD THOMPSON SAMPLING: A UNIFIED FRAMEWORK FOR LINEAR BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear contextual bandit is one of the most popular models in online decision-making with bandit feedback. Prior work has studied different variants of this model, e.g., misspecified, non-stationary, and multi-task/life-long linear contextual bandits. However, there is no single framework that can unify the algorithm design and analysis for these variants. In this paper, we propose a unified framework for linear contextual bandits based on feel-good Thompson sampling (Zhang, 2021). The algorithm derived from our framework achieves nearly minimax optimal regret in various settings and resolves the respective open problem in each setting. Specifically, let d be the dimension of the context and T be the length of the horizon, our algorithm achieves an $\widetilde{\mathcal{O}}(d\sqrt{ST})$ regret bound for non-stationary linear bandits with at most S switches, $\widetilde{\mathcal{O}}(d^{\frac{5}{6}}T^{\frac{2}{3}}P^{\frac{1}{3}})$ regret for non-stationary linear bandits with bounded path length P, and $\widetilde{\mathcal{O}}(d\sqrt{kT}+\sqrt{dkMT})$ regret for (generalized) lifelong linear bandits over M tasks that share an unknown representation of dimension k. We believe our framework will shed light on the design and analysis of other linear contextual bandit variants.

1 INTRODUCTION

Linear contextual bandit is one of the most popular models in online decision-making with a large, possibly infinite, action space. This bandit model has been widely studied in the past decade. One of the most successful approaches is based on the upper confidence bound (Auer, 2002). For example, LinUCB (Li et al., 2010) (or OFUL (Abbasi-Yadkori et al., 2011)) follows the optimism-in-the-face-of-uncertainty principle and chooses the best action within an elliptical confidence ball. The algorithm has been proved to be nearly minimax optimal by using the elliptical potential lemma to track the bonus term. With some modifications to the algorithm, one had generalized this algorithm to various settings, e.g., non-stationary linear bandits (Chen et al., 2019), multi-task linear bandits (Hu et al., 2021), to mention a few. The analyses for these generalizations require the corresponding modified elliptical potential lemma, which is, however, hard to derive in general. (One may refer to the technical note (Faury et al., 2021) which discusses the faults in the elliptical potential lemma for non-stationary linear bandits).

Another common approach for online decision-making is exponentially weighted sampling. By sampling from a distribution over actions based on their historical rewards, it gives rise to near-optimal policy-based algorithms for various settings such as the hedge algorithm (Littlestone & Warmuth, 1994) for prediction with expert advice. For contextual bandits, EXP4 (Auer et al., 2002) enjoys a regret bound of

$$\mathbb{E}[\mathrm{Regret}(T)] \leq \mathcal{O}(\sqrt{KT\log|\mathcal{H}|}),$$

where K is the number of actions, T is the length of the horizon, and $\mathcal{H}$ is the feasible policy set. Note that contextual policy-based algorithms usually allow the policy to take the round index as an input. The corresponding policy, namely "meta-policy", provides a natural way to deal with non-stationary environment. For instance, one can solve non-stationary expert problems using meta-experts by following different experts in different rounds (Herbster & Warmuth, 2004). With this idea, one can obtain the regret bound for a variety of bandit models by counting the number of policies $|\mathcal{H}|$, which is easy to do in general.

This motivates us to find a policy-based algorithm for linear contextual bandits. We note that EXP4 is not suitable for our purpose since its regret suffers a polynomial dependence on the number of actions, which can be unbounded in linear contextual bandits. This is due to the fact that EXP4 is designed for general reward functions and does not leverage the linear structure of linear bandits. Given this observation, we raise the following question:

Can we design an EXP4-type algorithm for linear contextual bandits?

In this paper, we answer the above question affirmatively. In detail, we propose **F**eel-**G**ood **T**hompson **S**ampling over **L**inear **P**olicies (FGTS.LP), which is a policy-based algorithm for linear contextual bandits. Our algorithm can be regarded as a policy-based adaption of feel-good Thompson sampling (Zhang, 2021) to linear bandits, while this adaption is nontrivial. Our algorithm enjoys a regret bound that is logarithmically dependent on the number of policies and polynomially dependent on the dimension of contexts. To be specific, we prove the following regret bound for FGTS.LP:

Theorem 1.1 (Regret Bound of FGTS.LP (informal))**.** *Let d be the dimension of the context, T be the length of the horizon, and $\mathcal{H}$ be the set of all feasible policy hypotheses. The regret of FGTS.LP is bounded by*

$$\mathbb{E}[\mathrm{Regret}(T)] \leq \mathcal{O}(\sqrt{dT \log N(\mathcal{H}, \epsilon)} + T\sqrt{d}\epsilon),$$

where ϵ is some hyperparameter and $N(\mathcal{H}, \epsilon)$ is the covering number of policy set which contains an ϵ-optimal policy.

The above theorem provides a general interface to analyze the performance of FGTS.LP in different settings. Following the idea of constructing meta-policies, FGTS.LP can deal with various linear contextual bandits. The results are highlighted as follows:

Theorem 1.2 (Regret Bound over Variants of Linear Bandits (informal))**.** *With specific modification for each setting, the regret of FGTS.LP is bounded as*

- $\mathbb{E}[\mathrm{Regret}(T)] \leq \widetilde{\mathcal{O}}(d\sqrt{T} + T\sqrt{d}\zeta)$ *for ζ-misspecified linear contextual bandits.*
- $\mathbb{E}[\mathrm{Regret}(T)] \leq \widetilde{\mathcal{O}}(d\sqrt{ST})$ *for non-stationary linear contextual bandits with at most S switches.*
- $\mathbb{E}[\mathrm{Regret}(T)] \leq \widetilde{\mathcal{O}}(d^{\frac{5}{6}} T^{\frac{2}{3}} P^{\frac{1}{3}})$ *for non-stationary linear contextual bandits with path length bounded by P.*
- $\mathbb{E}[\mathrm{Regret}(T)] \leq \widetilde{\mathcal{O}}(d\sqrt{kT} + \sqrt{dkMT})$ *for (generalized) lifelong linear contextual bandits over M tasks that share an unknown representation of dimension k.*

We note that the above results are all near-optimal which match or improve the state-of-the-art in the corresponding settings. To sum up, our contributions are:

- We propose a unified framework for design and analyze various linear contextual bandit models. Our framework is easy to interpret and enjoys near-optimal regret bound in different settings.
- We propose the first nearly minimax algorithm for non-stationary linear contextual bandits with a bounded number of switches.
- We propose a new algorithm for non-stationary linear contextual bandits with bounded path length. It is the first algorithm that achieves nearly minimax regret.
- We propose the first near-optimal algorithm for (generalized) lifelong linear contextual bandits. Its regret matches the state-of-the-art for multi-task linear contextual bandits (Hu et al., 2021), which is a special case of our model.

Notation. We use lower and upper case bold face letters to denote vectors and matrices respectively. We use $[k]$ to denote the set $\{1, 2, \cdots, k\}$. We denote the Euclidean norm of vector $\mathbf{x} \in \mathbb{R}^d$ by $\|\mathbf{x}\|_2$. For a matrix $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_k] \in \mathbb{R}^{d\times k}$, we define $\|\mathbf{A}\|_{2,\infty} = \max_{1\leq i\leq k} \|\mathbf{a}_i\|_2$. For two non-negative sequence $\{a_n\}, \{b_n\}$, we write $a_n \leq \mathcal{O}(b_n)$ if there exists an absolute constant $C > 0$ such that $a_n \leq Cb_n$ for all $n \geq 1$, and $a_n \leq \widetilde{\mathcal{O}}(b_n)$ if there exists an absolute constant k such that $a_n \leq \mathcal{O}(b_n \log^k b_n)$; we write $a_n \geq \Omega(b_n)$ if there exists an absolute constant $C > 0$

such that $a_n \geq Cb_n$ for all $n \geq 1$ and $a_n \geq \widetilde{\Omega}(b_n)$ if there exists absolute constant k such that $a_n \geq \Omega(b_n \log^{-k} b_n)$; we write $a_n = \Theta(b_n)$ if there exists absolute constants $0 < C_1 \leq C_2$ such that $C_1 b_n \leq a_n \leq C_2 b_n$ for all $n \geq 1$. For any set $\mathcal{C}$, we use $|\mathcal{C}|$ to denote its cardinality. We use $\log$ to denote $\log_e$ for short.

2 Related Work

Misspecified Linear Bandits. The misspecified linear bandits was first studied by Ghosh et al. (2017), and they proposed an algorithm that achieves sub-linear regret when the misspecification ζ is small. Lattimore & Szepesvari (2020) proposed an algorithm with an $\widetilde{\mathcal{O}}(d\sqrt{T} + T\sqrt{d}\zeta)$ regret but requiring the contexts to be stationary. The regret is nearly minimax optimal as they proved a matching lower bound in the paper. Later, Zanette et al. (2020) proposed a LinUCB-like algorithm with the same regret that gets rid of the requirement of stationary context. More recent works (Foster et al., 2020a; Krishnamurthy et al., 2021; Takemura et al., 2021) studied the problem when the misspecification level is unknown.

Non-Stationary Linear Bandits with Bounded Switches. Non-stationary online decision-making models with bounded switches have long been studied in the literature. Cesa-Bianchi et al. (1993) studied the problem in the full information experts setting, and proposed an algorithm with an expected regret bound $\widetilde{\mathcal{O}}(\sqrt{ST})$, where S is the number of switches. Later, a high probability bound was obtained by a a UCB-type algorithm (Auer, 2002). In the bandit feedback setting, EXP3.S (Auer et al., 2001) achieves a regret of $\widetilde{\mathcal{O}}(\sqrt{KST})$, where K is the number of actions. Borrowing the idea from EXP4, the algorithm has been extended to the contextual setting with the $\widetilde{\mathcal{O}}(\sqrt{KST \log |\Pi|})$ regret bound (Luo et al., 2018), where Π is the finite policy class. More recently, Luo et al. (2022) studied the non-stationary linear bandit setting where the action set is a unit ball and proposed a bandit-over-bandit approach that achieves a regret of $\widetilde{\mathcal{O}}(\sqrt{dST})$. There is no existing algorithms for non-stationary linear contextual bandits with bounded switches.

Non-Stationary Linear Bandits with Bounded Path Length. In recent years, non-Stationary linear bandits with bounded path length have received increasing attention. Various algorithms have been developed (Cheung et al., 2019; Chen et al., 2019; Zhao et al., 2020) in this setting. The key idea behind these algorithms is to progressively forget the past data, and the proof is based on the same kind of elliptical potential lemma. However, as discussed by Faury et al. (2021), there is a fault in the proof of the original paper (Cheung et al., 2019), and the corrected analysis can only get a regret bound of $\widetilde{\mathcal{O}}(d^{\frac{3}{4}} T^{\frac{3}{4}} P^{\frac{1}{4}})$. Meanwhile, the lower bound of this problem is $\Omega(T^{\frac{2}{3}} P^{\frac{1}{3}})$ (Faury et al., 2021). The algorithm derived from our framework closes the gap on T.

(Generalized) Lifelong Linear Bandits with Shared Representation. The multi-task linear bandit, in which the agent plays over a collection of tasks simultaneously, was first studied by Yang et al. (2020). They proposed an explore-then-commit algorithm that leverages the multi-task structure but requires a unit-ball action space. Their followed-up work (Yang et al., 2022) improved the regret of the algorithm and generalized it to the lifelong setting. The algorithm has a regret bound of $\widetilde{\mathcal{O}}(d\sqrt{kT} + k\sqrt{MT})$. Hu et al. (2021) studied the multi-task setting in linear contextual bandits and proposed a computationally inefficient algorithm with a regret of $\widetilde{\mathcal{O}}(d\sqrt{kT} + \sqrt{dkMT})$ using LinUCB-type approach. Recent work (Qin et al., 2022) studied the non-stationary lifelong linear bandits with a task diversity assumption. On the lower bound side, Yang et al. (2020) proved that any algorithm for multi-task linear bandits suffers at least an $\Omega(d\sqrt{kT} + k\sqrt{MT})$ regret.

3 The Proposed Framework

3.1 Problem Setup

We first introduce a framework which is able to cover variants of linear contextual bandits. Let d be the dimension of the context, and T be the lenght of horizon. Denote by $\mathcal{A} \subseteq \{\mathbf{a} \in \mathbb{R}^d : \|\mathbf{a}\|_2 \leq 1\}$ the action space and $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{A}}$ the reward function space. The contextual bandit can be described as a repeated game between an agent (the bandit algorithm) and the environment (the adversary). In each round $t = 1, \cdots, T$, the environment first picks a hidden reward function $f^{(t)} \in \mathcal{F}$ and draws an action set $\mathcal{A}^{(t)} \subseteq \mathcal{A}$. After observing the action set $\mathcal{A}^{(t)}$, the agent selects an action $\mathbf{a}^{(t)} \in \mathcal{A}^{(t)}$

and receives a stochastic reward

$$r^{(t)} = f^{(t)}(\mathbf{a}^{(t)}) + \xi^{(t)},$$

where $f^{(t)}(\mathbf{a}^{(t)})$ is the expected value of the observed reward $r^{(t)}$ and $\xi^{(t)}$ is a zero mean random noise satisfying $\mathbb{E}[\xi^{(t)}|\Omega^{(t-1)}, f^{(t)}, \mathcal{A}^{(t)}, \mathbf{a}^{(t)}] = 0$, where $\Omega^{(t)} = \{(f^{(\tau)}, \mathcal{A}^{(t)}, \mathbf{a}^{(\tau)}, r^{(\tau)})\}_{\tau=1}^{t}$ is the history in the first t rounds. The learning objective of the agent is to maximize the expected cumulative reward, or equivalently, to minimize the pseudo-regret

$$\text{Regret}(T) := \sum_{t=1}^{T} \Big(f^{(t)}(\mathbf{a}_*^{(t)}) - f^{(t)}(\mathbf{a}^{(t)}) \Big), \tag{1}$$

where $\mathbf{a}_*^{(t)} = \arg\max_{\mathbf{a} \in \mathcal{A}^{(t)}} f^{(t)}(\mathbf{a})$ is the optimal action in round t. We assume the reward functions can be approximated by linear functions.

Assumption 3.1. *There is a mapping $\boldsymbol{\theta} : \mathcal{F} \to \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq 1\}$ and a universal constant $\zeta \in [0, 1]$ such that*

$$\sup_{(f, \mathbf{a}) \in (\mathcal{F} \times \mathcal{A})} |\langle \mathbf{a}, \boldsymbol{\theta}(f) \rangle - f(\mathbf{a})| \leq \zeta.$$

Moreover, we assume $|f(\mathbf{a})| \leq 1$ for all $(f, \mathbf{a}) \in (\mathcal{F} \times \mathcal{A})$.

We call ζ the misspecification level of the model. For the special case $\zeta = 0$, the model reduces to standard linear bandits. We further assume the observed reward is universally bounded.

Assumption 3.2. *The observed reward always satisfies $|r^{(t)}| \leq 1$ for all $t \in [T]$.*

This assumption ensures the additive noise $\xi^{(t)}$ is bounded. Note that this assumption is not essential to the algorithm, our analysis can be naturally generalized to unbounded sub-Gaussian with constant variance as it showed by Zhang (2021). We use bounded noise to avoid introducing variance as an extra parameter.

The above framework is compatible with variants of linear contextual bandits. For example, one can formulate specific linear contextual bandits by restricting the environment to select reward sequences $(f^{(1)}, \cdots, f^{(T)})$ from the corresponding structural set. We will present some concrete instances in the sequel.

3.2 Specific Examples

3.2.1 Stationary Misspecified Linear Bandits

For stationary linear contextual bandits, the reward function is fixed before agent decides. Thus, the environment is restricted to select reward function sequences that has the same linear approximation across rounds.

Assumption 3.3. *There exists a vector $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ such that $\boldsymbol{\theta}(f^{(t)}) = \boldsymbol{\theta}_0$ for all $t \in [T]$.*

The reward function is allowed to be misspecified from linear functions in Assumption 3.1. It is easy to verify that our framework reduces to linear contextual bandits with misspecification (Foster et al., 2020a) under Assumptions 3.1- 3.3.

3.2.2 Non-Stationary Linear Bandits with Bounded Switches

For non-stationary linear bandits, we first consider the case where the reward function may change dramatically for a finite number of times (Auer et al., 2001; Luo et al., 2022). The agent is not told when and how the reward function switches, and the environment can schedule the change in an adversarial way. We describe this setting using the following assumption:

Assumption 3.4. *There exists a constant S such that $\sum_{t=2}^{T} \mathbb{1}[\boldsymbol{\theta}(f^{(t)}) \neq \boldsymbol{\theta}(f^{(t-1)})] \leq S$.*

We call S the number of switches. We call the problem formulated under Assumptions 3.1, 3.2 and 3.4 as the non-stationary linear bandits with bounded switches. We note that the non-stationary linear bandits with bounded switches is a harder problem than that with bounded path length, since there is a black-box reduction from non-stationary linear bandits with bounded path length to non-stationary misspecified linear bandits with bounded switches.

3.2.3 NON-STATIONARY LINEAR BANDITS WITH BOUNDED PATH LENGTH

We also consider another kind of non-stationary linear bandits where the reward function can drift slowly over time (Cheung et al., 2019). The agent does not know the evolution dynamic, while the environment can choose the reward function adversarially in each round. We characterize this setting by the following assumption.

Assumption 3.5. *There exists a constant P such that $\sum_{t=2}^{T} \|\boldsymbol{\theta}(f^{(t)}) - \boldsymbol{\theta}(f^{(t-1)})\|_2 \leq P$.*

We call P the path length. It is easy to verify that our framework under Assumptions 3.1, 3.2 and 3.5 reduces to non-stationary linear bandits with bounded path length (Faury et al., 2021).

3.2.4 (GENERALIZED) LIFELONG LINEAR BANDITS WITH SHARED REPRESENTATION

We consider the setting where the agents have to solve a collection of correlated tasks sequentially. Let M be the number of tasks. At the beginning of the process, the environment first chooses M tasks $f_1, \cdots, f_M$. In each round t, the environment draws a task index $m^{(t)} \in [M]$ and assign current reward function as $f^{(t)} = f_{m^{(t)}}$. Besides the action set $\mathcal{A}^{(t)}$, the agent is provided with the task index $m^{(t)}$ and selects the action $\mathbf{a}^{(t)}$.

It is worth noting that our lifelong learning setting is a generalized version of the lifelong learning setting studied in Yang et al. (2022). The original setting of lifelong learning, in which the same task only appears in one interval, is equivalent to the case where the same $m^{(t)}$ lies in a single interval. We consider this generalized setting since it also captures multi-task learning (Hu et al., 2021), where the tasks appear periodically. We make the following assumption.

Assumption 3.6. *There exists a hidden orthogonal matrix $\mathbf{B} \in \mathbb{R}^{d\times k}$ and a set of hidden vectors $\{\mathbf{w}_i\}_{i=1}^{M}$ with $\mathbf{w}_i \in \mathbb{R}^k$ such that $\boldsymbol{\theta}(f^{(t)}) = \mathbf{B}\mathbf{w}_{m^{(t)}}$ holds for all $t \in [T]$ where $\{m^{(t)}\}_{t=1}^{T}$ is some sequence with $m^{(t)} \in [M]$.*

In Assumption 3.6, $\mathbf{B}$ is called the linear feature extractor of the model. The assumption implies $[\boldsymbol{\theta}(f^{(1)}), \cdots, \boldsymbol{\theta}(f^{(T)})]$ is a low rank matrix which can be decomposed over the bases $\mathbf{B}$. This is a common assumption for representation learning over linear bandits in the literature (see e.g., (Hu et al., 2021; Yang et al., 2022)). The genearlized lifelong linear bandits is defined under our framework with Assumptions 3.1, 3.2 and 3.6.

4 FEEL-GOOD THOMPSON SAMPLING OVER LINEAR POLICIES

In this section, we present our main algorithm FGTS.LP (**F**eel-**G**ood **T**hompson **S**ampling over **L**inear **P**olicies) for solving linear context bandits under the framework described in Section 3. We first introduce some important concepts in the algorithm design.

4.1 LINEAR POLICY CLASS AND ITS COVERING

It is pretty annoying that the regret defined in (1) is related with dynamic variable $\mathcal{A}^{(t)}$. We introduce an essential concept in our algorithm. Instead of selecting an action $\mathbf{a}^{(t)}$, we let the bandit algorithm to select a *meta-policy* $\pi^{(t)}$, which decides how the agent take action based on the observed context over time. Formally, a meta-policy $\pi : [T] \times \mathcal{A} \to \mathbb{R}$ is a time-variant hypothesis for the expected reward of choosing any actions in any round, i.e., the expected reward of action $\mathbf{a} \in \mathcal{A}$ on round t is $\pi(t, \mathbf{a})$. Thus, the meta-policy suggests to play action $\mathbf{a}^{(t)} = \arg\max_{\mathbf{a}\in\mathcal{A}^{(t)}} \pi(t, \mathbf{a})$ when the action set $\mathcal{A}^{(t)}$ is realized on round t. Since the reward functions in our setting can be approximated by linear functions, we would focus on linear policies that can be approximated by linear functions:

Definition 4.1 (Linear Policy)**.** *A linear policy $\pi(\boldsymbol{\Theta})$ is a meta-policy parameterized by the matrix $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_T] \in \mathbb{R}^{d\times T}$ such that $\pi(\boldsymbol{\Theta})(t, \mathbf{a}) = \langle \mathbf{a}, \boldsymbol{\theta}_t \rangle$.*

Let $\mathcal{H}$ be the set of all feasible linear policies. Under certain assumptions within our framework, the linear policy set is restricted to a corresponding domain. For example, under Assumption 3.3, the linear policy set $\mathcal{H}$ is parameterized by matrices $\boldsymbol{\Theta}$ in which all columns are the same.

Algorithm 1 FGTS.LP

input: linear policies set $\mathcal{H}$; inverse temperature β; exploration parameter λ; covering radius ϵ.
construct an ϵ-net $\mathcal{H}_\epsilon$ of $\mathcal{H}$ such that $\log|\mathcal{H}_\epsilon| = \mathcal{O}(\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty}))$
for round $t = 1, \cdots, T$ **do**
  sample $\mathbf{\Theta}^{(t)} = [\boldsymbol{\theta}_1^{(t)}, \cdots, \boldsymbol{\theta}_T^{(t)}] \sim P^{(t)}(\mathbf{\Theta}) \propto \exp\big(-\beta \sum_{\tau=1}^{t-1} L^{(\tau)}(\mathbf{\Theta})\big)$ for (3) over $\mathcal{H}_\epsilon$
  take action $\mathbf{a}^{(t)} \leftarrow \arg\max_{\mathbf{a}\in\mathcal{A}^{(t)}} \langle \mathbf{a}, \boldsymbol{\theta}_t^{(t)} \rangle$
end for

Note that $\mathcal{H}$ is a set over continuous domain and thus of infinite size. To deal with the infinite number of linear policies, we the the approach of ϵ-net and covering number.

Definition 4.2 (ϵ-Net of Linear Policies)**.** *$\mathcal{H}_\epsilon$ is said to be a ϵ-net of $\mathcal{H}$ if any linear policy $\mathbf{\Theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_T] \in \mathcal{H}$ has an ϵ-approximation $\widetilde{\mathbf{\Theta}} = [\widetilde{\boldsymbol{\theta}}_1, \cdots, \widetilde{\boldsymbol{\theta}}_T] \in \mathcal{H}_\epsilon$ with $\|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\|_2 \le \epsilon$ for all $t \in [T]$. Moreover, we define the covering number of $\mathcal{H}$ (at scale ϵ) be the size of minimum $\mathcal{H}_\epsilon$, i.e., $N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty}) := \inf|\mathcal{H}_\epsilon|$.*

One can regard the metric entropy $\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty})$ as the effective dimension of the policy set.

4.2 The Algorithm

Now we are ready to present our main algorithm. The algorithm is an adaptation of the FGTS (Zhang, 2021). We present the pseudo-code in Algorithm 1.

At the beginning, the algorithm constructs an ϵ-net $\mathcal{H}_\epsilon$ over all feasible linear policies $\mathcal{H}$. In each round t, the algorithm first samples a policy $\mathbf{\Theta}^{(t)}$ according to the following exponential distribution

$$P^{(t)}(\mathbf{\Theta}) := \frac{\exp\big(-\beta \sum_{\tau=1}^{t-1} L^{(\tau)}(\mathbf{\Theta})\big)}{\sum_{\mathbf{\Theta}'\in\mathcal{H}_\epsilon} \exp\big(-\beta \sum_{\tau=1}^{t-1} L^{(\tau)}(\mathbf{\Theta}')\big)}, \tag{2}$$

where $\beta > 0$ is the inverse temperature and $L^{(t)} : \mathbb{R}^{d\times k} \to \mathbb{R}$ is a loss function defined as follows

$$L^{(t)}(\mathbf{\Theta}) = (r^{(t)} - \langle \mathbf{a}^{(t)}, \boldsymbol{\theta}_t \rangle)^2 - \lambda \max_{\mathbf{a}\in\mathcal{A}^{(t)}} \langle \mathbf{a}, \boldsymbol{\theta}_t \rangle. \tag{3}$$

Here $\lambda > 0$ is a tuning parameter which controls the exploration. The algorithm then chooses the optimal action $\mathbf{a}^{(t)}$ that maximizes the expected reward according to the chosen policy $\mathbf{\Theta}^{(t)}$ among the action set $\mathcal{A}^{(t)}$.

We note the loss function in (3) has two terms. The first term is the Thompson sampling term. It casts penalty based on the estimation error of the reward. Recall $r^{(\tau)}$ is a random variable with mean $f^{(\tau)}(\mathbf{a}^{(\tau)}) \approx \langle \mathbf{a}^{(\tau)}, \boldsymbol{\theta}(f^{(\tau)}) \rangle$, this term forces the algorithm to select policy close to the true policy $\boldsymbol{\theta}(f^{(\tau)})$. This encourages the algorithm to do exploitation. The second term is so called feel-good exploration term (Zhang, 2021). It favors the policy that grants large rewards for historical actions, which can be regarded as an bonus that encourages the algorithm to do exploration.

Note that Algorithm 1 needs to know T and $\mathcal{H}$ before hand. By restarting the algorithm periodically with the doubling trick, one can run the algorithm without knowing T in hindsight. However, it is unclear if the algorithm can be parameter-free for $\mathcal{H}$ since model selection for exponential weighted sampling algorithm remains an open question (Foster et al., 2020b). Furthermore, the algorithm is sampled over ϵ-net of $\mathcal{H}$, which is not necessarily computationally efficient in practice. So it is better to regard Algorithm 1 as an oracle-efficient algorithm provided the sampling oracle over $\mathcal{H}_\epsilon$.

4.3 Comparison with Existing Algorithms

Comparison with FGTS. One may see that FGTS.LP has the same structure as the original FGTS (Zhang, 2021). Both algorithms use exponentially weighted sampling over the same loss function. The main difference between the two algorithms is that the original FGTS samples over models (or equivalently, reward functions) rather than policies. Note that sampling over models is equivalent to choosing the stationary policies, which always select the same action given the same context. Thus,

the original FGTS cannot deal with non-stationary environments in which the reward functions change over time. In contrast, sampling over policies enables us to inspect the misspecification explicitly and thus analyze it meticulously. As a result, FGTS.LP can achieve nearly minimax regret bound for misspecified linear bandits and non-stationary linear bandits.

Comparison with EXP4. Another contextual bandit algorithm that uses exponentially weighted sampling is EXP4 (Auer et al., 2002). Using exponentially weighted sampling, one can corral a large set of policies achieving a low regret as long as the reward for each policy is revealed. However, the agent cannot observe the payoff for policies that are not chosen in the bandits setting, and has to estimate it. EXP4 uses an unbiased estimator to estimate the reward. However, the estimator has a large variance which is proportional to the number of actions. This causes a polynomial dependence on the size of the action set in the regret bound. As a comparison, FGTS.LP uses the expected reward reported by the policies but casts penalties on the policies that cannot estimate the payoff of chosen actions. Although this estimator may be biased in estimating the actual reward, the estimator will assigns a large value to the policy with a high return. So the regret analysis for exponentially weighted sampling (Littlestone & Warmuth, 1994) can still be applied to FGTS.LP. Since the variance of this estimator is much smaller and independent of the number of actions, FGTS.LP is able to get a regret bound that is independent of the size of the action set.

Comparison with LinUCB. In addition to Thompson sampling and posterior sampling, another common approach for linear contextual bandits is LinUCB (Li et al., 2010) (or OFUL (Abbasi-Yadkori et al., 2011)). The algorithm implements the optimism-in-face-of-uncertainty principle, which selects the policy with the largest optimistic reward among all possible policies that agree with the observed reward in history. Note that FGTS.LP tends to sample policy with large estimated reward on the historical data and small estimation error. Therefore, FGTS.LP can also be regarded as an implementation of the same principle. The difference is, rather than focusing on one action, FGTS.LP uses exponentially weighted sampling to select a policy. From the game theory perspective, the variants of linear contextual bandits can be regarded as a two-player game. It is known that a mixed strategy profile usually generates better rewards than a pure strategy profile against adversarial opponents (Fudenberg & Tirole, 1991). Thus, FGTS.LP will have a stronger ability to solve variants of linear contextual bandits.

5 Main Results

In this section, we present the theoretical results of our algorithm. The following theorem provides a regret guarantee for Algorithm 1 when applied to general linear bandits.

Theorem 5.1. *Suppose $\epsilon \in [0,1]$, set $\lambda = \Theta\big(\sqrt{\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty})/(dT) + (\epsilon+\zeta)^2/d}\big)$ and $\beta = \Theta(1)$. Under Assumptions 3.1 and 3.2, for any policy class $\mathcal{H}$, the expected regret of Algorithm 1 is bounded by*

$$\mathbb{E}[\mathrm{Regret}(T)] \leq \mathcal{O}\Big(\sqrt{dT \log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty})} + T\sqrt{d}(\epsilon+\zeta)\Big), \tag{4}$$

where the expectation is taken over all randomness of the learning algorithm and the data noise.

There are two terms on the RHS of (4), which we will explain separately. The first term depends on the metric entropy $\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty})$, which depicts the complexity of the policy set $\mathcal{H}$. The second term depends on the scale of the ϵ-net $\mathcal{H}_\epsilon$ as well as the misspecification level ζ. There is a trade-off between the first term and the second term driven by the scale of the ϵ-net: for large ϵ, the first term is small and the second term is large; for small ϵ, the first term will be large and the second term will be small.

With Theorem 5.1, one only needs to calculate the metric entropy $\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty})$ and select proper ϵ to minimize (4), i.e., ensuring $\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty}) \sim \sqrt{T}(\epsilon+\zeta)$. Since the metric entropy $\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty})$ is the effective dimension of the policy class, it is easy to calculate for different policy classes and derive the corresponding regret bound. As a result, Algorithm 1 can be viewed as a unified framework for linear contextual bandits.

5.1 Implications to Linear Bandit Variants

In this subsection, we show the implications of Theorem 5.1 on the specific examples of contextual linear bandit variants.

5.1.1 MISSPECIFIED LINEAR BANDITS

We start with the simple setting where the approximate linear function is stationary. In this setting, the ϵ-net of the linear policy class can be reduced to the ϵ-net of single reward function. Since the embedding is of dimension d, the covering number is exactly $\mathcal{O}(\epsilon^{-d})$:

Lemma 5.1 (Covering Number for Misspecified Linear Bandits)**.** *Under Assumptions 3.1 and 3.3, the metric entropy of linear policy class satisfies*

$$\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty}) \leq \mathcal{O}(d \log \epsilon^{-1}).$$

By choosing $\epsilon = T^{-1}$, the metric entropy can be bounded as $\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty}) \leq \mathcal{O}(d \log T)$. With Theorem 5.1, we directly get the following regret bound:

Corollary 5.1 (Upper Bound for Misspecified Linear Bandits)**.** *Under Assumptions 3.1, 3.2 and 3.3, the expected regret of FGTS.LP is bounded by*

$$\mathbb{E}[\text{Regret}(T)] \leq \mathcal{O}\Big(d\sqrt{T \log T} + T\sqrt{d}\zeta\Big).$$

The following lower bound shows that the regret upper bound in Corollary 5.1 is tight up to logarithmic factors:

Proposition 5.1 (Lower Bound for Misspecified Linear Bandits)**.** *Under Assumptions 3.1, 3.2 and 3.3, for any algorithm, there exists a bandit instance for which*

$$\mathbb{E}[\text{Regret}(T)] \geq \Omega\Big(d\sqrt{T} + T\sqrt{d/\log T}\zeta\Big).$$

5.1.2 NON-STATIONARY LINEAR BANDITS WITH BOUNDED SWITCHES

In this setting, we can construct the ϵ-net by enumerating the positions of switches and cover the reward function set after each switch. The set of reward functions after each switch has a covering number of $\mathcal{O}(\epsilon^{-d})$. So the set of reward functions with S switches has a covering number of $\mathcal{O}(\epsilon^{-dS})$. Moreover, the different positions of switches within T rounds can be bounded by $\mathcal{O}(T^S)$. With above facts, we can bound the size of ϵ-net:

Lemma 5.2 (Covering Number for Non-Stationary Linear Bandits with Bounded Switches)**.** *Under Assumptions 3.1 and 3.4, the metric entropy of the linear policy set satisfies*

$$\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty}) \leq \mathcal{O}(dS \log \epsilon^{-1} + S \log T).$$

Using Theorem 5.1 with $\epsilon = T^{-1}$, we immediately obtain the following regret bound:

Corollary 5.2 (Upper Bound for Non-Stationary Linear Bandits with Bounded Switches)**.** *Under Assumptions 3.1, 3.2 and 3.4, the expected regret of FGTS.LP is bounded by*

$$\mathbb{E}[\text{Regret}(T)] \leq \mathcal{O}\Big(d\sqrt{ST \log T} + T\sqrt{d}\zeta\Big).$$

The following lower bound shows FGTS.LP is nearly minimax optimal for non-stationary linear bandits with bounded switches:

Proposition 5.2 (Lower Bound for Non-Stationary Linear Bandits with Bounded Switches)**.** *Under Assumptions 3.1, 3.2 and 3.4, for any algorithm, there exists a bandit instance for which*

$$\mathbb{E}[\text{Regret}(T)] \geq \Omega\Big(d\sqrt{ST} + T\sqrt{d/\log T}\zeta\Big).$$

5.1.3 NON-STATIONARY LINEAR BANDITS WITH BOUNDED PATH LENGTH

Now we move to the more challenging setting where the parameters slowly drift over time. We would reduce the parameter drift to parameter switch. More specifically, a parameter switch is considered to happen only if the path length of parameter drift from last switch exceeds $\epsilon/2$. Since the total path length is bounded by P, there are no more than $2P\epsilon^{-1}$ switches. As a result, an $\epsilon/2$-net of the reward functions for non-stationary linear bandits with no more than $2P\epsilon^{-1}$ switches is an ϵ-net of the reward functions for non-stationary linear bandits with path length no more than P. This immediately gives the following result on its covering number:

Lemma 5.3 (Covering Number for Non-Stationary Linear Bandits with Bounded Path Length)**.** *Under Assumptions 3.1 and 3.5, the metric entropy of linear policies satisfies*

$$\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty}) \leq \mathcal{O}(d \log \epsilon^{-1} + dP\epsilon^{-1} \log \epsilon^{-1} + P\epsilon^{-1} \log T).$$

Using Theorem 5.1 with $\epsilon = \max\{T^{-\frac{1}{3}} d^{\frac{1}{3}} P^{\frac{1}{3}}, T^{-1}\}$, we get the following regret bound:

Corollary 5.3 (Upper Bound for Non-Stationary Linear Bandits with Bounded Path Length)**.** *Under Assumptions 3.1, 3.2 and 3.5, the expected regret of FGTS.LP is bounded by*

$$\mathbb{E}[\text{Regret}(T)] \leq \widetilde{\mathcal{O}}\Big(d^{\frac{5}{6}} T^{\frac{2}{3}} P^{\frac{1}{3}} + d\sqrt{T} + T\sqrt{d}\zeta\Big).$$

The following lower bound shows that the regret upper bound in Corollary 5.3 is nearly optimal:

Proposition 5.3 (Lower Bound for Non-Stationary Linear Bandits with Bounded Path Length)**.** *Under Assumptions 3.1, 3.2 and 3.5, for any algorithm, there exists a bandit instance for which*

$$\mathbb{E}[\text{Regret}(T)] \geq \Omega\Big(d^{\frac{2}{3}} T^{\frac{2}{3}} P^{\frac{1}{3}} + d\sqrt{T} + T\sqrt{d/\log T}\zeta\Big).$$

5.1.4 (Generalized) Lifelong Linear Bandits with Shared Representation

In this setting, every feasible linear policy can be described by a low-rank matrix that is the product of two matrices $\mathbf{B} \in \mathbb{R}^{d\times k}$ and $[\mathbf{w}_i]_{i=1}^M \in \mathbb{R}^{k\times M}$. We can construct the ϵ-net of linear policies using the product of two $\epsilon/2$-net over $\mathbf{B}$ and $[\mathbf{w}_i]_{i=1}^M$. Also, the dimension of linear policies is $dk + kM$. This implies the following result on its covering number:

Lemma 5.4 (Covering Number for Lifelong with Shared Representation)**.** *Under Assumptions 3.1 and 3.6, the metric entropy of linear policies satisfies*

$$\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty}) \leq \mathcal{O}\big(dk \log \epsilon^{-1} + kM \log \epsilon^{-1}\big).$$

Using Theorem 5.1 with $\epsilon = T^{-1}$, we obtain the following regret bound:

Corollary 5.4 (Upper Bound for Lifelong Linear Bandits with Shared Representation)**.** *Under Assumptions 3.1, 3.2 and 3.6, the expected regret of FGTS.LP is bounded by*

$$\mathbb{E}[\text{Regret}(T)] \leq \widetilde{\mathcal{O}}\Big(d\sqrt{kT} + \sqrt{dkMT} + T\sqrt{d}\zeta\Big).$$

If each task appears uniformly for T/M rounds, an algorithm that solves each task independently suffers a regret bound of $\widetilde{\Theta}(d\sqrt{MT})$ using a nearly minimax algorithm such as LinUCB (Li et al., 2010). In the case that $k \ll M$ and $k \ll d$, our algorithm saves a factor of $\sqrt{M/k}$ or $\sqrt{M/d}$, which shows it utilizes the underlying representation. Moreover, this regret bound matches the near-optimal algorithm for multi-task linear bandits (Hu et al., 2021), which is a special case of our setting. The following lower bound shows that our algorithm is near-optimal:

Proposition 5.4 (Lower Bound for Lifelong Linear Bandits with Shared Representation)**.** *Under Assumptions 3.1, 3.2 and 3.6, for any algorithm, there exists a bandit instance for which*

$$\mathbb{E}[\text{Regret}(T)] \geq \Omega\Big(d\sqrt{kT} + k\sqrt{MT} + T\sqrt{d/\log T}\zeta\Big).$$

6 Conclusion and Future Work

In this paper, we proposed a unified framework for linear contextual bandits based on feel-good Thompson sampling, which can cover different variants of linear contextual bandits. At the core of our algorithm is an adaption of feel-good Thompson sampling from reward function to policy class, which enables the algorithm to deal with time-varying environments. We showed that our algorithm can achieve near-optimal regret bounds for these variants of linear bandits, which resolve several open problems in the respective settings.

We notice that there is still a gap between the regret upper and lower bounds in terms of the dependence on the dimension d for non-stationary linear bandits with bounded path length. We leave it as a future work to close this gap.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvari. Improved algorithms for linear stochastic bandits. In *NIPS*, 2011.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *ICML*, 2013.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3:397–422, 2002.

Peter Auer, Yoav Freund, and Robert E. Schapire. The non-stochastic multi-armed bandit problem. 2001.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32:48–77, 2002.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Sham M. Kakade. Towards minimax policies for online linear optimization with bandit feedback. *ArXiv*, abs/1202.3079, 2012.

Nicolò Cesa-Bianchi, Yoav Freund, David P. Helmbold, David Haussler, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. In *STOC '93*, 1993.

Yifang Chen, Chung wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal, and parameter-free. *ArXiv*, abs/1902.00980, 2019.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Hedging the drift: Learning to optimize under non-stationarity. *ArXiv*, abs/1903.01461, 2019.

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

Louis Faury, Yoan Russac, Marc Abeille, and Clément Calauzènes. A technical note on non-stationary parametric bandits: Existing mistakes and preliminary solutions. In *ALT*, 2021.

Dylan J. Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *ICML*, 2020.

Dylan J. Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert E. Schapire. Practical contextual bandits with regression oracles. *ArXiv*, abs/1803.01088, 2018.

Dylan J. Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. *ArXiv*, abs/2107.05745, 2020a.

Dylan J. Foster, Akshay Krishnamurthy, and Haipeng Luo. Open problem: Model selection for contextual bandits. *ArXiv*, abs/2006.10940, 2020b.

Drew Fudenberg and Jean Tirole. *Game theory*. MIT press, 1991.

Avishek Ghosh, Sayak Ray Chowdhury, and Aditya Gopalan. Misspecified linear bandits. *ArXiv*, abs/1704.06880, 2017.

Nima Hamidi and Mohsen Bayati. On worst-case regret of linear thompson sampling. *ArXiv*, abs/2006.06790, 2020.

Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Machine Learning*, 32:151–178, 2004.

Jiachen Hu, Xiaoyu Chen, Chi Jin, Lihong Li, and Liwei Wang. Near-optimal representation learning for linear bandits and linear rl. In *ICML*, 2021.

Sanath Kumar Krishnamurthy, Vitor Hadad, and Susan Athey. Adapting to misspecification in contextual bandits with offline regression oracles. In *ICML*, 2021.

Tze Leung Lai and Herbert E. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. 2020.

Tor Lattimore and Csaba Szepesvari. Learning with good feature representations in bandits and in rl with a generative model. In *ICML*, 2020.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. *ArXiv*, abs/1003.0146, 2010.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Inf. Comput.*, 108: 212–261, 1994.

Haipeng Luo, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *COLT*, 2018.

Haipeng Luo, Mengxiao Zhang, Peng Zhao, and Zhi-Hua Zhou. Corralling a larger band of bandits: A case study on switching regret for linear bandits. *ArXiv*, abs/2202.06151, 2022.

Yuzhen Qin, Tommaso Menara, Samet Oymak, ShiNung Ching, and Fabio Pasqualetti. Non-stationary representation learning in sequential linear bandits. 2022.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Math. Oper. Res.*, 39:1221–1243, 2014.

Kei Takemura, Shinji Ito, Daisuke Hatano, Hanna Sumita, Takuro Fukunaga, Naonori Kakimura, and Ken-ichi Kawarabayashi. A parameter-free algorithm for misspecified linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3367–3375. PMLR, 2021.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, 2012.

Jiaqi Yang, Wei Hu, J. Lee, and Simon Shaolei Du. Provable benefits of representation learning in linear bandits. *ArXiv*, abs/2010.06531, 2020.

Jiaqi Yang, Qi Lei, Jason D. Lee, and Simon Shaolei Du. Nearly minimax algorithms for linear bandits with shared representation. *ArXiv*, abs/2203.15664, 2022.

Andrea Zanette, Alessandro Lazaric, Mykel J. Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *ICML*, 2020.

Tong Zhang. Feel-good thompson sampling for contextual bandits and reinforcement learning. *ArXiv*, abs/2110.00871, 2021.

Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. In *AISTATS*, 2020.

A Additional Related Work

	Existing Algorithm	FGTS.LP (This Paper)	Lower Bounds
Misspecified LCB	$\widetilde{\mathcal{O}}(d\sqrt{T}+T\sqrt{d}\zeta)$ (Zanette et al., 2020)	$\widetilde{\mathcal{O}}(d\sqrt{T}+T\sqrt{d}\zeta)$ (Corollary 5.1)	$\widetilde{\Omega}(d\sqrt{T}+T\sqrt{d}\zeta)$ (Lattimore & Szepesvari, 2020)
Non-stationary LCB (Bounded Switches)	–	$\widetilde{\mathcal{O}}(d\sqrt{ST})$ (Corollary 5.2)	$\widetilde{\Omega}(d\sqrt{ST})$ (Lemma E.5)
Non-stationary LCB (Bounded Length)	$\widetilde{\mathcal{O}}(d^{\frac{3}{4}}T^{\frac{3}{4}}P^{\frac{1}{4}})$ (Faury et al., 2021)	$\widetilde{\mathcal{O}}(d^{\frac{5}{6}}T^{\frac{2}{3}}P^{\frac{1}{3}})$ (Corollary 5.3)	$\widetilde{\Omega}(d^{\frac{2}{3}}T^{\frac{2}{3}}P^{\frac{1}{3}})$ (Cheung et al., 2019)
Multi-task LCB	$\widetilde{\mathcal{O}}(d\sqrt{kT}+\sqrt{dkMT})$ (Hu et al., 2021)	$\widetilde{\mathcal{O}}(d\sqrt{kT}+\sqrt{dkMT})$ (Corollary 5.4)	$\widetilde{\Omega}(d\sqrt{kT}+k\sqrt{MT})$ (Yang et al., 2020)
Lifelong LCB	–	$\widetilde{\mathcal{O}}(d\sqrt{kT}+\sqrt{dkMT})$ (Corollary 5.4)	$\widetilde{\Omega}(d\sqrt{kT}+k\sqrt{MT})$ (Yang et al., 2020)

Table 1: Summary of the results on variants of linear contextual bandits. d is the dimension of the context, T is the length of horizon, and ζ is the misspecification level for misspecified bandits. In addition, S is the number of switches and P is the path length for non-stationary bandits. Furthermore, M denotes the number of tasks and k denotes the representation dimension for multi-task and lifelong bandits.

Thompson Sampling. Thompson sampling is a classical algorithm that can be adapted to contextual bandits (Thompson, 1933). Russo & Roy (2014) proposed the first general theoretical guarantee for Thompson sampling. By utilizing the connection between Thompson sampling and optimistic policies, they showed Thompson sampling achieves a Bayesian regret bound of $\widetilde{\mathcal{O}}(d\sqrt{T})$ for linear bandits, where the underlying parameters are randomly drawn from some public distribution. For frequentist regret, Agrawal & Goyal (2013) showed a modified Thompson sampling in which the variance of the posterior distribution is inflated by $\widetilde{\mathcal{O}}(d)$ achieves a regret bound of $\widetilde{\mathcal{O}}(d^{\frac{3}{2}}\sqrt{T})$ on linear bandits. As shown by Hamidi & Bayati (2020), inflation is necessary for Thompson sampling to achieve sub-linear regret; thus, one cannot improve the regret bounds for original Thompson sampling. In comparison, feel-good Thompson sampling (Zhang, 2021) introduced an optimistic bonus on the loss function, making it possible to achieve a nearly minimax regret on the linear contextual bandits.

Exponentially Weighted Sampling. A classical technique for online learning is exponentially weighted sampling. This approach allows an algorithm to corral a collection of policies in small regret. By constructing unbiased estimators for unobserved actions, Auer et al. (2002) adapted the ideas to contextual bandits and proposed EXP4. Its regret suffers a polynomial dependence on the number of actions. For linear bandits with stationary contexts, the dependence can be reduced to logarithmic using an EXP2-type algorithm (Bubeck et al., 2012). The algorithm can also be regarded as an EXP3 with a modified estimator, which can be implemented efficiently under specific action sets using stochastic mirror descent.

UCB-Based Algorithms. The concept of upper confidence bound was first proposed by Lai & Robbins (1985) to solve multi-arm bandits. The algorithm can be adapted to linear contextual bandits with an infinite number of actions by analyzing the dynamics of the confidence set using an elliptical lemma (Dani et al., 2008; Li et al., 2010; Abbasi-Yadkori et al., 2011). However, it is hard to get the potential lemma for other function spaces, making it difficult to generalize the approach to a broader function class. A recent line of works showed the contextual bandits could be reduced to regression oracles (Foster et al., 2018; Foster & Rakhlin, 2020). This observation pointed out a way for UCB-based algorithms to work on general function classes. However, the algorithm suffers from a polynomial dependence on the number of actions, which is incapable of infinite action sets.

B Proof Sketch of Theorem 5.1

This section presents a proof sketch of our main Theorem 5.1. We note the proof is similar to the proof of Theorem 2 in Zhang (2021). The main difference comes from the usage of covering number of the policy set and our decoupling lemma under misspecification.

Let $\mathbf{\Theta}^* = [\boldsymbol{\theta}_1^*, \cdots, \boldsymbol{\theta}_T^*]$ be the optimal policy in the ϵ-net $\mathcal{H}_\epsilon$. For any linear policy $\mathbf{\Theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_T]$, denote $\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)}) := \langle \mathbf{a}^{(t)}, \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^* \rangle$ as the Bellman error and $\Delta_t^*(\mathbf{\Theta}) := \max_{\mathbf{a}\in\mathcal{A}^{(t)}} \langle \mathbf{a}, \boldsymbol{\theta}_t \rangle - \max_{\mathbf{a}\in\mathcal{A}^{(t)}} \langle \mathbf{a}, \boldsymbol{\theta}_t^* \rangle$ as the feel-good error. Since $\langle \mathbf{a}^{(t)}, \boldsymbol{\theta}_t \rangle = \max_{\mathbf{a}\in\mathcal{A}^{(t)}} \langle \mathbf{a}, \boldsymbol{\theta}_t \rangle$ according to definition of $\mathbf{a}^{(t)}$, we can decompose the expected regret on round t by

$$\mathop{\mathbb{E}}_{\mathbf{a}^{(t)}}[f^{(t)}(\mathbf{a}_*^{(t)}) - f^{(t)}(\mathbf{a}^{(t)})] \leq \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}, \mathbf{\Theta}^{(t)}} \Big[\Delta_t(\mathbf{\Theta}^{(t)}\mathbf{a}^{(t)}) - \Delta_t^*(\mathbf{\Theta}^{(t)})\Big] + 2(\epsilon + \zeta). \tag{5}$$

Besides, we construct a potential function following the classical analysis for exponentially weighted sampling,

$$\Phi^{(t)} = \frac{1}{\beta\lambda} \log \sum_{\mathbf{\Theta}\in\mathcal{H}_\epsilon} \exp\Big(\beta \sum_{\tau=1}^{t} \big(L^{(\tau)}(\mathbf{\Theta}^*) - L^{(\tau)}(\mathbf{\Theta})\big)\Big).$$

The boundary value of the potential function follows

$$\Phi^{(0)} - \Phi^{(T)} \leq \frac{\log|\mathcal{H}_\epsilon|}{\beta\lambda}. \tag{6}$$

The increment of potential function on round t satisfies

$$\mathop{\mathbb{E}}_{\mathbf{a}^{(t)}, r^{(t)}} \Big[\Phi^{(t)} - \Phi^{(t-1)}\Big] \leq \frac{1}{\beta\lambda} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}} \Big[\log \mathop{\mathbb{E}}_{r^{(t)}} \mathop{\mathbb{E}}_{\mathbf{\Theta}} \exp\Big(\beta\big(L^{(t)}(\mathbf{\Theta}^*) - L^{(t)}(\mathbf{\Theta})\big)\Big)\Big]. \tag{7}$$

Recall $L^{(t)}(\mathbf{\Theta}) = (r^{(t)} - \langle \mathbf{a}^{(t)}, \boldsymbol{\theta}_t \rangle)^2 - \lambda \max_{\mathbf{a}\in\mathcal{A}^{(t)}} \langle \mathbf{a}, \boldsymbol{\theta}_t \rangle$, the randomness on $r^{(t)}$ only casts an additive noise $-2\beta\xi^{(t)}\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)})$ to the term $\beta\big(L^{(t)}(\mathbf{\Theta}^*) - L^{(t)}(\mathbf{\Theta})\big)$ in the exponent. Since $\xi^{(t)}$ is constantly bounded and zero mean, the noise can be controlled by selecting proper β according to Hoeffding's lemma. With classical inequalities over logarithm and exponential, we can show the variation of potential function satisfies

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{a}^{(t)}, r^{(t)}} \Big[\Phi^{(t)} - \Phi^{(t-1)}\Big] \leq -\frac{1}{\lambda} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}, \mathbf{\Theta}} \Big[\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)})^2 + C_t(\mathbf{a}^{(t)})\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)})\Big] \\ + \mathop{\mathbb{E}}_{\mathbf{\Theta}} \Delta_t^*(\mathbf{\Theta}) + O\Big(\beta\lambda + \frac{\beta(\epsilon+\zeta)^2}{\lambda}\Big)\end{aligned} \tag{8}$$

where $C_t(\mathbf{a}^{(t)}) = f^{(t)} - \langle \mathbf{\Theta}^{(t)}, \boldsymbol{\theta}_t^* \rangle$ is the misspecification of $\mathbf{\Theta}^*$ with respect to ground truth $f^{(t)}$.

One can see that the first two terms on the RHS of (8) are very similar to the first two terms on the RHS of (5). But with one critical difference, the chosen policy $\mathbf{\Theta}^{(t)}$ in (5) is replaced by a random policy $\mathbf{\Theta}$ in (8): Although $\mathbf{\Theta}^{(t)}$ and $\mathbf{\Theta}$ have the same margin, $\mathbf{\Theta}^{(t)}$ is correlated with $\mathbf{a}^{(t)}$ but $\mathbf{\Theta}$ is independent from $\mathbf{a}^{(t)}$. As a result, $\mathbf{\Theta}^{(t)}\Delta_t^*(\mathbf{\Theta}^{(t)}) = \mathbb{E}_{\mathbf{\Theta}}\, \Delta_t^*(\mathbf{\Theta})$ but $\mathbb{E}_{\mathbf{a}^{(t)}, \mathbf{\Theta}^{(t)}}[\Delta_t(\mathbf{\Theta}^{(t)}, \mathbf{a}^{(t)})]$ is not necessary equal to $\mathbb{E}_{\mathbf{a}^{(t)}, \mathbf{\Theta}}[\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)})]$. Therefore, we cannot get a regret bound by combining (5), (6) and (8) directly.

This is where the decoupling lemma helps. Zhang (2021) proposed a lemma which implies

$$\mathop{\mathbb{E}}_{\mathbf{a}^{(t)}, \mathbf{\Theta}^{(t)}} \Delta_t(\mathbf{\Theta}^{(t)}, \mathbf{a}^{(t)}) \leq \frac{1}{\lambda} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}, \mathbf{\Theta}} \Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)})^2 + \mathcal{O}(\lambda d). \tag{9}$$

For our usage, in order to achieve optimal dependence with the covering radius and misspecification, we derived a decoupling lemma includes the corruptions. In particular, Lemma C.2 implies

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{a}^{(t)}, \mathbf{\Theta}^{(t)}} \Delta_t(\mathbf{\Theta}^{(t)}, \mathbf{a}^{(t)}) \leq \frac{1}{\lambda} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}, \mathbf{\Theta}} \Big[\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)})^2 + C_t(\mathbf{a}^{(t)})\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)})\Big] \\ + O\Big(\lambda d + (\epsilon+\zeta) + \frac{(\epsilon+\zeta)^2}{\lambda}\Big).\end{aligned} \tag{10}$$

By combining (5), (6), (8) and (10), with $\beta = \Theta(1)$, we conclude that

$$\mathbb{E}[\mathrm{Regret}^{(T)}] = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}}[f^{(t)}(\mathbf{a}_*^{(t)}) - f^{(t)}(\mathbf{a}^{(t)})] \leq O\Big(\lambda dT + (\epsilon+\zeta)T + \frac{\log|\mathcal{H}_\epsilon| + (\epsilon+\zeta)^2 T}{\lambda}\Big).$$

One can obtain the desired statement using $\lambda = \Theta\big(\sqrt{\log|\mathcal{H}_\epsilon|/(dT) + (\epsilon+\zeta)^2/d}\big)$.

We note our analysis uses a refined way to deal with the covering radius and misspecification. One can only get a regret bound of $\mathcal{O}(\sqrt{dT\log|\mathcal{H}_\epsilon|} + T\sqrt{d(\epsilon+\zeta)})$ if uses (9) and follows the proof in Zhang (2021) faithfully, which has a worse dependence on $(\epsilon+\zeta)$.

C FULL PROOF OF THEOREM 5.1

In this section, we will provide a complete proof of Theorem 5.1. The following decoupling lemma is a critical structural lemma for our analysis, which is a generalization of Lemma 2 in Zhang (2021).

Lemma C.1 (Decoupling Lemma)**.** *Let P be a joint distribution over two $\mathbb{R}^d$ space, i.e., $P \in \Delta(\mathbb{R}^d \times \mathbb{R}^d)$. For any constant $\lambda > 0$, we have*

$$\mathop{\mathbb{E}}_{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P}\langle \boldsymbol{\theta}, \boldsymbol{\phi}\rangle \le d\lambda + \frac{0.25}{\lambda} \mathop{\mathbb{E}}_{\substack{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P\\(\boldsymbol{\theta}',\boldsymbol{\phi}')\sim P}}\langle \boldsymbol{\theta}, \boldsymbol{\phi}'\rangle^2,$$

where $(\boldsymbol{\theta}, \boldsymbol{\phi})$ on LHS is a sample from P while $(\boldsymbol{\theta}, \boldsymbol{\phi})$ and $(\boldsymbol{\theta}', \boldsymbol{\phi}')$ on RHS are two independent samples from P.

Proof. Let $\boldsymbol{\Sigma} = \mathbb{E}_{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P}\boldsymbol{\theta}\boldsymbol{\theta}^\top$ be a matrix in $\mathbb{R}^{d\times d}$. Denote $\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_d$ as a set of orthogonal eigenvectors of $\boldsymbol{\Sigma}$. Let $s_i = \mathbb{E}_{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P}\langle \boldsymbol{\theta}, \boldsymbol{\xi}_i\rangle^2$. Then,

$$\begin{aligned}
\mathop{\mathbb{E}}_{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P}\langle \boldsymbol{\theta}, \boldsymbol{\phi}\rangle &= \mathop{\mathbb{E}}_{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P}\sum_{i=1}^d \langle \boldsymbol{\theta}, \boldsymbol{\xi}_i\rangle\langle \boldsymbol{\xi}_i, \boldsymbol{\phi}\rangle \\
&= \sum_{i=1}^d \mathop{\mathbb{E}}_{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P}\sqrt{\frac{2\lambda}{s_i}}\langle \boldsymbol{\theta}, \boldsymbol{\xi}_i\rangle \cdot \sqrt{\frac{0.5 s_i}{\lambda}}\langle \boldsymbol{\xi}_i, \boldsymbol{\phi}\rangle \\
&\le \sum_{i=1}^d \mathop{\mathbb{E}}_{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P}\Big(\frac{\lambda}{s_i}\langle \boldsymbol{\theta}, \boldsymbol{\xi}_i\rangle^2 + \frac{0.25 s_i}{\lambda}\langle \boldsymbol{\xi}_i, \boldsymbol{\phi}\rangle^2\Big), \qquad (11)
\end{aligned}$$

where the first equality follows from $\boldsymbol{\theta} = \sum_{i=1}^d \langle \boldsymbol{\theta}, \boldsymbol{\xi}_i\rangle \boldsymbol{\xi}_i$ and the inequality follows from $ab \le 0.5a^2 + 0.5b^2$. For the first term, we have

$$\sum_{i=1}^d \mathop{\mathbb{E}}_{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P}\frac{\lambda}{s_i}\langle \boldsymbol{\theta}, \boldsymbol{\xi}_i\rangle^2 = \sum_{i=1}^d \lambda = \lambda d. \qquad (12)$$

For the second term, it holds that

$$\begin{aligned}
\frac{0.25}{\lambda}\sum_{i=1}^d \mathop{\mathbb{E}}_{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P} s_i\langle \boldsymbol{\xi}_i, \boldsymbol{\phi}\rangle^2 &= \frac{0.25}{\lambda}\sum_{i=1}^d \mathop{\mathbb{E}}_{\substack{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P\\(\boldsymbol{\theta}',\boldsymbol{\phi}')\sim P}}\langle \boldsymbol{\theta}, \boldsymbol{\xi}_i\rangle^2\langle \boldsymbol{\xi}_i, \boldsymbol{\phi}'\rangle^2 \\
&= \frac{0.25}{\lambda}\sum_{i=1}^d\sum_{j=1}^d \mathop{\mathbb{E}}_{\substack{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P\\(\boldsymbol{\theta}',\boldsymbol{\phi}')\sim P}}\langle \boldsymbol{\theta}, \boldsymbol{\xi}_i\rangle\langle \boldsymbol{\xi}_i, \boldsymbol{\phi}'\rangle\langle \boldsymbol{\theta}, \boldsymbol{\xi}_j\rangle\langle \boldsymbol{\xi}_j, \boldsymbol{\phi}'\rangle \\
&= \frac{0.25}{\lambda}\mathop{\mathbb{E}}_{\substack{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P\\(\boldsymbol{\theta}',\boldsymbol{\phi}')\sim P}}\Big(\sum_{i=1}^d \langle \boldsymbol{\theta}, \boldsymbol{\xi}_i\rangle\langle \boldsymbol{\xi}_i, \boldsymbol{\phi}'\rangle\Big)^2 \\
&= \frac{0.25}{\lambda}\mathop{\mathbb{E}}_{\substack{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P\\(\boldsymbol{\theta}',\boldsymbol{\phi}')\sim P}}\langle \boldsymbol{\theta}, \boldsymbol{\phi}'\rangle^2, \qquad (13)
\end{aligned}$$

where the first equality follows from the definition of s_i, the second equality holds since $\mathbb{E}_{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P}\langle \boldsymbol{\theta}, \boldsymbol{\xi}_i\rangle\langle \boldsymbol{\theta}, \boldsymbol{\xi}_j\rangle = \boldsymbol{\xi}_i^\top\boldsymbol{\Sigma}\boldsymbol{\xi}_j = 0$ for all $i \ne j$ and the final equality follows from the definition of $\boldsymbol{\xi}_i$. By plugging (12) and (13) into (11), we can conclude that

$$\mathop{\mathbb{E}}_{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P}\langle \boldsymbol{\theta}, \boldsymbol{\phi}\rangle \le \lambda d + \frac{0.25}{\lambda}\mathop{\mathbb{E}}_{\substack{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P\\(\boldsymbol{\theta}',\boldsymbol{\phi}')\sim P}}\langle \boldsymbol{\theta}, \boldsymbol{\phi}'\rangle^2.$$

□

Using this lemma, we can bound the expectation of the inner product of two correlated random variables using the expectation of the inner product of two independent random variables with the corresponding margin. To provide a nearly minimax bound with optimal dependence on misspecification, we further propose a decoupling lemma that takes misspecification into account:

Lemma C.2 (Decoupling Lemma with Misspecification)**.** *Let P be a joint distribution over two $\mathbb{R}^d$ space, i.e., $P \in \Delta(\mathbb{R}^d \times \mathbb{R}^d)$. Let $C(\cdot) : \mathbb{R}^d \to \mathbb{R}$ be a function such that $|C(\boldsymbol{\theta})| \le \zeta$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$. For any constant $\lambda > 0$, we have*

$$\mathop{\mathbb{E}}_{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P}\langle \boldsymbol{\theta}, \boldsymbol{\phi}\rangle \le \lambda(d+1) + \frac{4\zeta^2}{\lambda} + 4\zeta + \frac{0.25}{\lambda} \mathop{\mathbb{E}}_{\substack{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P\\(\boldsymbol{\theta}',\boldsymbol{\phi}')\sim P}}\Big[\langle \boldsymbol{\theta}, \boldsymbol{\phi}'\rangle^2 + 8C(\boldsymbol{\theta})\cdot\langle \boldsymbol{\theta}, \boldsymbol{\phi}'\rangle\Big],$$

where $(\boldsymbol{\theta}, \boldsymbol{\phi})$ on LHS is a sample from P while $(\boldsymbol{\theta}, \boldsymbol{\phi})$ and $(\boldsymbol{\theta}', \boldsymbol{\phi}')$ on RHS are two independent samples from P.

Proof. Let P_+ be an auxiliary distribution over $\Delta(\mathbb{R}^{d+1}, \mathbb{R}^{d+1})$ in which each element is $(\boldsymbol{\theta}_+, \boldsymbol{\phi}_+) = ([\boldsymbol{\theta}^\top, C(\boldsymbol{\theta})]^\top, [\boldsymbol{\phi}^\top, 1]^\top)$ where $(\boldsymbol{\theta}, \boldsymbol{\phi}) \sim P$. For two independent samples $(\boldsymbol{\theta}_+, \boldsymbol{\phi}_+)$ and $(\boldsymbol{\theta}'_+, \boldsymbol{\phi}'_+)$, we have

$$\langle \boldsymbol{\theta}_+, \boldsymbol{\phi}_+\rangle = \langle \boldsymbol{\theta}, \boldsymbol{\phi}\rangle + 4C(\boldsymbol{\theta}), \tag{14}$$

$$\langle \boldsymbol{\theta}_+, \boldsymbol{\phi}'_+\rangle^2 = \langle \boldsymbol{\theta}, \boldsymbol{\phi}'\rangle^2 + 8C(\boldsymbol{\theta})\cdot\langle \boldsymbol{\theta}, \boldsymbol{\phi}'\rangle + 16C(\boldsymbol{\theta})^2. \tag{15}$$

Apply Lemma C.1 to distribution P_+, we have

$$\mathop{\mathbb{E}}_{(\boldsymbol{\theta}_+,\boldsymbol{\phi}_+)\sim P_+}\langle \boldsymbol{\theta}_+, \boldsymbol{\phi}_+\rangle \le \lambda(d+1) + \frac{0.25}{\lambda} \mathop{\mathbb{E}}_{\substack{(\boldsymbol{\theta}_+,\boldsymbol{\phi}_+)\sim P_+\\(\boldsymbol{\theta}'_+,\boldsymbol{\phi}'_+)\sim P_+}}\langle \boldsymbol{\theta}_+, \boldsymbol{\phi}'_+\rangle^2. \tag{16}$$

Plugging (14) and (15) into (16), we have

$$\mathop{\mathbb{E}}_{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P}\Big[\langle \boldsymbol{\theta}, \boldsymbol{\phi}\rangle + 4C(\boldsymbol{\theta})\Big] \le \lambda(d+1) + \frac{0.25}{\lambda} \mathop{\mathbb{E}}_{\substack{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P\\(\boldsymbol{\theta}',\boldsymbol{\phi}')\sim P}}\Big[\langle \boldsymbol{\theta}, \boldsymbol{\phi}'\rangle^2 + 8C(\boldsymbol{\theta})\cdot\langle \boldsymbol{\theta}, \boldsymbol{\phi}'\rangle + 16C(\boldsymbol{\theta})^2\Big].$$

Since $|C(\boldsymbol{\theta})| \le \zeta$ always holds, we can further conclude that

$$\mathop{\mathbb{E}}_{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P}\langle \boldsymbol{\theta}, \boldsymbol{\phi}\rangle \le \lambda(d+1) + \frac{4\zeta^2}{\lambda} + 4\zeta + \frac{0.25}{\lambda} \mathop{\mathbb{E}}_{\substack{(\boldsymbol{\theta},\boldsymbol{\phi})\sim P\\(\boldsymbol{\theta}',\boldsymbol{\phi}')\sim P}}\Big[\langle \boldsymbol{\theta}, \boldsymbol{\phi}'\rangle^2 + 8C(\boldsymbol{\theta})\cdot\langle \boldsymbol{\theta}, \boldsymbol{\phi}'\rangle\Big].$$

□

Denote $\Omega_-^{(t)} = \Omega^{(t-1)} \cup \{\mathcal{A}^{(t)}, f^{(t)}\}$ be the history before the agent chooses action on round t. Let $\boldsymbol{\Theta}^* = [\boldsymbol{\theta}_1^*, \cdots, \boldsymbol{\theta}_T^*]$ be the optimal policy within ϵ-net $\mathcal{H}_\epsilon$. For any linear policy $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_T]$, denote

$$\Delta_t(\boldsymbol{\Theta}, \mathbf{a}^{(t)}) := \langle \mathbf{a}^{(t)}, \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*\rangle, \Delta_t^*(\boldsymbol{\Theta}) := \max_{\mathbf{a}\in\mathcal{A}^{(t)}}\langle \mathbf{a}, \boldsymbol{\theta}_t\rangle - \max_{\mathbf{a}\in\mathcal{A}^{(t)}}\langle \mathbf{a}, \boldsymbol{\theta}_t^*\rangle.$$

Note Δ_t is referred to the Bellman error and Δ_t^* is referred to the feel-good error. The next lemma shows that we can decomposed the expected regret using these two notions:

Lemma C.3. *Under Assumption 3.1, the expected regret on round t satisfies*

$$\mathop{\mathbb{E}}_{\mathbf{a}^{(t)}|\Omega_-^{(t)}}\Big[f^{(t)}(\mathbf{a}_*^{(t)}) - f^{(t)}(\mathbf{a}^{(t)})\Big] \le \mathop{\mathbb{E}}_{\mathbf{a}^{(t)},\boldsymbol{\Theta}^{(t)}|\Omega_-^{(t)}}\Big[\Delta_t(\boldsymbol{\Theta}^{(t)}, \mathbf{a}^{(t)}) - \Delta_t^*(\boldsymbol{\Theta}^{(t)})\Big] + 2(\epsilon + \zeta).$$

Proof. For any action $\mathbf{a} \in \mathcal{A}$, we have

$$\langle \mathbf{a}, \boldsymbol{\theta}_t^*\rangle \le f^{(t)}(\mathbf{a}) + |f^{(t)}(\mathbf{a}) - \langle \mathbf{a}, \boldsymbol{\theta}(f^{(t)})\rangle| + |\langle \mathbf{a}, \boldsymbol{\theta}(f^{(t)})\rangle - \langle \mathbf{a}, \boldsymbol{\theta}_t^*\rangle| \le f^{(t)}(\mathbf{a}) + \epsilon + \zeta,$$

where the first inequality follows from triangle inequality and the second inequality holds follows from Assumption 3.1 and the definition of ϵ-net with $|\langle \mathbf{a}, \boldsymbol{\theta}(f^{(t)})\rangle - \langle \mathbf{a}, \boldsymbol{\theta}_t^*\rangle| \le \|\mathbf{a}\|_2 \cdot \|\boldsymbol{\theta}(f^{(t)}) - \boldsymbol{\theta}_t^*\|_2 \le \zeta$. Similarly, we have $\langle \mathbf{a}, \boldsymbol{\theta}_t^*\rangle \ge f^{(t)}(\mathbf{a}) - (\epsilon + \zeta)$. Thus,

$$\mathop{\mathbb{E}}_{\mathbf{a}^{(t)}|\Omega_-^{(t)}}\Big[f^{(t)}(\mathbf{a}_*^{(t)}) - f^{(t)}(\mathbf{a}^{(t)})\Big] \le \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}|\Omega_-^{(t)}}\Big[\langle \mathbf{a}_*^{(t)}, \boldsymbol{\theta}_t^*\rangle - \langle \mathbf{a}^{(t)}, \boldsymbol{\theta}_t^*\rangle\Big] + 2(\epsilon + \zeta). \tag{17}$$

Moreover, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{a}^{(t)}|\Omega_-^{(t)}}\Big[\langle \mathbf{a}_*^{(t)}, \boldsymbol{\theta}_t^*\rangle - \langle \mathbf{a}^{(t)}, \boldsymbol{\theta}_t^*\rangle\Big] \\
\leq &\mathop{\mathbb{E}}_{\mathbf{a}^{(t)}|\Omega_-^{(t)}}\Big[\max_{\mathbf{a}\in\mathcal{A}^{(t)}}\langle \mathbf{a}, \boldsymbol{\theta}_t^*\rangle - \langle \mathbf{a}^{(t)}, \boldsymbol{\theta}_t^*\rangle\Big] \\
= &\mathop{\mathbb{E}}_{\mathbf{a}^{(t)}|\Omega_-^{(t)}}\Big[\max_{\mathbf{a}\in\mathcal{A}^{(t)}}\langle \mathbf{a}, \boldsymbol{\theta}_t^*\rangle - \max_{\mathbf{a}\in\mathcal{A}^{(t)}}\langle \mathbf{a}, \boldsymbol{\theta}_t\rangle + \langle \mathbf{a}^{(t)}, \boldsymbol{\theta}_t\rangle - \langle \mathbf{a}^{(t)}, \boldsymbol{\theta}_t^*\rangle\Big] \\
= &\mathop{\mathbb{E}}_{\mathbf{a}^{(t)},\boldsymbol{\Theta}^{(t)}|\Omega_-^{(t)}}\Big[\Delta_t(\boldsymbol{\Theta}^{(t)}, \mathbf{a}^{(t)}) - \Delta_t^*(\boldsymbol{\Theta}^{(t)})\Big]. \qquad (18)
\end{aligned}$$

where the first equality follows from the fact that action $\mathbf{a}^{(t)}$ maximizes $\langle \mathbf{a}, \boldsymbol{\theta}_t\rangle$ and thus $\max_{\mathbf{a}\in\mathcal{A}^{(t)}}\langle \mathbf{a}, \boldsymbol{\theta}_t\rangle = \langle \mathbf{a}^{(t)}, \boldsymbol{\theta}_t\rangle$. By combining (17) and (18), we obtain the desired result. □

The following lemma shows a connection between Δ_t and a potential function.

Lemma C.4. *Define potential function*

$$\Phi^{(t)} := \frac{1}{\beta\lambda}\log\sum_{\boldsymbol{\Theta}\in\mathcal{H}_\epsilon}\exp\Big(\beta\sum_{\tau=1}^{t}\big(L^{(\tau)}(\boldsymbol{\Theta}^*) - L^{(\tau)}(\boldsymbol{\Theta})\big)\Big).$$

Let $C_t(\mathbf{a}^{(t)}) := f^{(t)}(\mathbf{a}^{(t)}) - \langle \mathbf{a}^{(t)}, \boldsymbol{\theta}_t^*\rangle$ be the misspecification of $\boldsymbol{\Theta}^*$ on action $\mathbf{a}^{(t)}$ with respect to ground truth $f^{(t)}$. For any $\epsilon \in [0,1]$, $\beta \in [0, 0.01]$ and $\lambda \in [0,1]$, under Assumptions 3.1 and 3.2, the expected increment of potential function on any round t satisfies

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{a}^{(t)},r^{(t)}|\Omega_-^{(t)}}\Big[\Phi^{(t)} - \Phi^{(t-1)}\Big] \leq &-\frac{0.25}{\lambda}\mathop{\mathbb{E}}_{\mathbf{a}^{(t)}|\Omega_-^{(t)}}\mathop{\mathbb{E}}_{\boldsymbol{\Theta}\sim P^{(t)}}\Big[\Delta_t(\boldsymbol{\Theta}, \mathbf{a}^{(t)})^2 + 8C_t(\mathbf{a}^{(t)})\Delta_t(\boldsymbol{\Theta}, \mathbf{a}^{(t)})\Big] \\
&+ \mathop{\mathbb{E}}_{\boldsymbol{\Theta}\sim P^{(t)}}\Delta_t^*(\boldsymbol{\Theta}) + 4\beta\lambda + \frac{8\beta}{\lambda}(\epsilon+\zeta)^2.
\end{aligned}$$

Proof. the expected increment of the potential function in round t conditional on the reward function and context in round t satisfies:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{a}^{(t)},r^{(t)}|\Omega_-^{(t)}}\Big[\Phi^{(t)} - \Phi^{(t-1)}\Big] \\
= &\frac{1}{\beta\lambda}\mathop{\mathbb{E}}_{\mathbf{a}^{(t)},r^{(t)}|\Omega_-^{(t)}}\left[\log\frac{\sum_{\boldsymbol{\Theta}\in\mathcal{H}_\epsilon}\exp\Big(\beta\sum_{\tau=1}^{t}\big(L^{(\tau)}(\boldsymbol{\Theta}^*) - L^{(\tau)}(\boldsymbol{\Theta})\big)\Big)}{\sum_{\boldsymbol{\Theta}\in\mathcal{H}_\epsilon}\exp\Big(\beta\sum_{\tau=1}^{t-1}\big(L^{(\tau)}(\boldsymbol{\Theta}^*) - L^{(\tau)}(\boldsymbol{\Theta})\big)\Big)}\right] \\
= &\frac{1}{\beta\lambda}\mathop{\mathbb{E}}_{\mathbf{a}^{(t)},r^{(t)}|\Omega_-^{(t)}}\left[\log\frac{\sum_{\boldsymbol{\Theta}\in\mathcal{H}_\epsilon}\exp\big(-\beta\sum_{\tau=1}^{t-1}L^{(\tau)}(\boldsymbol{\Theta})\big)\cdot\exp\big(\beta\big(L^{(t)}(\boldsymbol{\Theta}^*) - L^{(t)}(\boldsymbol{\Theta})\big)\big)}{\sum_{\boldsymbol{\Theta}\in\mathcal{H}_\epsilon}\exp\big(-\beta\sum_{\tau=1}^{t-1}L^{(\tau)}(\boldsymbol{\Theta})\big)}\right] \\
= &\frac{1}{\beta\lambda}\mathop{\mathbb{E}}_{\mathbf{a}^{(t)},r^{(t)}|\Omega_-^{(t)}}\Big[\log\mathop{\mathbb{E}}_{\boldsymbol{\Theta}\sim P^{(t)}}\exp\Big(\beta\big(L^{(t)}(\boldsymbol{\Theta}^*) - L^{(t)}(\boldsymbol{\Theta})\big)\Big)\Big] \\
\leq &\frac{1}{\beta\lambda}\mathop{\mathbb{E}}_{\mathbf{a}^{(t)}|\Omega_-^{(t)}}\Big[\log\mathop{\mathbb{E}}_{r^{(t)}|\Omega_-^{(t)},\mathbf{a}^{(t)}}\mathop{\mathbb{E}}_{\boldsymbol{\Theta}\sim P^{(t)}}\exp\Big(\beta\big(L^{(t)}(\boldsymbol{\Theta}^*) - L^{(t)}(\boldsymbol{\Theta})\big)\Big)\Big]. \qquad (19)
\end{aligned}$$

where the last equality follows from the construction of $P^{(t)}$ and the last inequality follows from Jensen's inequality. Moreover, we can decompose the loss function for any policy $\boldsymbol{\Theta}$ according to

$$\begin{aligned}
L^{(t)}(\boldsymbol{\Theta}) &= (r^{(t)} - \langle \mathbf{a}^{(t)}, \boldsymbol{\theta}_t\rangle)^2 - \lambda\max_{\mathbf{a}\in\mathcal{A}^{(t)}}\langle \mathbf{a}, \boldsymbol{\theta}_t\rangle \\
&= (f^{(t)}(\mathbf{a}^{(t)}) + \xi^{(t)} - \langle \mathbf{a}^{(t)}, \boldsymbol{\theta}_t\rangle)^2 - \lambda\max_{\mathbf{a}\in\mathcal{A}^{(t)}}\langle \mathbf{a}, \boldsymbol{\theta}_t\rangle \\
&= (C_t(\mathbf{a}^{(t)}) + \xi^{(t)} - \Delta_t(\boldsymbol{\Theta}, \mathbf{a}^{(t)}))^2 - \lambda\max_{\mathbf{a}\in\mathcal{A}^{(t)}}\langle \mathbf{a}, \boldsymbol{\theta}_t\rangle, \qquad (20)
\end{aligned}$$

where the first equality follows from the definition $r^{(t)} = f^{(t)}(\mathbf{a}^{(t)}) + \xi^{(t)}$ and the second equality follows from the definition of C_t. For the optimal policy $\mathbf{\Theta}^*$ in ϵ-net, one can see that

$$L^{(t)}(\mathbf{\Theta}^*) = (C_t(\mathbf{a}^{(t)}) + \xi^{(t)})^2 - \lambda \max_{\mathbf{a}\in\mathcal{A}^{(t)}} \langle \mathbf{a}, \boldsymbol{\theta}_t^* \rangle. \tag{21}$$

Combining (20) and (21), the terms in the exponent can be computed by

$$\begin{aligned} &\beta(L^{(t)}(\mathbf{\Theta}^*) - L^{(t)}(\mathbf{\Theta})) \\ &= -2\beta C_t(\mathbf{a}^{(t)})\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)}) - 2\beta\xi^{(t)}\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)}) - \beta\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)})^2 + \beta\lambda\Delta_t^*(\mathbf{\Theta}). \end{aligned} \tag{22}$$

Since $\xi^{(t)}$ is zero mean random variables with range 2, according to Hoeffding's lemma, we have

$$\mathop{\mathbb{E}}_{r^{(t)}|\Omega_-^{(t)}, a^{(t)}} \Big[\exp(-2\beta\xi^{(t)}\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)}) \Big] \leq \exp\Big(2\beta^2\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)})^2\Big). \tag{23}$$

In case that $\beta \in [0, 0.01]$, we have $2\beta^2 - \beta \leq -0.5\beta$. Plugging (22) and (23) into (19) gives

$$\begin{aligned} &\mathop{\mathbb{E}}_{\mathbf{a}^{(t)}, r^{(t)}|\Omega_-^{(t)}} \Big[\Phi^{(t)} - \Phi^{(t-1)} \Big] \\ &\leq \frac{1}{\beta\lambda} \mathop{\mathbb{E}}_{a^{(t)}|\Omega_-^{(t)}} \Bigg[\log \mathop{\mathbb{E}}_{\mathbf{\Theta}\sim P^{(t)}} \exp\Big(-2\beta C_t(\mathbf{a}^{(t)})\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)}) - 0.5\beta\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)})^2 + \beta\lambda\Delta_t^*(\mathbf{\Theta}) \Big) \Bigg]. \end{aligned}$$

According to Cauchy-Schwarz inequality, we can decompose RHS by

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}, r^{(t)}|\Omega_-^{(t)}} \Big[\Phi^{(t)} - \Phi^{(t-1)} \Big] \leq & \underbrace{\frac{0.5}{\beta\lambda} \mathop{\mathbb{E}}_{a^{(t)}|\Omega_-^{(t)}} \Bigg[\log \mathop{\mathbb{E}}_{\mathbf{\Theta}\sim P^{(t)}} \exp\Big(-4\beta C_t(\mathbf{a}^{(t)})\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)}) \Big) \Bigg]}_{I_1} \\ & + \underbrace{\frac{0.25}{\beta\lambda} \mathop{\mathbb{E}}_{a^{(t)}|\Omega_-^{(t)}} \Bigg[\log \mathop{\mathbb{E}}_{\mathbf{\Theta}\sim P^{(t)}} \exp\Big(-2\beta\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)})^2 \Big) \Bigg]}_{I_2} \\ & + \underbrace{\frac{0.25}{\beta\lambda} \mathop{\mathbb{E}}_{a^{(t)}|\Omega_-^{(t)}} \Bigg[\log \mathop{\mathbb{E}}_{\mathbf{\Theta}\sim P^{(t)}} \exp\Big(4\beta\lambda\Delta_t^*(\mathbf{\Theta}) \Big) \Bigg]}_{I_3}. \end{aligned} \tag{24}$$

For the first term,

$$\begin{aligned} I_1 &\leq \frac{0.5}{\beta\lambda} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}|\Omega_-^{(t)}} \Bigg[\log \mathop{\mathbb{E}}_{\mathbf{\Theta}\sim P^{(t)}} \Big[1 - 4\beta C_t(\mathbf{a}^{(t)})\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)}) + 16\beta^2(\epsilon+\zeta)^2 \Big] \Bigg] \\ &\leq -\frac{0.25}{\lambda} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}|\Omega_-^{(t)}} \mathop{\mathbb{E}}_{\mathbf{\Theta}\sim P^{(t)}} 8C_t(\mathbf{a}^{(t)})\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)}) + \frac{8\beta}{\lambda}(\epsilon+\zeta)^2 \end{aligned} \tag{25}$$

where the first inequality follows from $e^x \leq 1 + x + t^2$ for $|x| \leq t \leq 1$ with $|C_t(\mathbf{a}^{(t)})| \leq \epsilon + \zeta \leq 2$, $|\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)})| \leq 2$ and $\beta \in [0, 0.01]$, and the second inequality follows from $\log(1+x) \leq x$ for $x \geq -1$.

For the second term,

$$\begin{aligned} I_2 &\leq \frac{0.25}{\beta\lambda} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}|\Omega_-^{(t)}} \Bigg[\log \mathop{\mathbb{E}}_{\mathbf{\Theta}\sim P^{(t)}} \Big[1 - \beta\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)})^2 \Big] \Bigg] \\ &\leq -\frac{0.25}{\lambda} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}|\Omega_-^{(t)}} \mathop{\mathbb{E}}_{\mathbf{\Theta}\sim P^{(t)}} \Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)})^2, \end{aligned} \tag{26}$$

where the first inequality follows from $e^{-x} \leq 1 - 0.5x$ for $x \in [0, 1]$ with $|\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)})| \leq 2$ and $\beta \in [0, 0.01]$, and the second inequality follows from $\log(1 - x) \leq -x$ for all $x \leq 1$.

For the third term,

$$\begin{aligned} I_3 &\leq \frac{0.25}{\beta\lambda} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}|\Omega_-^{(t)}} \left[\log \mathop{\mathbb{E}}_{\mathbf{\Theta}\sim P^{(t)}} \left[1 + 4\beta\lambda\Delta_t^*(\mathbf{\Theta}) + 16\beta^2\lambda^2 \right] \right] \\ &\leq \mathop{\mathbb{E}}_{\mathbf{\Theta}\sim P^{(t)}} \Delta_t^*(\mathbf{\Theta}) + 4\beta\lambda, \end{aligned} \tag{27}$$

where the first inequality follows from $e^x \leq 1 + x + t^2$ for $|x| \leq t \leq 1$ with $|\Delta_t^*(\mathbf{\Theta})| \leq 2$, $\beta \in [0, 0.01]$ and $\lambda \in [0, 1]$, and the second inequality follows from $\log(1 + x) \leq x$ for $x \geq -1$.

Combining (25), (26) and (27) with (24), we have

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}, r^{(t)}|\Omega_-^{(t)}} \left[\Phi^{(t)} - \Phi^{(t-1)} \right] \leq -\frac{0.25}{\lambda} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}|\Omega_-^{(t)}} \mathop{\mathbb{E}}_{\mathbf{\Theta}\sim P^{(t)}} \left[\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)})^2 + 8C_t(\mathbf{a}^{(t)})\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)}) \right] \\ + \mathop{\mathbb{E}}_{\mathbf{\Theta}\sim P^{(t)}} \Delta_t^*(\mathbf{\Theta}) + 4\beta\lambda + \frac{8\beta}{\lambda}(\epsilon + \zeta)^2. \end{aligned}$$

□

The following Lemma upper bounds the total change of the potential:

Lemma C.5. *Under the settings of Lemma C.4, we have*

$$\Phi^{(0)} - \Phi^{(T)} \leq \frac{1}{\beta\lambda} \log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty}).$$

Proof. The statement can be proved by combining

$$\Phi^{(0)} = \frac{1}{\beta\lambda} \Big[\log \sum_{\mathbf{\Theta}\in\mathcal{H}_\epsilon} \exp(0) \Big] = \frac{1}{\beta\lambda} \log |\mathcal{H}_\epsilon|,$$

and

$$\Phi^{(T)} \geq \frac{1}{\beta\lambda} \mathop{\mathbb{E}}_{\Omega^{(T)}} \left[\log \max_{\mathbf{\Theta}\in\mathcal{H}_\epsilon} \exp\Big(\beta \sum_{\tau=1}^{T} (-L^{(\tau)}(\mathbf{\Theta}) + L^{(\tau)}(\mathbf{\Theta}^*)) \Big) \right] \geq 0,$$

where the second inequality follows from $\mathbf{\Theta}^* \in \mathcal{H}_\epsilon$. □

Note the chosen policy $\mathbf{\Theta}^{(t)}$ in Lemma C.3 is correlated with the selected action $\mathbf{a}^{(t)}$ while the random policy $\mathbf{\Theta}$ in Lemma C.4 is not. The following lemma presents a connection between these two notions.

Lemma C.6. *Let $C_t : \mathbb{R}^d \to \mathbb{R}$ be a function such that $|C_t(\mathbf{a})| \leq \epsilon + \zeta$ for any $\mathbf{a} \in \mathcal{A}$. Then,*

$$\mathop{\mathbb{E}}_{\mathbf{\Theta}^{(t)}|\Omega_-^{(t)}} \Delta_t^*(\mathbf{\Theta}^{(t)}) = \mathop{\mathbb{E}}_{\mathbf{\Theta}\sim P^{(t)}} \Delta_t^*(\mathbf{\Theta}),$$

and also

$$\begin{aligned} \mathop{\mathbb{E}}_{a^{(t)}, \mathbf{\Theta}^{(t)}|\Omega_-^{(t)}} \Delta_t(\mathbf{\Theta}^{(t)}, \mathbf{a}^{(t)}) \leq \frac{0.25}{\lambda} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}|\Omega_-^{(t)}} \mathop{\mathbb{E}}_{\mathbf{\Theta}\sim P^{(t)}} \left[\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)})^2 + 8C_t(\mathbf{a}^{(t)})\Delta_t(\mathbf{\Theta}, \mathbf{a}^{(t)}) \right] \\ + \lambda(d+1) + \frac{4}{\lambda}(\epsilon + \zeta)^2 + 4(\epsilon + \zeta). \end{aligned}$$

Proof. According to the construction of $P^{(t)}$, the conditional distribution $\mathbf{\Theta}^{(t)}|\Omega_-^{(t)}$ is identical to distribution $\mathbf{\Theta} \sim P^{(t)}$. This directly gives

$$\mathop{\mathbb{E}}_{\mathbf{\Theta}^{(t)}|\Omega_-^{(t)}} \Delta_t^*(\mathbf{\Theta}^{(t)}) = \mathop{\mathbb{E}}_{\mathbf{\Theta}\sim P^{(t)}} \Delta_t^*(\mathbf{\Theta}).$$

Moreover, apply Lemma C.2 to joint distribution $(\mathbf{a}^{(t)}, \boldsymbol{\theta}_t^{(t)} - \boldsymbol{\theta}_t^*)$ conditional on $\Omega_-^{(t)}$, we have

$$\mathop{\mathbb{E}}_{a^{(t)}, \boldsymbol{\theta}_t^{(t)} | \Omega_-^{(t)}} \langle \mathbf{a}^{(t)}, \boldsymbol{\theta}_t^{(t)} - \boldsymbol{\theta}_t^* \rangle \leq \frac{0.25}{\lambda} \mathop{\mathbb{E}}_{a^{(t)} | \Omega_-^{(t)}} \mathop{\mathbb{E}}_{\boldsymbol{\theta}_t \sim P^{(t)}} \left[\langle \mathbf{a}^{(t)}, \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^* \rangle^2 + 8C_t(\mathbf{a}^{(t)}) \langle \mathbf{a}^{(t)}, \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^* \rangle \right] + \lambda(d+1) + \frac{4}{\lambda}(\epsilon + \zeta)^2 + 4(\epsilon + \zeta).$$

The desired statement can be obtained using the definition of Δ_t. □

Now, we are ready to prove Theorem 5.1:

Proof of Theorem 5.1. We choose the parameters according to

$$\lambda := \sqrt{\frac{\log |\mathcal{H}_\epsilon| + (\epsilon + \zeta)^2 T}{dT}} = \mathcal{O}(1/d), \beta := 0.01. \tag{28}$$

Denote

$$\mathcal{X}_t := \frac{0.25}{\lambda} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)} | \Omega_-^{(t)}} \mathop{\mathbb{E}}_{\boldsymbol{\Theta} \sim P^{(t)}} \left[\Delta_t(\boldsymbol{\Theta}, \mathbf{a}^{(t)})^2 + 8C_t(\mathbf{a}^{(t)}) \Delta_t(\boldsymbol{\Theta}, \mathbf{a}^{(t)}) \right] - \mathop{\mathbb{E}}_{\boldsymbol{\Theta} \sim P^{(t)}} \Delta_t^*(\boldsymbol{\Theta}).$$

According to Lemma C.6, we have

$$\mathop{\mathbb{E}}_{\mathbf{a}^{(t)}, \boldsymbol{\Theta}^{(t)} | \Omega_-^{(t)}} \left[\Delta_t(\boldsymbol{\Theta}^{(t)}, \mathbf{a}^{(t)}) - \Delta_t^*(\boldsymbol{\Theta}^{(t)}) \right] \leq \mathcal{X}_t + \lambda(d+1) + \frac{4}{\lambda}(\epsilon + \zeta)^2 + 4(\epsilon + \zeta). \tag{29}$$

According to Lemma C.4, we have

$$\mathop{\mathbb{E}}_{\mathbf{a}^{(t)}, r^{(t)} | \Omega_-^{(t)}} \left[\Phi^{(t)} - \Phi^{(t-1)} \right] \leq -\mathcal{X}_t + 4\beta\lambda + \frac{8\beta}{\lambda}(\epsilon + \zeta)^2. \tag{30}$$

Combining (29) and (30) gives

$$\begin{aligned} &\mathop{\mathbb{E}}_{\mathbf{a}^{(t)}, \boldsymbol{\Theta}^{(t)} | \Omega_-^{(t)}} \left[\Delta_t(\boldsymbol{\Theta}^{(t)}, \mathbf{a}^{(t)}) - \Delta_t^*(\boldsymbol{\Theta}^{(t)}) \right] \\ \leq &\mathop{\mathbb{E}}_{\mathbf{a}^{(t)}, r^{(t)} | \Omega_-^{(t)}} \left[\Phi^{(t-1)} - \Phi^{(t)} \right] + \lambda(d + 1 + 4\beta) + \frac{4 + 8\beta}{\lambda}(\epsilon + \zeta)^2 + 4(\epsilon + \zeta). \end{aligned} \tag{31}$$

Moreover, we have

$$\begin{aligned} &\mathbb{E}[\mathrm{Regret}(T)] \\ = &\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)} | \Omega_-^{(t)}} \left[f^{(t)}(\mathbf{a}_*^{(t)}) - f^{(t)}(\mathbf{a}^{(t)}) \right] \\ \leq &\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}, \boldsymbol{\Theta}^{(t)} | \Omega_-^{(t)}} \left[\Delta_t(\boldsymbol{\Theta}^{(t)}, \mathbf{a}^{(t)}) - \Delta_t^*(\boldsymbol{\Theta}^{(t)}) \right] + 2(\epsilon + \zeta)T \\ \leq &\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\mathbf{a}^{(t)}, r^{(t)} | \Omega_-^{(t)}} \left[\Phi^{(t-1)} - \Phi^{(t)} \right] + \lambda(d + 1 + 4\beta)T + \frac{4 + 8\beta}{\lambda}(\epsilon + \zeta)^2 T + 6(\epsilon + \zeta)T \\ \leq &\frac{1}{\beta\lambda} \log |\mathcal{H}_\epsilon| + \lambda(d + 1 + 4\beta)T + \frac{4 + 8\beta}{\lambda}(\epsilon + \zeta)^2 T + 6(\epsilon + \zeta)T, \end{aligned}$$

where the first inequality follows from Lemma C.3, the second inequality follows from (31), and the last inequality follows from Lemma C.5. With the parameter defined in (28), we conclude that

$$\mathbb{E}[\mathrm{Regret}(T)] \leq O\Big(\sqrt{dT \log |\mathcal{H}_\epsilon|} + T\sqrt{d}(\epsilon + \zeta) \Big).$$

□

D PROOF OF COVERING NUMBER FROM SECTION 5.1

In this section, we prove covering number bounds for some specific linear bandit settings. We first introduce a classic result on the covering number of the unit ball.

Lemma D.1 (Lemma 5.2 in Vershynin (2012), Restated)**.** *The unit Euclidean ball B^d equipped with the Euclidean metric satisfies for every $\epsilon > 0$ that*

$$N(B^d, \epsilon, \|\cdot\|_2) \leq (1 + 2\epsilon^{-1})^d.$$

With this lemma, we can directly prove the covering number of stationary linear bandits.

Lemma 5.1 (Covering Number for Misspecified Linear Bandits). *Under Assumptions 3.1 and 3.3, the metric entropy of linear policies satisfies*

$$\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty}) \leq \mathcal{O}(d \log \epsilon^{-1}).$$

Proof. We will prove this lemma by constructing an ϵ-net for $\mathcal{H}$. Let B_ϵ^d be the minimum ϵ-net of unit Euclidean ball B^d in L_2 distance. Let $\mathcal{H}_\epsilon$ be the set that contains every matrix $[\widetilde{\boldsymbol{\theta}}, \cdots, \widetilde{\boldsymbol{\theta}}]$ for all $\widetilde{\boldsymbol{\theta}} \in B_\epsilon^d$. According to Lemma D.1, we have $|\mathcal{H}_\epsilon| = |B_\epsilon^d| \leq (1 + 2\epsilon^{-1})^d$.

We then show $\mathcal{H}_\epsilon$ is an ϵ-net of $\mathcal{H}$ under Assumption 3.3: For every linear policy $[\boldsymbol{\theta}, \cdots, \boldsymbol{\theta}]$, there exists $\widetilde{\boldsymbol{\theta}} \in B_\epsilon^d$ such that $\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\|_2 \leq \epsilon$ according to definition of B_ϵ^d. This shows $[\widetilde{\boldsymbol{\theta}}, \cdots, \widetilde{\boldsymbol{\theta}}] \in \mathcal{H}_\epsilon$ is an ϵ-approximation policy of $[\boldsymbol{\theta}, \cdots, \boldsymbol{\theta}]$ which implies $\mathcal{H}_\epsilon$ is indeed an ϵ-net of $\mathcal{H}$.

As a result, we can bound the metric entropy by

$$\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty}) \leq \log |\mathcal{H}_\epsilon| = \mathcal{O}(d \log \epsilon^{-1}).$$

□

One can naturally extend the proof to non-stationary linear bandits with bounded switches:

Lemma 5.2 (Covering Number for Non-Stationary Linear Bandits with Bounded Switches). *Under Assumptions 3.1 and 3.4, the metric entropy of linear policies satisfies*

$$\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty}) \leq \mathcal{O}(dS \log \epsilon^{-1} + S \log T).$$

Proof. We will prove this lemma by constructing an ϵ-net for $\mathcal{H}$. Let B_ϵ^d be the minimum ϵ-net of unit Euclidean ball B^d in L_2 distance. We build a set $\mathcal{H}_\epsilon$ by mapping any sequence $(t_1, \cdots, t_S)$ with $1 < t_1 < \cdots < t_S \leq T$ and $S + 1$ vectors $\widetilde{\boldsymbol{\theta}}_0, \cdots, \widetilde{\boldsymbol{\theta}}_S \in B_\epsilon^d$ to the matrix $[\widetilde{\boldsymbol{\theta}}_0, \cdots, \widetilde{\boldsymbol{\theta}}_0, \widetilde{\boldsymbol{\theta}}_1, \cdots, \widetilde{\boldsymbol{\theta}}_1, \cdots, \widetilde{\boldsymbol{\theta}}_S]$ where the first $\widetilde{\boldsymbol{\theta}}_i$ occurs at index t_i. t_i describes the round of i-th switch and $\widetilde{\boldsymbol{\theta}}_i$ characterize the parameter after the switch. The size of $\mathcal{H}_\epsilon$ can be bounded by

$$|\mathcal{H}_\epsilon| \leq \binom{T}{S} \cdot |B_\epsilon^d|^{S+1} \leq T^S \cdot (1 + 2\epsilon^{-1})^{d(S+1)},$$

where the first term is the number of possible (t_i) and the second term is the number of possible $\widetilde{\boldsymbol{\theta}}_i$.

We then show $\mathcal{H}_\epsilon$ is an ϵ-net of $\mathcal{H}$ under under Assumption 3.4: Every linear policy in $\mathcal{H}$ can also be written in form $[\boldsymbol{\theta}_0, \cdots, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_S]$ which can be covered by some $[\widetilde{\boldsymbol{\theta}}_0, \cdots, \widetilde{\boldsymbol{\theta}}_0, \widetilde{\boldsymbol{\theta}}_1, \cdots, \widetilde{\boldsymbol{\theta}}_1, \cdots, \widetilde{\boldsymbol{\theta}}_S]$ with the same switch locations and corresponding parameters. According to the definition of B_ϵ^d, $\|\boldsymbol{\theta}_i - \widetilde{\boldsymbol{\theta}}_i\| \leq \epsilon$ for every i and thus $\mathcal{H}_\epsilon$ is an ϵ-net of $\mathcal{H}$.

As a result, we can bound the metric entropy by

$$\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty}) \leq \log |\mathcal{H}_\epsilon| = \mathcal{O}(dS \log \epsilon^{-1} + S \log T).$$

□

For non-stationary linear bandits with bounded path length, we will show the ϵ-net for some bounded switches instances can be transformed into a desired covering:

Lemma 5.3 (Covering Number for Non-Stationary Linear Bandits with Bounded Path Length). *Under Assumptions 3.1 and 3.5, the metric entropy of linear policies satisfies*

$$\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty}) \leq \mathcal{O}(d \log \epsilon^{-1} + dP\epsilon^{-1} \log \epsilon^{-1} + P\epsilon^{-1} \log T).$$

Proof. Let $\widehat{\mathcal{H}}_{\epsilon/2}$ be the $\epsilon/2$-net covering of non-stationary bandits with no more than $2P\epsilon^{-1}$ (see Assumption 3.4). We will show $\widehat{\mathcal{H}}_{\epsilon/2}$ is an ϵ-net covering of non-stationary bandits with path length no more than P:

For some linear policy $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_T] \in \mathcal{H}$ such that $\sum_{t=1}^{T} \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}\| \leq P$, consider a sequence $1 = t_0 < t_1 < \cdots < t_M \leq T$ where t_i is the minimal index such that $\|\boldsymbol{\theta}_{t_i} - \boldsymbol{\theta}_{t_{i-1}}\| > \epsilon/2$ for all $i \geq 1$. Denote $\widehat{\boldsymbol{\Theta}} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{t_1}, \cdots, \boldsymbol{\theta}_{t_1}, \cdots, \boldsymbol{\theta}_{t_M}]$ as the linear policy, which quantizes the parameter drift in which the first $\boldsymbol{\theta}_{t_i}$ occurs at index t_i. Since the number of switch of $\widehat{\boldsymbol{\Theta}}$ is bounded by $M \leq P/(\epsilon/2) \leq 2P\epsilon^{-1}$, the $\epsilon/2$-net $\widehat{\mathcal{H}}_{\epsilon/2}$ contains an $\epsilon/2$ approximated policy of $\widehat{\boldsymbol{\Theta}}$. Denote the policy as $\widetilde{\boldsymbol{\Theta}}$.

Since deviation of each single round policy in the same slice is no more than $\epsilon/2$, we have $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{2,\infty} \leq \epsilon/2$. According to the definition of $\epsilon/2$-net, we also have $\|\widetilde{\boldsymbol{\Theta}} - \widehat{\boldsymbol{\Theta}}\|_{2,\infty} \leq \epsilon/2$. According to triangle inequality, we have $\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{2,\infty} \leq \epsilon$. This shows that $\boldsymbol{\Theta}$ always has an ϵ-approximation within $\widehat{\mathcal{H}}_{\epsilon/2}$ and thus $\widehat{\mathcal{H}}_{\epsilon/2}$ is indeed an ϵ-net of $\mathcal{H}$.

As a result, we can bound the metric entropy using Lemma 5.2

$$\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty}) \leq \log |\widehat{\mathcal{H}}_{\epsilon/2}| \leq \mathcal{O}(d \log \epsilon^{-1} + dP\epsilon^{-1} \log \epsilon^{-1} + P\epsilon^{-1} \log T),$$

where the extra $d \log \epsilon^{-1}$ term follows from the special case $S = 2P\epsilon^{-1} < 1$. □

Now we consider the lifelong linear bandits. Its covering number bound is essentially the covering number of low-rank matrices.

Lemma 5.4 (Covering Number for Lifelong Bandits with Shared Representation). *Under Assumptions 3.1 and 3.6, the metric entropy of linear policies satisfies*

$$\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty}) \leq \mathcal{O}(dk \log \epsilon^{-1} + kM \log \epsilon^{-1}).$$

Proof. We will prove this lemma by constructing an ϵ-net for $\mathcal{H}$. This is done by constructing an $\epsilon/2$-net for all possible feature extractor $\mathbf{B}$ and an $\epsilon/2$-net for all possible tasks specific vectors $\{\mathbf{w}_i\}_{i=1}^{M}$. In specific, we build a set $\mathcal{B}_{\epsilon/2}$ such that there exists $\widetilde{\mathbf{B}} \in \mathcal{B}_{\epsilon/2}$ such that $\|\mathbf{B} - \widetilde{\mathbf{B}}\|_2 \leq \epsilon/2$ and $\mathcal{W}_{\epsilon/2}$ such that there exists $\{\widetilde{\mathbf{w}}_i\}_{i=1}^{M} \in \mathcal{W}_{\epsilon/2}$ such that $\|\mathbf{w}_i - \widetilde{\mathbf{w}}_i\|_2 \leq \epsilon/2$ for all i. We generate the ϵ-net by mapping every pairs of $\widetilde{\mathbf{B}}$ and $\{\widetilde{\mathbf{w}}_i\}_{i=1}^{M}$ to matrix $[\widetilde{\boldsymbol{\theta}}^{(1)}, \cdots, \widetilde{\boldsymbol{\theta}}^{(T)}]$ where $\widetilde{\boldsymbol{\theta}}^{(t)} = \mathbf{B}\widetilde{\mathbf{w}}_{m^{(t)}}$.

In any round with $\boldsymbol{\theta}^{(t)} = \mathbf{B}\mathbf{w}_{m^{(t)}}$, the vector $\widetilde{\boldsymbol{\theta}}^{(t)} = \widetilde{\mathbf{B}}\widetilde{\mathbf{w}}_{m^{(t)}}$ with corresponding $\widetilde{\mathbf{B}}$ and $\widetilde{\mathbf{w}}_{m^{(t)}}$ in $\epsilon/2$-net satisfies

$$\begin{aligned}
\|\boldsymbol{\theta}^{(t)} - \widetilde{\boldsymbol{\theta}}^{(t)}\|_2 &= \|\mathbf{B}\mathbf{w}_{m^{(t)}} - \widetilde{\mathbf{B}}\widetilde{\mathbf{w}}_{m^{(t)}}\|_2 \\
&\leq \|\mathbf{B}\mathbf{w}_{m^{(t)}} - \mathbf{B}\widetilde{\mathbf{w}}_{m^{(t)}}\|_2 + \|\mathbf{B}\widetilde{\mathbf{w}}_{m^{(t)}} - \widetilde{\mathbf{B}}\widetilde{\mathbf{w}}_{m^{(t)}}\|_2 \\
&\leq \|\mathbf{B}\|_2 \|\mathbf{w}_{m^{(t)}} - \widetilde{\mathbf{w}}_{m^{(t)}}\|_2 + \|\mathbf{B} - \widetilde{\mathbf{B}}\|_2 \|\widetilde{\mathbf{w}}_{m^{(t)}}\|_2 \\
&\leq 1 \cdot \epsilon/2 + 1 \cdot \epsilon/2 \leq \epsilon,
\end{aligned}$$

where the second inequality follows from triangle inequality and the third inequality holds since $\mathbf{B}$ is orthogonal matrix. This shows any task can be ϵ-approximated and the generated set is indeed an ϵ-net. As a result, with Lemma D.1, we conclude that

$$\log N(\mathcal{H}, \epsilon, \|\cdot\|_{2,\infty}) \leq \log |\mathcal{B}_{\epsilon/2}| + \log |\mathcal{W}_{\epsilon/2}| \leq \mathcal{O}(dk \log \epsilon^{-1} + kM \log \epsilon^{-1}).$$

□

E LOWER BOUNDS IN SECTION 5.1

In this section, we give regret lower bounds for some specific linear bandit settings. We first present some existing result in the literature. It is easy to check the corresponding hard instance satisfies our formulation.

Lemma E.1. *(Theorem 24.1 in Lattimore & Szepesvári (2020), Restated) Under Assumptions 3.1, 3.2 and 3.3, for any algorithm, there exists a bandit instance for which*

$$\mathbb{E}[\text{Regret}(T)] \geq \Omega\Big(d\sqrt{T}\Big).$$

Lemma E.2. *(Theorem 1 in Cheung et al. (2019), Restated) Under Assumptions 3.1, 3.2 and 3.5, for any algorithm, there exists a bandit instance for which*

$$\mathbb{E}[\text{Regret}(T)] \geq \Omega\Big(d^{\frac{2}{3}}T^{\frac{2}{3}}P^{\frac{1}{3}}\Big).$$

Lemma E.3. *(Theorem 4 in Yang et al. (2020), Restated) Under Assumptions 3.1, 3.2 and 3.6, for any algorithm, there exists a bandit instance for which*

$$\mathbb{E}[\text{Regret}(T)] \geq \Omega\Big(d\sqrt{kT} + k\sqrt{MT}\Big).$$

Lemma E.4. *(Theorem F.1 in Lattimore & Szepesvari (2020), Restated) Under Assumptions 3.1 and 3.2, for any algorithm, there exists a bandit instance for which*

$$\mathbb{E}[\text{Regret}(T)] \geq \Omega\Big(T\sqrt{d/\log T}\zeta\Big).$$

Remark E.1. *In Lattimore & Szepesvari (2020), they proved a lower bound in the order of $\Omega(\min(T, K)\sqrt{d/\log(K)}\zeta)$, where K is the number of actions. Here we choose $K = T$ for simplicity.*

We further prove a lower bound for non-stationary linear bandits with bounded switches.

Lemma E.5. *Under Assumptions 3.1, 3.2 and 3.4, for any algorithm, there exists a bandit instance for which*

$$\mathbb{E}[\text{Regret}(T)] \geq \Omega\Big(d\sqrt{ST}\Big).$$

Proof. Without loss of generality, assume T is divided by S. Consider a bandit instance where the switch occurs at round $T/S, 2T/S, \cdots$. Note the reward functions after different switches are independent. Thus, according to Lemma E.1, any algorithm suffers $\Omega(d\sqrt{T/S})$ regret in any interval between two switches. Since there are S switches, for any algorithm, there exists a bandit instance for which $\mathbb{E}[\text{Regret}(T)] \geq \Omega(d\sqrt{ST})$. □

The lower bounds result in Section 5.1 can be obtained by combining above lemmas:

- Combining Lemma E.1 and Lemma E.4 gives Proposition 5.1.
- Combining Lemma E.5 and Lemma E.4 gives Proposition 5.2.
- Combining Lemma E.2, Lemma E.1, and Lemma E.4 gives Proposition 5.3.
- Combining Lemma E.3 and Lemma E.4 gives Proposition 5.4.