Best-of-Both-Worlds Linear Contextual Bandits

Masahiro Kato *mkato-csecon@g.ecc.u-tokyo.ac.jp*
The University of Tokyo

Shinji Ito *shinji@mist.i.u-tokyo.ac.jp*
The University of Tokyo
RIKEN AIP
NEC Corporation (affiliation upon submission)

Reviewed on OpenReview: *https://openreview.net/forum?id=aIG2RAtNuX&referrer*

Abstract

This study investigates the problem of K-armed linear contextual bandits, an instance of the multi-armed bandit problem, under an adversarial corruption. At each round, a decision-maker observes an independent and identically distributed context and then selects an arm based on the context and past observations. After selecting an arm, the decision-maker incurs a loss corresponding to the selected arm. The decision-maker aims to minimize the cumulative loss over the trial. The goal of this study is to develop a strategy that is effective in both stochastic and adversarial environments, with theoretical guarantees. We first formulate the problem by introducing a novel setting of bandits with adversarial corruption, referred to as the contextual adversarial regime with a self-bounding constraint. We assume linear models for the relationship between the loss and the context. Then, we propose a strategy that extends the `RealLinExp3` by Neu & Olkhovskaya (2020) and the Follow-The-Regularized-Leader (FTRL). The regret of our proposed algorithm is shown to be upper-bounded by $O\left(\min\left\{\frac{\log^3(T)}{\Delta_*}+\sqrt{\frac{C\log^3(T)}{\Delta_*}},\ \ \sqrt{T}\log^2(T)\right\}\right)$, where $T\in\mathbb{N}$ is the number of rounds, $\Delta_*>0$ is the constant minimum gap between the best and suboptimal arms for any context, and $C\in[0,T]$ is an adversarial corruption parameter. This regret upper bound implies $O\left(\frac{\log^3(T)}{\Delta_*}\right)$ in a stochastic environment and by $O\left(\sqrt{T}\log^2(T)\right)$ in an adversarial environment. We refer to our strategy as the `Best-of-Both-Worlds (BoBW) RealFTRL`, due to its theoretical guarantees in both stochastic and adversarial regimes.

1 Introduction

This study considers minimizing the cumulative regret in the multi-armed bandit (MAB) problem with contextual information. The MAB problem is a formulation of sequential decision-making. In this study, we develop an algorithm that utilizes side information called contextual information. We focus on linear contextual bandits and aim to design an algorithm that performs well in both stochastic and adversarial environments.

In our problem setting of contextual bandits, a decision-maker observes an independent and identically distributed (i.i.d.) context each round, draws an arm accordingly, and incurs a loss associated with the chosen arm. Additionally, we assume linear models between the loss and contexts, which is known as the linear contextual bandit problem. The contextual bandit problem is widely studied in fields such as sequential treatment allocation (Tewari & Murphy, 2017), personalized recommendations (Beygelzimer et al., 2011), and online advertising (Li et al., 2010). Based on these demands, existing studies explore the methods. For example, Abe & Long (1999) studies linear contextual bandits. Li et al. (2021) provides lower bounds. There are numerous other studies in this field (Chen et al., 2020; Ding et al., 2022).

The settings of linear contextual bandits are divided into stochastic, with fixed contextual and loss distributions, and adversarial environments, with fixed contexts but adversarially chosen losses[1]. Most existing studies focus on algorithms for either stochastic (Abe & Long, 1999; Rusmevichientong & Tsitsiklis, 2010; Chu et al., 2011; Abbasi-yadkori et al., 2011; Lattimore & Szepesvari, 2017) or adversarial linear contextual bandits (Neu & Olkhovskaya, 2020).

Thus, optimal algorithms typically differ between the stochastic and adversarial environments. However, a best-of-both-worlds framework exists, aiming for algorithms that are competent in both stochastic and adversarial environments (Bubeck & Slivkins, 2012; Seldin & Slivkins, 2014; Auer & Chiang, 2016; Seldin & Lugosi, 2017; Zimmert & Seldin, 2021; Lee et al., 2021). Building on existing work, we propose a best-of-both-worlds algorithm for stochastic and adversarial linear contextual bandits.

1.1 Main Contribution

In Section 2, we first introduce the setting of linear contextual bandits with adversarial corruption by defining the linear contextual adversarial regime with a self-bounding constraint. This setting is a generalization of the *adversarial regime with a self-bounding constraint* proposed by Zimmert & Seldin (2021). Under this regime, we bridge the stochastic and adversarial environments by an adversarial corruption parameter $C \geq 0$, where $C = 0$ corresponds to a stochastic environment and $C = T$ corresponds to an adversarial environment.

Then, in Section 3 inspired by the `RealLinEXP3` proposed by Neu & Olkhovskaya (2020) for adversarial contexts, our algorithm uses the Follow-the-Regularized-Leader (FTRL) approach to adapt well to the stochastic environment. Our algorithm design also follows existing studies in best-of-both-worlds (BoBW) studies, such as Ito et al. (2022). We refer to our algorithm as the `BoBW-RealFTRL`.

In Section 4, we show the following upper bound of the `BoBW-RealFTRL`:

$$O\Bigg(\min\Bigg\{\frac{K\log(T)\left(\log(T)+d\log(K)\right)\log(KT)}{\Delta_*}+\sqrt{\frac{CK\log(T)\left(\log(T)+d\log(K)\right)\log(KT)}{\Delta_*}},\\ \sqrt{\log(KT)TK\log(T)\left(\log(T)+d\log(K)\right)\log(KT)}\Bigg\}\Bigg)$$

where T is the number of rounds, d is the dimension of a context, and K is the number of arms, when there exists a constant minimum gap Δ_* between the conditional expected rewards of the best and suboptimal arms for any context when we consider a stochastic environment. Note that this regret upper bound holds both for stochastic and adversarial environments. As a specific case, under a stochastic environment, the regret upper bound is given as

$$O\left(\frac{K\log(T)\left(\log(T)+d\log(K)\right)\log(KT)}{\Delta_*}\right).$$

When there does not exist such a gap Δ_*, we show that the regret upper bound is given as

$$O\left(\sqrt{\log(KT)TK\log(T)\left(\log(T)+d\log(K)\right)\log(KT)}\right).$$

Note that our regret upper bound is $O\left(\min\left\{\frac{\log^3(T)}{\Delta_*}+\sqrt{\frac{C\log^3(T)}{\Delta_*}},\ \ \sqrt{T}\log^2(T)\right\}\right)$ when focusing on the order with respect to T. Furthermore, in a stochastic environment, the regret is upper bounded by $O\left(\frac{\log^3(T)}{\Delta_*}\right)$, and in an adversarial environment, the regret is upper bound by $O\left(\sqrt{T}\log^2(T)\right)$.

[1] We can define adversarial linear contextual bandits in different ways. For example, there are studies that consider contextual bandits with adversarial contexts and fixed losses (Chu et al., 2011; Abbasi-yadkori et al., 2011). On the other hand, several studies address contextual bandits with adversarial contexts and adversarial losses (Kanade & Steinke, 2014; Hazan et al., 2016). This study only focuses on contextual bandits with i.i.d. contexts and adversarial losses, which have been studied by Rakhlin & Sridharan (2016) and Syrgkanis et al. (2016).

In summary, we contribute to the field of linear contextual bandits by providing a BoBW algorithm. Our proposed algorithm and theoretical analysis draw partial inspiration from Neu & Olkhovskaya (2020) and Ito et al. (2022). However, we introduce several novel techniques to adapt these methodologies to our setting, such as specifying adaptation weights based on an estimation algorithm for the covariance matrix of contexts. It is important to note that our primary contribution lies not in theoretical advancements but in the proposition of a novel algorithm tailored to a unique setting. Our study significantly enriches the domains of linear contextual bandits and BoBW algorithms.

1.2 Related Work

In adversarial bandits, the `RealLinExp3`, the algorithm proposed by Neu & Olkhovskaya (2020), yields $O\left(\log(T)\sqrt{KdT}\right)$. In Table 1, we compare our regret upper bounds with the upper bounds of Neu & Olkhovskaya (2020).

Regret upper bounds in a stochastic setting are categorized into problem-dependent and problem-independent upper bounds, where the former utilizes some distributional information, such as the gap parameter Δ_*, to bound the regret, while the latter does not. Additionally, problem-dependent regret upper bounds in the stochastic bandits depend on the margin condition characterized by a parameter $\alpha \in [0, +\infty]$ (for the detailed definition, see Remark 1). Our case with Δ_* corresponds to a case with $\alpha = +\infty$. Note that in the adversarial bandits, the margin condition usually does not affect the upper bounds. Dani et al. (2008) proposes the `ConfidenceBAll`, and Abbasi-yadkori et al. (2011) proposes `OFUL`. They both present upper bound with and without the assumption of the existence of Δ_*[2]. As mentioned above, the regret upper bound under the assumption of the existence of Δ_* corresponds to a case with $\alpha = +\infty$ in the margin condition. In contrast, Goldenshluger & Zeevi (2013), Wang et al. (2018), and Bastani & Bayati (2020) propose algorithm in a case with $\alpha = 1$. Furthermore, Li et al. (2021) propose the ℓ_1-`ConfidenceBall` based algorithm whose upper bound tightly depends on unknown α. Note that their setting of linear contextual bandits slightly differs from the one presented in Neu & Olkhovskaya (2020). While Neu & Olkhovskaya (2020) considers a scenario where contexts remain constant across arms but the coefficients of linear models vary, other studies typically adopt a setting where each arm has its specific context, but the coefficients are the same. In this latter context, Abbasi-yadkori et al. (2011) demonstrates that their `OFUL` algorithm achieves an $O\left(d\log(T)/\Delta_*\right)$ regret upper bound. It is important to note that due to these differing settings, we do not compare compare their regret bounds to ours directly and omit it from Table 1.

There are several related studies for linear contextual bandits with adversarial corruption, including Lykouris & Vassilvtiskii (2018), Gupta et al. (2019), Zhao et al. (2021) and He et al. (2022). Lykouris & Vassilvtiskii (2018), Gupta et al. (2019), and Zhao et al. (2021) consider other corruption frameworks characterized by a constant $\widetilde{C} \in [0, T]$, which is different but related to our linear contextual adversarial regime with a self-bounding constraint. He et al. (2022) uses another constant $\widetilde{C}^\dagger \in [0, T]$ different but closely related to $\widetilde{C}$. For the detailed definitions, see Remark 2. The essential difference between our and their settings is the existence of the gap Δ_*. Furthermore, while our regret upper bound achieves the polylogarithmic order, those studies show roughly $\sqrt{T}$-order regret upper bounds. He et al. (2022) presents $\widetilde{O}\left(d\sqrt{K} + d\widetilde{C}^\dagger\right)$ regret under an adversarial corruption characterized by a constant $\widetilde{C}^\dagger > 0$.

The use of the FTRL approach for adversarial linear bandits is also independently explored by Liu et al. (2023) to relax the assumption used in Neu & Olkhovskaya (2020). In addition to the difference in contributions, while our algorithm utilizes the Shannon entropy in the regularization of the FTRL, Liu et al. (2023) employs the log-determinant barrier. We expect that combining these two methods will yield a BoBW algorithm with relaxed assumptions, and it is future work.

To establish our BoBW regret bounds, we utilize the self-bounding technique (Zimmert & Seldin, 2021; Wei & Luo, 2018), which yields poly-logarithmic regret bounds in stochastic environments. This is achieved by integrating regret upper bounds that are contingent on the arm-selection distributions q_t, and a lower

[2] Regret upper bounds with the assumption of the existence of Δ_* are called problem-dependent.

Table 1: Comparison of the regret. We compared the regret upper bound of our proposed `BoBW-RealFTRL` with the `RealLinExp3` (Neu & Olkhovskaya, 2020).

	Regret	Adversarial/Stochastic
Ours (`BoBW-RealFTRL`)	$O\left(\min\left\{\frac{D}{\Delta_*}+\sqrt{\frac{CD}{\Delta_*}}, \sqrt{\log(KT)TD}\right\}\right)$ where $D = K\log(T)\left(\log(T)+d\log(K)\right)\log(KT)$	Both
	$O\left(\sqrt{\log(KT)TD}\right)$	Adversarial
	$O\left(\frac{D}{\Delta_*}\right)$ $(C=0)$	Stochastic
`RealLinExp3`	$O\left(\log(T)\sqrt{KdT}\right)$	Adversarial

bound known as self-bound constraints. The q_t-dependent regret bounds are obtained using FTRL with a negative-entropy regularizer, which is also referred to as the *exponential weight* method.

Our approach includes an entropy-adaptive update rule for learning rates, originally developed for online learning in feedback graph contexts (Ito et al., 2022). This strategy has been proven effective in providing BoBW guarantees for exponential-weight-based algorithms across various sequential decision-making problems, such as multi-armed bandits (Jin et al., 2023), partial monitoring (Tsuchiya et al., 2023a), linear bandits (Kong et al., 2023), episodic Markov Decision Processes (MDPs) (Dann et al., 2023b), and sparse multi-armed bandits (Tsuchiya et al., 2023b). However, a common limitation of these results, stemming from the negative-entropy regularization, is the additional $\log T$ factors in the regret bounds.

Our obtained regret upper bounds are not $O(\log(T))$, and we assume the uniqueness of the optimal arm. A promising future direction to address these issues could be the exploration of alternative regularizers like Tsallis entropy. Utilizing such regularizers might allow us to tighten the regret upper bound and relax the assumption that the optimal arm is unique. Zimmert & Seldin (2021) proposes the use of Tsallis entropy, and Masoudian & Seldin (2021) enhances this approach, yielding tighter upper bounds with respect to T. We anticipate that by employing such regularizers, we could achieve $\log(T)$ or $\log^2(T)$ regret upper bounds, similar to the `OFUL` algorithm in Abbasi-yadkori et al. (2011). We expect that the regret upper bounds will be worse than the one in Abbasi-yadkori et al. (2011) due to the estimation of the inverse covariance matrix. While alternative regularization could be applicable in our context, it may require additional assumptions to derive desired regret upper bounds. Concerning the number of optimal arms, Ito (2021) and Jin et al. (2023) introduce BoBW algorithms for scenarios where the optimal arm is not unique in standard multi-armed bandits. However, the analysis is complex, and extensions to combinatorial semi-bandits or linear bandits, which are arguably simpler than contextual linear bandits, are yet to be seen. Thus, although their methods hold promise for linear contextual bandits, their straightforward application in our setting presents significant challenges.

Independently of us, Kuroki et al. (2024) proposes several FTRL-based BoBW algorithms for linear contextual bandits with adversarial corruptions. One of their proposed algorithms is also a BoBW algorithm based on FTRL with Shannon entropy, which coincides with our proposed algorithm. As another approach, they employ the black-box framework proposed by Dann et al. (2023a). They show that under the black-box framework, we can obtain a BoBW reduction of RealEXP3 with a regret of order $O(\log(T))$ in the stochastic regime. They also present another algorithm by combining the black-box framework with the continuous exponential weights algorithm in Olkhovskaya et al. (2023). While this algorithm incurs a regret of $O(\log^5(T))$ in the stochastic regime, it incurs a regret of $O(\log(T)dK\sqrt{\Lambda^*})$ in the adversarial regime, where Λ^* denotes the cumulative second moment of the losses incurred by the algorithm. Although the black-box framework provides a tight $O(\log(T))$ regret regarding T, limitations have been reported in practical implementation. As well as Kuroki et al. (2024) mentions, "FTRL with Shannon entropy regularization is a much more practical algorithm."

2 Problem Setting

Suppose that there are T rounds and K arms. In each round $t \in [T] \coloneqq \{1, 2, \dots, T\}$, a decision-maker observes a context $X_t \in \mathcal{X} \subset \mathbb{R}^d$, where $\mathcal{X}$ is a context space. Then, the decision-maker chooses an arm

$A_t \in [K] \coloneqq \{1, 2, \dots, K\}$ based on the context X_t and past observations. Each arm $a \in [K]$ is linked to a loss $\ell_t(a, X_t)$, which depends on X_t and round t. After choosing arm A_t in round t, the decision-maker incurs the loss $\ell_t(A_t, X_t)$. Our goal is to minimize the cumulative loss $\sum_{t=1}^{T} \ell_t(A_t, X_t)$. We introduce the setting in more detail in the following part.

Contextual distribution. Let a distribution of X_t be $\mathcal{D}$, which is invariant across $t \in [T]$. We also assume that $\mathcal{D}$ is known to the decision-maker.

Assumption 2.1 (Contextual distribution)**.** *The context X_t is an i.i.d. random variable, whose distribution $\mathcal{D}$ is* ***known*** *to the decision-maker, and the covariance matrix $\Sigma = \mathbb{E}\left[X_t X_t^\top\right]$ is positive definite with its smallest eigenvalue $\lambda_{\min} > 0$. We also assume that $\lambda_{\min}$ is* known*.*

2.1 Linear Contextual Bandits

This study assumes linear models between $\ell_t(a, X_t)$ and X_t as follows.

Assumption 2.2 (Linear models)**.** *For each $\ell_t(a, X_t)$, the following holds:*

$$\ell_t(a, X_t) = X_t^\top \theta_t(a) + \varepsilon_t(a),$$

where $\theta_t(a) \in \Theta$ is a d-dimensional parameter with a parameter space $\Theta \subset \mathbb{R}^d$, and $\varepsilon_t(a)$ is the error term independent of the sequence $\{X_t\}_{t\in[T]}$.

In each round $t \in [T]$, a context is sampled from $\mathcal{D}$ and the environment chooses $\{\theta_t(a)\}_{a\in[K]}$ based on the past observations $\mathcal{F}_{t-1} = (X_1, A_1, \ell_1(A_1, X_1), X_2, \dots, X_{t-1}, A_{t-1}, \ell_{t-1}(A_{t-1}, X_{t-1}))$.

For linear models and variables, we make the following assumptions.

Assumption 2.3 (Bounded variables)**.** *We assume the following:*

1. *There exists an universal constant $C_{\mathcal{X}} > 0$ such that for each $x \in \mathcal{X}$, $\|x\|_2 \le C_{\mathcal{X}}$ holds.*
2. *There exists an universal constant $C_\Theta > 0$ such that for each $\theta \in \Theta$, $\|\theta\|_2 \le C_\Theta$ holds.*
3. *There exists an universal constant $C_{\mathcal{E}} > 0$ such that $|\varepsilon_t(a)| \le C_{\mathcal{E}}$ holds.*

Under this assumption, there exists $C_\ell \coloneqq C(C_{\mathcal{X}}, C_\Theta, C_{\mathcal{E}})$ such that for all $\ell_t(a, x)$, the following holds for each $a \in [K]$ and $x \in \mathcal{X}$:

$$|\ell_t(a, x)| \le C_\ell.$$

For simplicity, we assume $C_\ell = 1$.

We consider a general framework where $\{\theta_t(a)\}_{a\in[K]}$ is generated in both stochastic and adversarial ways. See Section 2.5 for details.

2.2 Policy

We refer to a function that determines the arm draw as a policy. Let Π be a set of all possible policies $\pi : \mathcal{X} \to \mathcal{P} \coloneqq \left\{u = (u_1\ u_2\ \dots\ u_K)^\top \in [0,1]^K \mid \sum_{k=1}^{K} u_k = 1\right\}$. Let $\pi(a \mid x)$ be the a-th element of $\pi(x)$. The goal of the decision-maker is to minimize the cumulative loss $\sum_{t=1}^{T} \ell_t(A_t, X_t)$ incurred through T rounds by learning a policy $\pi \in \Pi$.

2.3 Data-Generating Process

In each round of a trial, the decision-maker first observes a context and then chooses an action based on the context and past observations obtained until the round. Specifically, we consider sequential decision-making with the following steps in each round $t \in [T]$:

1. The environment decides $\{\theta_t(a)\}_{a\in[K]}$ based on $\mathcal{F}_{t-1}$.
2. The decision-maker observes the context X_t, which is generated from a known distribution $\mathcal{D}$.
3. Based on the context X_t, the decision-maker chooses a policy $\pi_t(X_t) \in \mathcal{P}$.
4. The decision-maker chooses action $A_t \in [K]$ with probability $\pi_t(a \mid X_t)$.
5. The decision-maker incurs the loss $\ell_t(A_t, X_t)$.

The goal of the decision-maker is to choose actions in a way that the total loss is as small as possible.

2.4 Regret

This section provides the definition of the regret, a relative measure of the cumulative loss. We evaluate the performance of the decision or policy of the decision-maker by using regret. Let $\mathcal{R}$ be a set of all possible $\rho : \mathcal{X} \to [K]$. The quality of a decision by the decision-maker is measured by its total expected regret, defined as

$$R_T = \max_{\rho\in\mathcal{R}} \mathbb{E}\left[\sum_{t=1}^{T} \Big\{\ell_t(A_t, X_t) - \ell_t(\rho(X_t), X_t)\Big\}\right] = \max_{\rho\in\mathcal{R}} \mathbb{E}\left[\sum_{t=1}^{T} \Big\langle X_t, \theta_t(A_t) - \theta_t(\rho(X_t))\Big\rangle\right],$$

where the expectation is taken over the randomness of policies of the decision-maker, as well as the sequence of random contexts, $\{X_t\}_{t\in[T]}$, and losses, $\{\ell_t(\cdot, X_t)\}_{t\in[T]}$.

Let X_0 be an i.i.d. random variable from the distribution of X_t. Then, because X_t is an i.i.d. random variable from $\mathcal{D}$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \Big\langle X_t, \theta_t(\rho(X_t))\Big\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T} \Big\langle X_0, \theta_t(\rho(X_0))\Big\rangle\right] \geq \mathbb{E}\left[\min_{a\in[K]} \sum_{t=1}^{T} \Big\langle X_0, \mathbb{E}\left[\theta_t(a)\right]\Big\rangle\right].$$

Based on this inequality, we define an optimal policy a^*_T as

$$a^*_T(x) = \operatorname*{arg\,min}_{a\in[K]} \sum_{t=1}^{T} \Big\langle x, \mathbb{E}\left[\theta_t(a)\right]\Big\rangle.$$

Then, we have

$$R_T \leq \mathbb{E}\left[\sum_{t=1}^{T} \Big\langle X_t, \theta_t(A_t) - \theta_t(a^*_T(X_t))\Big\rangle\right].$$

Neu & Olkhovskaya (2020) refers to ρ as linear-classifier policies, while π_t is called stochastic policies. In our study, decision-makers compare their stochastic policies π_t to the optimal linear-classifier policy a^* using the regret.

2.5 Linear Contextual Adversarial Regime with a Self-Bounding Constraint

Then, we define the framework of a *linear contextual adversarial regime with a self-bounding constraint*, which is a generalization of adversarial and stochastic bandits.

Definition 2.4 (Linear contextual adversarial regime with a self-bounding constraint)**.** We say that an environment is in a linear contextual adversarial regime with a (Δ_*, C, T) self-bounding constraint for some $\Delta_*, C > 0$ if R_T is lower bounded as

$$R_T \geq \Delta_* \cdot \mathbb{E}\left[\sum_{t=1}^{T} \Big(1 - \pi_t(a^*_T(X_0) \mid X_0)\Big)\right] - C.$$

The contextual adversarial regime with a self-bounding constraint includes several important settings. Among them, we raise linear contextual bandits in stochastic bandits and adversarial bandits below. Both correspond to cases with $C = 0$ (linear contextual adversarial regime with a $(\Delta_*, 0, T)$ self-bounding constraint) and $C = 2T$ (linear contextual adversarial regime with a $(\Delta_*, 2T, T)$ self-bounding constraint), respectively. The parameter C determines the degree of an adversarial corruption: as C is larger, the adversarial corruption is also larger.

Example 1 (Linear contextual bandits in stochastic bandits.)**.** We consider a linear contextual adversarial regime with a $(\Delta_*, 0, T)$ self-bounding constraint, which corresponds to a stochastic environment. In stochastic bandits, the bandit models are fixed; that is, $\big(X_t, \ell_t(1, X_t), \dots, \ell_t(K, X_t)\big)$ are generated from a fixed distribution P_0. Let $\theta_1(a) = \cdots \theta_T(a) = \theta_0(a)$. Note that when considering stochastic bandits, we have $\mathbb{E}\left[\theta_t(a)\right] = \theta_0(a)$ and

$$a^*_T(x) = \operatorname*{arg\,min}_{a\in[K]} \sum_{t=1}^{T} \Big\langle x, \mathbb{E}\left[\theta_t(a)\right] \Big\rangle = \operatorname*{arg\,min}_{a\in[K]} \Big\langle x, \theta_0(a) \Big\rangle \qquad \forall x \in \mathcal{X}.$$

Let $a^*_T(x)$ be $a^*_0(x)$.

In this setting, we assume that for each P_0, there exist positive constraints Δ_* such that for all $x \in \mathcal{X}$,

$$\min_{b \neq a^*_0(x)} \Big\langle x, \theta_0(b) \Big\rangle - \Big\langle x, \theta_0(a^*_0(x)) \Big\rangle \geq \Delta_*. \tag{1}$$

Then, the regret can be lower bounded as $R_T \geq \Delta_* \cdot \mathbb{E}\left[\sum_{t=1}^{T} \Big(1 - \pi_t(a^*_0(X_0) \mid X_0)\Big)\right]$ (See Appendix A).

Example 2 (Linear contextual bandits in adversarial bandits)**.** We consider a linear contextual adversarial regime with a $(\Delta_*, 2T, T)$ self-bounding constraint, which corresponds to a stochastic environment In adversarial bandits, we do not assume any data-generating process for the $\ell_t(a, X_t)$, and the loss is decided to increase the regret based on the past observations $\mathcal{F}_{t-1}$.

Remark 1 (Margin conditions)**.** In linear contextual bandits, we often employ the *margin condition* to characterize the difficulty of the problem instance. The margin condition is defined as follows (Li et al., 2021): there exist Δ_*, C_1, a^*, and $\alpha \in [0, +\infty]$, such that for $h \in \left[C_1\sqrt{\frac{\log(d)}{T}}, \Delta_*\right]$,

$$\mathbb{P}\left(\Big\langle X_t, \theta_t(a^*)\Big\rangle \geq \min_{b\neq a^*} \Big\langle X_t, \theta_t(b)\Big\rangle - h\right) \leq \frac{1}{2}\left(\frac{h}{\Delta_*}\right)^{\alpha}.$$

Our definition of the linear contextual adversarial regime with a self-bounding constraint corresponds to a case with $\alpha = \infty$. In existing studies of linear contextual bandits under an adversarial environment, the margin condition with $\alpha = 0$ is assumed in Abbasi-yadkori et al. (2011), which demonstrates $O(\sqrt{T})$-regret. On the other hand, the margin condition with $\alpha \in (0, 1]$ is considered in works by Goldenshluger & Zeevi (2009; 2013), Wang et al. (2018), and Bastani & Bayati (2020). Goldenshluger & Zeevi (2009) establishes $O\left(T^{(1-\alpha)/2}\right)$-regret for $\alpha \in (0, 1)$, while Goldenshluger & Zeevi (2013) proves $O\left(d^3 \log(T)\right)$-regret for $\alpha = 1$. Furthermore, Abbasi-yadkori et al. (2011) explores a scenario with $\alpha = \infty$ and demonstrates $O\left(\log(T)/\Delta_*\right)$-regret. Li et al. (2021) presents a general lower bound that is applicable for various α. Extending our results to encompass more general values of α constitutes future work. We anticipate that an ideal BoBW algorithm would perform comparably to the existing studies, such as those by Abbasi-yadkori et al. (2011) and Goldenshluger & Zeevi (2009; 2013), in a stochastic environment.

Remark 2 (Linear contextual bandits with corruption in existing studies)**.** Lykouris & Vassilvtiskii (2018), Gupta et al. (2019), Zhao et al. (2021), and He et al. (2022) propose another definition of linear contextual contextual bandits with corruption. In their work, instead of our defined $\ell_t(a, X_t)$, they define a loss as

$$\widetilde{\ell}_t(a, X_t) = \ell_t(a, X_t) + \widetilde{c}_t(a),$$

where $\widetilde{c}_t(a)$ is an adversarial corruption term. For simplicity, let $\widetilde{c}_t(a) \in [-1, 1]$. In Zhao et al. (2021), the degree of the corruption is determined by $\widetilde{C} \in [0, T]$ defined as $\widetilde{C} = \sum_{t=1}^{T} \max_{a\in[K]} |c_t(a)|$. In Lykouris &

Vassilvtiskii (2018), Gupta et al. (2019), and He et al. (2022), the corruption level is determined by another parameter $\widetilde{C}^\dagger \in [0, T]$ defined as $\sum_{t=1}^T |c_t(A_t)|$. Here, $\widetilde{C} \geq \widetilde{C}^\dagger$ holds. Note that the adversarial corruption depends on A_t in Zhao et al. (2021), while the adversarial corruption is determined irrelevant to A_t in Lykouris & Vassilvtiskii (2018), Gupta et al. (2019), and He et al. (2022). Unlike ours, they do not assume the existence of Δ_* defined in equation 1. In this sense, our results and their results are complementary.

3 Algorithm

This section provides an algorithm for our defined problem. Our proposed algorithm is a generalization of the `RealLinEXP3` algorithm proposed by Neu & Olkhovskaya (2020). We extend the method by employing the Follow-The-Regularized-Leader (FTRL) approach with round-varying arm-drawing probabilities. Our design of the algorithm is also motivated by existing studies about Best-of-Both-Worlds (BoBW) algorithms in different Multi-Armed Bandit (MAB) problems, such as Ito et al. (2022).

In our setting, we first observe a context and then draw an arm based on that context. We consider stochastically drawing an arm. Therefore, in designing the algorithm, our interest lies in appropriately defining the arm-drawing probability. In the FTRL approach, we define this probability by utilizing an unbiased estimator of the loss function.

We refer to our proposed algorithm as `BoBW-RealFTRL` because it modifies the `RealLinEXP3` for a best-of-both-worlds algorithm using the FTRL framework. The pseudo-code is shown in Algorithm 1. In the following part, we explain the algorithm.

Unbiased loss estimator. For each $a \in [K]$, let us define an estimator of regression parameters as

$$\widehat{\theta}_t(a) \coloneqq \Sigma^\dagger_{t,a} \mathbb{1}[A_t = a] X_t \ell_t(A_t, X_t),$$

where $\Sigma^\dagger_{t,a}$ is an estimator of $\Sigma^{-1}_t \coloneqq \mathbb{E}\left[\mathbb{1}[A_t = a] X_t^\top X_t \mid \mathcal{F}_{t-1}\right]^{-1}$. Then, the loss can be estimated as

$$\widehat{\ell}_t(a, x) = \left\langle x, \widehat{\theta}_t(a) \right\rangle.$$

In analysis of adversarial bandits, the bias of $\widehat{\ell}_t(a, x)$ plays an important role. If $\Sigma^\dagger_{t,a} = \mathbb{E}\left[\mathbb{1}[A_t = a] X_t^\top X_t \mid \mathcal{F}_{t-1}\right]^{-1}$, then this loss estimator is unbiased for $x^\top \theta_0(a)$ because

$$\begin{aligned}\mathbb{E}\left[\widehat{\ell}_t(a, x) \mid \mathcal{F}_{t-1}\right] &= x^\top \mathbb{E}\left[\widehat{\theta}_t(a) \mid \mathcal{F}_{t-1}\right] = x^\top \Sigma^\dagger_{t,a} \mathbb{E}\left[\mathbb{1}[A_t = a] X_t \ell_t(A_t, X_t) \mid \mathcal{F}_{t-1}\right] \\ &= x^\top \Sigma^\dagger_{t,a} \mathbb{E}\left[\mathbb{1}[A_t = a] X_t \left\{X_t^\top \theta_0(a) + \varepsilon_t(a)\right\} \mid \mathcal{F}_{t-1}\right] = x^\top \theta_0(a).\end{aligned}$$

Note that in our algorithm, $\Sigma^\dagger_{t,a}$ is just an estimator of Σ^{-1}_t, and $\Sigma^\dagger_{t,a} = \Sigma^{-1}_t$ does not hold in general. Therefore, $\widehat{\ell}_t(a, x)$ is not unbiased. However, we can show that the bias can be ignored because it is sufficiently small to evaluate the regret in depth. We also define a vector of loss estimators as $\widehat{\ell}_t(x) = \left(\widehat{\ell}_t(1, x) \;\; \widehat{\ell}_t(2, x) \;\; \cdots \;\; \widehat{\ell}_t(K, x)\right)^\top$.

Estimation of Σ^{-1}_t. Our remaining task is to estimate Σ^{-1}_t. The difficulty of this task stems from the dependency on A_t, which varies across rounds. To address this issue, we employ the `Matrix Geometric Resampling` (MGR) proposed by Neu & Olkhovskaya (2020). The MGR assumes that we have access to the distribution $\mathcal{D}$ of X_t and estimates Σ^{-1}_t by using simulations. We introduce the algorithm in Algorithm 2.

In Algorithm 2, we define $W_{k,j,a}$ so that $\mathbb{E}[W_{k,j,a} \mid \mathcal{F}_{t-1}] = \Sigma_{t,a}$ holds for each $k \in \{1, 2, \dots, M_t\}$ and $j \in \{1, 2, \dots, k\}$. Let us define $V_{k,a} = \prod_{j=1}^k (I - \delta W_{k,j,a})$. Here, from the independence of the context $W_{k,j,a}$ from each other, we also have

$$\mathbb{E}[V_{k,a} \mid \mathcal{F}_{t-1}] = \mathbb{E}\left[\prod_{j=1}^k \left(I - \delta W_{k,j,a}\right) \mid \mathcal{F}_{t-1}\right] = \prod_{j=1}^k \left(I - \delta \mathbb{E}\left[W_{k,j,a} \mid \mathcal{F}_{t-1}\right]\right) = (I - \delta \Sigma_{t,a})^k.$$

Therefore, $\widehat{\Sigma}^{\dagger}_{t,a} = \delta I + \delta \sum_{k=1}^{M_t} V_{k,a}$ works as a good estimator of $\Sigma^{-1}_{t,a}$ on expectations when $M_t = \infty$ because

$$\mathbb{E}\left[\widehat{\Sigma}^{\dagger}_{t,a} \mid \mathcal{F}_{t-1}\right] = \delta I + \delta \sum_{k=1}^{\infty} (I - \delta\Sigma_{t,a})^k = \delta \sum_{k=0}^{\infty} (I - \delta\Sigma_{t,a})^k = \delta(\delta\Sigma_{t,a})^{-1} = \Sigma^{-1}_{t,a}.$$

holds, where we used $\sum_{k=0}^{\infty} A^k = (I - A)^{-1}$ for a diagonalizable square matrix A whose all eigenvalues are greater than -1 and less than 1. In implementation, M_t is usually finite, which causes an estimation error. Therefore, Neu & Olkhovskaya (2020) performs estimation error analysis of $\Sigma^{-1}_{t,a}$ with finite M_t in Lemma B.5.

Our proposed algorithm: `BoBW-RealFTRL`. Then, we define our policy, called the `BoBW-RealFTRL`, as

$$\pi_t(X_t) \coloneqq (1 - \gamma_t) q_t(X_t) + \frac{\gamma_t}{K}\iota, \tag{2}$$

where ι is a K-dimensional vector $\iota = (1\ 1\ \cdots\ 1)^\top$,

$$q_t(x) \in \operatorname*{arg\,min}_{q \in \Pi} \left\{ \sum_{s=1}^{t-1} \left\langle \widehat{\ell}_s(x), q(x) \right\rangle + \psi_t(q(x)) \right\} \quad \text{for } t \geq 2, \quad q_1(x) \coloneqq (1/K\ 1/K\ \cdots\ 1/K)^\top,$$

$$\psi_t(q(x)) \coloneqq -\beta_t H(q(x)), \quad H(q(x)) \coloneqq \sum_{a \in [K]} q(a \mid x) \log\left(\frac{1}{q(a \mid x)}\right), \tag{3}$$

$$\beta_{t+1} \coloneqq \beta_t + \frac{\beta_1}{\sqrt{1 + \left(\log(K)\right)^{-1} \sum_{s=1}^{t} H\left(q_s(X_s)\right)}}, \quad \beta_1 \coloneqq C_\Theta C_{\mathcal{X}}^2 \sqrt{K \log(T) \left(\frac{\log(T)}{\delta \lambda_{\min} \sqrt{\log(K)}} + d \right)}, \tag{4}$$

$$\widetilde{\gamma}_t \coloneqq \frac{K}{2\delta\lambda_{\min}\beta_t} \log(T) = \frac{C_\Theta C_{\mathcal{X}}^2 K}{\lambda_{\min}\beta_t} \log(T), \quad \gamma_t = \min\{1, \widetilde{\gamma}_t\}, \tag{5}$$

$$M_t \coloneqq (2\beta_t - 1), \quad \delta \coloneqq \frac{1}{2C_\Theta C_{\mathcal{X}}^2}.$$

Note that an FTRL algorithm with the Shannon entropy is a generalization of the exponential weighting algorithms. Indeed, $q_t(x)$ has an analytical solution given as

$$q_t(x) = \frac{\exp\left(-\sum_{s=1}^{t-1} \widehat{\ell}_s(a, x)/\beta_t\right)}{\sum_{b \in [K]} \exp\left(-\sum_{s=1}^{t-1} \widehat{\ell}_s(b, x)/\beta_t\right)}.$$

The parameters β_t and γ_t adaptively adjust the regularization and exploration based on the past observations up to the t-th round. Appropriately chosen fixed parameters for these can lead to the `RealLinEXP3` algorithm proposed by (Neu & Olkhovskaya, 2020), which is tailored for adversarial environments but does not achieve logarithmic regret upper bounds. Therefore, our algorithm extends `RealLinEXP3` and FTRL. In the realm of BoBW algorithms, FTRL-based methods are commonly utilized, linking our approach to existing literature. It is crucial to note the balance among β_t, γ_t, and M_t (a parameter in the MGR algorithm) is vital for ensuring desirable regret upper bounds. For instance, the exploration rate γ_t influences Σ_t^{-1}, and M_t is consequently affected as it is used in estimating Σ_t^{-1}. Since M_t also impacts the regret upper bound, parameter selection is meticulously conducted to tighten the bound considering the interplay among β_t, γ_t, and M_t.

4 Regret Analysis

This section provides upper bounds for the regret of our proposed `BoBW-RealFTRL` algorithm.

For notational simplicity, let us denote a^*_T by a^*. To derive upper bounds, we define the following quantities:

$$Q(a^* \mid x) = \sum_{t=1}^{T} \left\{ 1 - q_t\left(a^*(x) \mid x\right) \right\}, \qquad \overline{Q}(a^*) = \mathbb{E}\left[Q(a^* \mid X_0)\right].$$

Algorithm 1 `BoBW-RealFTRL`.

```
**Parameter:** Learning rate $\beta_1, \beta_2, \dots, \beta_T > 0$ and exploration parameter $\gamma_t \in (0, 1)$, which depend on $\lambda_{\min}$, $C_\Theta$, $C_{\mathcal{X}}$, $K$, $d$, and $T$.
**Initialization:** Set $\theta_0(a) = 0$ for all $a \in [K]$.
**for** $t = 1, \dots, T$ **do**
    Observe $X_t$.
    Draw $A_t \in [K]$ following the policy $\pi_t(X_t) \coloneqq (1 - \gamma_t) q_t(X_t) + \frac{\gamma_t}{K} \iota$ defined in equation 2.
    Observe the loss $\ell_t(A_t, X_t)$.
    Compute $\widehat{\theta}_t(a)$ for all $a \in [K]$.
**end for**
```

Algorithm 2 `Matrix Geometric Resampling` (Neu & Olkhovskaya, 2020).

```
**Input:** Context distribution $\mathcal{D}$, policy $\pi_t$, action $a \in [K]$.
**for** $k = 1, \dots, M_t$ **do**
    **for** $j = 1, \dots, k$ **do**
        Draw $X_{k,j} \sim \mathcal{D}$ and $V_{k,j} \sim \pi_t(\cdot \mid X_{k,j})$.
        Compute $W_{k,j,a} = \mathbb{1}[V_{k,j} = a] X_{k,j} X_{k,j}^\top$.
    **end for**
    Compute $V_{k,a} = \prod_{j=1}^{k} (I - \delta W_{k,j,a})$.
**end for**
**Return:** $\widehat{\Sigma}^\dagger_{t,a} = \delta I + \delta \sum_{k=1}^{M_t} V_{k,a}$.
```

Then, we show the following upper bound, which holds for general cases such as adversarial and stochastic environments. We show the proof in Sections B.1 and B.2.

Theorem 4.1 (General regret bounds)**.** *If the environment generates losses under the contextual adversarial regime with a self-bounding constraint (Definition 2.4), the* `BoBW-RealFTRL` *with* $\widehat{\Sigma}^\dagger_{t,a}$ *incurs the total regret*

$$R_T = O\left(\left(\frac{K\log(T)}{\beta_1}\left(\frac{\log(T)}{\delta\lambda_{\min}\sqrt{\log(K)}} + d\right) + \beta_1\sqrt{\log(K)}\right)\sqrt{\log(KT)}\max\left\{\overline{Q}^{1/2}(a^*), 1\right\}\right).$$

For each situation, such as adversarial environments and linear contextual adversarial regimes with a self-bounding constraint, we derive a specific upper bound.

First, from $\overline{Q}(a^*) \leq T$, the following regret bound holds without any assumptions on the loss; that is, it holds for an adversarial environment.

Corollary 4.2. *Assume the same conditions in Theorem 4.1. Then, under an adversarial environment, the regret satisfies*

$$R_T = O\left(\left(\frac{K\log(T)}{\beta_1}\left(\frac{\log(T)}{\delta\lambda_{\min}\sqrt{\log(K)}} + d\right) + \beta_1\sqrt{\log(K)}\right)\sqrt{\log(KT)}\sqrt{T}\right);$$

that is, from $\beta_1 = C_\Theta C_{\mathcal{X}}^2 \sqrt{K\log(T)\left(\frac{\log(T)}{\delta\lambda_{\min}\sqrt{\log(K)}} + d\right)}$,

$$R_T = O\left(\log(KT)\sqrt{K\log(K)T\log(T)\left(\frac{\log(T)}{\delta\lambda_{\min}\sqrt{\log(K)}} + d\right)}\right)$$

holds.

Furthermore, we derive a regret bound under the linear contextual adversarial regime with a self-bounding constraint.

Corollary 4.3 (Regret bounds under the linear contextual adversarial regime with a self-bounding constraint)**.** *Suppose that the same conditions in Theorem 4.1 hold. Then, under the contextual adversarial regime with self-bounding constraints, the regret satisfies*

$$R_T = O\Bigg(\left\{ \frac{K\log(T)}{\beta_1}\left(\frac{\log(T)}{\delta\lambda_{\min}\sqrt{\log(K)}} + d\right) + \beta_1\sqrt{\log(K)} \right\}^2 \log(KT)/\Delta_* \\ + \sqrt{C\left\{ \frac{K\log(T)}{\beta_1}\left(\frac{\log(T)}{\delta\lambda_{\min}\sqrt{\log(K)}} + d\right) + \beta_1\sqrt{\log(K)} \right\}^2 \log(KT)/\Delta_*}\Bigg);$$

that is, from $\beta_1 = C_\Theta C_{\mathcal{X}}^2\sqrt{K\log(T)\left(\frac{\log(T)}{\delta\lambda_{\min}\sqrt{\log(K)}} + d\right)}$,

$$R_T = O\left(\frac{K\log(K)\log(T)\left(\frac{\log(T)}{\delta\lambda_{\min}\sqrt{\log(K)}} + d\right)\log(KT)}{\Delta_*} + \sqrt{\frac{CK\log(K)\log(T)\left(\frac{\log(T)}{\delta\lambda_{\min}\sqrt{\log(K)}} + d\right)\log(KT)}{\Delta_*}} \right)$$

holds.

The result in Corollary 4.3 implies $R_T = O\left(\frac{\log^3(T)}{\Delta_*} + \sqrt{\frac{C\log^3(T)}{\Delta_*}}\right)$ regarding T.

Proof. From the definition of the contextual adversarial regime with a self-bounding constraint, we have

$$R_T \geq \Delta_* \cdot \mathbb{E}\left[\sum_{t=1}^T \Big(1 - \pi_t(a^*(X_0) \mid X_0)\Big)\right] - C = \Delta_* \cdot \overline{Q}(a^*) - C.$$

Therefore, from Lemma B.7, for any $\lambda > 0$, we have

$$\begin{aligned} R_T &= (1+\lambda)R_T - \lambda R_T \\ &= (1+\lambda)O\left(c\sqrt{\log(KT)}\sqrt{\sum_{t=1}^T \mathbb{E}\left[H(q_t(X_0))\right]}\right) - \lambda R_T \\ &\leq (1+\lambda)O\left(c\sqrt{\log(KT)}\sqrt{\sum_{t=1}^T \mathbb{E}\left[H(q_t(X_0))\right]}\right) - \lambda\Delta_* \cdot \overline{Q}(a^*) + \lambda C, \end{aligned}$$

where

$$c = \left(\frac{K\log(T)}{\beta_1}\left(\frac{\log(T)}{\delta\lambda_{\min}\sqrt{\log(K)}} + d\right) + \beta_1\sqrt{\log(K)}\right).$$

Here, as well as the proof of Theorem 4.1, from Lemma B.8, if $Q(a^* \mid x) \leq e$, we have $\sum_{t=1}^T H(q_t(x)) \leq e\log(KT)$ and otherwise, we have $\sum_{t=1}^T H(q_t(x)) \leq Q(a^* \mid x)\log(KT)$. Hence, we have $\sum_{t=1}^T H(q_t(x)) \leq \log(KT)\max\{e, Q(a^* \mid x)\}$. Here, to upper bound R_T, it is enough to only consider a case with $Q(a^* \mid x) \geq e$, and we obtain

$$R_T \leq (1+\lambda)O\left(c\sqrt{\log(KT)}\sqrt{\overline{Q}(a^*)\log(KT)}\right) - \lambda\Delta_* \cdot \overline{Q}(a^*) + \lambda C \leq \frac{O\left(\left\{(1+\lambda)c\right\}^2\sqrt{\log(KT)}\right)}{2\lambda\Delta_*} + \lambda\Delta_*.$$

where the second inequality follows from $a\sqrt{b} - \frac{c}{2}b \leq \frac{a^2}{c^2}$ holds for any $a, b, c > 0$. By choosing

$$\lambda = \sqrt{\frac{c^2\log(KT)}{\Delta_*}\Big/\left(\frac{c^2\log(KT)}{\Delta_*} + 2C\right)}.$$

Then, we obtain $R_T = O\left(c^2 \log(KT)/\Delta_* + \sqrt{Cc^2 \log(KT)/\Delta_*}\right)$. □

In the Appendix B, we show the proof procedure of Theorem 4.1.

5 Conclusion

We developed a BoBW algorithm for linear contextual bandits. Our proposed algorithm is based on the FTRL approach. In our theoretical analysis, we show that the upper bounds of the proposed algorithm are given as $O\left(\min\left\{\frac{D}{\Delta_*} + \sqrt{\frac{CD}{\Delta_*}}, \sqrt{\log(KT)TD}\right\}\right)$, where $D = K\log(T)\left(\log(T) + d\log(K)\right)\log(KT)$. This regret upper bound implies $O\left(\min\left\{\sqrt{\frac{TD}{\Delta_*}}, \sqrt{\log(KT)TD}\right\}\right)$ regret in an adversarial environment and $O\left(\frac{D}{\Delta_*}\right)$ regret in an adversarial environment and $O\left(\frac{D}{\Delta_*}\right)$ regret in a stochastic environment. This result also implies $O\left(\frac{\log^3(T)}{\Delta_*}\right)$ regret in a stochastic regime and $O\left(\sqrt{T}\log^2(T)\right)$ regret in an adversarial regime with respect to T.

There are four directions for future work in this study. The first direction is to develop an algorithm that does not require a contextual distribution while maintaining the BoBW property. We expect this extension can be accomplished by applying our proposed method to a method proposed by Liu et al. (2023), based on the FTRL approach with the log-determinant barrier. We note that standard linear contextual bandits in a stochastic environment do not require the contextual distribution to be known, but it is required for dealing with an adversarial environment.

The second direction is to provide lower bounds in our adversarial regimes. In existing studies, Li et al. (2021) provides a general upper bound that holds for a high-dimensional setting with various margin conditions. We can incorporate such results to derive a lower bound in our problem setting.

The third extension is to develop an algorithm that works for linear contextual bandits without assuming a specific minimum gap constant Δ_*. To address this issue, we might use the margin condition to generalize the minimum gap assumption. Lastly, tightening our regret upper bound is also an open problem.

References

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

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

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory (COLT)*, 2016.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1), 2020.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory (COLT)*, 2012.

Cheng Chen, Luo Luo, Weinan Zhang, Yong Yu, and Yijiang Lian. Efficient and robust high-dimensional linear contextual bandits. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

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

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *Annual Conference Computational Learning Theory (COLT)*, 2008.

Chris Dann, Chen-Yu Wei, and Julian Zimmert. A blackbox approach to best of both worlds in bandits and beyond. In Gergely Neu and Lorenzo Rosasco (eds.), *Conference on Learning Theory (COLT)*, volume 195, pp. 5503–5570. PMLR, 2023a.

Christoph Dann, Chen-Yu Wei, and Julian Zimmert. Best of both worlds policy optimization. In *International Conference on Machine Learning (ICML)*, 2023b.

Qin Ding, Cho-Jui Hsieh, and James Sharpnack. Robust stochastic linear contextual bandits under adversarial attacks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Alexander Goldenshluger and Assaf Zeevi. Woodroofe's one-armed bandit problem revisited. *The Annals of Applied Probability*, 19(4):1603–1633, 2009.

Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 2013.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory (COLT)*, 2019.

Elad Hazan, Tomer Koren, Roi Livni, and Yishay Mansour. Online learning with low rank experts. In *Conference on Learning Theory (COLT)*, 2016.

Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Conference on Learning Theory*, volume 134, pp. 2552–2583, 2021.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Tiancheng Jin, Junyan Liu, and Haipeng Luo. Improved best-of-both-worlds guarantees for multi-armed bandits: Ftrl with general regularizers and multiple optimal arms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Varun Kanade and Thomas Steinke. Learning hurdles for sleeping experts. *ACM Transactions on Computation Theory*, 6(3), 2014.

Fang Kong, Canzhe Zhao, and Shuai Li. Best-of-three-worlds analysis for linear bandits with follow-the-regularized-leader algorithm. In *Conference on Learning Theory (COLT)*, 2023.

Yuko Kuroki, Alberto Rumi, Taira Tsuchiya, Fabio Vitale, and Nicolò Cesa-Bianchi. Best-of-both-worlds algorithms for linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238, pp. 1216–1224. PMLR, 2024.

Tor Lattimore and Csaba Szepesvari. The End of Optimism? An Asymptotic Analysis of Finite-Armed Linear Bandits. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, and Xiaojin Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *International Conference on Machine Learning (ICML)*, 2021.

Ke Li, Yun Yang, and Naveen N. Narisetty. Regret lower bound and optimal algorithm for high-dimensional contextual linear bandit. *Electronic Journal of Statistics*, 15(2):5652 – 5695, 2021.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *International Conference on World Wide Web (WWW)*, 2010.

Haolin Liu, Chen-Yu Wei, and Julian Zimmert. Bypassing the simulator: Near-optimal adversarial linear contextual bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Thodoris Lykouris and Sergei Vassilvtiskii. Competitive caching with machine learned advice. In *International Conference on Machine Learning (ICML)*, 2018.

Saeed Masoudian and Yevgeny Seldin. Improved analysis of the tsallis-inf algorithm in stochastically constrained adversarial bandits and stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, volume 134, pp. 3330–3350, 2021.

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

Julia Olkhovskaya, Jack Mayo, Tim van Erven, Gergely Neu, and Chen-Yu Wei. First- and second-order bounds for adversarial linear contextual bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Alexander Rakhlin and Karthik Sridharan. Bistro: An efficient relaxation-based method for contextual bandits. In *International Conference on Machine Learning (ICML)*, 2016.

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

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory (COLT)*, 2017.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning (ICML)*, 2014.

Vasilis Syrgkanis, Haipeng Luo, Akshay Krishnamurthy, and Robert E Schapire. Improved regret bounds for oracle-based adversarial contextual bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Ambuj Tewari and Susan A. Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health: Sensors, Analytic Methods, and Applications*, pp. 495–517, 2017.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Best-of-both-worlds algorithms for partial monitoring. In *International Conference on Algorithmic Learning Theory (ALT)*, 2023a.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Stability-penalty-adaptive follow-the-regularized-leader: Sparsity, game-dependency, and best-of-both-worlds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

Xue Wang, Mingcheng Wei, and Tao Yao. Minimax concave penalized multi-armed bandit model with high-dimensional covariates. In *International Conference on Machine Learning (ICML)*, 2018.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference on Learning Theory (COLT)*, 2018.

Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Linear contextual bandits with adversarial corruptions, 2021. URL https://openreview.net/forum?id=Wz-t1oOTWa.

Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(1), 2021.

A Details of Example 1

When $\min_{b\neq a_0^*(x)} \Big\langle x, \theta_0(b)\Big\rangle - \Big\langle x, \theta_0(a_0^*(x))\Big\rangle \geq \Delta_*$ holds for all $x \in \mathcal{X}$, we have

$$\begin{aligned}
R_T &= \mathbb{E}\left[\sum_{t=1}^T \ell_t(A_t, X_t) - \sum_{t=1}^T \ell_t(a_0^*(X_t), X_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \sum_{a\in[K]} X_t\Big(\theta_0^*(a) - \theta_0^*(a_0^*(X_t))\Big)\pi_t(a \mid X_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \sum_{a\neq a_0^*(X_t)} X_t\Big(\theta_0^*(a) - \theta_0^*(a_0^*(X_t))\Big)\pi_t(a \mid X_t)\right] \\
&\geq \Delta_* \cdot \mathbb{E}\left[\sum_{t=1}^T \sum_{a\neq a_0^*(X_t)} \pi_t(a \mid X_t)\right] \\
&\geq \Delta_* \cdot \mathbb{E}\left[\sum_{t=1}^T \Big(1 - \pi_t(a_0^*(X_0) \mid X_0)\Big)\right]
\end{aligned}$$

B Proof of Theorem 4.1

B.1 Preliminaries for the Proof of Theorem 4.1

Let X_0 be a sample from the context distribution $\mathcal{D}$ independent of $\mathcal{F}_T$. Let $D_t(p, q)$ denote the Bregman divergence of $p.q \in \Pi$ with respect to ψ_t; that is,

$$D_t(p, q) = \psi_t(p) - \psi_t(q) - \Big\langle \nabla\psi_t(q), p - q\Big\rangle.$$

Let us define $\pi^* \in \Pi$ as $\pi^*(a^*(x) \mid x) = 1$ and $\pi^*(a \mid x) = 0$ for all $a \in [K]\backslash\{a^*(x)\}$.

Then, the following lemma holds. The proof is shown in Appendix C

Lemma B.1. *If A_t is chosen as our proposed method, the regret is bounded by*

$$\begin{aligned}
R_T \leq \mathbb{E}\Bigg[\sum_{t=1}^T \Big\{\gamma_t &+ \Big\langle \widehat{\ell}_t(X_0), q_t(X_0) - q_{t+1}(X_0)\Big\rangle \\
&- D_t(q_{t+1}(X_0), q_t(X_0)) + \psi_t(q_{t+1}(X_0)) - \psi_{t+1}(q_{t+1}(X_0))\Big\} \\
&+ \psi_{T+1}(\pi^*(X_0)) - \psi_1(q_1(X_0))\Bigg] + 2\sum_{t=1}^T \max_{a\in[K]} \left|\mathbb{E}\left[\langle X_t, \theta_t(a) - \widehat{\theta}_t(a)\rangle\right]\right|.
\end{aligned}$$

To show Lemma B.1, we use the following proposition from Neu & Olkhovskaya (2020).

Proposition B.2. *Suppose that $\pi_t \in \mathcal{F}_{t-1}$ and that $\mathbb{E}\left[\widehat{\theta}_t(a)|\mathcal{F}_{t-1}\right] = \theta_t(a)$ for all t, a hold. Then, the following holds:*

$$\mathbb{E}\left[\sum_{t=1}^T \sum_{a\in[K]} \Big(\pi_t(a \mid X_t) - \pi^*(a \mid X_t)\Big)\Big\langle X_t, \theta_t(a)\Big\rangle\right] = \mathbb{E}\left[\sum_{t=1}^T \sum_{a\in[K]} \Big(\pi_t(a \mid X_0) - \pi^*(a \mid X_0)\Big)\Big\langle X, \widehat{\theta}_t(a)\Big\rangle\right].$$

This proposition plays an important role throughout this study.

Bounding the stability term. For the stability term $\left\langle \widehat{\ell}_t(X_0), q_t(X_0) - q_{t+1}(X_0) \right\rangle - D_t(q_{t+1}(X_0), q_t(X_0))$, we use the following proposition from Ito et al. (2022).

Proposition B.3 (From Lemma 8 in Ito et al. (2022))**.** *If ψ_t is given as equation 3, for any $\ell : \mathcal{X} \to \mathbb{R}^K$ and $p, q \in \Pi$, we have*

$$\left\langle \ell_t(x), p(x) - q(x) \right\rangle - D_t(q(x), p(x)) \leq \beta_t \sum_{a \in [K]} p(a \mid x) \xi \left(\frac{\ell_t(a, x)}{\beta_t} \right).$$

for any $x \in \mathcal{X}$, where $\xi(x) \coloneqq \exp(-x) + x - 1$.

For $\widehat{\ell}_t(a, x)$, if $\frac{\widehat{\ell}_t(a,x)}{\beta_t} \geq -1$ holds, then Proposition B.3 implies

$$\left\langle \widehat{\ell}_t(x), q_t(x) - q_{t+1}(x) \right\rangle - D_t(q_{t+1}(x), q_t(x)) \leq \frac{1}{\beta_t} \sum_{a \in [K]} \pi_t(a \mid x) \widehat{\ell}_t^2(a, x).$$

For the RHS, we apply the following proposition from Neu & Olkhovskaya (2020).

Proposition B.4 (From Lemma 6 in Neu & Olkhovskaya (2020))**.** *For each $t \in [T]$, our strategy satisfies*

$$\mathbb{E}\left[\sum_{a \in [K]} \pi_t(a \mid X_0) \widehat{\ell}_t^2(a, X_0) \mid \mathcal{F}_{t-1} \right] \leq 3Kd.$$

Estimation error of the design matrix. Next, we bound $\sum_{t=1}^T \max_{a \in [K]} \left| \mathbb{E}\left[\langle X_t, \theta_t(a) - \widehat{\theta}_t(a) \rangle \right] \right|$. An upper bound of $\sum_{t=1}^T \max_{a \in [K]} \left| \mathbb{E}\left[\langle X_t, \theta_t(a) - \widehat{\theta}_t(a) \rangle \right] \right|$ is given as the following lemma.

Lemma B.5. *For all t such that $\gamma_t = \widetilde{\gamma}_t$, we have $\left| \mathbb{E}\left[\langle X_t, \theta_t(a) - \widehat{\theta}_t(a) \rangle \right] \right| \leq C_{\mathcal{X}} C_{\Theta} / T$.*

Proof of Lemma B.5. From Lemma 5 in Neu & Olkhovskaya (2020), we have $\left| \mathbb{E}\left[\langle X_t, \theta_t(a) - \widehat{\theta}_t(a) \rangle \right] \right| \leq C_{\mathcal{X}} C_{\Theta} \exp\left(-\frac{\gamma_t \delta}{K} \lambda_{\min} M_t \right)$. Then, when $\gamma_t = \widetilde{\gamma}_t$, we have

$$\begin{aligned} \exp\left(-\frac{\gamma_t \delta}{K} \lambda_{\min} M_t \right) &= \exp\left(-\frac{K \log(T)}{\delta \lambda_{\min} \cdot 2\beta_t} \frac{\delta \lambda_{\min}}{K} M_t \right) \\ &\leq \exp\left(-\frac{K \log(T)}{\delta \lambda_{\min} \cdot (2\beta_t - 1)} \frac{\delta \lambda_{\min}}{K} M_t \right) = \exp\left(-\log(T) \right) = \frac{1}{T}, \end{aligned}$$

where recall that we defined $M_t = 2\beta_t - 1$. □

B.2 Proof of Theorem 4.1

Then, we show the following lemma. The proof is shown in Appendix D.

Lemma B.6. *The regret for the* `BoBW-RealFTRL` *with $\widehat{\Sigma}^\dagger_{t,a}$ is bounded as*

$$R_T \leq \mathbb{E}\left[\sum_{t=1}^T \left\{ \gamma_t + \frac{3Kd}{\beta_t} + \left(\beta_{t+1} - \beta_t \right) H(q_{t+1}(X_0)) \right\} \right] + \beta_1 \log(K) + 2C_{\mathcal{X}} C_{\Theta}.$$

From this result, we obtain the following lemma. We provide the proof in Appendix E

Lemma B.7. *Assume the conditions in Theorem B.6. Suppose that β_t and γ_t satisfy equation 4 and equation 5. Then, we have*

$$R_T \leq \bar{c} \sqrt{\mathbb{E}\left[\sum_{t=1}^T H(q_t(X_0)) \right]} + 2T_0 + 2C_{\mathcal{X}} C_{\Theta},$$

where

$$\overline{c} = O\left(\frac{K\log(T)}{\beta_1}\left(\frac{\log(T)}{\delta\lambda_{\min}\sqrt{\log(K)}} + d\right) + \beta_1\sqrt{\log(K)}\right),$$

$$T_0 \coloneqq \left\lceil \frac{K\log(T)}{2\delta\lambda_{\min}\beta_1\sqrt{\log(K)}}\sqrt{\sum_{s=1}^{T} H\big(q_s(X_s)\big)}\right\rceil.$$

Next, we consider bounding $\sum_{t=1}^{T} H(q_t(x))$ by $Q(a^* \mid x)$ as shown in the following lemma.

Lemma B.8 (From Lemma 4 in Ito et al. (2022))**.** *For any $a^* : \mathcal{X} \to [K]$, the following holds:*

$$\sum_{t=1}^{T} H(q_t(x)) \leq Q(a^* \mid x)\log\left(\frac{eKT}{Q(a^* \mid x)}\right),$$

where e is Napier's constant.

By using the above lemmas and propositions, we prove Theorem 4.1.

Proof of Theorem 4.1. From Lemma B.8, if $Q(a^* \mid x) \leq e$, we have $\sum_{t=1}^{T} H(q_t(x)) \leq e\log(KT)$ and otherwise, we have $\sum_{t=1}^{T} H(q_t(x)) \leq Q(a^* \mid x)\log(KT)$. Hence, we have $\sum_{t=1}^{T} H(q_t(x)) \leq \log(KT)\max\{e, Q(a^* \mid x)\}$. From Lemma B.7, we have

$$\begin{aligned}R_T &\leq \overline{c}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[H(q_t(X_0))\right]} + 2C_{\mathcal{X}}C_{\Theta} \\ &= O\left(\left(\frac{K\log(T)}{\beta_1}\left(\frac{\log(T)}{\delta\lambda_{\min}\sqrt{\log(K)}} + d\right) + \beta_1\sqrt{\log(K)}\right)\sqrt{\log(KT)}\max\left\{\overline{Q}^{1/2}, 1\right\}\right).\end{aligned}$$

□

C Proof of Lemma B.1

Let us define

$$\widehat{R}_T(x) \coloneqq \sum_{t=1}^{T}\sum_{a\in[K]}\Big(\pi_t(a \mid x) - \pi^*(a \mid x)\Big)\left\langle x, \widehat{\theta}_t(a)\right\rangle.$$

Then, the following holds:

$$R_T \leq \mathbb{E}\left[\widehat{R}_T(X_0)\right] + 2\sum_{t=1}^{T}\max_{a\in[K]}\left|\mathbb{E}\left[\langle X_t, \theta_t(a) - \widehat{\theta}_t(a)\rangle\right]\right|.$$

Then, we prove Lemma B.1 as follows.

Proof of Lemma B.1. From the definition of our algorithm, we have

$$\begin{aligned}R_T &= \mathbb{E}\left[\sum_{t=1}^{T}\ell_t(A_t, X_t) - \sum_{t=1}^{T}\ell_t(a^*_T, X)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T}\langle \ell_t(X_t), \pi_t(X_t) - \pi^*(X_t)\rangle\right]\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\sum_{t=1}^{T} \langle \ell_t(X_t), q_t(X_t) - \pi^*(X_t)\rangle + \sum_{t=1}^{T} \gamma_t \left\langle \ell_t(X_t), \frac{1}{K}\iota - q_t(X_t)\right\rangle\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \ell_t(X_t), q_t(X_t) - \pi^*(X_t)\right\rangle + C_\ell \sum_{t=1}^{T} \gamma_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \ell_t(X_0), q_t(X_0) - \pi^*(X_0)\right\rangle + C_\ell \sum_{t=1}^{T} \gamma_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \widehat{\ell}_t(X_0), q_t(X_0) - \pi^*(X_0)\right\rangle + C_\ell \sum_{t=1}^{T} \gamma_t\right] + \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \ell_t(X_0) - \widehat{\ell}_t(X_0), q_t(X_0) - \pi^*(X_0)\right\rangle\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \widehat{\ell}_t(X_0), q_t(X_0) - \pi^*(X_0)\right\rangle + C_\ell \sum_{t=1}^{T} \gamma_t\right] + 2\sum_{t=1}^{T} \max_{a\in[K]} \left|\mathbb{E}\left[\left\langle X_0, \theta_t(a) - \widehat{\theta}_t(a)\right\rangle\right]\right|. \qquad (6)
\end{aligned}
$$

Then, from the definitions of q_t, for each $x \in \mathcal{X}$, we also have

$$
\begin{aligned}
&\sum_{t=1}^{T} \left\langle \widehat{\ell}_t(x), \pi^*(x)\right\rangle + \psi_{T+1}(\pi^*(x)) \\
&\geq \sum_{t=1}^{T} \left\langle \widehat{\ell}_t(x), q_{T+1}(x)\right\rangle + \psi_{T+1}(q_{T+1}(x)) - \psi_{T+1}(q_{T+1}(x)) \\
&\quad + \psi_{T+1}(\pi^*(x)) - \langle \nabla\psi_t(q_{T+1}(x)), \pi^*(x) - q_{T+1}(x)\rangle \\
&= \sum_{t=1}^{T} \left\langle \widehat{\ell}_t(x), q_{T+1}(x)\right\rangle + \psi_{T+1}(q_{T+1}(x)) + D_{T+1}(\pi^*(x), q_{T+1}(x)),
\end{aligned}
$$

where we used that $\langle \nabla\psi_t(q_{T+1}(x)), \pi^*(x) - q_{T+1}(x)\rangle \geq 0$ holds for a convex function ψ_t. Then, it holds that

$$
\begin{aligned}
&\sum_{t=1}^{T} \left\langle \widehat{\ell}_t(x), \pi^*(x)\right\rangle + \psi_{T+1}(\pi^*(x)) \\
&\geq \sum_{t=1}^{T} \left\langle \widehat{\ell}_t(x), q_{T+1}(x)\right\rangle + D_{T+1}(\pi^*(x), q_{T+1}(x)) + \psi_{T+1}(q_{T+1}(x)) \\
&\geq \sum_{t=1}^{T} \left\langle \widehat{\ell}_t(x), q_{T+1}(x)\right\rangle + \psi_T(q_T(x)) \\
&\quad + D_T(q_{T+1}(x), q_T(x)) + D_{T+1}(\pi^*(x), q_{T+1}(x)) - \psi_T(q_{T+1}(x)) + \psi_{T+1}(q_{T+1}(x)) \\
&= \sum_{t=1}^{T-1} \left\langle \widehat{\ell}_t(x), q_{T+1}(x)\right\rangle + \psi_T(q_T(x)) \\
&\quad + \left\langle \widehat{\ell}_T(x), q_{T+1}(x)\right\rangle + D_T(q_{T+1}(x), q_T(x)) + D_{T+1}(\pi^*(x), q_{T+1}(x)) - \psi_T(q_{T+1}(x)) + \psi_{T+1}(q_{T+1}(x)) \\
&\geq \sum_{t=1}^{T-1} \left\langle \widehat{\ell}_t(x), q_T(x)\right\rangle + \psi_T(q_T(x)) \\
&\quad + \left\langle \widehat{\ell}_T(x), q_{T+1}(x)\right\rangle + D_T(q_{T+1}(x), q_T(x)) + D_{T+1}(\pi^*(x), q_{T+1}(x)) - \psi_T(q_{T+1}(x)) + \psi_{T+1}(q_{T+1}(x)) \\
&\geq \sum_{t=1}^{T} \left\langle \widehat{\ell}_t(x), q_{t+1}(x)\right\rangle + \sum_{t=1}^{T} D_t(q_{t+1}(x), q_t(x)) - \sum_{t=1}^{T} \left\{\psi_t(q_{t+1}(x)) - \psi_{t+1}(q_{t+1}(x))\right\} + \psi_1(q_1(x)).
\end{aligned}
$$

Therefore, we have

$$
\sum_{t=1}^{T} \left\langle \widehat{\ell}_t(x), q_t(x) - \pi^*(x)\right\rangle
$$

$$
\begin{aligned}
&\leq \sum_{t=1}^{T} \left\{ \left\langle \widehat{\ell}_t(x), q_t(x) - q_{t+1}(x) \right\rangle - D_t(q_{t+1}(x), q_t(x)) + \psi_t(q_{t+1}(x)) - \psi_{t+1}(q_{t+1}(x)) \right\} \\
&\qquad + \psi_{T+1}(\pi^*(x)) - \psi_1(q_1(x)).
\end{aligned}
$$

Combining this with equation 6, we obtain

$$
\begin{aligned}
R_T \leq \mathbb{E}\Bigg[&\sum_{t=1}^{T} \left\{ \left\langle \widehat{\ell}_t(X_0), q_t(X_0) - q_{t+1}(X_0) \right\rangle - D_t(q_{t+1}(x), q_t(X_0)) + \psi_t(q_{t+1}(X_0)) - \psi_{t+1}(q_{t+1}(X_0)) \right\} \\
&+ \psi_{T+1}(\pi^*(X_0)) - \psi_1(q_1(X_0)) + C_\ell \sum_{t=1}^{T} \gamma_t \Bigg] + 2 \sum_{t=1}^{T} \max_{a \in [K]} \left| \mathbb{E}\left[\left\langle X_0, \theta_t(a) - \widehat{\theta}_t(a) \right\rangle \right] \right|.
\end{aligned}
$$

□

D Proof of Lemma B.6

Proof of Lemma B.6. From Lemma B.1, we have

$$
\begin{aligned}
R_T \leq \mathbb{E}\Bigg[&\sum_{t=1}^{T} \Big(\gamma_t + \left\langle \widehat{\ell}_t(X_0, d), \pi_t(X_0) - q_{t+1}(X_0) \right\rangle - D_t(q_{t+1}(X_0), \pi_t(X_0)) \\
&\qquad\qquad + \psi_t(q_{t+1}(X_0)) - \psi_{t+1}(q_{t+1}(X_0)) \Big) + \psi_{T+1}(\pi^*(X_0)) - \psi_1(q_1(x)) \Bigg] \\
&+ 2 \sum_{t=1}^{T} \max_{a \in [K]} \left| \mathbb{E}\left[\langle X_t, \theta_t(a) - \widehat{\theta}_t(a) \rangle \right] \right|.
\end{aligned}
$$

First, we show

$$
\mathbb{E}\left[\left\langle \widehat{\ell}_t(X_0), \pi_t(X_0) - q_{t+1}(X_0) \right\rangle - D_t(q_{t+1}(X_0), \pi_t(X_0)) \right] \leq \frac{3Kd}{\beta_t}. \tag{7}
$$

To show this, we confirm $\frac{\widehat{\ell}_t(a,x)}{\beta_t} \geq -1$, which is necessary to derive an upper bound from Proposition B.3. We have

$$
\begin{aligned}
\frac{1}{\beta_t} \cdot \left\langle X_0, \widehat{\theta}_t(a) \right\rangle &= \frac{1}{\beta_t} \cdot X_0^\top \widehat{\Sigma}_{t,a}^\dagger X_t \left\langle X_t, \theta_{t,a} \right\rangle \mathbb{1}[A_t = a] \geq -\frac{C_\ell}{\beta_t} \cdot \left| X_0^\top \widehat{\Sigma}_{t,a}^\dagger X_t \right| \\
&\geq -\frac{1}{\beta_t} C_\Theta C_{\mathcal{X}}^2 \left\| \widehat{\Sigma}_{t,a}^\dagger \right\|_{\mathrm{op}} \geq -\frac{1}{\beta_t} C_\Theta C_{\mathcal{X}}^2 \delta \left(1 + \sum_{k=1}^{M_t} \| V_{k,a} \|_{\mathrm{op}} \right) = -\frac{1}{2\beta_t}(M_t + 1),
\end{aligned}
$$

where we used that $\delta = \frac{1}{2C_\Theta C_{\mathcal{X}}^2}$. Here, recall that we defined M_t as $2\beta_t - 1$. Therefore, $\frac{\widehat{\ell}_t(a,x)}{\beta_t} = -1$ holds. Then, we have

$$
\begin{aligned}
&\left\langle \widehat{\ell}_t(x), \pi_t(x) - q_{t+1}(x) \right\rangle - D_t(q_{t+1}(x), \pi_t(x)) \\
&\leq \beta_t \sum_{a \in [K]} \pi_t(a \mid x) \xi \left(\frac{\widehat{\ell}_t(a,x)}{\beta_t} \right) \leq \frac{1}{\beta_t} \sum_{a \in [K]} \pi_t(a \mid x) \widehat{\ell}_t^2(a,x).
\end{aligned}
$$

Then, from Proposition B.4, we have equation 7.

From $\psi_t(q(x)) = -\beta_t H(q(x))$, we have

$$
\sum_{t=1}^{T} \left(\psi_t(q_{t+1}(x)) - \psi_{t+1}(q_{t+1}(x)) \right) + \psi_{T+1}(\pi^*(x)) - \psi_1(q_1(x))
$$

$$\leq \sum_{t=1}^{T} (\beta_{t+1} - \beta_t) H(q_{t+1}(x)) + \beta_1 \log(K).$$

From $\beta_{t+1} = \beta_t + \frac{\beta_1}{\sqrt{1+\left(\log(K)\right)^{-1}\sum_{s=1}^{t} H\left(q_s(X_s)\right)}}$, we obtain

$$\beta_t = \beta_1 + \sum_{u=1}^{t-1} \frac{\beta_1}{\sqrt{1+\left(\log(K)\right)^{-1}\sum_{s=1}^{u} H\left(q_s(X_s)\right)}} \geq \frac{t\beta_1}{\sqrt{1+\left(\log(K)\right)^{-1}\sum_{s=1}^{t} H\left(q_s(X_s)\right)}}.$$

Here, we have

$$\begin{aligned}\widetilde{\gamma}_t = \frac{K}{2\delta\lambda_{\min}\beta_t}\log(T) &\leq \frac{K}{2\delta\lambda_{\min}t\beta_1}\log(T)\sqrt{1+\left(\log(K)\right)^{-1}\sum_{s=1}^{t} H\left(q_s(X_s)\right)} \\ &\leq \frac{K}{2\delta\lambda_{\min}t\beta_1\sqrt{\log(K)}}\log(T)\sqrt{\sum_{s=1}^{T} H\left(q_s(X_s)\right)}.\end{aligned}$$

Therefore, $\widetilde{\gamma}_t \leq 1$ holds for all $t \geq T_0$ such that

$$T_0 = \left\lceil \frac{K\log(T)}{2\delta\lambda_{\min}\beta_1\sqrt{\log(K)}}\sqrt{\sum_{s=1}^{T} H\left(q_s(X_s)\right)} \right\rceil.$$

From Lemma B.5, if there exists $T_0 \in \mathbb{N}$ such that $\gamma_t = \widetilde{\gamma}_t$ holds for all $t \geq T_0$, we have

$$\sum_{t=t=1}^{T} \max_{a\in[K]} \left| \mathbb{E}\left[\langle X_t, \theta_t(a) - \widehat{\theta}_t(a)\rangle\right] \right| \leq T_0 + \sum_{t=1}^{T} C_{\mathcal{X}} C_{\Theta} \frac{1}{T} = 2T_0 + C_{\mathcal{X}} C_{\Theta},$$

where we used $\langle X_t, \theta_t(a) - \widehat{\theta}_t(a)\rangle \leq 2$ □

E Proof of Lemma B.7

Proof of Lemma B.7. Firstly, we note that the following equality holds:

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T} (\beta_{t+1} - \beta_t) H(q_{t+1}(X_{t+1}))\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} (\beta_{t+1} - \beta_t) \mathbb{E}\left[H(q_{t+1}(X_{t+1})) \mid \mathcal{F}_t\right]\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} (\beta_{t+1} - \beta_t) \mathbb{E}\left[H(q_{t+1}(X_0)) \mid \mathcal{F}_t\right]\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} (\beta_{t+1} - \beta_t) H(q_{t+1}(X_0))\right]\end{aligned}$$

We show the following two inequalities:

$$\sum_{t=1}^{T} \left(\gamma_t + \frac{3Kd}{\beta_t}\right) = O\left(\frac{K\log(T)}{\beta_1}\left(\frac{\log(T)}{\delta\lambda_{\min}\sqrt{\log(K)}} + d\right)\sqrt{\sum_{t=1}^{T} H\left(q_t(X_t)\right)}\right) \tag{8}$$

$$\sum_{t=1}^{T} \left(\beta_{t+1} - \beta_t\right) H(q_{t+1}(X_{t+1})) = O\left(\beta_1 \sqrt{\log(K)} \sqrt{\sum_{t=1}^{T} H(q_t(X_t)}\right). \tag{9}$$

First, we show equation 8. From $\gamma_t = \min\left\{1, \frac{K}{4\delta\lambda_{\min}\beta_t}\log(T)\right\}$, we obtain

$$\sum_{t=1}^{T}\left(\gamma_t + \frac{3Kd}{\beta_t}\right) \leq \sum_{t=1}^{T}\left(\frac{K}{2\delta\lambda_{\min}\beta_t}\log(T) + \frac{3Kd}{\beta_t}\right) = \left(\frac{K}{2\frac{1}{2C_\Theta C_{\mathcal{X}}^2}\delta\lambda_{\min}}\log(T) + 3Kd\right)\sum_{t=1}^{T}\frac{1}{\beta_t}.$$

From $\beta_{t+1} = \beta_t + \frac{\beta_1}{\sqrt{1+\big(\log(K)\big)^{-1}\sum_{s=1}^{t} H\big(q_s(X_s)\big)}}$, we obtain

$$\beta_t = \beta_1 + \sum_{u=1}^{t-1}\frac{\beta_1}{\sqrt{1+\big(\log(K)\big)^{-1}\sum_{s=1}^{u} H\big(q_s(X_s)\big)}} \geq \frac{t\beta_1}{\sqrt{1+\big(\log(K)\big)^{-1}\sum_{s=1}^{t} H\big(q_s(X_s)\big)}}.$$

Therefore, we have

$$\sum_{t=1}^{T}\frac{1}{\beta_t} \leq \sum_{t=1}^{T}\frac{\sqrt{1+\big(\log(K)\big)^{-1}\sum_{s=1}^{t} H\big(q_s(X_s)\big)}}{t\beta_1} \leq \frac{1+\log(T)}{\beta_1}\sqrt{1+\big(\log(K)\big)^{-1}\sum_{s=1}^{T} H\big(q_s(X_s)\big)}.$$

By using $H\big(q_1(x)\big) = \log(K)$, we obtain

$$\sum_{t=1}^{T}\left(\gamma_t + \frac{3Kd}{\beta_t}\right) = O\left(\frac{K\log(T)}{\beta_1}\left(\frac{\log(T)}{\delta\lambda_{\min}\sqrt{\log(K)}} + d\right)\sqrt{\sum_{t=1}^{T} H\big(q_{t+1}(X_s)\big)}\right).$$

Next, we show equation 9. From the definitions of β_t and γ_t, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\beta_{t+1}-\beta_t\right) H(q_{t+1}(X_{t+1})) = \sum_{t=1}^{T}\frac{\beta_1}{\sqrt{1+\big(\log(K)\big)^{-1}\sum_{s=1}^{t} H\big(q_s(X_s)\big)}} H(q_{t+1}(X_{t+1}))\\
&= 2\beta_1\sqrt{\log(K)}\sum_{t=1}^{T}\frac{H\big(q_{t+1}(X_{t+1})\big)}{\sqrt{\log(K)+\sum_{s=1}^{t} H\big(q_s(X_s)\big)}+\sqrt{\log(K)+\sum_{s=1}^{t} H\big(q_s(X_s)\big)}}\\
&\leq 2\beta_1\sqrt{\log(K)}\sum_{t=1}^{T}\frac{H\big(q_{t+1}(X_{t+1})\big)}{\sqrt{\log(K)+\sum_{s=1}^{t+1} H\big(q_s(X_s)\big)}+\sqrt{\log(K)+\sum_{s=1}^{t} H\big(q_s(X_s)\big)}}\\
&\leq 2\beta_1\sqrt{\log(K)}\sum_{t=1}^{T}\frac{H\big(q_{t+1}(X_{t+1})\big)}{\sqrt{\sum_{s=1}^{t+1} H\big(q_s(X_s)\big)}+\sqrt{\sum_{s=1}^{t} H\big(q_s(X_s)\big)}}\\
&= 2\beta_1\sqrt{\log(K)}\sum_{t=1}^{T}\frac{H(q_{t+1}(X_{t+1}))}{H(q_{t+1}(X_{t+1}))}\left\{\sqrt{\sum_{s=1}^{t+1} H\big(q_s(X_s)\big)} - \sqrt{\sum_{s=1}^{t} H\big(q_s(X_s)\big)}\right\}\\
&= 2\beta_1\sqrt{\log(K)}\sum_{t=1}^{T}\left\{\sqrt{\sum_{s=1}^{t+1} H\big(q_s(X_s)\big)} - \sqrt{\sum_{s=1}^{t} H\big(q_s(X_s)\big)}\right\}\\
&= 2\beta_1\sqrt{\log(K)}\left\{\sqrt{\sum_{s=1}^{T+1} H\big(q_s(X_s)\big)} - \sqrt{H\big(q_1(X_1)\big)}\right\}
\end{aligned}$$

$$\leq 2\beta_1 \sqrt{\log(K)} \sqrt{\sum_{s=1}^{T} H\big(q_s(X_s)\big)},$$

where we used $\sqrt{H(q_{T+1}(X_{T+1}))} \leq \sqrt{H(q_1(X_1))}$.

From $\beta_{t+1} = \beta_t + \frac{\beta_1}{\sqrt{1+\big(\log(K)\big)^{-1}\sum_{s=1}^{t} H\big(q_s(X_s)\big)}}$, we obtain

$$\beta_t = \beta_1 + \sum_{u=1}^{t-1} \frac{\beta_1}{\sqrt{1+\big(\log(K)\big)^{-1}\sum_{s=1}^{u} H\big(q_s(X_s)\big)}} \geq \frac{t\beta_1}{\sqrt{1+\big(\log(K)\big)^{-1}\sum_{s=1}^{t} H\big(q_s(X_s)\big)}}.$$

Inequalities equation 8 and equation 9 combined with the inequality in Lemma B.6 yield

$$\begin{aligned}
R_T &\leq \mathbb{E}\left[\sum_{t=1}^{T}\left\{\gamma_t + \frac{3Kd}{\beta_t} + (\beta_{t+1}-\beta_t)\, H(q_{t+1}(X_0))\right\}\right] + \beta_1 \log(K) + 2T_0 + 2C_{\mathcal{X}}C_{\Theta} \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left\{\gamma_t + \frac{3Kd}{\beta_t} + (\beta_{t+1}-\beta_t)\, H(q_{t+1}(X_{t+1}))\right\}\right] + \beta_1 \log(K) + 2T_0 + 2C_{\mathcal{X}}C_{\Theta} \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left\{O\left(\frac{K\log(T)\big(\log(T)+\delta\lambda_{\min}d\big)}{\beta_1\delta\lambda_{\min}\sqrt{\log(K)}}\sqrt{\sum_{s=1}^{T} H\big(q_{t+1}(X_s)\big)}\right) + O\left(\beta_1\sqrt{\log(K)}\sqrt{\sum_{t=1}^{T} H(q_t(X_t)}\right)\right\}\right] \\
&\quad + \beta_1\log(K) + 2T_0 + 2C_{\mathcal{X}}C_{\Theta} \\
&= \sum_{t=1}^{T}\left\{O\left(\frac{K\log(T)\big(\log(T)+\delta\lambda_{\min}d\big)}{\beta_1\delta\lambda_{\min}\sqrt{\log(K)}}\sqrt{\sum_{s=1}^{T}\mathbb{E}\left[H\big(q_{t+1}(X_0)\big)\right]\,)}\right) + O\left(\beta_1\sqrt{\log(K)}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[H(q_t(X_t)\right]}\right)\right\} \\
&\quad + \beta_1\log(K) + 2T_0 + 2C_{\mathcal{X}}C_{\Theta}.
\end{aligned}$$

Thus, we obtain the regret bound in Lemma B.7. □