Best-of-Both-Worlds Algorithms for Linear Contextual Bandits

Yuko Kuroki *[1], Alberto Rumi†[2,1], Taira Tsuchiya‡[3], Fabio Vitale §[1], and Nicolò Cesa-Bianchi ¶[2,4]

[1]CENTAI Institute, Turin, Italy
[2]Università degli Studi di Milano, Milan, Italy
[3]The University of Tokyo, Tokyo, Japan
[4]Politecnico di Milano, Milan, Italy

December 27, 2023

Abstract

We study best-of-both-worlds algorithms for K-armed linear contextual bandits. Our algorithms deliver near-optimal regret bounds in both the adversarial and stochastic regimes, without prior knowledge about the environment. In the stochastic regime, we achieve the polylogarithmic rate $\frac{(dK)^2 \mathrm{polyln}(dKT)}{\Delta_{\min}}$, where $\Delta_{\min}$ is the minimum suboptimality gap over the d-dimensional context space. In the adversarial regime, we obtain either the first-order $\widetilde{\mathcal{O}}(dK\sqrt{L^*})$ bound, or the second-order $\widetilde{\mathcal{O}}(dK\sqrt{\Lambda^*})$ bound, where L^* is the cumulative loss of the best action and Λ^* is a notion of the cumulative second moment for the losses incurred by the algorithm. Moreover, we develop an algorithm based on FTRL with Shannon entropy regularizer that does not require the knowledge of the inverse of the covariance matrix, and achieves a polylogarithmic regret in the stochastic regime while obtaining $\widetilde{\mathcal{O}}\big(dK\sqrt{T}\big)$ regret bounds in the adversarial regime.

1 Introduction

Because of their relevance in practical applications, contextual bandits are a fundamental model of sequential decision-making with partial feedback. In particular, linear contextual bandits [Abe and Long, 1999, Auer, 2002], in which contexts are feature vectors and the loss is a linear function of the context, are among the most studied variants of contextual bandits. Traditionally, contextual bandits (and, in particular, their linear variant) have been investigated under stochastic assumptions on the generation of rewards. Namely, the loss of each action is a fixed and unknown linear function of the context to which some zero-mean noise is added. For this setting, efficient and nearly optimal algorithms, like OFUL [Abbasi-Yadkori et al., 2011] and a contextual variant of Thompson Sampling [Agrawal and Goyal, 2013], have been proposed in the past.

Recently, Neu and Olkhovskaya [2020] introduced an adversarial variant of linear contextual bandits, where there are K arms and the linear loss associated with each arm is adversarially chosen in each round. They prove an upper bound on the regret of order $\sqrt{dKT}$ disregarding logarithmic factors, where d is the dimensionality of contexts and T is the time horizon. A matching lower bound $\Omega\big(\sqrt{dKT}\big)$ for this model is implied by the results of Zierahn et al. [2023]. The upper bound has been recently extended by Olkhovskaya et al. [2023], who show first and second-order regret bounds respectively of the order of $K\sqrt{dL^*}$ and $K\sqrt{d\Lambda^*}$ (again disregarding log factors), where L^* is cumulative loss of the best action and Λ^* is a notion of cumulative second moment for the losses incurred by the algorithm.

*yuko.kuroki@centai.eu
†alberto.rumi@unimi.it
‡tsuchiya@mist.i.u-tokyo.ac.jp
§fabio.vitale@centai.eu
¶nicolo.cesa-bianchi@unimi.it

Table 1: A comparison of regret bounds for linear contextual bandits. $\widetilde{\mathcal{O}}$ ignores (poly)logarithmic factors. The $\sqrt{C}$ column specifies whether in the corrupted stochastic regime the algorithm achieves the optimal $\sqrt{C}$ dependence on the corruption level $C \geq 0$. For the bound in the adversarial regime, we omit additive terms polylogarithmic in T. See Section 2 for a formal definition of the quantities appearing in the bounds.

reference	stochastic	adversarial	$\sqrt{C}$	$\mathbf{\Sigma}^{-1}$
Neu and Olkhovskaya [2020]	–	$\mathcal{O}\left(\sqrt{TK\max\left\{d, \frac{\ln T}{\lambda_{\min}(\mathbf{\Sigma})}\right\}\ln(K)}\right)$	–	Unknown
Olkhovskaya et al. [2023]	–	$\widetilde{\mathcal{O}}\left(K\sqrt{d\Lambda^*}\right)$	–	Known
Olkhovskaya et al. [2023]	–	$\widetilde{\mathcal{O}}\left(K\sqrt{dL^*}\right)$	–	Known
Zierahn et al. [2023]	–	$\mathcal{O}\left(\sqrt{TK\max\left\{d, \frac{\ln T}{\lambda_{\min}(\mathbf{\Sigma})}\right\}\ln(K)}\right)$	–	Unknown
Proposition 8	$\mathcal{O}\left(\frac{K^2}{\Delta_{\min}}\left(d+\frac{1}{\lambda_{\min}(\mathbf{\Sigma})}\right)^2\ln(K)\ln T\right)$	$\mathcal{O}\left(\sqrt{TK^2\left(d+\frac{1}{\lambda_{\min}(\mathbf{\Sigma})}\right)^2\ln(K)}\right)$	✓	Known
Theorem 1	$\mathcal{O}\left(\frac{(dK)^2}{\Delta_{\min}}\ln^2(dKT)\ln^3 T\right)$	$\widetilde{\mathcal{O}}\left(dK\sqrt{\Lambda^*}\right)$	✓	Known
Corollary 1	$\mathcal{O}\left(\frac{(dK)^2}{\Delta_{\min}}\ln^2(dKT)\ln^3 T\right)$	$\widetilde{\mathcal{O}}\left(dK\sqrt{\min\{L^*, \overline{\Lambda}\}}\right)$	✓	Known
Theorem 2	$\mathcal{O}\left(\frac{K}{\Delta_{\min}}\left(d+\frac{\ln T}{\lambda_{\min}(\mathbf{\Sigma})}\right)\ln(KT)\ln T\right)$	$\mathcal{O}\left(\sqrt{TK\left(d+\frac{\ln T}{\lambda_{\min}(\mathbf{\Sigma})}\right)\ln(T)\ln(K)}\right)$	✓	Unknown

The above model of K-armed linear contextual bandits has also been studied in a stochastic setting—see, e.g., [Bastani et al., 2021]. By reducing K-armed linear contextual bandits to linear bandits, and applying the gap-dependent bound of OFUL [Abbasi-Yadkori et al., 2011], one can show a regret bound of the order of $\frac{dK}{\Delta_{\min}}\ln(T)$ for the stochastic setting, ignoring logarithmic factors in K and d, where $\Delta_{\min}$ is the minimum sub-optimality gap over the context space.

In this work, we address the problem of obtaining *best-of-both-worlds* (BoBW) bounds for K-armed linear contextual bandits: namely, the problem of designing algorithms simultaneously achieving good regret bounds in both the adversarial and stochastic regimes without any prior knowledge about the environment. Starting from the seminal work of Bubeck and Slivkins [2012], Seldin and Slivkins [2014] for K-armed bandits, there is a growing interest in BoBW results [Seldin and Lugosi, 2017, Wei and Luo, 2018, Zimmert and Seldin, 2021]. Various bounds have been established for different models, including linear bandits [Lee et al., 2021, Kong et al., 2023, Ito and Takemura, 2023b,a], contextual bandits [Pacchiano et al., 2022, Dann et al., 2023], K-armed bandits with feedback graphs [Ito et al., 2022, Rouyer et al., 2022], combinatorial semi-bandits [Zimmert et al., 2019, Ito, 2021], episodic MDPs [Jin et al., 2021], to name a few. However, known BoBW results for contextual bandits are not satisfying. The algorithm of Dann et al. [2023] essentially relies on Exp4, which is computationally expensive when the class of policies is large. In this paper, we devise the first BoBW algorithms for K-armed linear contextual bandits that, among other advantages, can be implemented in time polynomial in d and K. Next, we list the main contributions of this work.

Contributions. We introduce the first BoBW algorithms for K-armed linear contextual bandits. In the stochastic regime, our algorithms achieve the (poly)logarithmic rate $\frac{(dK)^2\text{polyln}(dKT)}{\Delta_{\min}}$. In the adversarial regime, we obtain either a first-order $\widetilde{\mathcal{O}}(dK\sqrt{L^*})$ bound, or a second order $\widetilde{\mathcal{O}}(dK\sqrt{\Lambda^*})$ bound (Theorem 1 and Corollary 1). We also propose a simpler and more efficient algorithm based on the follow-the-regularized-leader (FTRL) framework, that simultaneously achieves polylogarithmic regret in the stochastic regime and $\widetilde{\mathcal{O}}(dK\sqrt{T})$ regret in the adversarial regime (Theorem 2), without prior knowledge of the inverse of the contextual covariance matrix $\mathbf{\Sigma}$. Our proposed algorithms are also applicable to the corrupted stochastic regime.

Techniques. Our data-dependent bounds are based on the black-box framework proposed by Dann et al. [2023], who provide a unified algorithm transforming a bandit algorithm for the adversarial regime into a BoBW algorithm. Directly adapting to our setting the results for contextual bandits with finite policy classes in their work involves a prohibitive computational cost, since it is known that the number of policies to consider in the adversarial regime is of order $\left(TK^{-2}d^{-1}\right)^{Kd}$ [Allen-Zhu et al., 2018, Olkhovskaya et al., 2023]. Within the same framework, we may also apply the Exp3-type algorithm of Neu and Olkhovskaya [2020]. However, this only results in zero-order (i.e., not data-dependent) regret bounds $\mathcal{O}(\sqrt{T})$—see Proposition 8 in Appendix E.3. In order to obtain data-dependent guarantees, we instead apply the continuous exponential weights algorithm for adversarial linear contextual bandits recently investigated by Olkhovskaya et al. [2023]. In particular, we show that it is possible to choose the learning rates so as to fulfill the data-dependent stability condition required in Dann et al. [2023] for applying their black-box framework.

The data-dependent bounds achieved by the black-box approach are favorable in the sense that the algorithm performs well when there is an action achieving a small cumulative loss or the loss has a small variance. However, this approach may have limitations as it requires knowledge of the inverse of the covariance matrix $\mathbf{\Sigma}^{-1}$ and may not be practical to implement. To overcome this issue, we show how FTRL with Shannon entropy regularization—which is a much more practical algorithm—can be run with an estimate of $\mathbf{\Sigma}^{-1}$ computed using *Matrix Geometric Resampling* (MGR) of Neu and Bartók [2013], Neu and Bartók [2016], thus avoiding the advance knowledge of $\mathbf{\Sigma}^{-1}$. In order to construct this algorithm, we rely on an adaptive learning rate framework for obtaining BoBW guarantees in FTRL with Shannon entropy regularization, proposed in Ito et al. [2022] and later used in Tsuchiya et al. [2023a,b], Kong et al. [2023]. The difference from their work is that while they crucially rely on the unbiasedness of the loss estimator, we need to deal with the *biased* loss estimator that comes from the use of the covariance matrix estimation in MGR. Neu and Olkhovskaya [2020] and Zierahn et al. [2023] applied FTRL+MGR, which allows controlling the bias of the loss estimator, but they focused only on the adversarial regime. Moreover, their methods only attain a sub-optimal regret bound $\mathcal{O}(\sqrt{T})$ in the stochastic regime. The derivation of our bounds for K-armed linear contextual bandits requires nontrivial scheduling of the learning rates and of the adaptive mixing rates of exploration. With these techniques, we successfully provide the first BoBW bounds for K-armed linear contextual bandits without knowing $\mathbf{\Sigma}^{-1}$.

Table 1 summarizes our results in the context of the previous literature. The upper bound of Zierahn et al. [2023] is for a combinatorial contextual setting where the action space satisfies $\mathcal{A} \subseteq \{0,1\}^K$ and we assume $\max_{a\in\mathcal{A}} \|a\|_1 \leq 1$. The best known lower bound for the adversarial or distribution-free setting is $\Omega\left(\sqrt{dKT}\right)$ also due to Zierahn et al. [2023], see Appendix C.

Related work. Despite the vast literature on contextual bandits [Chu et al., 2011, Syrgkanis et al., 2016, Rakhlin and Sridharan, 2016, Zhao et al., 2021, Ding et al., 2022, He et al., 2022], only a few data-dependent bounds have been proven since the question was posed by Agarwal et al. [2017a]. The first result is by Allen-Zhu et al. [2018], but the algorithm is not applicable to a large class of policies. Foster and Krishnamurthy [2021] obtained first-order bounds for stochastic losses via an efficient regression-based algorithm. Recently Olkhovskaya et al. [2023] proved first- and second-order bounds for stochastic contexts but adversarial losses. Yet, BoBW bounds are not addressed in these studies. There are some BoBW results in the model selection problem [Pacchiano et al., 2020, 2022, Agarwal et al., 2017b, Cutkosky et al., 2021, Lee et al., 2021, Wei et al., 2022]. In particular, Pacchiano et al. [2022] achieved the first BoBW high-probability regret bound for general contextual linear bandits. However, the algorithm achieving this result has a running time linear in the number of policies, which makes it intractable for infinite policy classes. A more detailed review of related works can be found in Appendix B.

2 Problem Statement

Given a K-action set $[K] := \{1, 2, \ldots, K\}$, a context space of a full-dimensional compact set $\mathcal{X} \subseteq \mathbb{R}^d$, and a distribution $\mathcal{D}$ over $\mathcal{X}$, our learning protocol can be described as follows. At each time step $t = 1, 2, \ldots, T$:

- For each action $a \in [K]$, the environment chooses a loss vector $\boldsymbol{\theta}_{t,a} \in \mathbb{R}^d$
- Independently of the choice of loss vectors $\boldsymbol{\theta}_{t,a}$ for $a \in [K]$, the environment draws the context vector $X_t \in \mathcal{X}$ from the context distribution $\mathcal{D}$ unknown to the learner
- The learner observes context X_t and chooses action $A_t \in [K]$
- The learner incurs and observes the loss $\ell_t(X_t, A_t)$.

Assumptions. Like previous works on adversarial linear contextual bandits [Neu and Olkhovskaya, 2020, Olkhovskaya et al., 2023, Zierahn et al., 2023] and linear bandits [Lee et al., 2021, Dann et al., 2023], we make the following assumptions:

- The distribution $\mathcal{D}$ from which contexts X are independently drawn satisfies $\mathbb{E}[XX^\top] = \mathbf{\Sigma} \succ 0$;
- $\|X\|_2 \le 1$ $\mathcal{D}$-almost surely;
- $\|\boldsymbol{\theta}_{t,a}\|_2 \le 1$ for all $a \in [K]$ and $t \in [T]$;
- $\ell_t(\boldsymbol{x}, a) \in [-1, 1]$ for all $\boldsymbol{x} \in \mathcal{X}$, $a \in [K]$, and $t \in [T]$.

Further conditions on the loss functions $\ell_t(\boldsymbol{x}, a)$ as well as the loss vectors $\boldsymbol{\theta}_{t,a}$ for each $a \in [K]$ and t are defined in each regime as follows.

Adversarial regime. The loss function is defined by $\ell_t(X_t, a) := \langle X_t, \boldsymbol{\theta}_{t,a}\rangle$, where $\boldsymbol{\theta}_{t,a}$ is chosen by an oblivious adversary for all a and t.

Stochastic regime. The loss function is defined by $\ell_t(X_t, a) := \langle X_t, \boldsymbol{\theta}_a\rangle + \varepsilon_t(X_t, a)$, where $\boldsymbol{\theta}_a$ for each action a is fixed and unknown, and $\varepsilon_t(X_t, a)$ is independent and bounded zero-mean noise.

Corrupted stochastic regime. The loss function is defined by $\ell_t(X_t, a) := \langle X_t, \boldsymbol{\theta}_{t,a}\rangle + \varepsilon_t(X_t, a)$, where $\varepsilon_t(X_t, a)$ is independent and bounded zero-mean noise and the vectors $\boldsymbol{\theta}_{t,a}$ are such that there exist fixed and unknown vectors $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K$ and an unknown constant $C > 0$ for which $\sum_{t=1}^T \max_{a\in[K]} \|\boldsymbol{\theta}_{t,a} - \boldsymbol{\theta}_a\|_2 \le C$ holds. Note that $C = 0$ corresponds to the stochastic regime and $C = \Theta(T)$ corresponds to the adversarial regime with additional zero-mean noise.

Let Π be the set of all deterministic policies $\pi : \mathcal{X} \to [K]$ mapping contexts to actions. We define $\pi^* \in \Pi$ as the optimal policy:

$$\pi^*(\boldsymbol{x}) := \operatorname*{arg\,min}_{a\in[K]} \mathbb{E}\left[\sum_{t=1}^T \ell_t(\boldsymbol{x}, a)\right] \quad \forall \boldsymbol{x} \in \mathcal{X}, \tag{1}$$

where the expectation is taken over the randomness by loss functions. Then, the learner's goal is to minimize the total expected regret against the optimal policy π^*:

$$R_T = \mathbb{E}\left[\sum_{t=1}^T \Big(\ell_t(X_t, A_t) - \ell_t(X_t, \pi^*(X_t))\Big)\right], \tag{2}$$

where the expectation is taken over the learner's randomness as well as the sequence of random contexts and loss functions.

In the (corrupted) stochastic regime, given $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K$, let $\Delta_{\min}(\boldsymbol{x}) := \min_{a\ne\pi^*(\boldsymbol{x})}\langle \boldsymbol{x}, \boldsymbol{\theta}_a - \boldsymbol{\theta}_{\pi^*(\boldsymbol{x})}\rangle$ for all $\boldsymbol{x} \in \mathcal{X}$. Then, we define the minimum sub-optimality gap by $\Delta_{\min} := \min_{\boldsymbol{x}\in\mathcal{X}} \Delta_{\min}(\boldsymbol{x}) > 0$.

We denote the cumulative loss incurred by the optimal policy by $L^* := \mathbb{E}\big[\sum_{t=1}^T \ell_t(X_t, \pi^*(X_t))\big]$ and the cumulative variance of a policy choosing actions $A_1, A_2, \dots$ with respect to a predictable loss sequence $\boldsymbol{m}_{t,a} \in \mathbb{R}^d$ for action a by $\Lambda^* := \mathbb{E}\big[\sum_{t=1}^T (\ell_t(X_t, A_t) - \langle X_t, \boldsymbol{m}_{t,A_t}\rangle)^2\big]$. We use $\overline{\Lambda} := \mathbb{E}\big[\sum_{t=1}^T (\ell_t(X_t, A_t) - \langle X_t, \overline{\boldsymbol{\theta}}\rangle)^2\big]$ with $\overline{\boldsymbol{\theta}} := \frac{1}{TK}\sum_{t=1}^T\sum_{a=1}^K \boldsymbol{\theta}_{t,a}$.

Additional notation. We denote by $\mathbb{E}_X[\cdot]$ the expectation over a random variable (r.v.) X. We denote by $\mathbb{E}_X[\cdot|Y]$ the expectation over X conditioned on Y. When we write $\mathbb{E}[X] \cdot \mathbb{E}[X|Y]$, we take the expectation conditioned on Y with respect to all sources of randomness in X. We denote by $\mathcal{F}_t = \sigma(X_s, A_s, \forall s \le t)$ the filtration generated by all the random variables X_s and the set of actions A_s, for each $s \le t$. Then we write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_{t-1}]$. For any semi-definite matrix $\mathbf{B} \in \mathbb{R}^{d\times d}$, we use $\lambda_{\min}(\mathbf{B})$ to denote the smallest eigenvalue of $\mathbf{B}$, and write $\|\boldsymbol{u}\|_{\mathbf{B}} = \sqrt{\boldsymbol{u}^\top \mathbf{B}\boldsymbol{u}}$ for $\boldsymbol{u} \in \mathbb{R}^d$. We also define the *probabilistic policy* mapping each context $\boldsymbol{x}$ to a probability distribution $\pi(\cdot \mid \boldsymbol{x})$ over $[K]$ (i.e., an element of the simplex $\Delta([K])$). For the analysis of data-dependent bounds, we use the notion $\xi_{t,a} := (\ell_t(X_t, a) - \langle X_t, \boldsymbol{m}_{t,a}\rangle) \in \mathbb{R}$ with a loss predictor $\boldsymbol{m}_{t,a}$ for $t \in [T]$ and $a \in [K]$. We write $\mathbb{1}[\cdot]$ to denote the indicator function.

3 Follow-the-Regularized-Leader

Following the existing BoBW algorithms, we rely on the FTRL framework. Given context X_t, we consider the FTRL predictor in $\Delta([K])$ defined as

$$p_t(\cdot|X_t) \in \mathop{\arg\min}_{r\in\Delta([K])} \left\{ \sum_{s=1}^{t-1} \langle r, \widehat{\boldsymbol{\ell}}_s(X_t)\rangle + \psi_t(r) \right\},$$

where $\widehat{\boldsymbol{\ell}}_s(X_t) := \left(\langle X_t, \widehat{\boldsymbol{\theta}}_{s,1}\rangle, \ldots, \langle X_t, \widehat{\boldsymbol{\theta}}_{s,K}\rangle\right)^\top \in \mathbb{R}^K$, and $\widehat{\boldsymbol{\theta}}_{t,a}$ is an estimator of the linear loss $\boldsymbol{\theta}_{t,a} \in \mathbb{R}^d$. We use the (negative) Shannon entropy $\psi_t(r) = -\frac{H(r)}{\eta_t}$ as the regularizer, where H is the Shannon entropy and $\eta_t > 0$ is a learning rate. It is well known that $p_t(\cdot|X_t)$ is equivalent to the Exp3-type prediction

$$p_t(a|X_t) = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1}\langle X_t, \widehat{\boldsymbol{\theta}}_{s,a}\rangle\right)}{\sum_{b\in[K]} \exp\left(-\eta_t \sum_{s=1}^{t-1}\langle X_t, \widehat{\boldsymbol{\theta}}_{s,b}\rangle\right)}. \tag{3}$$

The learner's policy $\pi_t(\cdot|X_t) \in \Delta([K])$ that selects the next action usually combines $p_t(\cdot|X_t)$ with some exploration strategy to control the variance of the loss estimates.

We next introduce the Optimistic FTRL (OFTRL) framework [Rakhlin and Sridharan, 2013]. In OFTRL, a loss predictor $\boldsymbol{m}_{t,a} \in \mathbb{R}^d$ for each action a is available to the learner at the beginning of each round t. OFTRL can be viewed as adding $\boldsymbol{m}_{t,a}$ to the objective as a guess for the next loss vector. The OFTRL prediction $p_t(\cdot|X_t)$ is then defined as

$$\mathop{\arg\min}_{r\in\Delta([K])} \left\{ \sum_{s=1}^{t-1} \langle r, \widehat{\boldsymbol{\ell}}_s(X_t)\rangle + \langle r, \boldsymbol{m}_t(X_t)\rangle + \psi_t(r) \right\},$$

where $\boldsymbol{m}_t(X_t) := \left(\langle X_t, \boldsymbol{m}_{t,1}\rangle, \ldots, \langle X_t, \boldsymbol{m}_{t,K}\rangle\right) \in \mathbb{R}^K$.

In the following sections, we apply OFTRL in Theorem 1 exploiting the predicted loss $\boldsymbol{m}_t(X_t)$ to achieve first- and second-order regret bounds, and in Theorem 2, we apply FTRL to obtain a worst-case regret bound in the adversarial regime, while guaranteeing the polylogarithmic regret in the stochastic regime.

4 Data-dependent Bounds

In this section, we discuss how the reduction framework is adapted to K-armed linear contextual bandits. We design an algorithm, MWU-LC, that satisfies the data-dependent stability conditions (Proposition 1), so that we can use it as a base algorithm in the black-box reduction of Dann et al. [2023] and obtain the desired BoBW bound for arbitrary $\boldsymbol{m}_{t,a}$ (Theorem 1). By choosing the appropriate loss predictor $\boldsymbol{m}_{t,a}$, we also show how to simultaneously achieve first- and second-order bounds (Corollary 1).

MWU-LC (Algorithm 1) is an instance of OFTRL using a multiplicative weight update. Notably, such an approach has been taken by Ito et al. [2020] for adversarial linear bandits where they use truncated distribution techniques to make an unbiased loss estimator stable. Recently, Olkhovskaya et al. [2023] extended Ito et al. [2020] to the adversarial K-armed linear contextual bandits. MWU-LC is built upon the algorithm of Olkhovskaya et al. [2023], but in a setting where a loss is observed with some probability q_t. The design of the learning rate is significantly modified in order to achieve BoBW bounds. In particular, we show that MWU-LC achieves a stability condition called *data-dependent importance-weighting stability* (see Definition 4 in Appendix E).

Additional assumptions. If the density function $h : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ has a convex support and $\ln(h(y))$ for $y \in \mathbb{R}^d$ is a concave function on the support, we call the distribution *log-concave*. As in Olkhovskaya et al. [2023], we assume that (i) context distribution $\mathcal{D}$ is log-concave and its support is known to the learner, and (ii) the learner has access to $\boldsymbol{\Sigma}^{-1}$, the inverse of the covariance matrix of contexts. However, these assumptions will be both dropped in Section 5. We assume that loss predictors satisfy $\langle X_t, \boldsymbol{m}_{t,a}\rangle \in [-1, 1]$ for all t and $a \in [K]$. Finally, when we discuss first-order regret bounds, we assume $0 \leq \ell_t(X_t, a) \leq 1$ for all t and $a \in [K]$, which is a standard assumption to ensure that $L^* = \mathbb{E}\left[\sum_{t=1}^T \ell_t(X_t, \pi^*(X_t))\right] \geq 0$.

Continuous MWU method. The learner has access to a loss predictor $\boldsymbol{m}_{t,a} \in \mathbb{R}^d$ for each action a at round t, also called the hint vector. The learner computes the density $p_t(\cdot|X_t)$ supported on $\Delta([K])$ and based on the continuous exponential weights $w_t(\cdot|X_t)$:

$$w_t(r|X_t) := \exp\left(-\eta_t\left(\sum_{s=1}^{t-1}\langle r, \widehat{\boldsymbol{\ell}}_s(X_t)\rangle + \langle r, \boldsymbol{m}_t(X_t)\rangle\right)\right),$$
$$p_t(r|X_t) := \frac{w_t(r|X_t)}{\int_{\Delta([K])} w_t(y|X_t)\ dy}, \tag{4}$$

where $r \in \Delta([K])$, $\eta_t > 0$ is a learning rate, and $\widehat{\boldsymbol{\theta}}_{s,a}$ is the unbiased estimate for the loss vectors $\boldsymbol{\theta}_{s,a}$, which will be determined later.

For the rejection sampling step in Line 5-6, we use the following covariance matrix $\overline{\boldsymbol{\Sigma}}_{t,a} \in \mathbb{R}^{d\times d}$:

$$\overline{\boldsymbol{\Sigma}}_{t,a} := \mathbb{E}_{X,y_t\sim p_t(\cdot|X)}\left[y_t(a)^2 XX^\top|\mathcal{F}_{t-1}\right] \ . \tag{5}$$

The number of steps required for the rejection sampling is $\mathcal{O}(1)$, which can be shown via the concentration property of the log-concave distribution (e.g., Lemma 1 of Ito et al. [2020]) and the log-concavity of $\mathcal{D}$. The truncated distribution $\widetilde{p}_t(\cdot|X_t)$ of $p_t(\cdot|X_t)$ is defined as:

$$\widetilde{p}_t(r|X_t) := \frac{p_t(r|X_t)\mathbb{1}\left[\sum_{a=1}^K r_a^2\|X_t\|^2_{\overline{\boldsymbol{\Sigma}}^{-1}_{t,a}} \le dK\widetilde{\gamma}_t^2\right]}{\mathbb{P}_{y\sim p_t(\cdot|X_t)}\left[\sum_{a=1}^K y_a^2\|X_t\|^2_{\overline{\boldsymbol{\Sigma}}^{-1}_{t,a}} \le dK\widetilde{\gamma}_t^2\right]}$$

for $r \in \Delta([K])$, where $\widetilde{\gamma}_t > 1$ is the truncation level to be specified soon. Thus, $Q_t \in \Delta([K])$ is sampled from the truncated distribution $\widetilde{p}_t(\cdot|X_t)$ and the learner chooses an action $A_t \sim Q_t$. The probability that the learner can observe a loss, $q_t \in (0,1]$ (calculated in Algorithm 5 in Appendix E), is given to the base algorithm in the reduction framework. If the learner observes a loss, then upd_t is set to 1, otherwise upd_t is set to 0. Then MWU-LC constructs an unbiased estimator $\widehat{\boldsymbol{\theta}}_{t,a}$ of $\boldsymbol{\theta}_{t,a}$ for each $a \in [K]$ as follows:

$$\widehat{\boldsymbol{\theta}}_{t,a:} = \boldsymbol{m}_{t,a} + \frac{\mathsf{upd}_t}{q_t} Q_t(a)\widetilde{\boldsymbol{\Sigma}}^{-1}_{t,a} X_t \xi_{t,a}\mathbb{1}[A_t = a], \tag{6}$$

where $\xi_{t,a} = (\ell_t(X_t,a) - \langle X_t, \boldsymbol{m}_{t,a}\rangle)$ and $\widetilde{\boldsymbol{\Sigma}}_{t,a} \in \mathbb{R}^{d\times d}$ is given by:

$$\widetilde{\boldsymbol{\Sigma}}_{t,a} := \mathbb{E}_X\left[Q_t(a)^2 XX^\top|\mathcal{F}_{t-1}\right] \ . \tag{7}$$

For MWU-LC with update probability q_t, we design a novel update rule for the learning rate $\eta_t > 0$ as follows:

$$\eta_t := \left(\frac{800dK\widetilde{\gamma}_t^2}{\min_{j\le t} q_j} + \sum_{j=1}^{t-1}\frac{\beta_j}{q_j}\right)^{-\frac{1}{2}}, \tag{8}$$

where we set $\beta_t := 16\widetilde{\gamma}_t^2\xi_{t,A_t}^2$ and $\widetilde{\gamma}_t := 4\ln(10dKt)$ for $t \in [T]$.

Theoretical results. The following proposition implies that MWU-LC satisfies the data-dependent importance-weighting stability. The proof is provided in Appendix F.

Proposition 1. *Assume that $\overline{\boldsymbol{\Sigma}}_{t,a}$ in (5) and $\overline{\boldsymbol{\Sigma}}_{t,a}$ in (7) are known to the learner at each round t and action a. Given an adaptive sequence of weights $q_1, q_2, \ldots \in (0,1]$, suppose that MWU-LC observes the feedback in round t with probability q_t. Let $R(\tau, a^*) = \mathbb{E}\left[\sum_{t=1}^{\tau}\ell_t(X_t,A_t) - \ell_t(X_t,a^*)\right]$ for round $\tau \in [1,T]$ and comparator action $a^* \in [K]$. Let $\kappa(d,K,T) = 32Kd\ln(10dK\tau)\ln(\tau)$. Then, for any τ and a^*, the regret $R(\tau,a^*)$ of MWU-LC is bounded by*

$$\kappa(d,K,T)\left(\sqrt{\mathbb{E}\left[\sum_{t=1}^{\tau}\frac{\mathsf{upd}_t\xi_{t,A_t}^2}{q_t^2}\right]} + \mathbb{E}\left[\frac{\sqrt{50dK}}{\min_{j\le\tau} q_j}\right]\right).$$

Algorithm 1: Continuous MWU (MWU-LC)

Input : Set of K arms

Receive update probability q_t;
for $t = 1, 2, \ldots, T$ **do**
    Observe X_t;
    do
        Draw $Q \sim p_t(\cdot|X_t)$ defined in (4)
    while $\sum_{a=1}^{K} Q(a)^2 \|X_t\|^2_{\overline{\mathbf{\Sigma}}^{-1}_{t,a}} \le dK\widetilde{\gamma}_t^2$;
    $Q_t \leftarrow Q \in \Delta([K])$;
    Choose an action $A_t \sim Q_t$;
    With probability q_t, observe the loss $\ell_t(X_t, A_t)$ as a feedback;
    Compute $\widehat{\boldsymbol{\theta}}_{t,a}$ for $a \in [K]$ as in (6);
    Update $p_t(\cdot|X_t)$ as in (4);
    Update η_t as in (8);

Owing to Proposition 1, if MWU-LC is run with the black-box reduction procedure (Algorithms 3 and 5 in Appendix E) as a base algorithm, we obtain the following BoBW guarantee.

Theorem 1. *Assume that $\overline{\mathbf{\Sigma}}_{t,a}$ in (5) and $\overline{\mathbf{\Sigma}}_{t,a}$ in (7) are known to the learner at each round t and action a. Let $\kappa_1(d,K,T) = K^2 d^2 \ln^2(dKT)\ln^2(T)$ and $\kappa_2(d,K,T) = (dK)^{3/2}\ln(dKT)\ln(T)$ be problem-dependent constants. Combining the base algorithm MWU-LC (Algorithm 1) with Algorithms 3 and 5, it holds that*

$$R_T = \mathcal{O}\left(\sqrt{\kappa_1(d,K,T)\Lambda^* \ln^2 T} + \kappa_2(d,K,T)\ln^2(T)\right)$$

in the adversarial regime and

$$R_T = \mathcal{O}\left(\frac{\kappa_1(d,K,T)\ln(T)}{\Delta_{\min}} + \sqrt{\frac{\kappa_1(d,K,T)\ln TC}{\Delta_{\min}}} + \kappa_2(d,K,T)\ln(T)\ln\frac{C}{\Delta_{\min}}\right)$$

in the corrupted stochastic regime.

For a concrete choice of $\boldsymbol{m}_{t,a}$ for each $a \in [K]$, which in turn determines Λ^*, we utilize the online optimization method. For any positive semi-definite matrix $\mathbf{S} \in \mathbb{R}^{d\times d}$, define the predictor $\boldsymbol{m}_{t,a}$ as a vector in $\mathcal{M} := \{\boldsymbol{m} \in \mathbb{R}^d \mid \langle \boldsymbol{x}, \boldsymbol{m}\rangle \le 1,\ \forall \boldsymbol{x} \in \mathcal{X}\}$ that minimizes the following expression:

$$\|\boldsymbol{m}\|^2_{\mathbf{S}} + \sum_{j=1}^{t-1} \mathbb{1}[A_j = a]\left(\langle \boldsymbol{\theta}_{j,a} - \boldsymbol{m}, X_j\rangle\right)^2 \tag{9}$$

Based on Ito et al. [2020], we construct $\mathbf{S}$ via the barycentric spanner for $\mathcal{X}$ [Awerbuch and Kleinberg, 2004], which is given by (26) in Appendix F. Then, we show the following corollary using $\mathbf{S}$, which implies that we obtain the regret bound depending on $\sqrt{\min\{L^*, \overline{\Lambda}\}}$, see Section 2 for a definition of $\overline{\Lambda}$.

Corollary 1. *Let $\boldsymbol{m}_{t,a}$ at each $t \in [T]$ and $a \in [K]$ be given by the minimizer of (9). Then, under the same assumptions as Theorem 1 and for any $\boldsymbol{m}^* \in \mathcal{M}$, R_T is bounded by*

$$\widetilde{\mathcal{O}}\left(Kd\sqrt{\min\left\{L^*, \mathbb{E}\Big[\sum_{t=1}^{T}\langle X_t, \boldsymbol{\theta}_{t,A_t} - \boldsymbol{m}^*\rangle^2\Big]\right\}} + K^2 d^2\right)$$

for the adversarial regime, and is the same regret as Theorem 1 for the corrupted stochastic regime.

Algorithm 2: FTRL with Shannon entropy (FTRL-LC)

Input : Arms $[K]$

Initialization: Set $\widetilde{\boldsymbol{\theta}}_{0,a} = \mathbf{0}$ for all $a \in [K]$. Initialize η_1 and γ_1 by (13). Set $M_1 \leftarrow 1$.

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

  Observe X_t;

  Compute $p_t(\cdot|X_t)$ by FTRL in (10) with regularizer $\psi_t(r) = -\frac{1}{\eta_t} H(r)$;

  Set
$$\pi_t(a|X_t) \leftarrow (1-\gamma_t)p_t(a|X_t) + \gamma_t \frac{1}{K}; \tag{11}$$

  Sample an action $A_t \sim \pi_t(\cdot|X_t)$;

  Observe the loss $\ell_t(X_t, A_t)$ and compute $\widetilde{\boldsymbol{\theta}}_{t,a}$ for all $a \in [K]$ using (12);

  Update η_t and γ_t by (13);

  Update M_t by (14);

Remark 1. *Although the first-order bound is obtained by just setting $\boldsymbol{m}_{t,a} = \mathbf{0}$ (see Corollary 2 in Appendix F), computing the minimizer of (9) as $\boldsymbol{m}_{t,a}$ allows a single algorithm to achieve first- and second-order bounds simultaneously. Compared with Olkhovskaya et al. [2023], our results only have an additional factor $\sqrt{d}$ in the adversarial regime while also providing gap-dependent polylogarithmic regret in the (corrupted) stochastic regime.*

We just saw how our first approach in this section achieves theoretical advantages and a polynomial-time running time due to the log-concavity of $\mathcal{D}$. However, removing the prior knowledge of $\boldsymbol{\Sigma}^{-1}$ seems challenging, as computation of (5) and (7) involves expectation depending on both $\mathcal{D}$ and a learner's policy. Moreover, the continuous exponential weights incur a high (yet polynomial) sampling cost, resulting in $\mathcal{O}\big((K^5 + \ln T)g_{\boldsymbol{\Sigma}_t}\big)$ per round running time, where $g_{\boldsymbol{\Sigma}_t}$ is the time to construct the covariance matrix for each round (see Section 3.3 in Olkhovskaya et al. [2023] or Section 4.4 in Ito et al. [2020] for detailed discussion). To address these issues, we next devise a simpler solution using FTRL instead of relying on the reduction framework.

5 Uuknown $\boldsymbol{\Sigma}^{-1}$ Case

We present a computationally efficient algorithm, called FTRL-LC, based on FTRL with negative Shannon entropy. This algorithm does not require knowledge of $\boldsymbol{\Sigma}^{-1}$, and only needs access to context distribution $\mathcal{D}$ and minimum eigenvalue $\lambda_{\min}(\boldsymbol{\Sigma})$.

Proposed method. Recall that, given context X_t, FTRL computes the probability vector $p_t(\cdot|X_t) \in \Delta([K])$ as follows:

$$p_t(\cdot|X_t) := \mathop{\arg\min}_{r \in \Delta([K])} \left\{ \sum_{s=1}^{t-1} \langle r, \widetilde{\boldsymbol{\ell}}_s(X_t) \rangle + \psi_t(r) \right\}, \tag{10}$$

where $\psi_t \colon \Delta([K]) \to \mathbb{R}$ is the convex regularizer, $\widetilde{\boldsymbol{\ell}}_s(X_t) := (\langle X_t, \widetilde{\boldsymbol{\theta}}_{s,1}\rangle, \ldots, \langle X_t, \widetilde{\boldsymbol{\theta}}_{s,K}\rangle) \in \mathbb{R}^K$, and $\widetilde{\boldsymbol{\theta}}_{s,a} \in \mathbb{R}^d$ is the (possibly biased) estimator for $\boldsymbol{\theta}_{s,a}$. Then, the policy $\pi_t(\cdot|X_t)$ that selects the action A_t is defined by mixing the probability vector $p_t(\cdot|X_t)$ with uniform exploration, where the adaptive mixture rate $\gamma_t \in [0, 1/2]$ is defined later in (13). For the regularizer in (10), we use the (negative) Shannon entropy $\psi_t(r) = -\frac{1}{\eta_t} H(r)$ as introduced in Section 3, where the learning rate $\eta_t > 0$ will be specified later. The pseudo-code of FTRL-LC is given in Algorithm 2.

Loss estimation. Here we describe the method for estimating $\boldsymbol{\theta}_{t,a}$. Given the covariance matrix $\boldsymbol{\Sigma}_{t,a} := \mathbb{E}_t[\mathbb{1}[A_t = a] X_t X_t^\top]$, it is known that we can construct the unbiased estimator $\widehat{\boldsymbol{\theta}}_{t,a}$ defined by

$$\widehat{\boldsymbol{\theta}}_{t,a} := \boldsymbol{\Sigma}_{t,a}^{-1} X_t \ell_t(X_t, A_t) \mathbb{1}[A_t = a], \quad \forall a \in [K].$$

While this estimate is unbiased, $\mathbb{E}_t[\widehat{\boldsymbol{\theta}}_{t,a}] = \boldsymbol{\theta}_{t,a}$, computing this estimator is computationally inefficient as its construction requires computing the inverse of the $d \times d$ covariance matrix $\boldsymbol{\Sigma}_{t,a}$. Such a heavy computation requiring time equal to $\mathcal{O}(d^3)$ is prohibitive when $d \gg 1$. Furthermore, this estimation approach assumes that the covariance matrix is known in advance, which is not the case in most real-world scenarios.

To avoid such practical problems, we consider relying on the approach of *Matrix Geometric Resampling* (MGR) developed by Neu and Bartók [2013], Neu and Bartók [2016] and later used in Neu and Olkhovskaya [2020], Zierahn et al. [2023]. The MGR procedure, detailed in Appendix G.1, has $M_t > 0$ iterations and outputs $\widehat{\boldsymbol{\Sigma}}^{+}_{t,a}$ as the estimate of $\boldsymbol{\Sigma}^{-1}_{t,a}$. MGR can be implemented in $\mathcal{O}(M_t K d + K d^2)$ time [Neu and Olkhovskaya, 2020]. Using $\widehat{\boldsymbol{\Sigma}}^{+}_{t,a}$, we can define the estimator of $\boldsymbol{\theta}_{t,a}$ by

$$\widetilde{\boldsymbol{\theta}}_{t,a} := \widehat{\boldsymbol{\Sigma}}^{+}_{t,a} X_t \ell_t(X_t, A_t) \mathbb{1}[A_t = a], \quad \forall a \in [K]. \tag{12}$$

However, $\widehat{\boldsymbol{\Sigma}}^{+}_{t,a}$ is biased in general when $M_t > 0$ is finite, implying that the estimator $\widetilde{\boldsymbol{\theta}}_{t,a}$ in (12) may be biased (although $\mathbb{E}_t[\widehat{\boldsymbol{\Sigma}}^{+}_{t,a}] = \boldsymbol{\Sigma}^{-1}_{t,a}$ when $M_t \to \infty$). This biasedness needs to be handled when designing the learning rate $(\eta_t)_t$ for FTRL.

Learning rate. To achieve BoBW guarantees while dealing with a biased estimator, we need to design a learning rate η_t and a mixture rate γ_t achieving $\mathcal{O}(\sqrt{T})$ regret in the adversarial regime and $\mathcal{O}(\mathrm{poly}(\ln T))$ regret in the stochastic regime. To achieve this goal, we define the learning rate and mixture rate as follows:

$$\begin{aligned}
\beta'_{t+1} &= \beta'_t + \frac{c'_1}{\sqrt{1 + (\ln K)^{-1} \sum_{s=1}^{t} H(p_s(\cdot|X_s))}}\,, \\
\beta_t &= \max\left\{2, c'_2 \ln T, \beta'_t\right\}\,, \\
\eta_t &= \frac{1}{\beta_t}\,,\ \gamma_t = \alpha_t \cdot \eta_t,\ \alpha_t = \frac{4K\ln(t)}{\lambda_{\min}(\boldsymbol{\Sigma})}\,,
\end{aligned} \tag{13}$$

where $c'_1 = \sqrt{\left(3Kd + \frac{2K\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)\frac{\ln T}{\ln K}}$, $c'_2 = \frac{8K}{\lambda_{\min}(\boldsymbol{\Sigma})}$, and we set $\beta'_1 = c'_1 \geq 1$. These definitions ensure $0 \leq \gamma_t \leq 1/2$ and $0 < \eta_t \leq 1/2$.

Unlike the existing algorithms, which are designed for the adversarial regime and use a fixed number of iterations of MGR (i.e., $M_t = M$ for some $M > 0$ at all $t \in [T]$ [Neu and Olkhovskaya, 2020, Zierahn et al., 2023]), determining M_t adaptively is also crucial to prove BoBW guarantees. We set M_t at round $t > 1$ to

$$M_t = \left\lceil \frac{4K}{\gamma_t \lambda_{\min}(\boldsymbol{\Sigma})} \ln(t) \right\rceil \ (\geq 1). \tag{14}$$

Theoretical results. Here, we formally state the main result and sketch a summary of the key analysis to guarantee the regret upper bound. The complete proof of Theorem 2 and the following lemmas can be found in Appendix G.

Theorem 2. *Let $c_4 = \mathcal{O}\big(\frac{K\ln(K)}{\lambda_{\min}(\boldsymbol{\Sigma})}\ln(T)\big)$ be a problem-dependent constant. The regret R_T of FTRL-LC (Algorithm 2) for the adversarial regime is bounded by*

$$R_T = \mathcal{O}\left(\sqrt{T\left(d + \frac{\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})}\right) K \ln(K)\ln(T)} + c_4\right).$$

For the stochastic regime, the regret is bounded by

$$R_T = \mathcal{O}\left(\frac{K}{\Delta_{\min}}\left(d + \frac{\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)\ln(KT)\ln T\right) =: R_T^{\mathrm{sto}},$$

and for the corrupted stochastic regime, the regret is bounded by

$$R_T = \mathcal{O}\left(R_T^{\mathrm{sto}} + \sqrt{C R_T^{\mathrm{sto}}}\right).$$

Our bound achieves $\widetilde{\mathcal{O}}\left(\sqrt{TK\max\left\{d, \frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})}\right\}}\right)$ recovering the best-known result in the adversarial regime [Neu and Olkhovskaya, 2020, Zierahn et al., 2023] up to log-factors when $T \geq \frac{K^2}{\lambda_{\min}(\boldsymbol{\Sigma})^2}$ and has a performance comparable to $\frac{dK}{\Delta_{\min}}\ln(T)$ in the stochastic regime. In the corrupted stochastic regime, we have the desired dependence of $\sqrt{C}$ for the corruption level $C > 0$.

Regret analysis. For the sake of simplicity, in our analysis we introduce a variant of our bandit problem that we call *auxiliary game*, where the context vector $\boldsymbol{x} \in \mathcal{X}$ does not change over time, and for each trial $t \in [T]$ the incurred loss is obtained replacing $\boldsymbol{\theta}_{t,a}$ by a (possibly biased) loss vector estimator $\widetilde{\boldsymbol{\theta}}_{t,a}$ as follows. Let $\widetilde{\boldsymbol{\theta}}_{t,a} \in \mathbb{R}^d$ be an estimator of the loss vector $\boldsymbol{\theta}_{t,a}$ with bias $\boldsymbol{b}_{t,a} \in \mathbb{R}^d$ and $a \in [K]$. Suppose that the learner's action A_t is selected by a probabilistic policy $\pi_t(\cdot|\boldsymbol{x}) \in \Delta([K])$. Then, the regret in the auxiliary game against the comparator $\pi^*(\boldsymbol{x})$ defined in (1) for the estimated loss is defined as

$$\widetilde{R}_T(\boldsymbol{x}) := \mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,A_t}\rangle - \langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,\pi^*(\boldsymbol{x})}\rangle\right]. \tag{15}$$

As in Neu and Olkhovskaya [2020], Olkhovskaya et al. [2023], Zierahn et al. [2023], we define a *ghost sample* $X_0 \sim \mathcal{D}$, which is drawn independently of the entire interaction history, i.e., X_0 is independent of any of $X_1, \ldots, X_t$ used to construct the loss estimators $\widetilde{\boldsymbol{\theta}}_{t,a}$. With this notation, it is known that R_T is bounded as follows (see Eq.(6) in Neu and Olkhovskaya [2020] and Lemma 7 in Appendix D):

$$R_T \leq \mathbb{E}[\widetilde{R}_T(X_0)] + 2\sum_{t=1}^{T} \max_{a\in[K]} \left|\mathbb{E}[\langle X_t, \boldsymbol{b}_{t,a}\rangle]\right|.$$

Thanks to this upper bound, it suffices to bound the regret of the auxiliary game and control the bias. To do so, we start with Lemma 1, which can be proven via the standard analysis of FTRL with Shannon entropy while taking the context into account.

Lemma 1. *Suppose that* $\max_{\boldsymbol{x}\in\mathcal{X}} |\eta_t\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle| \leq 1$ *holds, and* A_t *is chosen by* $\pi_t(\cdot|\boldsymbol{x})$ *defined by* (11) *for* $\boldsymbol{x} \in \mathcal{X}$. *Then, we have*

$$\begin{aligned}\widetilde{R}_T(\boldsymbol{x}) \leq & \sum_{t=1}^{T} (\beta_{t+1} - \beta_t)\, H(p_{t+1}(\cdot|\boldsymbol{x})) + \beta_1 \ln K \\ & + \sum_{t=1}^{T} \eta_t \sum_{a=1}^{K} \pi_t(a|\boldsymbol{x})\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle^2 + U(\boldsymbol{x}),\end{aligned} \tag{16}$$

where $U(\boldsymbol{x}) = \sum_{t=1}^{T} \gamma_t \sum_{a\neq\pi^*(\boldsymbol{x})} \frac{1}{K}\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle$ *is the regret due to the uniform exploration.*

We next state the following lemma, showing that our careful parameter tuning allows us to bound the RHS of (16).

Lemma 2. *Suppose that* $\eta_t \leq \frac{1}{2}$, $\gamma_t = \alpha_t \cdot \eta_t$, *and set* M_t *as in* (14). *Then, it holds that* (i) $|\mathbb{E}_t[\langle X_t, \widetilde{\boldsymbol{\theta}}_{t,a} - \widehat{\boldsymbol{\theta}}_{t,a}\rangle]| \leq \exp\left(-\frac{\gamma_t\lambda_{\min}(\boldsymbol{\Sigma})M_t}{2K}\right) \leq 1/t^2$ *and* (ii) $|\eta_t\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle| \leq 1, \quad \forall \boldsymbol{x} \in \mathcal{X}$.

Thanks to (ii), the requirement of Lemma 1 is met by our parameter tuning. The statement (i) is useful to bound the penalty term caused by the biased $\widetilde{\boldsymbol{\theta}}_{t,a}$, i.e., $\mathbb{E}[U(X_0)]$ and $\sum_{t=1}^{T} \max_{a\in[K]} |\mathbb{E}[\langle X_t, \boldsymbol{b}_{t,a}\rangle]|$.

From Lemma 1, we can derive Lemma 3 providing an upper bound on the expected regret of the auxiliary game dependent on the sum of the Shannon entropy over $[T]$.

Lemma 3 (Entropy-dependent regret bound for the auxiliary game)**.** *Let* $X_0 \sim \mathcal{D}$ *be a ghost sample drawn independently of the entire interaction history. Let* $\kappa = c_1'\sqrt{\ln K} + \frac{\left(3Kd + \frac{2K\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)\ln T}{c_1'\sqrt{\ln K}}$. *If* A_t *is chosen by* $\pi_t(\cdot|X_0)$ *defined by* (11) *for* X_0, *then, the expected regret of the auxiliary game* $\mathbb{E}[\widetilde{R}_T(X_0)]$ *is bounded by*

$$\mathcal{O}\left(\kappa\sqrt{\mathbb{E}\left[\sum_{t=1}^{T} H(p_t(\cdot|X_0))\right]} + \frac{K\ln K}{\lambda_{\min}(\boldsymbol{\Sigma})}\ln T\right).$$

We introduce the following notation for the further analysis: Let $\varrho_0(\pi^*) := \sum_{t=1}^{T}(1-p_t(\pi^*(X_0)|X_0))$ and $\varrho_{(X_t)_{t=1}^T}(\pi^*) := \sum_{t=1}^{T}(1-p_t(\pi^*(X_t)|X_t))$. Now, we are ready to sketch the proof of Theorem 2.

Proof Sketch of Theorem 2. For the adversarial regime, by the fact that $H(p_t(\cdot|X_0)) \leq \ln K$, we immediately have the desired regret bound from the above lemmas. To analyze the corrupted stochastic regime we start with a lower bound on the regret. We can show that $R_T \geq \frac{\Delta_{\min}}{2}\mathbb{E}[\varrho_{(X_t)_{t=1}^T}(\pi^*)]-2C$ from the definition of the stochastic regime with adversarial corruption (Lemma 21 in Appendix G). For the upper bound depending on $\varrho_0(\pi^*)$, we use the inequality of $\sum_{t=1}^{T} H(p_t(\cdot|X_0)) \leq \varrho_0(\pi^*)\ln\frac{\mathrm{e}KT}{\varrho_0(\pi^*)}$ (Lemma 22 in Appendix G). When $\varrho_0(\pi^*) < \mathrm{e}$, then we have the desired bound trivially from this inequality. In the case of $\varrho_0(\pi^*) \geq \mathrm{e}$, using $\mathbb{E}\left[\sum_{t=1}^{T} H(p_t(\cdot|X_0))\right] \leq \mathbb{E}[\varrho_0(\pi^*)]\ln(KT)$, we have $R_T = \widetilde{\mathcal{O}}(\mathrm{poly}(\ln T)\cdot\mathbb{E}[\varrho_0(\pi^*)]+c_4)$, where c_4 is a problem-dependent constant. Here, we use the fact that X_0 and X_t follows the same distribution to show $\mathbb{E}[\varrho_{(X_t)_{t=1}^T}(\pi^*)] = \mathbb{E}[\varrho_0(\pi^*)]$ (Lemma 20 in Appendix G). Then, the final part can be done via standard self-bounding techniques. Plugging the above upper and lower bound on R_T into $R_T = (1+\lambda)R_T - \lambda R_T$ for $\lambda \in (0,1]$, taking the worst-case with respect to $\mathbb{E}[\varrho_0(\pi^*)]$, and then optimizing $\lambda \in (0,1]$ completes the proof for the corrupted stochastic regime. □

6 Conclusion

We proposed algorithms for K-armed linear contextual bandits to achieve the BoBW guarantees. The first approach is to use a continuous MWU method with a reduction framework, thereby attaining either first- or second-order regret bounds in the adversarial regime and polylogarithmic regret in the (corrupted) stochastic regime. We also designed a simpler FTRL with Shannon entropy that does not require the knowledge $\mathbf{\Sigma}^{-1}$ and achieves the worst-case regret in the adversarial regime without sacrificing the polylogarithmic regret in the (corrupted) stochastic regime.

For future directions, it is important to develop a computationally efficient algorithm that can achieve data-dependent bounds without relying on knowledge of $\mathbf{\Sigma}^{-1}$. Even without this knowledge, the FTRL-LC algorithm achieved the optimal worst-case regret up to log factors in the adversarial regime. However, in the stochastic regime, additional $\ln(T)$ and $\ln(KT)$ terms arise due to MGR and Shannon entropy, respectively. An additional log factor is also common when using FTRL with Shannon entropy in other bandit settings. Therefore, it would be interesting to explore alternative regularizers. Another direction is to extend the current results to the contextual combinatorial bandit setting.

References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 2312–2320, 2011.

N. Abe and P. M. Long. Associative reinforcement learning using linear probabilistic concepts. In *Proc. of International Conference on Machine Learning (ICML)*, pages 3–11, 1999.

A. Agarwal, A. Krishnamurthy, J. Langford, H. Luo, and S. R. E. Open problem: First-order regret bounds for contextual bandits. In *Proc. of Annual Conference on Learning Theory (COLT)*, pages 4–7, 2017a.

A. Agarwal, H. Luo, B. Neyshabur, and R. E. Schapire. Corralling a band of bandit algorithms. In *Proc. of Annual Conference on Learning Theory (COLT)*, pages 12–38, 2017b.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proc. of International Conference on Machine Learning (ICML)*, pages 127–135, 2013.

Z. Allen-Zhu, S. Bubeck, and Y. Li. Make the minority great again: First-order regret bound for contextual bandits. In *Proc. of International Conference on Machine Learning (ICML)*, pages 186–194, 2018.

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

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

B. Awerbuch and R. D. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proc. of ACM Symp. on Theory of Computing (STOC)*, page 45–53, 2004.

H. Bastani, M. Bayati, and K. Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.

I. Bogunovic, A. Krause, and J. Scarlett. Corruption-tolerant gaussian process bandit optimization. In *Proc. of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1071–1081, 2020.

I. Bogunovic, A. Losalka, A. Krause, and J. Scarlett. Stochastic linear bandits robust to adversarial attacks. In *Proc. of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 991–999, 2021.

S. Bubeck and A. Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory*, pages 42.1–42.23, 2012.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proc. of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 208–214, 2011.

A. Cutkosky, C. Dann, A. Das, C. Gentile, A. Pacchiano, and M. Purohit. Dynamic balancing for model selection in bandits and rl. In *Proc. of International Conference on Machine Learning (ICML)*, pages 2276–2285, 2021.

C. Dann, C.-Y. Wei, and J. Zimmert. A blackbox approach to best of both worlds in bandits and beyond. In *Proc. of Annual Conference on Learning Theory (COLT)*, pages 5503–5570, 2023.

Q. Ding, C.-J. Hsieh, and J. Sharpnack. Robust stochastic linear contextual bandits under adversarial attacks. In *Proc. of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 7111–7123, 2022.

D. Foster and A. Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *Proc. of International Conference on Machine Learning (ICML)*, pages 3199–3210, 2020.

D. Foster, A. Agarwal, M. Dudik, H. Luo, and R. Schapire. Practical contextual bandits with regression oracles. In *Proc. of International Conference on Machine Learning (ICML)*, pages 1539–1548, 2018.

D. J. Foster and A. Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 18907–18919, 2021.

D. J. Foster, C. Gentile, M. Mohri, and J. Zimmert. Adapting to misspecification in contextual bandits. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 11478–11489, 2020.

E. Garcelon, B. Rozière, L. Meunier, J. Tarbouriech, O. Teytaud, A. Lazaric, and M. Pirotta. Adversarial attacks on linear contextual bandits. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 14362–14373, 2020.

A. Gupta, T. Koren, and K. Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Proc. of Annual Conference on Learning Theory (COLT)*, pages 1562–1578, 2019.

J. He, D. Zhou, T. Zhang, and Q. Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 34614–34625, 2022.

S. Ito. Hybrid regret bounds for combinatorial semi-bandits and adversarial linear bandits. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 2654–2667, 2021.

S. Ito and K. Takemura. An exploration-by-optimization approach to best of both worlds in linear bandits. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

S. Ito and K. Takemura. Best-of-three-worlds linear bandit algorithm with variance-adaptive regret bounds. In *Proc. of Annual Conference on Learning Theory (COLT)*, pages 2653–2677, 2023b.

S. Ito, S. Hirahara, T. Soma, and Y. Yoshida. Tight first- and second-order regret bounds for adversarial linear bandits. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 2028–2038, 2020.

S. Ito, T. Tsuchiya, and J. Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 28631–28643, 2022.

T. Jin, L. Huang, and H. Luo. The best of both worlds: stochastic and adversarial episodic mdps with unknown transition. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 20491–20502, 2021.

K.-S. Jun, L. Li, Y. Ma, and J. Zhu. Adversarial attacks on stochastic bandits. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 3640–3649, 2018.

Y. Kang, C.-J. Hsieh, and T. Lee. Robust lipschitz bandits to adversarial corruptions. *arXiv preprint arXiv:2305.18543*, 2023.

F. Kong, C. Zhao, and S. Li. Best-of-three-worlds analysis for linear bandits with follow-the-regularized-leader algorithm. In *Proc. of Annual Conference on Learning Theory (COLT)*, pages 657–673, 2023.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

C.-W. Lee, H. Luo, C.-Y. Wei, M. Zhang, and X. Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *Proc. of International Conference on Machine Learning (ICML)*, pages 6142–6151, 2021.

Y. Li, Y. Wang, and Y. Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. In *Proc. of Annual Conference on Learning Theory (COLT)*, pages 2173–2174, 2019.

F. Liu and N. Shroff. Data poisoning attacks on stochastic bandits. In *Proc. of International Conference on Machine Learning (ICML)*, pages 4042–4050, 2019.

T. Lykouris, V. Mirrokni, and R. Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proc. of ACM Symp. on Theory of Computing (STOC)*, pages 114–122, 2018.

G. Neu and G. Bartók. An efficient algorithm for learning with semi-bandit feedback. In S. Jain, R. Munos, F. Stephan, and T. Zeugmann, editors, *Proc. of International Conference on Algorithmic Learning Theory (ALT)*, pages 234–248, 2013.

G. Neu and G. Bartók. Importance weighting without importance weights: An efficient algorithm for combinatorial semi-bandits. *Journal of Machine Learning Research (JMLR)*, 17(1):5355–5375, 2016. ISSN 1532-4435.

G. Neu and J. Olkhovskaya. Efficient and robust algorithms for adversarial linear contextual bandits. In *Proc. of Annual Conference on Learning Theory (COLT)*, pages 3049–3068, 2020.

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

A. Pacchiano, M. Phan, Y. Abbasi Yadkori, A. Rao, J. Zimmert, T. Lattimore, and C. Szepesvari. Model selection in contextual stochastic bandit problems. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 10328–10337, 2020.

A. Pacchiano, C. Dann, and C. Gentile. Best of both worlds model selection. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 1883–1895, 2022.

A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *Proc. of Annual Conference on Learning Theory (COLT)*, pages 993–1019, 2013.

A. Rakhlin and K. Sridharan. Bistro: An efficient relaxation-based method for contextual bandits. In *Proc. of International Conference on Machine Learning (ICML)*, pages 1977–1985, 2016.

C. Rouyer, D. van der Hoeven, N. Cesa-Bianchi, and Y. Seldin. A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 35035–35048, 2022.

Y. Seldin and G. Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the 2017 Conference on Learning Theory*, pages 1743–1759, 2017.

Y. Seldin and A. Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *Proc. of International Conference on Machine Learning (ICML)*, pages 1287–1295, 2014.

V. Syrgkanis, A. Krishnamurthy, and R. Schapire. Efficient algorithms for adversarial contextual learning. In *Proc. of International Conference on Machine Learning (ICML)*, pages 2159–2168, 2016.

T. Tsuchiya, S. Ito, and J. Honda. Best-of-both-worlds algorithms for partial monitoring. In *Proc. of International Conference on Algorithmic Learning Theory (ALT)*, pages 1484–1515, 2023a.

T. Tsuchiya, S. Ito, and J. Honda. Stability-penalty-adaptive follow-the-regularized-leader: Sparsity, game-dependency, and best-of-both-worlds. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2023b.

C.-Y. Wei and H. Luo. More adaptive algorithms for adversarial bandits. In *Proc. of Annual Conference on Learning Theory (COLT)*, pages 1263–1291, 2018.

C.-Y. Wei, C. Dann, and J. Zimmert. A model selection approach for corruption robust reinforcement learning. In *Proc. of International Conference on Algorithmic Learning Theory (ALT)*, pages 1043–1096, 2022.

C. Ye, W. Xiong, Q. Gu, and T. Zhang. Corruption-robust algorithms with uncertainty weighting for nonlinear contextual bandits and Markov decision processes. In *Proc. of International Conference on Machine Learning (ICML)*, pages 39834–39863, 2023.

H. Zhao, D. Zhou, and Q. Gu. Linear contextual bandits with adversarial corruptions, 2021.

L. Zierahn, D. van der Hoeven, N. Cesa-Bianchi, and G. Neu. Nonstochastic contextual combinatorial bandits. In *Proc. of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 8771–8813, 2023.

J. Zimmert and Y. Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. In *Proc. of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 467–475, 2021.

J. Zimmert, H. Luo, and C.-Y. Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *Proc. of International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 7683–7692, 2019.

A Notation

In this appendix, we provide Table 2 summarizing the most important notations used in the paper.

Table 2: Notations.

Symbol	Meaning
$[K] := \{1, 2, \ldots, K\}$	Finite action set
$d \in \mathbb{N}$	Dimension of loss vectors and contexts
$\mathcal{X} \subseteq \mathbb{R}^d$	A context space of a full-dimensional compact set
$\mathcal{D} \in \Delta(\mathcal{X})$	Context distribution over $\mathcal{X}$
$\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$	Covariance matrix of contexts, $\mathbb{E}_{X\sim\mathcal{D}}[XX^\top]$
$\boldsymbol{\theta}_{t,a} \in \mathbb{R}^d$	Loss vector of action $a \in [K]$ at round $t \in [T]$
$\boldsymbol{\theta}_a \in \mathbb{R}^d$	Fixed and unknown vectors of action $a \in [K]$ at round $t \in [T]$ (corrupted and stochastic regime)
$C \in [0, T]$	Corruption level, upper bound of $\sum_{t=1}^T \max_{a\in[K]} \|\boldsymbol{\theta}_{t,a} - \boldsymbol{\theta}_a\|_2$
$\pi(\cdot \mid \boldsymbol{x}) \in \Delta([K])$	Probabilistic policy mapping each context $\boldsymbol{x}$ to a probability distribution
Π	Set of all deterministic policies $\pi : \mathcal{X} \to [K]$
$\pi^* \in \Pi$	Optimal policy
$\Delta_{\min} > 0$	Minimum sub-optimal gap over a context space, $\min_{\boldsymbol{x}\in\mathcal{X}} \min_{a\neq\pi^*(\boldsymbol{x})} \langle \boldsymbol{x}, \boldsymbol{\theta}_a - \boldsymbol{\theta}_{\pi^*(\boldsymbol{x})}\rangle$
$\boldsymbol{m}_{t,a} \in \mathbb{R}^d$	Loss predictor for action $a \in [K]$ and $t \in [T]$
L^*	$\mathbb{E}\big[\sum_{t=1}^T \ell_t(X_t, \pi^*(X_t))\big]$
Λ^*	$\mathbb{E}\big[\sum_{t=1}^T (\ell_t(X_t, A_t) - \langle X_t, \boldsymbol{m}_{t,A_t}\rangle)^2\big]$
$\overline{\Lambda}$	$\mathbb{E}\big[\sum_{t=1}^T (\ell_t(X_t, A_t) - \langle X_t, \overline{\boldsymbol{\theta}}\rangle)^2\big]$ with $\overline{\boldsymbol{\theta}} := \frac{1}{TK}\sum_{t=1}^T\sum_{a=1}^K \boldsymbol{\theta}_{t,a}$.
$\xi_{t,a} \in \mathbb{R}$	$(\ell_t(X_t, a) - \langle X_t, \boldsymbol{m}_{t,a}\rangle)$ with a loss predictor $\boldsymbol{m}_{t,a}$ for for action $a \in [K]$ and $t \in [T]$
$\widehat{\boldsymbol{\theta}}_{t,a} \in \mathbb{R}^d$	Unbiased estimator for $\boldsymbol{\theta}_{t,a}$ for $a \in [K]$ and $t \in [T]$
$\widehat{\boldsymbol{\ell}}_s(X_t) \in \mathbb{R}^K$	Estimated loss vector for X_t at round $t \in [T]$, $\big(\langle X_t, \widehat{\boldsymbol{\theta}}_{s,1}\rangle, \ldots, \langle X_t, \widehat{\boldsymbol{\theta}}_{s,K}\rangle\big)$
$\boldsymbol{m}_t(X_t) \in \mathbb{R}^K$	Predicted loss vector for X_t at round $t \in [T]$, $\big(\langle X_t, \boldsymbol{m}_{t,1}\rangle, \ldots, \langle X_t, \boldsymbol{m}_{t,K}\rangle\big)$
$\widehat{R}_T(\boldsymbol{x})$	Regret of auxiliary game for context $\boldsymbol{x}$ and unbiased loss estimator $\widehat{\boldsymbol{\theta}}_{t,a}$ at round t, $\mathbb{E}\left[\sum_{t=1}^T \langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{t,A_t}\rangle - \langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{t,\pi^*(\boldsymbol{x})}\rangle\right]$
$\widetilde{\boldsymbol{\theta}}_{t,a} \in \mathbb{R}^d$	Biased estimator for $\boldsymbol{\theta}_{t,a}$ for $a \in [K]$ and $t \in [T]$
$\widetilde{\boldsymbol{\ell}}_s(X_t) \in \mathbb{R}^K$	Estimated loss vector for X_t at round $t \in [T]$, $\big(\langle X_t, \widetilde{\boldsymbol{\theta}}_{s,1}\rangle, \ldots, \langle X_t, \widetilde{\boldsymbol{\theta}}_{s,K}\rangle\big)$
$\widetilde{R}_T(\boldsymbol{x})$	Regret of auxiliary game for context $\boldsymbol{x}$ and loss estimator $\widetilde{\boldsymbol{\theta}}_{t,a}$ at round t, $\mathbb{E}\left[\sum_{t=1}^T \langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,A_t}\rangle - \langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,\pi^*(\boldsymbol{x})}\rangle\right]$

B Additional Related Work

There is another line of research dedicated to studying the problem of model selection. A few notable works in this area include Pacchiano et al. [2020, 2022], Agarwal et al. [2017b], Cutkosky et al. [2021], Lee et al. [2021], Wei et al. [2022]. Among these, Pacchiano et al. [2022] addressed the general contextual linear bandit problem with a nested policy class. They achieved the first high probability regret bound, recovering the result of Agarwal et al. [2017b] in the adversarial regime, and attained a gap-dependent bound in the stochastic regime. They also showed a lower bound for the stochastic regime, indicating that a perfect model selection among m logarithmic rate learners is impossible. Formally, this implies that the optimal dependence of the complexity parameter for the largest policy class cannot be improved over a quadratic, i.e., $\frac{R(\Pi_m)^2 \ln T}{\Delta_{\min}}$, where $R(\Pi_m)$ is the complexity parameter for the largest policy class. In their best-of-both-worlds model selection algorithm, the base learners aggregated by the meta-algorithm are required to satisfy anytime high-probability regret guarantees in the adversarial regime, along with notions of high probability stability and action space extendability. Although a high-probability variant of Exp4 of Auer et al. [2002] could be a viable option as a base learner to meet these requirements, its running time, however, is generally linear in the number of policies. This makes it intractable for an infinite policy class of $\pi : \mathcal{X} \to [K]$, where $\mathcal{X} \subseteq \mathbb{R}^d$. Leaving aside the computational issues, Pacchiano et al. [2022] have not addressed data-dependent bounds in the adversarial regime, nor have the corrupted regime been explicitly investigated.

Since Lykouris et al. [2018] first proposed the stochastic K-armed bandits with adversarial corruptions, different problem settings including contextual bandits, have been well-studied in the literature. Zhao et al. [2021], Ding et al. [2022], He et al. [2022] extended the model studied in Abbasi-Yadkori et al. [2011] under the corruption framework by Lykouris et al. [2018] for the linear contextual bandits. For further extensions, Bogunovic et al. [2020] introduced the kernelized MAB problem. Ye et al. [2023] recently studied nonlinear contextual bandits and Markov Decision Processes, and Kang et al. [2023] introduced Lipschitz bandits in the presence of adversarial corruptions. We also mention a few works of Jun et al. [2018], Liu and Shroff [2019], Garcelon et al. [2020], Bogunovic et al. [2021] in this line of research that studied a different adversary model, where the adversary may add the corruption after observing the learner's action A_t. Garcelon et al. [2020] examined several attack scenarios and showed that a malicious adversary could manipulate a linear contextual bandit algorithm for the adversary's benefit. It is also notable that regret can be defined in different ways, taking into account losses after corruption or losses without corruption. However, the difference between the two definitions is negligible, at most $O(C)$, where C is the corruption level. For a more detailed discussion on these different notions of regret, refer to Gupta et al. [2019], Ito [2021].

Algorithms for linear contextual bandits that provide regret guarantees have been developed with various assumptions on the losses and contexts. The stochastic linear contextual bandit is the most extensively studied model among them. Here, the context in each round can be arbitrarily generated while an unknown loss (reward) vector is fixed over time [Chu et al., 2011, Abbasi-Yadkori et al., 2011, Li et al., 2019]. Efficient computational techniques have also been developed to take advantage of the availability of a regression oracle [Foster et al., 2018]. Foster et al. [2020] studied the misspecified linear contextual bandit problem for infinite actions with an online regression oracle. In addition, Foster and Rakhlin [2020] extended oracle-based algorithms for a general function class.

Despite the rich history of contextual bandits literature we described above, few results have been known for data-dependent bounds as the question was posed by Agarwal et al. [2017a]. Allen-Zhu et al. [2018] first affirmatively solved this question for adversarial losses and contexts. However, their algorithm only works for a moderate number of policies. Foster and Krishnamurthy [2021] provided the first optimal and efficient reduction from contextual bandits to online regression with the cross-entropy loss, thereby achieving a first-order regret guarantee, but the loss function is assumed to be fixed over time. The work of Olkhovskaya et al. [2023] first achieved the first- and second-order bounds for adversarial losses and i.i.d contexts case. The critical difference between the above-mentioned work and our study is that these have not investigated the BoBW guarantee.

C Lower Bound

An algorithm is said to be *orthogonal* if it does not use the information from rounds in which $X_t \neq X_s$ for $s < t$ to make a prediction at round t [Zierahn et al., 2023]. For the class of orthogonal algorithms, Zierahn et al. [2023] proved the following regret lower bound for the combinatorial full-bandit setting in the adversarial regime. In the combinatorial full-bandit setting, the action space satisfies $\mathcal{A} \subseteq \{0,1\}^K$ and $\max_{a \in \mathcal{A}} \|a\|_1 \leq S$.

Proposition 2 (Theorem 19 in Zierahn et al. [2023])**.** *Suppose $T \geq dSK$ and $K \geq 2S$. In the combinatorial full-bandit setting, any orthogonal algorithm satisfies*

$$R_T \geq \frac{S^{3/2}\sqrt{dKT}}{16(192 + 96\ln(T))}.$$

In their proof of the lower bound, they construct the S instances of n-armed bandit problems for $n = \frac{K}{S} \in \mathbb{N}$. Therefore, the statement for $S = 1$ implies the lower bound for the K-armed contextual bandit case:

$$R_T = \Omega(\sqrt{dKT}),$$

which we are interested in. Also note that both FTRL-LC and MWU-LC with $\boldsymbol{m}_{t,a} = 0$ are orthogonal, as formally stated in the following lemmas.

Lemma 4. *Suppose that $\mathcal{X}$ consist of only basis vectors i.e., $\mathcal{X} = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K\}$, and pick some $t \in [T]$. Let $X_{t'} \neq X_t$ and let $a \in [K]$. Then, $\langle X_{t'}, \widetilde{\boldsymbol{\theta}}_{t,a} \rangle = 0$ holds for the biased estimator $\widetilde{\boldsymbol{\theta}}_{t,a}$ in (12) of FTRL-LC, and $\langle X_{t'}, \widehat{\boldsymbol{\theta}}_{t,a} \rangle = 0$ holds for the unbiased estimator $\widehat{\boldsymbol{\theta}}_{t,a}$ with $\boldsymbol{m}_{t,a} = \boldsymbol{0}$ in (6) of MWU-LC.*

Proof of Lemma 4. We follow the proof of Lemma 17 in Zierahn et al. [2023]. First consider $\widetilde{\boldsymbol{\theta}}_{t,a}$ in (12). Let $\widehat{\boldsymbol{\Sigma}}^{+}_{t,a}$ be a sample of the MGR (Algorithm 7) with M-iteration and it can be written as

$$\widehat{\boldsymbol{\Sigma}}^{+}_{t,a} = \rho \sum_{k=0}^{M} \prod_{j=1}^{k} (\mathbf{I} - \rho \mathbf{B}_{k,a}).$$

Notice that $\mathbf{B}_{k,a} = \mathbb{1}[A(k) = a] X(k) X(k)^\top$ is diagonal since $\mathcal{D}$ has the support of $\mathcal{X} = \{\boldsymbol{e}_1, \dots, \boldsymbol{e}_K\}$. So as $\widehat{\boldsymbol{\Sigma}}^{+}_{t,a}$ for all $a \in [K]$. Let $X_{t'} = \boldsymbol{e}_i$ and $X_t = \boldsymbol{e}_j$ and pick $a \in [K]$. Then we see that

$$\begin{aligned}
\langle X_{t'}, \widetilde{\boldsymbol{\theta}}_{t,a} \rangle &= \boldsymbol{e}_i^\top \widetilde{\boldsymbol{\theta}}_{t,a} \\
&= \boldsymbol{e}_i^\top \widehat{\boldsymbol{\Sigma}}^{+}_{t,a} X_t \ell_t(X_t, A_t) \mathbb{1}[A_t = a] \\
&= \boldsymbol{e}_i^\top \widehat{\boldsymbol{\Sigma}}^{+}_{t,a} \boldsymbol{e}_j \ell_t(X_t, A_t) \mathbb{1}[A_t = a] \\
&= (\widehat{\boldsymbol{\Sigma}}^{+}_{t,a})_{i,j} \ell_t(X_t, A_t) \mathbb{1}[A_t = a],
\end{aligned}$$

concluding that $\langle X_{t'}, \widetilde{\boldsymbol{\theta}}_{t,a} \rangle = 0$ if $i \neq j$.

Next, we consider $\widehat{\boldsymbol{\theta}}_{t,a}$ in (6), where $\widetilde{\boldsymbol{\Sigma}}^{-1}_{t,a}$ is given by (7) and $\xi_{t,a} = (\ell_t(X_t, a) - \langle X_t, \boldsymbol{m}_{t,a} \rangle)$. By a similar discussion, we have

$$\begin{aligned}
\langle X_{t'}, \widehat{\boldsymbol{\theta}}_{t,a} \rangle &= \boldsymbol{e}_i^\top \widehat{\boldsymbol{\theta}}_{t,a} \\
&= \boldsymbol{e}_i^\top \Big(\boldsymbol{m}_{t,a} + \frac{\mathsf{upd}_t}{q_t} Q_t(a) \widetilde{\boldsymbol{\Sigma}}^{-1}_{t,a} X_t \xi_{t,a} \mathbb{1}[A_t = a] \Big) \\
&= \boldsymbol{e}_i^\top \Big(\boldsymbol{m}_{t,a} + \frac{\mathsf{upd}_t}{q_t} Q_t(a) \widetilde{\boldsymbol{\Sigma}}^{-1}_{t,a} \boldsymbol{e}_j \xi_{t,a} \mathbb{1}[A_t = a] \Big) \\
&= \boldsymbol{m}_{t,a}(i) + (\widetilde{\boldsymbol{\Sigma}}^{-1}_{t,a})_{i,j} \frac{\mathsf{upd}_t}{q_t} Q_t(a) \xi_{t,a} \mathbb{1}[A_t = a].
\end{aligned}$$

Therefore, we conclude that $\langle X_{t'}, \widehat{\boldsymbol{\theta}}_{t,a} \rangle = 0$ if $i \neq j$ and $\boldsymbol{m}_{t,a} = \boldsymbol{0}$, since $\widetilde{\boldsymbol{\Sigma}}_{t,a}$ is diagonal in this case. □

Lemma 5. *Suppose that $\mathcal{X}$ consist of only basis vectors i.e., $\mathcal{X} = \{\boldsymbol{e}_1, \dots, \boldsymbol{e}_K\}$. Also, suppose that in round t, the context is a basis vector in the direction $i \in [K]$. Then, in FTRL-LC and MWU-LC with $\boldsymbol{m}_{t,a} = \boldsymbol{0}$ for each $a \in [K]$, the observation obtained in round t does not affect the algorithm's prediction in all subsequent rounds such that the context is a basis vector in direction $j \neq i$.*

Proof of Lemma 5. Lemma 4 implies that when $X_t = \boldsymbol{e}_i$, then $p_t(\cdot|X_t)$ in (11) in FTRL-LC can be written as

$$p_t(\cdot|\boldsymbol{e}_i) = \operatorname*{arg\,min}_{r \in \Delta([K])} \left\{ \sum_{s<t: X_s = \boldsymbol{e}_i} \langle r, \widetilde{\boldsymbol{\ell}}_s(\boldsymbol{e}_i) \rangle + \psi_t(r) \right\},$$

where $\widetilde{\boldsymbol{\ell}}_s(\boldsymbol{e}_i := (\langle \boldsymbol{e}_i, \widetilde{\boldsymbol{\theta}}_{s,1} \rangle, \dots, \langle \boldsymbol{e}_i, \widetilde{\boldsymbol{\theta}}_{s,K} \rangle)^\top \in \mathbb{R}^K$. Also, we can write $w_t(r|X_t)$ for $r \in \Delta([K])$ in (4) of MWU-LC as

$$w_t(r|\boldsymbol{e}_i) = \exp\left(-\eta_t \sum_{a \in [K]} r_a \left\langle \boldsymbol{e}_i, \sum_{s<t: X_s = \boldsymbol{e}_i} \widehat{\boldsymbol{\theta}}_{s,a} + \boldsymbol{m}_{t,a} \right\rangle \right),$$

where $\boldsymbol{m}_{t,a} = \boldsymbol{0}$. These equations mean that both algorithms do not use the information at round $s < t$ wherein $X_t \neq X_s$. □

D Useful Lemmas

This section presents some known results from existing literature, such as basic regret bounds in FTRL and basic regret decompositions often used for K-armed linear contextual bandits.

D.1 Analysis of FTRL

We introduce a standard FTRL analysis (e.g. Exercise 28.12 of Lattimore and Szepesvári 2020) when it is applied to K-armed linear contextual bandits with a fixed context $\boldsymbol{x} \in \mathcal{X}$. The following Lemma 6 will be used to analyze the regret of the auxiliary game given by (16) in Lemma 1.

The Bregman divergence from $p \in \Delta([K])$ to $q \in \Delta([K])$ is defined as

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

Lemma 6. *Let $p_t(\cdot|\boldsymbol{x})$ be a FTRL prediction with loss estimators $\widetilde{\boldsymbol{\theta}}_{t,a}$ for each $a \in [K]$, which is given by (10) with any convex regularizer $\psi_t(\cdot)$. Suppose that A_t is chosen by $\pi_t(\cdot|\boldsymbol{x}) := (1-\gamma_t)p_t(\cdot|\boldsymbol{x}) + \gamma_t \frac{1}{K}$, where $\gamma_t \in [0,1]$ is the mixture rate. Then, for any context $\boldsymbol{x} \in \mathcal{X}$, we have*

$$\begin{aligned}
&\mathbb{E}_{A_t}\left[\sum_{t=1}^{T}\left(\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,A_t}\rangle - \langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,\pi^*(\boldsymbol{x})}\rangle\right)\right] \\
&\leq \sum_{t=1}^{T}\left(\psi_t(p_{t+1}(\cdot|\boldsymbol{x})) - \psi_{t+1}(p_{t+1}(\cdot|\boldsymbol{x}))\right) + \psi_{T+1}(\pi^*(\cdot|\boldsymbol{x})) - \psi_1(p_1(\cdot|\boldsymbol{x})) \\
&\quad + \sum_{t=1}^{T}(1-\gamma_t)\left(\left\langle p_t(\cdot|\boldsymbol{x}) - p_{t+1}(\cdot|\boldsymbol{x})), \widetilde{\boldsymbol{\ell}}_t(\boldsymbol{x})\right\rangle - D_t(p_{t+1}(\cdot|\boldsymbol{x}), p_t(\cdot|\boldsymbol{x}))\right) + U(\boldsymbol{x}),
\end{aligned}$$

where $U(\boldsymbol{x}) = \sum_{t=1}^{T}\gamma_t\left\langle \frac{1}{K}\mathbf{1} - \pi^*(\cdot|\boldsymbol{x}), \widetilde{\boldsymbol{\ell}}_t(\boldsymbol{x})\right\rangle$, and $\pi^*(a|\boldsymbol{x}) = 1$ if $a = \pi^*(\boldsymbol{x})$ otherwise 0.

Proof of Lemma 6. From the definition of the auxiliary game and the design of the algorithm, for any $\boldsymbol{x} \in \mathcal{X}$, we have

$$\begin{aligned}
\mathbb{E}_{A_t}\left[\sum_{t=1}^{T}\left(\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,A_t}\rangle - \langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,\pi^*(\boldsymbol{x})}\rangle\right)\right] &= \sum_{t=1}^{T}\sum_{a\in[K]}(\pi_t(a|\boldsymbol{x}) - \pi^*(a|\boldsymbol{x}))\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle \\
&= \sum_{t=1}^{T}(1-\gamma_t)\sum_{a\in[K]}(p_t(a|\boldsymbol{x}) - \pi^*(a|\boldsymbol{x}))\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle + \sum_{t=1}^{T}\gamma_t\left\langle \frac{1}{K}\mathbf{1} - \pi^*(\cdot|\boldsymbol{x}), \widetilde{\boldsymbol{\ell}}_t(\boldsymbol{x})\right\rangle \\
&= \sum_{t=1}^{T}(1-\gamma_t)\sum_{a\in[K]}(p_t(a|\boldsymbol{x}) - \pi^*(a|\boldsymbol{x}))\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle + U(\boldsymbol{x}).
\end{aligned}$$

By the standard analysis of FTRL (see, e.g., Exercise 28.12 of Lattimore and Szepesvári 2020), the first term in the RHS above is further bounded as

$$\begin{aligned}
&\sum_{t=1}^{T}(1-\gamma_t)\sum_{a\in[K]}(p_t(a|\boldsymbol{x}) - \pi^*(a|\boldsymbol{x}))\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle \\
&\leq \sum_{t=1}^{T}(1-\gamma_t)\left(\left\langle p_t(\cdot|\boldsymbol{x}) - p_{t+1}(\cdot|\boldsymbol{x})), \widetilde{\boldsymbol{\ell}}_t(\boldsymbol{x})\right\rangle - D_t(p_{t+1}(\cdot|\boldsymbol{x}), p_t(\cdot|\boldsymbol{x}))\right) \\
&\quad + \sum_{t=1}^{T}\left(\psi_t(p_{t+1}(\cdot|\boldsymbol{x})) - \psi_{t+1}(p_{t+1}(\cdot|\boldsymbol{x}))\right) + \psi_{T+1}(\pi^*(\cdot|\boldsymbol{x})) - \psi_1(p_1(\cdot|\boldsymbol{x})).
\end{aligned}$$

Combining the above arguments completes the proof. ∎

D.2 Fundamental bounds for K-armed linear contextual bandits

First, we introduce a fundamental regret decomposition using the auxiliary game in (15).

Lemma 7 (c.f. Equation (6) of Neu and Olkhovskaya [2020]). *Let $X_0 \sim \mathcal{D}$ be a ghost sample drawn independently from the entire interaction history. Then we have*

$$R_\tau \leq \mathbb{E}[\widetilde{R}_\tau(X_0)] + 2\sum_{t=1}^{\tau} \max_{a\in[K]} |\mathbb{E}[\langle X_t, \boldsymbol{b}_{t,a}\rangle]|$$

Next, we introduce the following lemma for analysis related to a ghost sample X_0, which will be used to prove Proposition 7 and Lemma 3.

Lemma 8 (c.f. Lemma 6 in Neu and Olkhovskaya [2020]). *Let $X_0 \sim \mathcal{D}$ be a ghost sample drawn independently from the entire interaction history. Suppose that X_t is satisfying $\|X_t\|_2 \leq 1$, and $0 < \rho \leq \frac{1}{2}$. Then, for any time step t and an estimator $\widetilde{\boldsymbol{\theta}}_{t,a}$, we have*

$$\mathbb{E}_t\left[\sum_{a=1}^{K} \pi_t(a|X_0)\langle X_0, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle^2\right] \leq 3Kd. \tag{17}$$

Lastly, we introduce the following lemma, which will be used to prove Lemma 2 to control the biased term caused by MGR procedure.

Lemma 9 (c.f. Lemma 5 in Neu and Olkhovskaya [2020]). *Let $\widehat{\boldsymbol{\theta}}_{t,a} = \boldsymbol{\Sigma}_{t,a}^{-1} X_t \ell_t(X_t, A_t)\mathbb{1}[A_t = a]$ for all $a \in [K]$, and let $\widetilde{\boldsymbol{\theta}}_{t,a} = \widehat{\boldsymbol{\Sigma}}_{t,a}^{+} X_t \ell_t(X_t, A_t)\mathbb{1}[A_t = a]$ for all $a \in [K]$ where $\widehat{\boldsymbol{\Sigma}}_{t,a}^{+}$ is obtained by MGR with $\rho = \frac{1}{2}$ of Algorithm 7. Then, we have*

$$\left|\mathbb{E}[\langle X_t, \widetilde{\boldsymbol{\theta}}_{t,a} - \widehat{\boldsymbol{\theta}}_{t,a}\rangle|\mathcal{F}_{t-1}]\right| \leq \exp\left(-\frac{\gamma_t \lambda_{\min}(\boldsymbol{\Sigma})}{2K} M_t\right).$$

E Appendix for Reduction Approach

We summarize the known results of the black-box reduction framework of Dann et al. [2023], when it is adapted to our K-armed linear contextual bandit problem, although Dann et al. [2023] provided for several other different problem settings. Then, as a naive adaption of Dann et al. [2023], we describe a base algorithm for K-armed linear contextual bandits with adaptive learning rates and provide its analysis, resulting in Proposition 8. For notational convenience, we use $R(\tau, a^*)$ to denote the pseudo-regret of $\mathbb{E}\left[\sum_{t=1}^{\tau} \ell_t(X_t, A_t) - \ell_t(X_t, a^*)\right]$ for round $\tau \in [1, T]$ and comparator action $a^* \in [K]$ fixed in hindsight. All the pseudo-codes of reduction algorithms are also detailed in this appendix to make the paper self-contained.

E.1 Zero-order bound via reduction framework

Inspired by the techniques of model selections, the reduction approach of Dann et al. [2023] relies on an algorithm satisfying the following condition, called *α-local-self-bounding condition* (LSB).

Definition 1 (α-local-self-bounding condition or α-LSB, Adaption of Definition 4 of [Dann et al., 2023]). *We say an algorithm satisfies the α-local-self-bounding condition if it takes a candidate action $\widehat{a} \in [K]$ as input and has the following pseudo-regret guarantee for any stopping time $\tau \in [1, T]$ and for any $a^* \in [K]$:*

$$R(\tau, a^*) \leq \min\left\{c_0^{1-\alpha}\mathbb{E}[\tau]^\alpha, (c_1 \ln T)^{1-\alpha}\mathbb{E}\left[\sum_{t=1}^{\tau}(1 - \mathbb{1}[a^* = \widehat{a}]p_t(a^*|X_t))\right]^\alpha\right\} + c_2 \ln T, \tag{18}$$

where c_0, c_1, c_2 are problem dependent constants and $p_t(a^*|X_t)$ is the probability choosing a^* at round t.

For a reduction procedure, detailed in Algorithm 3, that turns any LSB algorithm into a best-of-both-world algorithm, its BoBW guarantees are stated in the following proposition.

Proposition 3 (Adaption of Theorem 6 of Dann et al. [2023]). *If an algorithm $\mathcal{L}$ satisfies α-LSB with (c_0, c_1, c_2), then the regret of Algorithm 3 with $\mathcal{L}$ as the base algorithm is upper bounded by $\mathcal{O}(c_0^{1-\alpha}T^\alpha + c_2 \ln^2(T))$ in the adversarial regime and by $\mathcal{O}(c_1 \ln(T)\Delta_{\min}^{-\frac{\alpha}{1-\alpha}} + (c_1 \ln T)^{1-\alpha}(C\Delta_{\min}^{-1})^\alpha + c_2 \ln(T)\ln(C\Delta_{\min}^{-1}))$ in the corrupted stochastic regime.*

Algorithm 3: BoBW via local-self-bounding (LSB) algorithm, Adaption of Algorithm 1 in Dann et al. [2023]

Input : LSB algorithm $\mathcal{L}$

$T_1 \leftarrow 0; \quad T_0 \leftarrow -c_2 \ln T$;

$\widehat{A}_1 \sim \textbf{unif}([K])$, $t \leftarrow 1$;

for $k = 1, 2, \dots$ **do**

  Initialize $\mathcal{L}$ with candidate action $\widehat{A}_k$;

  Set the number of pulls $N_k(a)$ for all $a \in [K]$;

  for $t = T_k + 1, T_k + 2, \dots$ **do**

    Observe X_t;

    Choose action A_t according to $\mathcal{L}$, and advance $\mathcal{L}$ by one step;

    $N_k(a_t) \leftarrow N_k(a_t) + 1$;

    if $t - T_k \geq 2(T_k - T_{k-1})$ *and* $\exists a \in [K] \setminus \{\widehat{A}_k\}$ *such that* $N_k(a) \geq \frac{t-T_k}{2}$ **then**

      $\widehat{A}_{k+1} \leftarrow a$;

      $T_{k+1} \leftarrow t$;

      break

Algorithm 4: LSB via Corral, Adaption of Algorithm 2 in Dann et al. [2023]

Input : candidate action $\widehat{a} \in [K]$, $\frac{1}{2}$-iw-stable algorithm $\mathcal{B}$ over $[K] \setminus \{\widehat{a}\}$ with constants c_1 and c_2

Define: $\psi_t(q) = -\frac{2}{\eta_t} \sum_{i=1}^{2} \sqrt{q_i} + \frac{1}{\beta} \sum_{i=1}^{2} \ln \frac{1}{q_i}$

$B_0 = 0$;

for $t = 1, 2, \dots$ **do**

  Observe X_t;

  Compute

$$\overline{q}_t \leftarrow \operatorname*{arg\,min}_{q \in \Delta([2])} \left\{ \left\langle q, \sum_{\tau=1}^{t-1} z_\tau - \begin{bmatrix} 0 \\ B_{t-1} \end{bmatrix} + \psi_t(q) \right\rangle \right\}, \quad q_t \leftarrow \left(1 - \frac{1}{2t^2}\right) \overline{q}_t + \frac{1}{4t^2} \mathbf{1}$$

  with $\eta_t \leftarrow \frac{1}{\sqrt{t} + 8\sqrt{c_1}}$, $\beta = \frac{1}{8c_2}$;

  Sample $i_t \sim q_t$;

  if $i_t = 1$ **then**

    Choose $A_t = \widehat{a}$ and observe $\ell_t(X_t, A_t)$;

  else

    Choose A_t according to base algorithm $\mathcal{B}$ and observe $\ell_t(X_t, A_t)$;

  Define $z_{t,i} \leftarrow \frac{(\ell_t(X_t, A_t) + 1)\mathbb{1}[i_t = i]}{q_{t,i}} - 1$ and $B_t \leftarrow \sqrt{c_1 \sum_{\tau=1}^{t} \frac{1}{q_{\tau,2}}} + \frac{c_2}{\min_{\tau \leq t} q_{\tau,2}}$;

Since algorithms satisfying the LSB condition are not common, Dann et al. [2023] further introduced the notion of the *importance-weighting stability* (iw-stable), and presented a variant of Corral algorithm (Algorithm 4) [Agarwal et al., 2017b] that runs over a candidate action $\widehat{a}$ and an importance-weighting stable algorithm $\mathcal{B}$ over the action set $[K] \setminus \{\widehat{a}\}$.

Definition 2 (iw-stable, Adaption of Definition 8 of Dann et al. [2023])**.** *Given an adaptive sequence of weights $q_1, q_2, \ldots \in (0, 1]$, suppose that the feedback in round t is observed with probability q_t. Then, an algorithm is $\frac{1}{2}$-importance-weighting stable if it obtains the following pseudo-regret guarantee for any stopping time $\tau \in [1, T]$ and any $a^* \in [K]$:*

$$R(\tau, a^*) \leq \mathbb{E}\left[\sqrt{c_1 \sum_{t=1}^{\tau} \frac{1}{q_t}} + \frac{c_2}{\min_{t \leq \tau} q_t}\right]. \tag{19}$$

Proposition 4 (Theorem 11 of Dann et al. [2023])**.** *If an algorithm $\mathcal{B}$ is $\frac{1}{2}$-iw-stable with constant (c_1, c_2), then Algorithm 4 with $\mathcal{B}$ as the base algorithm satisfies $\frac{1}{2}$-LSB with constants $(\overline{c}_0, \overline{c}_1, \overline{c}_2)$, where $\overline{c}_0 = \overline{c}_1 = \mathcal{O}(c_1)$ and $\overline{c}_2 = \mathcal{O}(c_2)$.*

E.2 First- and second-order bounds via reduction framework

Next, we introduce a reduction scheme that can also be adapted to obtain a data-dependent bound relying on a notion of *data-dependent local self-bounding* (dd-LSB) [Dann et al., 2023], when it is applied to our setting. In order to make the paper self-contained, we detail the pseudo-code of a Corral algorithm (Algorithm 6 of Dann et al. [2023]) in Algorithm 5.

Definition 3 (dd-LSB, Definition 20 of Dann et al. [2023])**.** *An algorithm is said to be dd-LSB (data-dependent LSB) if it takes a candidate action $\widehat{a} \in \mathcal{A}$ as input and satisfies the following pseudo-regret guarantee for any stopping time $\tau \in [1, T]$ and action $a^* \in [K]$,*

$$R(\tau, a^*) \leq \sqrt{c_1 \ln(T) \mathbb{E}\left[\sum_{t=1}^{\tau}\left(\sum_{a \in [K]} (p_t(a|X_t)\xi_{t,a}^2 - \mathbb{1}[a^* = \widehat{a}] p_t(a^*|X_t)^2 \xi_{t,a^*}^2)\right)\right]} + c_2 \ln T$$

where c_1, c_2 are problem-dependent constants and $p_t(a^*|X_t)$ is the probability choosing a^* at round t.

The performance of an algorithm with dd-LSB condition is guaranteed as the following proposition.

Proposition 5 (Theorem 23 of Dann et al. [2023])**.** *If an algorithm $\mathcal{L}$ satisfies dd-LSB, then the regret of Algorithm 3 with $\mathcal{L}$ as the base algorithm is upper bounded by $O(\sqrt{c_1 \mathbb{E}\left[\sum_{t=1}^{T} \xi_{t,A_t}^2\right] \ln^2 T} + c_2 \ln^2(T))$ in the adversarial regime and by $O\left(\frac{c_1 \ln(T)}{\Delta_{\min}} + \sqrt{\frac{c_1 \ln TC}{\Delta_{\min}}} + c_2 \ln(T) \ln(C\Delta_{\min}^{-1})\right)$ in the corrupted stochastic regime.*

To achieve the dd-LSB condition, Dann et al. [2023] also proposed a variant of *Corral algorithm* of Agarwal et al. [2017b], which is detailed in Algorithm 5. This Corral algorithm is run over two base algorithms with refined weights (q_t): one is to play the current candidate action $\widehat{a}$ and the other is an algorithm with the *data-dependent-importance-weighting-stable* (dd-iw-stable) condition over the action set of $\mathcal{A} \setminus \{\widehat{a}\}$, given in Definition 4. It is guaranteed that the Corral algorithm (Algorithm 5) satisfies the dd-LSB condition when a base algorithm is dd-iw-stable, formally stated in Proposition 6.

Definition 4. *[dd-iw-stable, Adaption of Definition 21 of Dann et al. [2023]] Given an adaptive sequence of weights $q_1, q_2, \ldots \in (0, 1]$, suppose that the feedback in round t is observed with probability q_t. Then, an algorithm is $\frac{1}{2}$-dd-iw-stable (data-dependent-iw-stable) if it satisfies the following pseudo-regret guarantee for any stopping time $\tau \in [1, T]$ and for any $a^* \in [K]$:*

$$R(\tau, a^*) \leq \sqrt{c_1 \mathbb{E}\left[\sum_{t=1}^{\tau} \frac{\mathsf{upd}_t \cdot \xi_{t,A_t}^2}{q_t^2}\right]} + \mathbb{E}\left[\frac{c_2}{\min_{t \leq \tau} q_t}\right],$$

where $\mathsf{upd}_t = 1$ if feedback is observed in round t and $\mathsf{upd}_t = 0$ otherwise.

Proposition 6 (Theorem 22 of Dann et al. [2023])**.** *If a base algorithm $\mathcal{B}$ is $\frac{1}{2}$-dd-iw-stable with constants (c_1, c_2), then Algorithm 5 with $\mathcal{B}$ satisfies $\frac{1}{2}$-dd-LSB with constants $(\bar{c}_1, \bar{c}_2)$ where $\bar{c}_1 = \mathcal{O}(c_1)$ and $\bar{c}_2 = \mathcal{O}(\sqrt{c_1} + \sqrt{c_2})$.*

Algorithm 5: dd-LSB via Corral, Adaption of Algorithm 6 in Dann et al. [2023]

Input : candidate action $\widehat{a} \in [K]$, $\frac{1}{2}$-iw-stable algorithm $\mathcal{B}$ over $[K] \setminus \{\widehat{a}\}$ with constants (c_1, c_2)

Define: $\psi(q) := \sum_{i=1}^{2} \ln \frac{1}{q_i}$, $\quad B_0 := 0$;

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

- Observe X_t;
- Let $\mathcal{B}$ output an action $\widetilde{A}_t$;
- Receive predictors $\boldsymbol{m}_{t,a}$ for all $a \in [K]$, and set $y_{t,1} = \langle X_t, \boldsymbol{m}_{t,\widehat{a}} \rangle$ and $y_{t,2} = \langle X_t, \boldsymbol{m}_{t,\widetilde{A}_t} \rangle$;
- Compute

$$\overline{q}_t \leftarrow \mathop{\arg\min}_{q \in \Delta_2} \left\{ \left\langle q, \sum_{\tau=1}^{t-1} z_\tau + y_t - \begin{bmatrix} 0 \\ B_{t-1} \end{bmatrix} \right\rangle + \frac{1}{\eta_t} \psi(q) \right\}, \quad q_t \leftarrow \left(1 - \frac{1}{2t^2}\right) \overline{q}_t + \frac{1}{4t^2} \mathbf{1},$$

  where $\eta_t \leftarrow \frac{1}{4} (\ln T)^{\frac{1}{2}} \left(\sum_{\tau=1}^{t-1} (\mathbb{1}[i_\tau = i] - q_{\tau,i})^2 \xi_{\tau,A_\tau}^2 + (c_1 + c_2^2) \ln T \right)^{-\frac{1}{2}}$;
- Sample $i_t \sim q_t$;
- **if** $i_t = 1$ **then**
  - Choose $A_t = \widehat{a}$ and observe $\ell_t(X_t, A_t)$;
- **else**
  - Choose $A_t = \widetilde{A}_t$ and observe $\ell_t(X_t, A_t)$;
- Define $z_{t,i} \leftarrow \frac{(\ell_t(X_t, A_t) - y_{t,i}) \mathbb{1}[i_t = i]}{q_{t,i}} + y_{t,i}$ and $B_t \leftarrow \sqrt{c_1 \sum_{\tau=1}^{t} \frac{\xi_{t,A_t}^2 \mathbb{1}[i_\tau = 2]}{q_{\tau,2}^2}} + \frac{c_2}{\min_{\tau \le t} q_{\tau,2}}$;

E.3 Naive adaption

As we discussed in Appendix E.1, the work of Dann et al. [2023] devised a black-box reduction framework to obtain a zero-order regret bound in the adversarial regime as well as the regret in the form of $\frac{\ln T}{\Delta_{\min}}$ in the (corrupted) stochastic regime. In this section, we demonstrate that a basic Exp3-type algorithm with an adaptive learning rate satisfies the importance-weighting stability (Definition 2), where its pseudocode is detailed in Algorithm 6. Specifically, the base algorithm is built upon RealLinExp3 in Neu and Olkhovskaya [2020], but we assume that $\boldsymbol{\Sigma}^{-1}$ is known to the learner.

Proposition 7 (iw-stable condition of Adaptive-RealLinExp3 as a base algorithm)**.** *Assume that $\boldsymbol{\Sigma}^{-1}$ is known to the learner. Then, RealLinExp3 with adaptive learning rate (Algorithm 6) for K-armed linear contextual bandits is $\frac{1}{2}$-importance-weighting stable, where $c_1 = \mathcal{O}\left(\ln(K) K^2 \left(d + \frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)^2\right)$ and $c_2 = \frac{K \ln K}{\lambda_{\min}(\boldsymbol{\Sigma})}$.*

The proof of Proposition 7 will be stated soon. Using Propositions 3, 4, and 7, we have the following proposition.

Proposition 8 (BoBW reduction with a base algorithm of RealLinExp3)**.** *Assume that $\boldsymbol{\Sigma}^{-1}$ is known to the learner. Combining Algorithms 3, 4 and 6 results in the following the regret bound: for the adversarial regime,*

$$R_T = \mathcal{O}\left(\sqrt{c_1 T} + c_2 \ln^2 T\right),$$

Algorithm 6: RealLinExp3 with adaptive learning rate (ADAPTIVE-REALLINEXP3)

Input : Arms $[K]$

Receive update probability $q_t \in (0, 1]$;

Let $\eta_t \leftarrow \min\left\{\sqrt{\frac{\ln K}{\sum_{s=1}^{t} \frac{1}{q_s}}}, \frac{1}{2c}\min_{s \le t} q_s\right\}, \quad \gamma_t \leftarrow \frac{c \cdot \eta_t}{q_t}$, where $c = \frac{K}{\lambda_{\min}(\boldsymbol{\Sigma})}$;

Initialization: Set $\widehat{\boldsymbol{\theta}}_{0,i} = \mathbf{0}$ for all $i \in [K]$;

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

Observe X_t, and for all $a \in [K]$, set

$$p_t(a|X_t) = \exp\left(-\eta_t \sum_{s=1}^{t-1} \langle X_t, \widehat{\boldsymbol{\theta}}_{s,a}\rangle\right);$$

Sample an action A_t from the policy defined as

$$\pi_t(a|X_t) = (1 - \gamma_t)\frac{p_t(a|X_t)}{\sum_{b \in [K]} p_t(b|X_t)} + \gamma_t \frac{1}{K};$$

With probability q_t, observe the loss $\ell_t(X_t, a_t)$ (in this case, set $\mathsf{upd}_t = 1$, otherwise set $\mathsf{upd}_t = 0$);

Compute $\widehat{\boldsymbol{\theta}}_{t,a} = \frac{\mathsf{upd}_t}{q_t}\boldsymbol{\Sigma}_{t,a}^{-1} X_t \ell_t(X_t, A_t)\mathbb{1}[A_t = a]$ for all $a \in [K]$;

and for the corrupted stochastic regime,

$$R_T = \mathcal{O}\left(\frac{c_1 \ln T}{\Delta_{\min}} + \sqrt{\frac{c_1 \ln T}{\Delta_{\min}} C} + c_2 \ln(T) \ln\left(\frac{C}{\Delta_{\min}}\right)\right),$$

where $c_1 = \mathcal{O}\left(\ln(K)K^2\left(d + \frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)^2\right)$ *and* $c_2 = \frac{K \ln K}{\lambda_{\min}(\boldsymbol{\Sigma})}$.

Proposition 8 implies that we obtain desired BoBW bounds if the learner access to $\boldsymbol{\Sigma}_{t,a}^{-1} := \mathbb{E}_t[\mathbb{1}[A_t = a]X_t X_t^\top]$ for computing the unbiased estimator $\widehat{\boldsymbol{\theta}}_{t,a}$ at each round t and $a \in [K]$. However, it only gives the zero-order bound in the adversarial regime. To obtain data-dependent bounds we use a continuous MWU approach as described in Section 4. Importantly, removing the prior knoweges of $\boldsymbol{\Sigma}_{t,a}^{-1}$ is addressed in Section 5. In what follows, we state the proof of Proposition 7.

Proof of Proposition 7. While $\pi^* \in \Pi$ is a deterministic policy, we will also write it using the notations of a probabilistic policy: Let $\pi^*(a|\boldsymbol{x}) = 1$ if $a = \pi^*(\boldsymbol{x})$ otherwise 0 for $a \in [K]$, and $\boldsymbol{x} \in \mathcal{X}$. Let $X_0 \sim \mathcal{D}$ be a ghost sample chosen independently from the entire history. Then, we have

$$\mathbb{E}_t[\langle X_t, \boldsymbol{\theta}_{t,\pi(X_t)}\rangle] = \mathbb{E}_t[\langle X_0, \boldsymbol{\theta}_{t,\pi(X_0)}\rangle].$$

We define $\widehat{R}_T(\boldsymbol{x})$ as the regret of auxiliary game for context $\boldsymbol{x}$ and unbiased loss estimator $\widehat{\boldsymbol{\theta}}_{t,a}$ at round t:

$$\widehat{R}_T(\boldsymbol{x}) := \mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{t,A_t}\rangle - \langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{t,\pi^*(\boldsymbol{x})}\rangle\right]. \tag{20}$$

Using this property and unbiased estimator $\widehat{\boldsymbol{\theta}}_{t,a}$, as also analyzed in Lemma 3 in Olkhovskaya et al. [2023], we have

$$\begin{aligned}
R_\tau &= \mathbb{E}\left[\sum_{t=1}^{\tau}\Big(\ell_t(X_t, A_t) - \ell_t(X_t, \pi^*(X_t))\Big)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{\tau}\Big(\ell_t(X_0, A_t) - \ell_t(X_0, \pi^*(X_0))\Big)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{\tau}\Big(\langle X_0, \widehat{\boldsymbol{\theta}}_{t,A_t}\rangle - \langle X_0, \widehat{\boldsymbol{\theta}}_{t,\pi^*(X_0)}\rangle\Big)\right]. \qquad (21)
\end{aligned}$$

Then, by the definition of $\widehat{R}_T(\boldsymbol{x})$ in (20), RHS of (21) can be written as $\mathbb{E}[\widehat{R}_\tau(X_0)]$:

$$\mathbb{E}[\widehat{R}_\tau(X_0)] = \sum_{t=1}^{\tau}\mathbb{E}_t\left[\sum_{a\in[K]}(\pi_t(a|X_0) - \pi^*(a|X_0))\langle X_0, \widehat{\boldsymbol{\theta}}_{t,a}\rangle\right].$$

We begin with the following lemma using a basic FTRL analysis.

Lemma 10. *For any context $\boldsymbol{x} \in \mathcal{X}$, and suppose that $\widehat{\boldsymbol{\theta}}_{t,a}$ satisfies $|\eta_t\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{t,a}\rangle| \le 1$. Then, for any time step τ, we have*

$$\mathbb{E}[\widehat{R}_\tau(\boldsymbol{x})] \le 2\sum_{t=1}^{\tau}\mathbb{E}_t[\gamma_t] + \mathbb{E}\left[\frac{\ln K}{\eta_\tau}\right] + \sum_{t=1}^{\tau}\mathbb{E}_t\left[\eta_t\sum_{a=1}^{K}\pi_t(a|\boldsymbol{x})\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{t,a}\rangle^2\right]$$

Proof of Lemma 10. Since $\pi_t(a|\boldsymbol{x}) = (1-\gamma_t)p_t(a|\boldsymbol{x}) + \gamma_t\frac{1}{K}$ where we recall that $p_t(a|\boldsymbol{x})$ is given in (3):

$$p_t(a|\boldsymbol{x}) = \frac{\exp\left(-\eta_t\sum_{s=1}^{t-1}\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{s,a}\rangle\right)}{\sum_{b\in[K]}\exp\left(-\eta_t\sum_{s=1}^{t-1}\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{s,b}\rangle\right)} \text{ for } a \in [K],$$

we see that

$$\begin{aligned}
&\mathbb{E}[\widehat{R}_\tau(\boldsymbol{x})] = \sum_{t=1}^{\tau}\mathbb{E}_t\left[\sum_{a\in[K]}(\pi_t(a|\boldsymbol{x}) - \pi^*(a|\boldsymbol{x}))\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{t,a}\rangle\right] \\
&\le \sum_{t=1}^{\tau}\mathbb{E}_t\left[(1-\gamma_t)\sum_{a\in[K]}(p_t(a|\boldsymbol{x}) - \pi^*(a|\boldsymbol{x}))\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{t,a}\rangle\right] + \sum_{t=1}^{\tau}\mathbb{E}_t\left[\frac{\gamma_t}{K}\sum_{a\in[K]}(\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{t,a}\rangle - \langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{t,\pi^*(x)}\rangle)\right].
\end{aligned}$$

As discussed in Section 3, $p_t(\cdot|\boldsymbol{x})$ can also be described as the FTRL with negative Shannon entropy:

$$p_t(\cdot|\boldsymbol{x}) \in \mathop{\arg\min}_{p\in\Delta([K])}\left\{\sum_{s=1}^{t-1}\langle p, \widehat{\boldsymbol{\ell}}_s(\boldsymbol{x})\rangle + \psi_t(p)\right\}, \qquad (22)$$

where $\psi_t(p) = -\frac{1}{\eta_t}H(p) = \frac{1}{\eta_t}\sum_{a\in[K]}p_a\ln p_a$. By a standard FTRL analysis as in Lemma 6 and similar analysis of derivation of (46) in Lemma 1, we have

$$\begin{aligned}
&\sum_{t=1}^{\tau}\mathbb{E}_t\left[(1-\gamma_t)\sum_{a\in[K]}(p'_t(a|\boldsymbol{x}) - \pi^*(a|\boldsymbol{x}))\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{t,a}\rangle\right] \\
&\le \sum_{t=1}^{\tau}\mathbb{E}_t\left[\eta_t\sum_{a=1}^{K}\pi_t(a|\boldsymbol{x})\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{t,a}\rangle^2\right] + \mathbb{E}\left[\frac{\ln K}{\eta_t}\right].
\end{aligned}$$

Since

$$\sum_{t=1}^{\tau}\mathbb{E}_t\left[\frac{\gamma_t}{K}\sum_{a\in[K]}(\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{t,a}\rangle - \langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{t,\pi^*(x)}\rangle)\right] = \sum_{t=1}^{\tau}\mathbb{E}_t\left[\frac{\gamma_t}{K}\sum_{a\in[K]}(\langle \boldsymbol{x}, \boldsymbol{\theta}_{t,a}\rangle - \langle \boldsymbol{x}, \boldsymbol{\theta}_{t,\pi^*(x)}\rangle)\right] \le 2\sum_{t=1}^{\tau}\mathbb{E}_t[\gamma_t]$$

by $|\langle \boldsymbol{x}, \boldsymbol{\theta}_{t,a}\rangle| \le 1$, combining above equalites gives the desired result. □

We next introduce the following lemma, which is implied by Lemma 8, for known $\mathbf{\Sigma}_{t,a}^{-1}$ and unbiased estimator $\widehat{\boldsymbol{\theta}}_{t,a}$ in Line 8 of Algorithm 6.

Lemma 11. *Let $X_0 \sim \mathcal{D}$ be a ghost sample chosen independently from the entire interaction history. Then for any time step t, we have*

$$\mathbb{E}_t\left[\sum_{a=1}^{K} \pi_t(a|X_0)\langle X_0, \widehat{\boldsymbol{\theta}}_{t,a}\rangle^2\right] \le \frac{\sum_{a=1}^{K} \mathbb{E}_t[\mathrm{tr}(\mathbf{\Sigma}_{t,a}\mathbf{\Sigma}_{t,a}^{-1}\mathbf{\Sigma}_{t,a}\mathbf{\Sigma}_{t,a}^{-1})]}{q_t} \le \frac{3Kd}{q_t}. \tag{23}$$

Then, we are ready to prove Proposition 7. We first show $|\eta_t\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{t,a}\rangle| \le 1$.

$$\begin{aligned} |\eta_t\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{t,a}\rangle| = \eta_t|\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{t,a}\rangle| &= \eta_t\left|\boldsymbol{x}^\top \frac{\mathsf{upd}_t}{q_t}\mathbf{\Sigma}_{t,a}^{-1}X_t\ell_t(X_t,a)\mathbb{1}[A_t=a]\right| \le \frac{\eta_t}{q_t}|\boldsymbol{x}^\top\mathbf{\Sigma}_{t,a}^{-1}X_t| \\ &\le \frac{\eta_t}{q_t}\|\mathbf{\Sigma}_{t,a}^{-1}\|_{\mathrm{op}}\cdot\max_{\boldsymbol{x}\in\mathcal{X}}\|\boldsymbol{x}\|^2 \le \frac{\eta_t}{q_t}\frac{1}{\lambda_{\min}(\mathbf{\Sigma}_{t,a})} \le \frac{\eta_t}{q_t}\frac{K}{\lambda_{\min}(\mathbf{\Sigma})\gamma_t} \le 1, \end{aligned} \tag{24}$$

where we used $\ell_t(X_t,a) \le 1$ in the first inequality, $\lambda_{\min}(\mathbf{\Sigma}_{t,a}) \ge \frac{\gamma_t\lambda_{\min}(\mathbf{\Sigma})}{K}$ in the forth inequality, and the definition of $\gamma_t = \frac{\eta_t K}{q_t\lambda_{\min}(\mathbf{\Sigma})}$ in the last inequality.

Next, we will give the bound of $\sum_{t=1}^{\tau}\frac{\eta_t}{q_t}$. Since we have

$$\sum_{t=1}^{\tau}\frac{1}{q_t}\frac{1}{\sqrt{\sum_{s=1}^{t}\frac{1}{q_s}}} \le 2\sum_{t=1}^{\tau}\frac{\frac{1}{q_t}}{\sqrt{\sum_{s=1}^{t}\frac{1}{q_s}}+\sqrt{\sum_{s=1}^{t-1}\frac{1}{q_s}}} = 2\sum_{t=1}^{\tau}\left(\sqrt{\sum_{s=1}^{t}\frac{1}{q_s}}-\sqrt{\sum_{s=1}^{t-1}\frac{1}{q_s}}\right) = 2\sqrt{\sum_{s=1}^{\tau}\frac{1}{q_s}}$$

and using the definition of η_t, we obtain

$$\sum_{t=1}^{\tau}\frac{\eta_t}{q_t} \le \sqrt{\ln K}\sum_{t=1}^{\tau}\frac{1}{q_t}\sqrt{\frac{1}{\sum_{s=1}^{t}\frac{1}{q_s}}} \le \sqrt{4\ln K\sum_{t=1}^{\tau}\frac{1}{q_t}}. \tag{25}$$

Furthermore, by the definition of η_t, it is easy to see that

$$\frac{1}{\eta_\tau} \le \sqrt{\frac{\sum_{t=1}^{\tau}\frac{1}{q_t}}{\ln K}} + \frac{c}{\min_{t\le\tau} q_t}.$$

Therefore, by combining the above inequalities, we have for any $a^* \in [K]$ and $\tau \in [T]$,

$$\begin{aligned} R(\tau, a^*) = \mathbb{E}\left[\sum_{t=1}^{\tau}\Big(\ell_t(X_t,A_t)-\ell_t(X_t,a^*)\Big)\right] &= \mathbb{E}\left[\sum_{t=1}^{\tau}\Big(\ell_t(X_0,A_t)-\ell_t(X_0,a^*)\Big)\right] \\ &\le \mathbb{E}\left[\sum_{t=1}^{\tau}\Big(\ell_t(X_0,A_t)-\ell_t(X_0,\pi^*(X_0))\Big)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{\tau}\Big(\langle X_0,\widehat{\boldsymbol{\theta}}_{t,A_t}\rangle-\langle X_0,\widehat{\boldsymbol{\theta}}_{t,\pi^*(X_0)}\rangle\Big)\right] \\ &= \mathbb{E}\left[\widehat{R}_T(X_0)\right] \\ &\le 2\sum_{t=1}^{\tau}\mathbb{E}_t\left[\gamma_t\right] + \mathbb{E}\left[\frac{\ln K}{\eta_\tau}\right] + \sum_{t=1}^{\tau}\mathbb{E}_t\left[\eta_t\sum_{a=1}^{K}\pi_t(a|X_0)\langle X_0,\widehat{\boldsymbol{\theta}}_{t,a}\rangle^2\right] \\ &\le 2c\cdot\mathbb{E}\left[\sum_{t=1}^{\tau}\frac{\eta_t}{q_t}\right] + \mathbb{E}\left[\frac{\ln K}{\eta_\tau}\right] + 3Kd\cdot\mathbb{E}\left[\sum_{t=1}^{\tau}\frac{\eta_t}{q_t}\right] \\ &\le (2c+3Kd)\sqrt{4\ln K\sum_{t=1}^{\tau}\frac{1}{q_t}} + \sqrt{\ln K\sum_{t=1}^{\tau}\frac{1}{q_t}} + \frac{2c\ln K}{\min_{t\le\tau} q_t} \end{aligned}$$

$$\leq \sqrt{(4(2c+3Kd)^2+1)\ln K \sum_{t=1}^{\tau} \frac{1}{q_t} + \frac{2c\ln K}{\min_{t\leq\tau} q_t}}$$

$$\leq \sqrt{36K^2\left(d+\frac{1}{\lambda_{\min}(\mathbf{\Sigma})}\right)^2 \ln(K)\sum_{t=1}^{\tau}\frac{1}{q_t} + \frac{\frac{2K}{\lambda_{\min}(\mathbf{\Sigma})}\ln K}{\min_{t\leq\tau} q_t}},$$

where the first and second equalities follow from the property of X_0 and the fact that $\widehat{\boldsymbol{\theta}}_{t,a}$ is unbiased for all t and a, the first inequality follows from the definition of the optimal policy $\pi^*(X_0)$, the second inequality follows from Lemma 10, and third inequality follows from the definition γ_t and Lemma 11, the fourth inequality follows from (25) and the definition of η_t. Lastly, we have the statement plugging in the definition of $c = \frac{K}{\lambda_{\min}(\mathbf{\Sigma})}$.

□

F Appendix for Data-dependent Bounds

In this section, we describe how to find a positive semidefinite matrix $\mathbf{S} \in \mathbb{R}^{d\times d}$ to compute a loss predictor $\boldsymbol{m}_{t,a}$ in (9) for each round t and $a \in [K]$, and provide omitted proofs for both Corollary 1 and Proposition 1. Combining Proposition 5, 6, and Proposition 1 immediately implies Theorem 1.

F.1 Concrete choice for a loss predictor

As in Ito et al. [2020], if we have the prior knowledge of the support of $\mathcal{D}$, i.e., context space $\mathcal{X}$, we can find a appropriate matrix $\mathbf{S}$ such that $\|\boldsymbol{m}^*\|_{\mathbf{S}}^2 = \mathcal{O}(d)$ for any vector $\boldsymbol{m}^* \in \mathcal{M}$, and $\max_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x}\|_{\mathbf{S}^{-1}}^2 = \mathcal{O}(d)$. $\mathcal{X}_{\text{span}} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d\} \subseteq \mathcal{X}$ is said to be 2-barycentric spanner for $\mathcal{X}$ if each $\boldsymbol{x} \in \mathcal{X}$ can be expressed as linear combination of elements in $\mathcal{X}_{\text{span}}$ with coefficients in $[-2, 2]$. Define $\mathbf{S} \in \mathbb{R}^{d\times d}$ as

$$\mathbf{M} = (\boldsymbol{x}_1\boldsymbol{x}_2\cdots\boldsymbol{x}_d), \quad \mathbf{S} = \mathbf{M}\mathbf{M}^\top = \sum_{i=1}^{d} \boldsymbol{x}_i\boldsymbol{x}_i^\top. \tag{26}$$

Then, for $\boldsymbol{m} \in \mathcal{M}$, we can easily confirm $\|\boldsymbol{m}\|_{\mathbf{S}}^2 = \boldsymbol{m}^\top \left(\sum_{i=1}^{d} \boldsymbol{x}_i\boldsymbol{x}_i^\top\right)\boldsymbol{m} \leq d$ and $\|\boldsymbol{x}\|_{\mathbf{S}^{-1}}^2 = \boldsymbol{x}^\top \left(\mathbf{M}^{-1}\right)^\top \mathbf{M}^{-1}\boldsymbol{x} = \boldsymbol{u}^\top\boldsymbol{u} \leq 4d$ using some $\boldsymbol{u} \in [-2, 2]^d$ such that $\boldsymbol{x} = \mathbf{M}\boldsymbol{u}$. Due to Proposition 2.4 in Awerbuch and Kleinberg [2004], computation of 2-barycentric spanner for $\mathcal{X}$ can be done in polynomial time, making $O(d^2 \ln d)$-call for linear optimization oracle over $\mathcal{X}$.

F.2 Proof of Corollary 1

We prove Corollary 1 based on Lemma 12 with a concrete choice of a loss predictor. Lemma 12 provides the upper bound of $\mathbb{E}\left[\sum_{t=1}^{T} \xi_{t,A_t}^2\right]$ if we choose $\boldsymbol{m}_{t,a}$ by (9).

Lemma 12. *Let $\mathcal{M} := \{\boldsymbol{m} \in \mathbb{R}^d \mid \langle \boldsymbol{x}, \boldsymbol{m}\rangle \leq 1,\ \forall \boldsymbol{x} \in \mathcal{X}\}$. For $a \in [K]$ and any positive semi-definite matrix $\mathbf{S} \in \mathbb{R}^{d\times d}$, define the predictor $\boldsymbol{m}_{t,a}$ as*

$$\boldsymbol{m}_{t,a} \in \operatorname*{arg\,min}_{\boldsymbol{m}\in\mathcal{M}} \left\{ \|\boldsymbol{m}\|_{\mathbf{S}}^2 + \sum_{j=1}^{t-1} \mathbb{1}[A_j = a]\left(\langle \boldsymbol{\theta}_{j,a} - \boldsymbol{m}, X_j\rangle\right)^2 \right\}.$$

Then, for any $\boldsymbol{m}^* \in \mathcal{M}$, it holds that

$$\mathbb{E}\left[\sum_{t=1}^{T} \xi_{t,A_t}^2\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \left(\langle \boldsymbol{\theta}_{t,A_t} - \boldsymbol{m}^*, X_t\rangle\right)^2\right] + K\|\boldsymbol{m}^*\|_S^2 + 8Kd\ln\left(1 + \frac{T}{d}\max_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x}\|_{\mathbf{S}^{-1}}^2\right),$$

where $\xi_{t,A_t} = (\langle \boldsymbol{\theta}_{t,A_t} - \boldsymbol{m}_{t,A_t}, X_t\rangle)$.

Proof of Lemma 12. The proof can be shown in a proof similar to Lemma 3 of Ito et al. [2020] and Theorem 11.7 of Cesa-Bianchi and Lugosi [2006], by carefully considering contexts and definition of the predictor of $\boldsymbol{m}_{t,a}$.

For any $a \in [K]$ and any $\boldsymbol{m}^* \in \mathcal{M}$, we first need to show

$$\sum_{t=1}^{T} \mathbb{1}[A_t = a] \langle \boldsymbol{\theta}_{t,a} - \boldsymbol{m}_{t,a}, X_t \rangle^2 \leq \sum_{t=1}^{T} \mathbb{1}[A_t = a] \left(\langle \boldsymbol{\theta}_{t,a} - \boldsymbol{m}^*, X_t \rangle \right)^2 + \|\boldsymbol{m}^*\|_S^2 + 8 \sum_{t=1}^{T} \|X_t\|^2_{\mathbf{G}_t^{-1}} . \tag{27}$$

From this, we have that

$$\sum_{t=1}^{T} \sum_{a \in [K]} \mathbb{1}[A_t = a] \langle \boldsymbol{\theta}_{t,a} - \boldsymbol{m}_{t,a}, X_t \rangle^2 \leq \sum_{t=1}^{T} \sum_{a \in [K]} \mathbb{1}[A_t = a] \langle \boldsymbol{\theta}_{t,a} - \boldsymbol{m}^*, X_t \rangle^2 + K \|\boldsymbol{m}^*\|_S^2 + 8K \sum_{t=1}^{T} \|X_t\|^2_{\mathbf{G}_t^{-1}} .$$

Therefore, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{\theta}_{t,A_t} - \boldsymbol{m}_{t,A_t}, X_t \rangle^2\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{A_t \sim Q_t}[\langle \boldsymbol{\theta}_{t,A_t} - \boldsymbol{m}_{t,A_t}, X_t \rangle^2]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \sum_{a \in [K]} \mathbb{1}[A_t = a] \langle \boldsymbol{\theta}_{t,a} - \boldsymbol{m}_{t,a}, X_t \rangle^2\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \sum_{a \in [K]} \mathbb{1}[A_t = a] \langle \boldsymbol{\theta}_{t,a} - \boldsymbol{m}^*, X_t \rangle^2 + K \|\boldsymbol{m}^*\|_S^2 + 8K \sum_{t=1}^{T} \|X_t\|^2_{\mathbf{G}_t^{-1}}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{\theta}_{t,A_t} - \boldsymbol{m}^*, X_t \rangle^2\right] + K \|\boldsymbol{m}^*\|_S^2 + 8K \mathbb{E}\left[\sum_{t=1}^{T} \|X_t\|^2_{\mathbf{G}_t^{-1}}\right] .
\end{aligned} \tag{28}$$

For $t = 0, 1, \ldots, T$, we define convex functions $f_t : \mathcal{M} \to \mathbb{R}$ and $F_t : \mathcal{M} \to \mathbb{R}$ as follows:

$$\begin{aligned}
f_0(\boldsymbol{m}) &= \frac{1}{2} \|\boldsymbol{m}\|_{\mathbf{S}}^2, \\
f_t(\boldsymbol{m}) &= \frac{1}{2} \mathbb{1}[A_t = a] \left(\langle \boldsymbol{\theta}_{t,a} - \boldsymbol{m}, X_t \rangle \right)^2 && (t \in [T]), \\
F_t(\boldsymbol{m}) &= \sum_{j=0}^{t} f_j(\boldsymbol{m}) && (t \in \{0, 1, \ldots, T\}).
\end{aligned}$$

Then, the definition of $\boldsymbol{m}_{t,a}$ in (9) can be rewritten as:

$$\boldsymbol{m}_{t,a} \in \operatorname*{argmin}_{\boldsymbol{m} \in \mathcal{M}} F_{t-1}(\boldsymbol{m}). \tag{29}$$

By applying this fact repeatedly, we can derive the following for arbitrary $\boldsymbol{m}^* \in \mathcal{M}$.

$$\begin{aligned}
F_T(\boldsymbol{m}^*) &\geq F_T(\boldsymbol{m}_{T+1,a}) = F_{T-1}(\boldsymbol{m}_{T+1,a}) + f_T(\boldsymbol{m}_{T+1,a}) \geq F_{T-1}(\boldsymbol{m}_{t,a}) + f_T(\boldsymbol{m}_{T+1,a}) \\
&= f_{T-2}(\boldsymbol{m}_{t,a}) + f_{T-1}(\boldsymbol{m}_{t,a}) + f_T(\boldsymbol{m}_{T+1,a}) \geq \cdots \geq f_0(\boldsymbol{m}_{1,a}) + \sum_{t=1}^{T} f_t(\boldsymbol{m}_{T+1,a}) \\
&\geq \sum_{t=1}^{T} f_t(\boldsymbol{m}_{T+1,a}) .
\end{aligned}$$

From this, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{1}[A_t = a] \left(\langle \boldsymbol{\theta}_{t,a} - \boldsymbol{m}_{t,a}, X_t\rangle\right)^2 - \sum_{t=1}^{T} \mathbb{1}[A_t = a] \left(\langle \boldsymbol{\theta}_{t,a} - \boldsymbol{m}^*, X_t\rangle\right)^2 \\
&= 2\sum_{t=1}^{T} f_t\left(\boldsymbol{m}_{t,a}\right) - 2\sum_{t=1}^{T} f_t\left(\boldsymbol{m}^*\right) \\
&= 2\sum_{t=1}^{T} f_t\left(\boldsymbol{m}_{t,a}\right) - 2\left(F_T\left(\boldsymbol{m}^*\right) - f_0\left(\boldsymbol{m}^*\right)\right) \leq 2 f_0\left(\boldsymbol{m}^*\right) + 2\sum_{t=1}^{T}\left(f_t\left(\boldsymbol{m}_{t,a}\right) - f_t\left(\boldsymbol{m}_{T+1,a}\right)\right) \\
&= \|\boldsymbol{m}^*\|_S^2 + 2\sum_{t=1}^{T}\left(f_t\left(\boldsymbol{m}_{t,a}\right) - f_t\left(\boldsymbol{m}_{T+1,a}\right)\right)
\end{aligned}$$

We next show

$$f_t\left(\boldsymbol{m}_{t,a}\right) - f_t\left(\boldsymbol{m}_{T+1,a}\right) \leq 4\left\|X_t\right\|_{\mathbf{G}_t^{-1}}^2,$$

where we define positive semi-definite matrices $\mathbf{G}_t \in \mathbb{R}^{d\times d}$ for $t = 0, 1, \ldots, T$ by

$$\mathbf{G}_t = \mathbf{S} + \sum_{j=1}^{t} X_j X_j^\top.$$

For positive definite matrix $\mathbf{S}$, $f_0(\boldsymbol{m})$ is strongly convex with respect to the norm $\|\boldsymbol{u}\|_{\mathbf{S}}^2$. Also note that $f_t(\boldsymbol{m})$ for $t \in [T]$ is a convex function. Therefore, F_t is $\mathbf{G}_t$-strongly convex, i.e., it holds for any $\boldsymbol{m}, \boldsymbol{m}' \in \mathcal{M}$ that

$$F_t\left(\boldsymbol{m}'\right) \geq F_t(\boldsymbol{m}) + \langle \nabla F_t(\boldsymbol{m}), \boldsymbol{m}' - \boldsymbol{m}\rangle + \|\boldsymbol{m}' - \boldsymbol{m}\|_{\mathbf{G}_t}^2. \tag{30}$$

Further, (29) implies that

$$\left\langle \nabla F_{t-1}\left(\boldsymbol{m}_{t,a}\right), \boldsymbol{m} - \boldsymbol{m}_{t,a}\right\rangle \geq 0 \tag{31}$$

for any $\boldsymbol{m} \in \mathcal{M}$ and $t \in [T]$. From (30) and this inequality, we can show that

$$\begin{aligned}
&f_t\left(\boldsymbol{m}_{t,a}\right) - f_t\left(\boldsymbol{m}_{T+1,a}\right) \\
&= F_t\left(\boldsymbol{m}_{t,a}\right) - F_t\left(\boldsymbol{m}_{T+1,a}\right) - F_{t-1}\left(\boldsymbol{m}_{t,a}\right) + F_{t-1}\left(\boldsymbol{m}_{T+1,a}\right) \\
&\leq \left\langle \nabla F_t\left(\boldsymbol{m}_{t,a}\right), \boldsymbol{m}_{t,a} - \boldsymbol{m}_{T+1,a}\right\rangle - \left\|\boldsymbol{m}_{t,a} - \boldsymbol{m}_{T+1,a}\right\|_{\mathbf{G}_t}^2 + \left\langle \nabla F_{t-1}\left(\boldsymbol{m}_{T+1,a}\right), \boldsymbol{m}_{T+1,a} - \boldsymbol{m}_{t,a}\right\rangle \\
&\leq \left\langle \nabla F_t\left(\boldsymbol{m}_{t,a}\right) - \nabla F_{t-1}\left(\boldsymbol{m}_{t,a}\right), \boldsymbol{m}_{t,a} - \boldsymbol{m}_{T+1,a}\right\rangle \\
&\qquad + \left\langle \nabla F_{t-1}\left(\boldsymbol{m}_{T+1,a}\right) - \nabla F_t\left(\boldsymbol{m}_{T+1,a}\right), \boldsymbol{m}_{T+1,a} - \boldsymbol{m}_{t,a}\right\rangle - \left\|\boldsymbol{m}_{t,a} - \boldsymbol{m}_{T+1,a}\right\|_{\mathbf{G}_t}^2 \\
&= \left\langle \nabla f_t\left(\boldsymbol{m}_{t,a}\right), \boldsymbol{m}_{t,a} - \boldsymbol{m}_{T+1,a}\right\rangle - \left\|\boldsymbol{m}_{t,a} - \boldsymbol{m}_{T+1,a}\right\|_{\mathbf{G}_t}^2 - \left\langle \nabla f_t\left(\boldsymbol{m}_{T+1,a}\right), \boldsymbol{m}_{T+1,a} - \boldsymbol{m}_{t,a}\right\rangle \\
&= \left\langle \nabla f_t\left(\boldsymbol{m}_{t,a}\right) + \nabla f_t\left(\boldsymbol{m}_{T+1,a}\right), \boldsymbol{m}_{t,a} - \boldsymbol{m}_{T+1,a}\right\rangle - \left\|\boldsymbol{m}_{t,a} - \boldsymbol{m}_{T+1,a}\right\|_{\mathbf{G}_t}^2 \\
&\leq \left\|\nabla f_t\left(\boldsymbol{m}_{t,a}\right) + \nabla f_t\left(\boldsymbol{m}_{T+1,a}\right)\right\|_{\mathbf{G}_t^{-1}} \left\|\boldsymbol{m}_{t,a} - \boldsymbol{m}_{T+1,a}\right\|_{\mathbf{G}_t} - \left\|\boldsymbol{m}_{t,a} - \boldsymbol{m}_{T+1,a}\right\|_{\mathbf{G}_t}^2 \\
&\leq \frac{1}{4}\left\|\nabla f_t\left(\boldsymbol{m}_{t,a}\right) + \nabla f_t\left(\boldsymbol{m}_{T+1,a}\right)\right\|_{\mathbf{G}_t^{-1}}^2 = \frac{1}{4}\left\|\left(\langle \boldsymbol{m}_{t,a} - \boldsymbol{\theta}_{t,a}, X_t\rangle + \langle \boldsymbol{m}_{T+1,a} - \boldsymbol{\theta}_{t,a}, X_t\rangle\right) X_t\right\|_{\mathbf{G}_t^{-1}}^2 \\
&\leq 4\left\|X_t\right\|_{\mathbf{G}_t^{-1}}^2,
\end{aligned}$$

where the first and second inequalities follow from (30) and (31) respectively, the third inequality follows from the Cauchy-Schwarz inequality, the forth inequality follows from the fact that $a^2 - ab + b^2/4 = (a - b/2)^2 \geq 0$ for $a, b \in \mathbb{R}$. Therefore, we obtain

$$\sum_{t=1}^{T} \mathbb{1}[A_t = a] \left\langle \boldsymbol{\theta}_{t,a} - \boldsymbol{m}_{t,a}, X_t\right\rangle^2 - \sum_{t=1}^{T} \mathbb{1}[A_t = a] \left\langle \boldsymbol{\theta}_{t,a} - \boldsymbol{m}^*, X_t\right\rangle^2 \leq \|\boldsymbol{m}^*\|_S^2 + 8\sum_{t=1}^{T}\left\|X_t\right\|_{\mathbf{G}_t^{-1}}^2,$$

which is (27). We next show

$$\sum_{t=1}^{T} \|X_t\|_{\mathbf{G}_t^{-1}}^2 \leq d \ln \left(1 + \frac{T}{d} \max_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x}\|_{\mathbf{S}^{-1}}^2\right). \tag{32}$$

Using Lemma 11.11 and similar analysis of Theorem 11.7 in Cesa-Bianchi and Lugosi [2006], we have

$$\begin{aligned}
&\ln\det \mathbf{G}_t - \ln\det \mathbf{G}_{t-1} = -\left(\ln\det\left(\mathbf{G}_t - X_t X_t^\top\right) - \ln\det \mathbf{G}_t\right) \\
&= -\ln\det\left(\mathbf{G}_t^{-\frac{1}{2}}\left(\mathbf{G}_t - X_t X_t^\top\right)\mathbf{G}_t^{-\frac{1}{2}}\right) = -\ln\det\left(I - \mathbf{G}_t^{-\frac{1}{2}} X_t X_t^\top \mathbf{G}_t^{-\frac{1}{2}}\right) \\
&= -\ln\left(1 - \left\|\mathbf{G}_t^{-\frac{1}{2}} X_t\right\|_2^2\right) \geq \left\|\mathbf{G}_t^{-\frac{1}{2}} X_t\right\|_2^2 = \|X_t\|_{\mathbf{G}_t^{-1}}^2,
\end{aligned}$$

where the forth equality holds since the matrix $\left(I - \mathbf{G}_t^{-\frac{1}{2}} X_t X_t^\top \mathbf{G}_t^{-\frac{1}{2}}\right)$ has eigenvalues $\lambda_1' = 1 - \left\|\mathbf{G}_t^{-\frac{1}{2}} X_t\right\|_2^2$ and $\lambda_2' = \lambda_3' = \cdots = \lambda_d' = 1$, and the inequality follows from $\ln(1+y) \leq y$ for $y > -1$. Therefore, we obtain

$$X_t^\top \mathbf{G}_t^{-1} X_t \leq \ln \frac{\det \mathbf{G}_t}{\det \mathbf{G}_{t-1}}.$$

Let $\lambda_1, \lambda_2, \ldots, \lambda_d \geq 0$ be eigenvalues of $\sum_{t=1}^T \mathbf{S}^{-\frac{1}{2}} X_t X_t^\top \mathbf{S}^{-\frac{1}{2}}$. Then, we have

$$\sum_{t=1}^{T} \|X_t\|_{\mathbf{G}_t^{-1}}^2 \leq \ln\det \mathbf{G}_t - \ln\det \mathbf{G}_0 = \ln\det\left(I + \sum_{t=1}^{T} \mathbf{S}^{-\frac{1}{2}} X_t X_t^\top \mathbf{S}^{-\frac{1}{2}}\right) = \sum_{i=1}^{d} \ln\left(1 + \lambda_i\right).$$

Since we have $\sum_{i=1}^d \lambda_i = \mathrm{tr}\left(\sum_{t=1}^T \mathbf{S}^{-\frac{1}{2}} X_t X_t^\top \mathbf{S}^{-\frac{1}{2}}\right) = \sum_{t=1}^T \|X_t\|_{\mathbf{S}^{-1}}^2 \leq T \max_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x}\|_{\mathbf{S}^{-1}}^2$, it holds that $\sum_{i=1}^d \ln\left(1+\lambda_i\right) \leq d\ln\left(1 + \frac{T}{d}\max_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x}\|_{\mathbf{S}^{-1}}^2\right)$ which gives us (32). Combining it with (28), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{\theta}_{t,A_t} - \boldsymbol{m}_{t,A_t}, X_t\rangle^2\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{\theta}_{t,A_t} - \boldsymbol{m}^*, X_t\rangle^2\right] + K\|\boldsymbol{m}^*\|_S^2 + 8K\mathbb{E}\left[\sum_{t=1}^{T} \|X_t\|_{\mathbf{G}_t^{-1}}^2\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{\theta}_{t,A_t} - \boldsymbol{m}^*, X_t\rangle^2\right] + K\|\boldsymbol{m}^*\|_S^2 + 8Kd\ln\left(1 + \frac{T}{d}\max_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x}\|_{\mathbf{S}^{-1}}^2\right),
\end{aligned}$$

which concludes the proof. □

We are ready to prove Corollary 1.

Proof of Corollary 1. Since we choose $\mathbf{S}$ by (26), it holds that $\|\boldsymbol{m}^*\|_{\mathbf{S}}^2 = \mathcal{O}(d)$ and $\max_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x}\|_{\mathbf{S}^{-1}}^2 = \mathcal{O}(d)$.

Then, by Lemma 12 and Theorem 1, it holds that for any $\boldsymbol{m}^* \in \mathcal{M}$,

$$
\begin{aligned}
R_T &= \mathcal{O}\left(\sqrt{\kappa_1(d,K,T)\mathbb{E}\left[\sum_{t=1}^T \xi_{t,A_t}^2\right]\ln^2 T} + \kappa_2(d,K,T)\ln^2(T)\right) \\
&= \mathcal{O}\left(\sqrt{\kappa_1(d,K,T)\left(\mathbb{E}\left[\sum_{t=1}^T (\langle \boldsymbol{\theta}_{t,A_t} - \boldsymbol{m}^*, X_t\rangle)^2\right] + K\|\boldsymbol{m}^*\|_S^2 + Kd\ln\left(1+\frac{T}{d}\max_{\boldsymbol{x}\in\mathcal{X}}\|\boldsymbol{x}\|^2_{\mathbf{S}^{-1}}\right)\right)\ln^2 T} + \kappa_2(d,K,T)\ln^2(T)\right) \\
&= \mathcal{O}\left(\sqrt{\kappa_1(d,K,T)\left(\mathbb{E}\left[\sum_{t=1}^T (\langle \boldsymbol{\theta}_{t,A_t} - \boldsymbol{m}^*, X_t\rangle)^2\right] + Kd\ln(1+T)\right)\ln^2 T} + \kappa_2(d,K,T)\ln^2(T)\right) \\
&= \mathcal{O}\left(Kd\ln(dKT)\ln^2(T)\sqrt{\left(\mathbb{E}\left[\sum_{t=1}^T (\langle \boldsymbol{\theta}_{t,A_t} - \boldsymbol{m}^*, X_t\rangle)^2\right] + Kd\ln(T)\right)} + (dK)^{3/2}\ln(dKT)\ln^3(T)\right) \\
&= \mathcal{O}\left(Kd\ln(dKT)\ln^2(T)\sqrt{\mathbb{E}\left[\sum_{t=1}^T (\langle \boldsymbol{\theta}_{t,A_t} - \boldsymbol{m}^*, X_t\rangle)^2\right]} + (dK)^{3/2}\ln^{3/2}(dKT)\ln^3(T)\right) \\
&= \widetilde{\mathcal{O}}\left(Kd\sqrt{\mathbb{E}\left[\sum_{t=1}^T (\langle \boldsymbol{\theta}_{t,A_t} - \boldsymbol{m}^*, X_t\rangle)^2\right]} + (dK)^{3/2}\right),
\end{aligned}
$$

in the adversarial regime. Therefore, we obtain

$$
R_T = \widetilde{\mathcal{O}}\left(Kd\sqrt{\mathbb{E}\left[\sum_{t=1}^T (\langle \boldsymbol{\theta}_{t,A_t} - \boldsymbol{m}^*, X_t\rangle)^2\right]} + (dK)^{3/2}\right) = \widetilde{\mathcal{O}}\left(Kd\sqrt{\Lambda} + (dK)^{3/2}\right). \tag{33}
$$

On the other hand, for $\boldsymbol{m}^* = \mathbf{0}$, we also have that

$$
\begin{aligned}
\frac{R_T}{\widehat{c}} &\leq \sqrt{\kappa_1(d,K,T)\left(\mathbb{E}\left[\sum_{t=1}^T (\langle X_t, \boldsymbol{\theta}_{t,A_t}\rangle)^2\right] + Kd\ln(1+T)\right)\ln^2 T} + \kappa_2(d,K,T)\ln^2(T) \\
&\leq \sqrt{\mathbb{E}\left[\sum_{t=1}^T \langle X_t, \boldsymbol{\theta}_{t,A_t}\rangle\right] + Kd\ln(1+T)}\sqrt{\kappa_1(d,K,T)\ln^2 T} + \kappa_2(d,K,T)\ln^2(T) \\
&\leq \sqrt{\kappa_1(d,K,T)\ln^2 T}\sqrt{\mathbb{E}\left[\sum_{t=1}^T \langle X_t, \boldsymbol{\theta}_{t,A_t}\rangle\right]} + \sqrt{\kappa_1(d,K,T)\ln^2 T}\sqrt{Kd\ln(1+T)} + \kappa_2(d,K,T)\ln^2(T),
\end{aligned}
$$

where $\widehat{c}$ is a universal constant and the second inequality follows from $0 \leq \mathbb{E}[\langle X_t, \boldsymbol{\theta}_{t,A_t}\rangle] \leq 1$. By the definition of $R_T = \mathbb{E}\left[\sum_{t=1}^T \langle X_t, \boldsymbol{\theta}_{t,A_t}\rangle\right] - L^*$, and solving the quadratic inequality for $\mathbb{E}\left[\sum_{t=1}^T \langle X_t, \boldsymbol{\theta}_{t,A_t}\rangle\right]$, we obtain

$$
\begin{aligned}
&\sqrt{\mathbb{E}\left[\sum_{t=1}^T \langle X_t, \boldsymbol{\theta}_{t,A_t}\rangle\right]} \\
&< \frac{\sqrt{\kappa_1(d,K,T)\ln^2 T} + \sqrt{\kappa_1(d,K,T)\ln^2 T + 4\left(L^* + \sqrt{\kappa_1(d,K,T)\ln^2 T}\sqrt{Kd\ln(1+T)} + \kappa_2(d,K,T)\ln^2(T)\right)}}{2} \\
&\leq \sqrt{\frac{\kappa_1(d,K,T)\ln^2 T + \kappa_1(d,K,T)\ln^2 T + 4\left(L^* + \sqrt{\kappa_1(d,K,T)\ln^2 T}\sqrt{Kd\ln(1+T)} + \kappa_2(d,K,T)\ln^2(T)\right)}{2}}.
\end{aligned}
$$

This indicates that

$$\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\langle X_t,\boldsymbol{\theta}_{t,A_t}\rangle\right]} = \mathcal{O}\left(\sqrt{L^* + K^2d^2\ln^2(dKT)\ln^4(T)}\right)$$
$$= \widetilde{\mathcal{O}}\left(\sqrt{L^* + K^2d^2}\right).$$

Therefore, in this case, we obtain

$$R_T = \mathcal{O}\left(\sqrt{\kappa_1(d,K,T)\ln^2 T}\sqrt{L^* + K^2d^2\ln^2(dKT)\ln^4(T)} + \sqrt{\kappa_1(d,K,T)\ln^2 T}\sqrt{Kd\ln(1+T)} + \kappa_2(d,K,T)\ln^2(T)\right)$$
$$= \mathcal{O}\left(Kd\ln(dKT)\ln^2(T)\sqrt{L^*} + K^2d^2\ln^2(dKT)\ln^4(T)\right)$$
$$= \widetilde{\mathcal{O}}\left(Kd\sqrt{L^*} + K^2d^2\right).$$

From (33) and this, we conclude that

$$R_T = \widetilde{\mathcal{O}}\left(Kd\sqrt{\min\{L^*,\overline{\Lambda}\}} + K^2d^2\right).$$

□

We note that instead computing $\boldsymbol{m}_{t,a}$ in (9) at round t for each $a \in [K]$, we can still get the first-order regret bound in the adversarial regime, just by setting $\boldsymbol{m}_{t,a} = \mathbf{0} \in \mathbb{R}^d$.

Corollary 2. *Let $\kappa_1(d,K,T) = \mathcal{O}\left(K^2d^2\ln^2(dKT)\ln^2(T)\right)$ and $\kappa_2(d,K,T) = \mathcal{O}\left((dK)^{3/2}\ln(dKT)\ln(T)\right)$. Combining Algorithms 1, 3, and 5 results in the following the regret bound*

$$R_T = \mathcal{O}\left(\sqrt{\kappa_1(d,K,T)\ln^2 T\cdot L^*} + \ln T^{3/2}\kappa_1(d,K,T)^{3/4} + \kappa_2(d,K,T)\ln^2(T)\right).$$

in the adversarial regime, and

$$R_T = \mathcal{O}\left(\frac{\kappa_1(d,K,T)\ln(T)}{\Delta_{\min}} + \sqrt{\frac{\kappa_1(d,K,T)\ln TC}{\Delta_{\min}}} + \kappa_2(d,K,T)\ln(T)\ln(C\Delta_{\min}^{-1})\right)$$

in the corrupted stochastic regime.

Proof of Corollary 2. Taking $\boldsymbol{m}_{t,a} = \mathbf{0}$ in Theorem 1, for a universal constant $\widehat{c} > 0$, we have

$$\frac{R_T}{\widehat{c}} \le \sqrt{\kappa_1(d,K,T)\ln^2 T}\cdot\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}(\langle X_t,\boldsymbol{\theta}_{t,A_t}\rangle)^2\right]} + \kappa_2(d,K,T)\ln^2(T)$$
$$\le \sqrt{\kappa_1(d,K,T)\ln^2 T}\cdot\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\langle X_t,\boldsymbol{\theta}_{t,A_t}\rangle\right]} + \kappa_2(d,K,T)\ln^2(T),$$

where the second inequality follows from $0 \le \ell_t(X_t, A_t) \le 1$. By the definition of $R_T = \mathbb{E}\left[\sum_{t=1}^{T}\langle X_t,\boldsymbol{\theta}_{t,A_t}\rangle\right] - L^*$, and solving the quadratic inequality for $\mathbb{E}\left[\sum_{t=1}^{T}\langle X_t,\boldsymbol{\theta}_{t,A_t}\rangle\right]$, we obtain $\mathbb{E}\left[\sum_{t=1}^{T}\langle X_t,\boldsymbol{\theta}_{t,A_t}\rangle\right] = \mathcal{O}(L^* + \ln T\sqrt{\kappa_1(d,K,T)})$. Therefore, we have

$$R_T = \mathcal{O}\left(\sqrt{\kappa_1(d,K,T)\ln^2 T\cdot L^*} + (\ln T)^{3/2}\kappa_1(d,K,T)^{3/4} + \kappa_2(d,K,T)\ln^2(T)\right),$$

which completes the proof.

□

F.3 Proof of Proposition 1

Before we state the proof, we introduce the concentration property of a log-concave distribution, which is proved in Lemma 1 in Ito et al. [2020].

Lemma 13 (Lemma 1, Ito et al. [2020])**.** *If y follows a log-concave distribution p over $\mathbb{R}^d$ and $\mathbb{E}_{y\sim p}[yy^\top] \preceq I$, we have*

$$\Pr[\|y\|_2^2 \geq d\alpha^2] \leq d\exp(1-\alpha)$$

for arbitrary $\alpha \geq 0$.

In order to proceed with further analysis, we introduce several definitions. For a probability vector $r \in \Delta([K])$ and d-dimensional context $\boldsymbol{x} \in \mathcal{X}$, we denote the dK-dimentional vector $\boldsymbol{z}(r, \boldsymbol{x}) := (r_1 \cdot \boldsymbol{x}^\top, \ldots, r_K \cdot \boldsymbol{x}^\top)^\top \in \mathbb{R}^{dK}$. We define the $dK \times dK$ matrix $\overline{\boldsymbol{\Sigma}_{\mathbf{b}}}(t) := \mathrm{diag}_{a\in[K]}(\overline{\boldsymbol{\Sigma}}_{t,a}) \in \mathbb{R}^{dK} \times \mathbb{R}^{dK}$ as a block diagonal arrangement of the covariance matrices per arm, where $\overline{\boldsymbol{\Sigma}}_{t,a}$ is given in (5). Similarly, we also define $\widetilde{\boldsymbol{\Sigma}_{\mathbf{b}}}(t) := \mathrm{diag}_{a\in[K]}(\widetilde{\boldsymbol{\Sigma}}_{t,a}) \in \mathbb{R}^{dK} \times \mathbb{R}^{dK}$, where $\widetilde{\boldsymbol{\Sigma}}_{t,a}$ is given in (7). Using these notation, we can rewrite the $\widetilde{p}_t(r|\boldsymbol{x})$ for $r \in \Delta([K])$ and a context $\boldsymbol{x} \in \mathcal{X}$ as follows:

$$\widetilde{p}_t(r|\boldsymbol{x}) = \frac{p_t(r|\boldsymbol{x})\mathbb{1}\left[\sum_{a=1}^K r_a^2 \|\boldsymbol{x}\|^2_{\overline{\boldsymbol{\Sigma}}_{t,a}^{-1}} \leq dK\widetilde{\gamma}_t^2\right]}{\mathbb{P}_{y\sim p_t(\cdot|\boldsymbol{x})}\left[\sum_{a=1}^K y_a^2 \|\boldsymbol{x}\|^2_{\overline{\boldsymbol{\Sigma}}_{t,a}^{-1}} \leq dK\widetilde{\gamma}_t^2\right]} = \frac{p_t(r|\boldsymbol{x})\mathbb{1}\left[\|\boldsymbol{z}(r,\boldsymbol{x})\|^2_{\overline{\boldsymbol{\Sigma}_{\mathbf{b}}}^{-1}(t)} \leq dK\widetilde{\gamma}_t^2\right]}{\mathbb{P}_{y\sim p_t(\cdot|\boldsymbol{x})}\left[\|\boldsymbol{z}(y,\boldsymbol{x})\|^2_{\overline{\boldsymbol{\Sigma}_{\mathbf{b}}}^{-1}(t)} \leq dK\widetilde{\gamma}_t^2\right]}. \tag{34}$$

For a context $\boldsymbol{x}$, we define $Q_t(\boldsymbol{x})$ as a sample generated from $\widetilde{p}_t(\cdot|\boldsymbol{x})$ in (34), and define $Q(\boldsymbol{x})$ as a sample generated from $p_t(\cdot|\boldsymbol{x})$ in (4) wherein X_t is replaced with $\boldsymbol{x}$. Let $\boldsymbol{\theta}_t := (\boldsymbol{\theta}_{t,1}^\top, \ldots, \boldsymbol{\theta}_{t,K}^\top)^\top \in \mathbb{R}^{dK}$, and let its estimate be $\widehat{\boldsymbol{\theta}}_t := (\widehat{\boldsymbol{\theta}}_{t,1}^\top, \ldots, \widehat{\boldsymbol{\theta}}_{t,K}^\top)^\top \in \mathbb{R}^{dK}$. We denote $\boldsymbol{m}_t := (\boldsymbol{m}_{t,1}^\top, \ldots, \boldsymbol{m}_{t,K}^\top)^\top \in \mathbb{R}^{dK}$.

For notational convenience, we also define random vector $Z(\boldsymbol{x})^\top \in \mathbb{R}^{dK}$ for context $\boldsymbol{x}$ and $Q(\boldsymbol{x}) \sim p_t(\cdot|\boldsymbol{x})$ as:

$$Z(\boldsymbol{x}) := \boldsymbol{z}(Q(\boldsymbol{x}), \boldsymbol{x}) = (Q_1(\boldsymbol{x}) \cdot \boldsymbol{x}^\top, \ldots, Q_K(\boldsymbol{x}) \cdot \boldsymbol{x}^\top)^\top \in \mathbb{R}^{dK}.$$

And, we define $\widetilde{Z}_t(\boldsymbol{x})$ for context $\boldsymbol{x}$ and $Q_t(\boldsymbol{x}) \sim \widetilde{p}_t(\cdot|\boldsymbol{x})$ as:

$$\widetilde{Z}_t(\boldsymbol{x}) := \boldsymbol{z}(Q_t(\boldsymbol{x}), \boldsymbol{x}) = (Q_{t,1}(\boldsymbol{x}) \cdot \boldsymbol{x}^\top, \ldots, Q_{t,K}(\boldsymbol{x}) \cdot \boldsymbol{x}^\top)^\top \in \mathbb{R}^{dK}.$$

For the optimal policy $\pi^* \in \Pi$ and context $\boldsymbol{x}$, we define $Z^*(\boldsymbol{x})$ as:

$$Z^*(\boldsymbol{x}) := (\mathbf{0}^\top, \ldots, \boldsymbol{x}^\top, \ldots, \mathbf{0}^\top)^\top \in \mathbb{R}^{dK},$$

where the term of $\boldsymbol{x}$ is placed on $\pi^*(\boldsymbol{x})$-th element and $\mathbf{0} \in \mathbb{R}^d$ is placed on other elements. Finally for the uniform distribution over K-action $\mu_0 = (\frac{1}{K}, \ldots, \frac{1}{K})$, and context $\boldsymbol{x}$, we define $\overline{Z}(\boldsymbol{x})$ as:

$$\overline{Z}(\boldsymbol{x}) := \left(\frac{1}{K}\boldsymbol{x}^\top, \ldots, \frac{1}{K}\boldsymbol{x}^\top\right)^\top \in \mathbb{R}^{dK}.$$

Proof of Proposition 1. Using the above notations, the regret can be decomposed as:

$$\begin{aligned} R_\tau &= \mathbb{E}\left[\sum_{t=1}^{\tau}(\ell_t(X_t, A_t) - \ell_t(X_t, \pi^*(X_t)))\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{\tau}\left\langle \widetilde{Z}_t(\boldsymbol{x}) - Z^*(X_t), \boldsymbol{\theta}_t\right\rangle\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{\tau}\left\langle \widetilde{Z}_t(X_t) - Z(X_t), \boldsymbol{\theta}_t\right\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{\tau}\langle Z(X_t) - Z^*(X_t), \boldsymbol{\theta}_t\rangle\right]. \end{aligned} \tag{35}$$

Following the idea of the auxiliary game as presented in (15), for the optimal policy $\pi^* \in \Pi$, the unbiased estimate of loss vectors $\widehat{\boldsymbol{\theta}}_t$, and a fixed context $\boldsymbol{x} \in \mathcal{X}$, we define

$$\widehat{R}_\tau(\boldsymbol{x}) := \sum_{t=1}^{\tau} \mathbb{E}_t \left[\left\langle Z(\boldsymbol{x}) - Z^*(\boldsymbol{x}), \widehat{\boldsymbol{\theta}}_t \right\rangle \right].$$

Let $X_0 \sim \mathcal{D}$ be a ghost sample drawn independently from the entire interaction history. Then we have

$$\mathbb{E}\left[\sum_{t=1}^{\tau} \langle Z(X_t) - Z^*(X_t), \boldsymbol{\theta}_t \rangle \right] = \mathbb{E}\left[\sum_{t=1}^{\tau} \left\langle Z(X_0) - Z^*(X_0), \widehat{\boldsymbol{\theta}}_t \right\rangle \right] = \mathbb{E}[\widehat{R}_\tau(X_0)], \tag{36}$$

where we used the property of unbiased estimates $\widehat{\boldsymbol{\theta}}_t$ and the fact that X_0 is independent of any past history to constract $\widehat{\boldsymbol{\theta}}_t$.

For further analysis, we introduce some lemmas from the prior analysis. The following lemmas hold for our unbiased estimator $\widehat{\boldsymbol{\theta}}$ and definitions of $\widetilde{\gamma}_t$ and η_t, since we sample $Q(\boldsymbol{x})$ from the distribution $p_t(\cdot|\boldsymbol{x})$ defined in (4) and $Q_t(\boldsymbol{x})$ from the truncated distribution $\widetilde{p}_t(\cdot|\boldsymbol{x})$ defined in (34) for context $\boldsymbol{x}$. We begin with Lemma C.1 of Olkhovskaya et al. [2023], implying that $Z(\boldsymbol{x})$ follows a log-concave distribution under the assumption that the underlying context distribution $\mathcal{D}$ is log-concave.

Lemma 14 (c.f. Lemma C.1 of Olkhovskaya et al. [2023]). *Suppose that $\boldsymbol{z}(q, \boldsymbol{x}) = \sum_{a \in [K]} q_a \varphi(\boldsymbol{x}, a)$ for $q \in \Delta([K])$ and $\varphi(\boldsymbol{x}, a) = (\mathbf{0}^\top, \ldots, \boldsymbol{x}^\top, \ldots, \mathbf{0})$ such that $\boldsymbol{x}$ is on the a-th co-ordinate and $Q(\boldsymbol{x}) \sim p(\cdot|\boldsymbol{x})$ for log-concave $p(\cdot|\boldsymbol{x})$. If $X \sim p_X(\cdot)$ and $p_X(\cdot)$ is log-concave and $Z(X) = \boldsymbol{z}(Q(X), X)$, then $Z(X)$ also follows a log-concave distribution.*

To see that the first term of $\mathbb{E}\left[\sum_{t=1}^{\tau} \left\langle \widetilde{Z}_t(X_t) - Z(X_t), \boldsymbol{\theta}_t \right\rangle\right]$ in (35) is a constant, we make use of Lemma C.2 in Olkhovskaya et al. [2023], which is the analog of Lemma 4 Ito et al. [2020]. This lemma implies that $\widetilde{Z}_t(X_t)$ is close to $Z(X_t)$, and also provides a useful relation between covariance matrices $\overline{\boldsymbol{\Sigma}_{\mathbf{b}}}(t)$ and $\widetilde{\boldsymbol{\Sigma}_{\mathbf{b}}}(t)$. The log-concavity of $Z(X_t)$ is crucial in the proof to utilize its concentration property stated in Lemma 1 of Ito et al. [2020] (Lemma 13).

Lemma 15 (c.f. Lemma C.2 in Olkhovskaya et al. [2023]). *Suppose that $\widetilde{\gamma}_t \geq 4\ln(10dKt)$ and $\langle (r_1 \cdot \boldsymbol{x}^\top, \ldots, r_K \cdot \boldsymbol{x}^\top), \boldsymbol{\theta}_t \rangle \in [-1, 1]$ for any t, a policy $r \in \Delta([K])$ and context $X_t \in \mathcal{X}$. Then, we have*

$$\left| \mathbb{E}_t \left[\left\langle \widetilde{Z}_t(X_t) - Z(X_t), \boldsymbol{\theta}_t \right\rangle \right] \right| \leq \frac{1}{2t^2}.$$

Further, we have

$$\frac{3}{4} \overline{\boldsymbol{\Sigma}_{\mathbf{b}}}(t) \preceq \widetilde{\boldsymbol{\Sigma}_{\mathbf{b}}}(t) \preceq \frac{4}{3} \overline{\boldsymbol{\Sigma}_{\mathbf{b}}}(t). \tag{37}$$

Next, we introduce Lemma 4.4 in Olkhovskaya et al. [2023], the analog of Lemma 5 in Ito et al. [2020], which can be shown via standard the OMD analysis [Rakhlin and Sridharan, 2013].

Lemma 16 (c.f. Lemma 4.4 in Olkhovskaya et al. [2023]). *Assume that $\eta_{t+1} \leq \eta_t$ for all t, let μ_0 be a uniform distribution over $[K]$ and $\psi(y) = \exp(y) - y - 1$. Then, the regret $\widehat{R}_\tau(\boldsymbol{x})$ for fixed $\boldsymbol{x} \in \mathcal{X}$ of Algorithm 1 almost surely satisfies*

$$\widehat{R}_\tau(\boldsymbol{x}) \leq \frac{1}{\tau} \sum_{t=1}^{\tau} \left\langle \overline{Z}(\boldsymbol{x}) - Z^*(\boldsymbol{x}), \widehat{\boldsymbol{\theta}}_t \right\rangle + \frac{K \ln \tau}{\eta_\tau} + \sum_{t=1}^{\tau} \frac{1}{\eta_t} \mathbb{E}_t \left[\psi(-\eta_t \langle Z(\boldsymbol{x}), \widehat{\boldsymbol{\theta}}_t - \boldsymbol{m}_t \rangle) \right]. \tag{38}$$

Next, we introduce Lemma 6 of Ito et al. [2020] to evaluate the third term of RHS of (38).

Lemma 17 (Lemma 6 in Ito et al. [2020]). *If y follows a log-concave distribution over $\mathbb{R}$ and if $\mathbb{E}[y^2] \leq \frac{1}{100}$, we have*

$$\mathbb{E}[\psi(y)] \leq \mathbb{E}[y^2] + 30 \exp\left(-\frac{1}{\sqrt{\mathbb{E}[y^2]}} \right) \leq 2\mathbb{E}[y^2] \text{ where } \psi(x) = \exp(y) - y - 1.$$

Now, we start by evaluating the term $\mathbb{E}_t\left[(-\eta_t\langle Z(X_0), \widehat{\boldsymbol{\theta}}_t - \boldsymbol{m}_t\rangle)^2\right]$. We recall that the definition of $\widehat{\boldsymbol{\theta}}_{t,a}$ is given by

$$\widehat{\boldsymbol{\theta}}_{t,a} := \boldsymbol{m}_{t,a} + \frac{\mathsf{upd}_t}{q_t} Q_{t,a}(X_t)\widetilde{\boldsymbol{\Sigma}}_{t,a}^{-1} X_t \xi_{t,a} \mathbb{1}[A_t = a],$$

where $\xi_{t,a} := (\ell_t(X_t, a) - \langle X_t, \boldsymbol{m}_{t,a}\rangle)$. Then, we have that

$$\begin{aligned}
&\mathbb{E}_{Q(X_0)\sim p_t(\cdot|X_0),\mathsf{upd}_t\sim q_t}\left[(-\eta_t\langle Z(X_0), \widehat{\boldsymbol{\theta}}_t - \boldsymbol{m}_t\rangle)^2 \mid \mathcal{F}_{t-1}\right]\\
&= \mathbb{E}_{\mathsf{upd}_t\sim q_t}\left[\eta_t^2 \frac{\mathsf{upd}_t^2}{q_t^2}\mathbb{E}_t\left[\xi_{t,A_t}^2 Z(X_t)^\top \widetilde{\boldsymbol{\Sigma}_{\mathbf{b}}}^{-1}(t) Z(X_0) Z(X_0)^\top \widetilde{\boldsymbol{\Sigma}_{\mathbf{b}}}^{-1}(t) Z(X_t)\right]\right]\\
&= \eta_t^2 \mathbb{E}_{\mathsf{upd}_t\sim q_t}\left[\frac{\mathsf{upd}_t^2}{q_t^2}\mathbb{E}_t\left[\xi_{t,A_t}^2 Z(X_t)^\top \widetilde{\boldsymbol{\Sigma}_{\mathbf{b}}}^{-1}(t) \overline{\boldsymbol{\Sigma}_{\mathbf{b}}}(t) \widetilde{\boldsymbol{\Sigma}_{\mathbf{b}}}^{-1}(t) Z(X_t)\right]\right]\\
&\le \frac{4}{3}\eta_t^2 \mathbb{E}_{\mathsf{upd}_t\sim q_t}\left[\frac{\mathsf{upd}_t^2}{q_t^2}\mathbb{E}_t\left[\xi_{t,A_t}^2 Z(X_t)^\top \widetilde{\boldsymbol{\Sigma}_{\mathbf{b}}}^{-1}(t) \widetilde{\boldsymbol{\Sigma}_{\mathbf{b}}}(t) \widetilde{\boldsymbol{\Sigma}_{\mathbf{b}}}^{-1}(t) Z(X_t)\right]\right]\\
&= \frac{4}{3}\eta_t^2 \mathbb{E}_{\mathsf{upd}_t\sim q_t}\left[\frac{\mathsf{upd}_t^2}{q_t^2}\mathbb{E}_t\left[\xi_{t,A_t}^2 Z(X_t)^\top \widetilde{\boldsymbol{\Sigma}_{\mathbf{b}}}^{-1}(t) Z(X_t)\right]\right]\\
&\le 2\eta_t^2 \mathbb{E}_{\mathsf{upd}_t\sim q_t}\left[\frac{\mathsf{upd}_t^2}{q_t^2}\mathbb{E}_t\left[\xi_{t,A_t}^2 Z(X_t)^\top \overline{\boldsymbol{\Sigma}_{\mathbf{b}}}^{-1}(t) Z(X_t)\right]\right]\\
&= 2\eta_t^2 \mathbb{E}_{\mathsf{upd}_t\sim q_t}\left[\frac{\mathsf{upd}_t^2}{q_t^2}\mathbb{E}_t\left[\xi_{t,A_t}^2 \|Z(X_t)\|^2_{\overline{\boldsymbol{\Sigma}_{\mathbf{b}}}^{-1}(t)}\right]\right]\\
&\le \frac{2dK\eta_t^2\widetilde{\gamma}_t^2}{q_t}\mathbb{E}_t\left[\xi_{t,A_t}^2\right]\\
&\le \frac{1}{100},
\end{aligned} \tag{39}$$

where the first and second inequalities follow from Lemma 15, the third inequality follows from Line 6 in Algorithm 1 of $\|Z(X_t)\|^2_{\overline{\boldsymbol{\Sigma}_{\mathbf{b}}}^{-1}(t)} \le dK\widetilde{\gamma}_t^2$, and we used $\eta_t \le \frac{2\sqrt{q_t}}{\sqrt{800dK}\widetilde{\gamma}_t}$ and the assumptions that $|\ell_t(X_t, A_t)| \le 1$ and $|\langle X_t, \boldsymbol{m}_{t,A_t}\rangle| \le 1$ in the last inequality. Then using Lemma 17 for $y = -\eta_t\langle Z(X_0), \widehat{\boldsymbol{\theta}}_t - \boldsymbol{m}_t\rangle$ and (39), we obtain

$$\begin{aligned}
\frac{1}{\eta_t}\mathbb{E}\left[\psi(-\eta_t\langle Z(X_0), \widehat{\boldsymbol{\theta}}_t - \boldsymbol{m}_t\rangle)\right] &\le \frac{2}{\eta_t}\mathbb{E}\left[(-\eta_t\langle Z(X_0), \widehat{\boldsymbol{\theta}}_t - \boldsymbol{m}_t\rangle)^2\right]\\
&\le \frac{4dK\eta_t\widetilde{\gamma}_t^2}{q_t}\mathbb{E}_t\left[\xi_{t,A_t}^2\right].
\end{aligned} \tag{40}$$

From the fact that $(r_1 \cdot \boldsymbol{x}, \dots, r_K \cdot \boldsymbol{x})^\top \boldsymbol{\theta}_t \in [-1, 1]$ for any t, $r \in \Delta([K])$ and $\boldsymbol{x} \in \mathcal{X}$, we also see that the first term of RHS in (38) is bounded by a constant:

$$\mathbb{E}\left[\frac{1}{\tau}\sum_{t=1}^{\tau}\left\langle \overline{Z}(X_0) - Z^*(X_0), \widehat{\boldsymbol{\theta}}_t\right\rangle\right] = \frac{1}{\tau}\sum_{t=1}^{\tau}\left\langle \overline{Z}(X_0) - Z^*(X_0), \boldsymbol{\theta}_t\right\rangle \le 2. \tag{41}$$

Now, we are ready to prove the main statement. For any stopping time $\tau \in [1, T]$ and $a^* \in [K]$, we have that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{\tau}(\ell_t(X_t, a_t) - \ell_t(X_t, a^*))\right]\\
&\le \mathbb{E}\left[\sum_{t=1}^{\tau}(\ell_t(X_t, a_t) - \ell_t(X_t, \pi^*(X_t)))\right]\\
&= \mathbb{E}\left[\sum_{t=1}^{\tau}\langle Z_t(X_t) - Z(X_t), \boldsymbol{\theta}_t\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{\tau}\langle Z(X_t) - Z^*(X_t), \boldsymbol{\theta}_t\rangle\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\sum_{t=1}^{\tau} \langle Z_t(X_t) - Z(X_t), \boldsymbol{\theta}_t \rangle\right] + \mathbb{E}[\widehat{R}_\tau(X_0, \pi^*)] \\
&\le \sum_{t=1}^{\tau} \frac{1}{2\tau^2} + \mathbb{E}\left[\frac{1}{\tau}\sum_{t=1}^{\tau}\left\langle \overline{Z}(X_0) - Z^*(X_0), \widehat{\boldsymbol{\theta}}_t \right\rangle\right] + \mathbb{E}\left[\frac{K\ln\tau}{\eta_\tau}\right] + \mathbb{E}\left[\sum_{t=1}^{\tau}\frac{1}{\eta_t}\mathbb{E}_t\left[\psi(-\eta_t\langle Z(X_0), \widehat{\boldsymbol{\theta}}_t - \boldsymbol{m}_t\rangle)\right]\right] \\
&\le 3 + \mathbb{E}\left[\frac{K\ln\tau}{\eta_\tau}\right] + \mathbb{E}\left[\sum_{t=1}^{\tau}\frac{1}{\eta_t}\mathbb{E}_t\left[\psi(-\eta_t\langle Z(X_0), \widehat{\boldsymbol{\theta}}_t - \boldsymbol{m}_t\rangle)\right]\right] \\
&\le 3 + \mathbb{E}\left[\frac{K\ln\tau}{\eta_\tau}\right] + \mathbb{E}\left[\sum_{t=1}^{\tau}\frac{4dK\eta_t\widetilde{\gamma}_t^2}{q_t}\mathbb{E}_t\left[\xi_{t,A_t}^2\right]\right],
\end{aligned}
$$

where we use Lemma 16 in the first inequality and we use (41) in the second inequality and we use (40) in the last inequality.

Recall that $\beta_t := 16\widetilde{\gamma}_t^2\xi_{t,A_t}^2$, and $\widetilde{\gamma}_t = 4\ln(10dKt)$. Also, recall that the learning rate η_t is defined as follows:

$$
\eta_t = \frac{1}{\sqrt{\frac{800dK\widetilde{\gamma}_t^2}{\min_{j\le t} q_j} + \sum_{j=1}^{t-1}\frac{\beta_j}{q_j}}}.
$$

We also define η'_t as follows:

$$
\eta'_t := \frac{1}{\sqrt{\frac{800dK\widetilde{\gamma}_{t-1}^2}{\min_{j\le t-1} q_j} + \sum_{j=1}^{t-1}\frac{\beta_j}{q_j}}}.
$$

Using $\frac{x}{2\sqrt{y}} \le \sqrt{y} - \sqrt{y-x}$ for $x = \frac{\beta_t}{q_t}, y = \frac{800dK\widetilde{\gamma}_t^2}{\min_{j\le t} q_j} + \sum_{j=1}^{t}\frac{\beta_j}{q_j}$, we have that

$$
\begin{aligned}
\frac{\frac{\beta_t}{q_t}}{2\sqrt{\frac{800dK\widetilde{\gamma}_t^2}{\min_{j\le t} q_j} + \sum_{j=1}^{t}\frac{\beta_j}{q_j}}} &\le \sqrt{\frac{800dK\widetilde{\gamma}_t^2}{\min_{j\le t} q_j} + \sum_{j=1}^{t}\frac{\beta_j}{q_j}} - \sqrt{\frac{800dK\widetilde{\gamma}_t^2}{\min_{j\le t} q_j} + \sum_{j=1}^{t-1}\frac{\beta_j}{q_j}} \\
&\le \sqrt{\frac{800dK\widetilde{\gamma}_t^2}{\min_{j\le t} q_j} + \sum_{j=1}^{t}\frac{\beta_j}{q_j}} - \sqrt{\frac{800dK\widetilde{\gamma}_{t-1}^2}{\min_{j\le t-1} q_j} + \sum_{j=1}^{t-1}\frac{\beta_j}{q_j}} \\
&= \frac{1}{\eta'_{t+1}} - \frac{1}{\eta'_t}, \qquad (42)
\end{aligned}
$$

where we used $\frac{\widetilde{\gamma}_t}{\min_{j\le t} q_j} \ge \frac{\widetilde{\gamma}_{t-1}}{\min_{j\le t-1} q_j}$ in the second inequality. Summing up over $t = 1, \ldots, \tau$ gives

$$
\sum_{t=1}^{\tau}\left(\frac{1}{\eta'_{t+1}} - \frac{1}{\eta'_t}\right) = \frac{1}{\eta'_{\tau+1}} - \frac{1}{\eta'_1} \le \frac{1}{\eta'_{\tau+1}} = \sqrt{\frac{800dK\widetilde{\gamma}_\tau^2}{\min_{j\le\tau} q_j} + \sum_{j=1}^{\tau}\frac{\beta_j}{q_j}}. \qquad (43)
$$

Therefore, using the definition of η_t and β_t, we have that

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{\tau}\frac{4dK\eta_t\widetilde{\gamma}_t^2}{q_t}\xi_{t,A_t}^2\right] = \mathbb{E}\left[\sum_{t=1}^{\tau}\frac{dK\beta_t}{4q_t}\eta_t\right] = \mathbb{E}\left[\sum_{t=1}^{\tau}\frac{dK\beta_t}{4q_t}\frac{1}{\sqrt{\frac{800dK\widetilde{\gamma}_t^2}{\min_{j\le t} q_j} + \sum_{j=1}^{t-1}\frac{\beta_j}{q_j}}}\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{\tau}\frac{dK\beta_t}{2q_t}\frac{1}{\sqrt{\frac{800dK\widetilde{\gamma}_t^2}{\min_{j\le t} q_j} + \sum_{j=1}^{t}\frac{\beta_j}{q_j}}}\right] \le dK\mathbb{E}\left[\sum_{t=1}^{\tau}\left(\frac{1}{\eta'_{t+1}} - \frac{1}{\eta'_t}\right)\right] \le dK\mathbb{E}\left[\sqrt{\frac{800dK\widetilde{\gamma}_\tau^2}{\min_{j\le\tau} q_j} + \sum_{j=1}^{\tau}\frac{\beta_j}{q_j}}\right] \\
&= dK\mathbb{E}\left[\sqrt{\frac{800dK\widetilde{\gamma}_\tau^2}{\min_{j\le\tau} q_j} + \sum_{t=1}^{\tau}\frac{16\widetilde{\gamma}_t^2\xi_{t,A_t}^2}{q_t}}\right]
\end{aligned}
$$

$$\leq 4dK\widetilde{\gamma}_\tau \mathbb{E}\left[\sqrt{\frac{50dK}{\min_{j\leq\tau} q_j} + \sum_{t=1}^{\tau} \frac{\xi_{t,A_t}^2}{q_t}}\right]$$

$$= 16dK\ln(10dK\tau)\sqrt{\frac{50dK}{\min_{j\leq\tau} q_j} + \mathbb{E}\left[\sum_{t=1}^{\tau} \frac{\mathsf{upd}_t \xi_{t,A_t}^2}{q_t^2}\right]},$$

where we used $\frac{\beta_t}{q_t} \leq \frac{800dK\widetilde{\gamma}_t^2}{\min_{j\leq t} q_j}$ in the first inequality and the second inequality follows from (42), and the third inequality follows from (43).

Next we evaluate the term $\mathbb{E}\left[\frac{K\ln\tau}{\eta_\tau}\right]$.

$$\mathbb{E}\left[\frac{K\ln\tau}{\eta_\tau}\right] \leq K\ln\tau\mathbb{E}\left[\sqrt{\frac{800dK\widetilde{\gamma}_\tau^2}{\min_{j\leq\tau} q_j} + \sum_{t=1}^{\tau}\frac{16\widetilde{\gamma}_t^2\xi_{t,a}^2}{q_t}}\right]$$

$$\leq 16K\ln(\tau)\ln(10dK\tau)\cdot\sqrt{\frac{50dK}{\min_{j\leq\tau} q_j} + \mathbb{E}\left[\sum_{t=1}^{\tau}\frac{\mathsf{upd}_t \xi_{t,A_t}^2}{q_t^2}\right]}$$

Therefore, we conclude that

$$\mathbb{E}\left[\sum_{t=1}^{\tau}(\ell_t(X_t,a_t) - \ell_t(X_t,a^*))\right]$$

$$\leq 16K\ln(10dK\tau)\left(\ln(\tau)+d\right)\cdot\sqrt{\frac{50dK}{\min_{j\leq\tau} q_j} + \mathbb{E}\left[\sum_{t=1}^{\tau}\frac{\mathsf{upd}_t \xi_{t,A_t}^2}{q_t^2}\right]} + 3$$

$$\leq 32Kd\ln(10dK\tau)\ln(\tau)\left(\sqrt{\mathbb{E}\left[\sum_{t=1}^{\tau}\frac{\mathsf{upd}_t \xi_{t,A_t}^2}{q_t^2}\right]} + \mathbb{E}\left[\frac{\sqrt{50dK}}{\min_{j\leq\tau} q_j}\right]\right).$$

□

Remark 2. *We omitted the proof of Theorem 1 since using Proposition 5 and 6 , and the dd-iw-stable condition proved in Proposition 1 immediately implies Theorem 1.*

G Appendix for FTRL-LC (Algorithm 2)

In this appendix, we describe the detailed procedure of MGR and all the technical proof for analysis of FTRL-LC.

G.1 Matrix geometric resampling

We detail the whole procedure of MGR in Algorithm 7 [Neu and Bartók, 2013, Neu and Bartók, 2016, Neu and Olkhovskaya, 2020]. MGR takes inputs of context distribution $\mathcal{D}$, policy π_t, action $a \in [K]$, number of iterations M_t, and constant ρ, and outputs $\widehat{\mathbf{\Sigma}}_{t,a}^{+} = \rho\mathbf{I} + \rho\sum_{k=1}^{M_t}\mathbf{A}_{k,a}$ as the estimate of the inverse of the covariance matrix $\mathbf{\Sigma}_{t,a}^{-1}$. In this work, we set $\rho = \frac{1}{2}$.

G.2 Useful lemma for the entropy term

First, we introduce the following lemma, which implies that the definition of β_t' based on entropy terms is crucial in the analysis for FTRL with Shannon entropy regularizer. The proof follows the similar argument as Proposition 1 of Ito et al. [2022].

Algorithm 7: Matrix Geometric Resampling (MGR) [Neu and Olkhovskaya, 2020]

Input : Context distribution $\mathcal{D}$, policy π_t, action $a \in [K]$, number of iterations M_t, constant $\rho = \frac{1}{2}$

for $k = 1, 2, \ldots, M_t$ **do**

  Draw $X(k) \sim \mathcal{D}$ and $A(k) \sim \pi_t(\cdot|X(k))$

  Compute $\mathbf{B}_{k,a} = \mathbb{1}[A(k) = a] X(k) X(k)^\top$

  Compute $\mathbf{A}_{k,a} = \Pi_{j=1}^{k} (\mathbf{I} - \rho \mathbf{B}_{k,a})$

Output : $\widehat{\boldsymbol{\Sigma}}^{+}_{t,a} = \rho \mathbf{I} + \rho \sum_{k=1}^{M_t} \mathbf{A}_{k,a}$

Lemma 18. *Let β'_t be updated by (13) for each round t. Then for a ghast sample X_0, we have*

$$\mathbb{E}\left[\sum_{t=1}^{T} \left(\beta'_{t+1} - \beta'_t\right) H(p_{t+1}(\cdot|X_0))\right] = \mathcal{O}\left(c'_1 \sqrt{\ln K} \sqrt{\sum_{t=1}^{T} \mathbb{E}\left[H(p_t(\cdot|X_0))\right]}\right).$$

Proof of Lemma 18. From our definition of β'_t, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \left(\beta'_{t+1} - \beta'_t\right) H(p_{t+1}(\cdot|X_0))\right] = \mathbb{E}\left[\sum_{t=1}^{T} \frac{c'_1}{\sqrt{1 + (\ln K)^{-1} \sum_{s=1}^{t} H(p_s(\cdot|X_s))}} H(p_{t+1}(\cdot|X_0))\right]$$

$$= 2c'_1 \sqrt{\ln K} \mathbb{E}\left[\sum_{t=1}^{T} \frac{H(p_{t+1}(\cdot|X_0))}{\sqrt{4 \ln K + 4 \sum_{s=1}^{t} H(p_s(\cdot|X_s))}}\right]$$

$$= 2c'_1 \sqrt{\ln K} \mathbb{E}\left[\sum_{t=1}^{T} \frac{H(p_{t+1}(\cdot|X_0))}{\sqrt{\ln K + \sum_{s=1}^{t} H(p_s(\cdot|X_s))} + \sqrt{\ln K + \sum_{s=1}^{t} H(p_s(\cdot|X_s))}}\right]$$

$$\leq 2c'_1 \sqrt{\ln K} \mathbb{E}\left[\sum_{t=1}^{T} \frac{H(p_{t+1}(\cdot|X_0))}{\sqrt{\sum_{s=1}^{t+1} H(p_s(\cdot|X_s))} + \sqrt{\sum_{s=1}^{t} H(p_s(\cdot|X_s))}}\right],$$

where in the last step we used the fact that $H(p_s(\cdot|X_s)) \leq H(p_1(\cdot|X_1)) = \ln K$. Using the property that $\mathbb{E}_{X_{t+1} \sim \mathcal{D}}[H(p_{t+1}(\cdot|X_{t+1}))|\mathcal{F}_t] = \mathbb{E}_{X_0 \sim \mathcal{D}}[H(p_{t+1}(\cdot|X_0))|\mathcal{F}_t]$, we have

$$2c'_1 \sqrt{\ln K} \mathbb{E}\left[\sum_{t=1}^{T} \frac{H(p_{t+1}(\cdot|X_0))}{\sqrt{\sum_{s=1}^{t+1} H(p_s(\cdot|X_s))} + \sqrt{\sum_{s=1}^{t} H(p_s(\cdot|X_s))}}\right]$$

$$= 2c'_1 \sqrt{\ln K} \mathbb{E}\left[\sum_{t=1}^{T} \frac{H(p_{t+1}(\cdot|X_0)) \left(\sqrt{\sum_{s=1}^{t+1} H(p_s(\cdot|X_s))} - \sqrt{\sum_{s=1}^{t} H(p_s(\cdot|X_s))}\right)}{H(p_{t+1}(\cdot|X_{t+1}))}\right]$$

$$= 2c'_1 \sqrt{\ln K} \mathbb{E}\left[\sum_{t=1}^{T} \left(\sqrt{\sum_{s=1}^{t+1} H(p_s(\cdot|X_s))} - \sqrt{\sum_{s=1}^{t} H(p_s(\cdot|X_s))}\right)\right]$$

$$= 2c'_1 \sqrt{\ln K} \mathbb{E}\left[\left(\sqrt{\sum_{s=1}^{T+1} H(p_s(\cdot|X_s))} - \sqrt{H(p_1(\cdot|X_1))}\right)\right]$$

$$\leq 2c'_1 \sqrt{\ln K} \mathbb{E}\left[\sqrt{\sum_{s=1}^{T} H(p_s(\cdot|X_s))}\right],$$

where in the last step we again used the fact that $H(p_s(\cdot|X_s)) \leq H(p_1(\cdot|X_1)) = \ln K$. Hence, again using the fact that X_0 and X_t follows the same distribution $\mathcal{D}$ and the linearity of the expectation, we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\beta'_{t+1}-\beta'_t\right)H(p_{t+1}(\cdot|X_0))\right] = \mathcal{O}\left(c'_1\sqrt{\ln K}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[H(p_t(\cdot|X_t))\right]}\right) = \mathcal{O}\left(c'_1\sqrt{\ln K}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[H(p_t(\cdot|X_0))\right]}\right),$$

which concludes the proof.

□

G.3 Proof of Lemma 1

The proof follows the standard analysis of FTRL with the negative Shannon entropy.

Proof of Lemma 1. By Lemma 6, for any context $\boldsymbol{x} \in \mathcal{X}$, we have

$$\begin{aligned}
\widetilde{R}_T(\boldsymbol{x}) &= \mathbb{E}_{A_t}\left[\sum_{t=1}^{T}\left(\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,A_t}\rangle - \langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,\pi^*(\boldsymbol{x})}\rangle\right)\right] \\
&\leq \sum_{t=1}^{T}\left(\psi_t(p_{t+1}(\cdot|\boldsymbol{x})) - \psi_{t+1}(p_{t+1}(\cdot|\boldsymbol{x}))\right) + \psi_{T+1}(\pi^*(\cdot|\boldsymbol{x})) - \psi_1(p_1(\cdot|\boldsymbol{x})) \\
&\quad + \sum_{t=1}^{T}(1-\gamma_t)\left(\left\langle p_t(\cdot|\boldsymbol{x}) - p_{t+1}(\cdot|\boldsymbol{x})), \widetilde{\boldsymbol{\ell}}_t(\boldsymbol{x})\right\rangle - D_t(p_{t+1}(\cdot|\boldsymbol{x}), p_t(\cdot|\boldsymbol{x}))\right) + U(\boldsymbol{x}). \qquad (44)
\end{aligned}$$

We first bound the stability term $\left\langle p_t(\cdot|\boldsymbol{x}) - p_{t+1}(\cdot|\boldsymbol{x})), \widetilde{\boldsymbol{\ell}}_t(\boldsymbol{x})\right\rangle - D_t(p_{t+1}(\cdot|\boldsymbol{x}), p_t(\cdot|\boldsymbol{x}))$. Since the function $f(q) = \sum_{a\in[K]}(p_t(a|\boldsymbol{x}) - q(a))\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle - D_t(q, p_t(\cdot|\boldsymbol{x}))$ is concave with respect to $q \in \Delta([K])$, its maximum solution is obtained by computing the point where its derivative is equal to zero. For each $a \in [K]$, we have

$$\frac{\partial}{\partial q(a)}\left(\sum_{a\in[K]}(p_t(a|\boldsymbol{x}) - q(a))\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle - D_t(q, p_t(\cdot|\boldsymbol{x}))\right) = -\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle - \frac{1}{\eta_t}(\ln q(a) - \ln p_t(a|\boldsymbol{x})),$$

and thus the maximum solution is obtained for $q^*(a) = p_t(a|\boldsymbol{x})\exp(-\eta_t\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle)$. Hence, we can show

$$\begin{aligned}
&\sum_{a\in[K]}(p_t(a|\boldsymbol{x}) - p_{t+1}(a|\boldsymbol{x}))\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle - D_t(p_{t+1}(\cdot|\boldsymbol{x}), p_t(\cdot|\boldsymbol{x})) \\
&\leq \sum_{a\in[K]}(p_t(a|\boldsymbol{x}) - q^*(a))\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle - D_t(q^*, p_t(\cdot|\boldsymbol{x})) \\
&= \sum_{a\in[K]}\left(\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle(p_t(a|\boldsymbol{x}) - q^*(a)) - \frac{1}{\eta_t}(q^*(a)\ln p_t(a|\boldsymbol{x}) - p_t(a|\boldsymbol{x})\ln p_t(a|\boldsymbol{x}) - (\ln p_t(a|\boldsymbol{x}) + 1)(q^*(a) - p_t(a|\boldsymbol{x})))\right) \\
&= \sum_{a\in[K]}\left(\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle p_t(a|\boldsymbol{x}) + \frac{1}{\eta_t}(q^*(a) - p_t(a|\boldsymbol{x}))\right) \\
&= \frac{1}{\eta_t}\sum_{a\in[K]}p_t(a|\boldsymbol{x})\left(\exp(-\eta_t\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle) + \eta_t\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle - 1\right). \qquad (45)
\end{aligned}$$

Using the inequality $\exp(-x) \leq 1 - x + x^2$ that holds for any $x \geq -1$ and the assumption that $|\eta_t\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle| \leq 1$, we can bound the RHS of (45) is bounded as

$$\frac{1}{\eta_t}\sum_{a\in[K]}p_t(a|\boldsymbol{x})\left(\exp(-\eta_t\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle) + \eta_t\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle - 1\right) \leq \eta_t\sum_{a\in[K]}p_t(a|\boldsymbol{x})\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle^2,$$

implying that

$$(1-\gamma_t)\sum_{a\in[K]}(p_t(a|\boldsymbol{x})-p_{t+1}(a|\boldsymbol{x}))\langle \boldsymbol{x},\widetilde{\boldsymbol{\theta}}_{t,a}\rangle - D_t(p_{t+1}(\cdot|\boldsymbol{x}),p_t(\cdot|\boldsymbol{x})) \le (1-\gamma_t)\eta_t\sum_{a\in[K]}p_t(a|\boldsymbol{x})\langle \boldsymbol{x},\widetilde{\boldsymbol{\theta}}_{t,a}\rangle^2.$$

Since $p_t(a|\boldsymbol{x})=\frac{1}{1-\gamma_t}(\pi_t(a|\boldsymbol{x})-\frac{\gamma_t}{K})$ from the definition of $\pi_t(a|\boldsymbol{x})$, we obtain

$$(1-\gamma_t)\sum_{a\in[K]}(p_t(a|\boldsymbol{x})-p_{t+1}(a|\boldsymbol{x}))\langle \boldsymbol{x},\widetilde{\boldsymbol{\theta}}_{t,a}\rangle - D_t(p_{t+1}(\cdot|\boldsymbol{x}),p_t(\cdot|\boldsymbol{x})) \le \eta_t\sum_{a\in[K]}\pi_t(a|\boldsymbol{x})\langle \boldsymbol{x},\widetilde{\boldsymbol{\theta}}_{t,a}\rangle^2. \tag{46}$$

For the penalty term, using $0\le H(p)\le \ln K$ that holds for any $p\in\Delta([K])$, we can show

$$\begin{aligned}\sum_{t=1}^T(\psi_t(p_{t+1}(\cdot|\boldsymbol{x}))-\psi_{t+1}(p_{t+1}(\cdot|\boldsymbol{x})))+\psi_{T+1}(\pi^*(\cdot|\boldsymbol{x}))-\psi_1(p_1(\cdot|\boldsymbol{x}))&\\ \le\sum_{t=1}^T(\beta_{t+1}-\beta_t)\,H(p_{t+1}(\cdot|\boldsymbol{x}))+\beta_1\ln K.&\end{aligned} \tag{47}$$

Combining (44), (46), and (47) completes the proof of Lemma 1. □

G.4 Proof of Lemma 2

Proof of Lemma 2. Let $\|\cdot\|_{\mathrm{op}}$ be the operator norm of any positive semi-definite matrix. Recall that definitions of the baised estimator $\widetilde{\boldsymbol{\theta}}_{t,a}=\widehat{\boldsymbol{\Sigma}}^{+}_{t,a}X_t\ell_t(X_t,A_t)\mathbb{1}[A_t=a]$ and unbiased estimator $\widehat{\boldsymbol{\theta}}_{t,a}=\boldsymbol{\Sigma}^{-1}_{t,a}X_t\ell_t(X_t,A_t)\mathbb{1}[A_t=a]$. The first statements of (i) can be shown by using these definitions and adapting a similar analysis for Lemma 5 in Neu and Olkhovskaya [2020] (Lemma 9). For $\widehat{\boldsymbol{\Sigma}}^{+}_{t,a}$, the output of MGR procedure in Algorithm 7 with $\rho=\frac{1}{2}$, we have $\mathbb{E}_t[\mathbf{A}_{t,a}]=\mathbb{E}_t\left[\Pi_{j=1}^k\left(I-\rho\mathbf{B}_{k,a}\right)\right]=\left(I-\frac{1}{2}\boldsymbol{\Sigma}_{t,a}\right)^k$ for each $a\in[K]$. Then, it gives $\mathbb{E}_t[\widehat{\boldsymbol{\Sigma}}^{+}_{t,a}]=\frac{1}{2}\sum_{k=0}^{M_t}\left(I-\frac{1}{2}\boldsymbol{\Sigma}^{-1}_{t,a}\right)^k=\boldsymbol{\Sigma}^{-1}_{t,a}-\left(I-\frac{1}{2}\boldsymbol{\Sigma}_{t,a}\right)^{M_t}\boldsymbol{\Sigma}^{-1}_{t,a}$. Using these expressions, for the biased estimator $\widehat{\boldsymbol{\theta}}_{t,a}$ of each action $a\in[K]$, we have that

$$\begin{aligned}\mathbb{E}_t[\widetilde{\boldsymbol{\theta}}_{t,a}]&=\mathbb{E}_t[\widehat{\boldsymbol{\Sigma}}^{+}_{t,a}X_t\ell_t(X_t,a)\mathbb{1}[A_t=a]]\\&=\mathbb{E}_t[\widehat{\boldsymbol{\Sigma}}^{+}_{t,a}]\mathbb{E}_t[X_t\langle X_t,\boldsymbol{\theta}_{t,a}\rangle\mathbb{1}[A_t=a]]\\&=\mathbb{E}_t[\widehat{\boldsymbol{\Sigma}}^{+}_{t,a}]\mathbb{E}_t[X_tX_t^\top\mathbb{1}[A_t=a]]\cdot\boldsymbol{\theta}_{t,a}\\&=\mathbb{E}_t[\widehat{\boldsymbol{\Sigma}}^{+}_{t,a}]\boldsymbol{\Sigma}_{t,a}\boldsymbol{\theta}_{t,a}\\&=\left(\boldsymbol{\Sigma}^{-1}_{t,a}-\left(I-\frac{1}{2}\boldsymbol{\Sigma}_{t,a}\right)^{M_t}\boldsymbol{\Sigma}^{-1}_{t,a}\right)\boldsymbol{\Sigma}_{t,a}\boldsymbol{\theta}_{t,a}\\&=\boldsymbol{\theta}_{t,a}-\left(I-\frac{1}{2}\boldsymbol{\Sigma}_{t,a}\right)^{M_t}\boldsymbol{\theta}_{t,a},\end{aligned}$$

implying that

$$\mathbb{E}_t[\widetilde{\boldsymbol{\theta}}_{t,a}-\widehat{\boldsymbol{\theta}}_{t,a}]=-\left(I-\frac{1}{2}\boldsymbol{\Sigma}_{t,a}\right)^{M_t}\boldsymbol{\theta}_{t,a}.$$

Therefore, we obtain

$$\begin{aligned}\mathbb{E}_t[\langle X_t,\widetilde{\boldsymbol{\theta}}_{t,a}-\widehat{\boldsymbol{\theta}}_{t,a}\rangle]\le\|X_t\|_2\|\boldsymbol{\theta}_{t,a}\|_2\left\|\left(I-\frac{1}{2}\boldsymbol{\Sigma}_{t,a}\right)^{M_t}\right\|_{\mathrm{op}}&\le\left\|\left(I-\frac{1}{2}\boldsymbol{\Sigma}_{t,a}\right)^{M_t}\right\|_{\mathrm{op}}\\ \le\left(1-\frac{\gamma_t\lambda_{\min}(\boldsymbol{\Sigma})}{2K}\right)\le\exp\left(-\frac{\gamma_t\lambda_{\min}(\boldsymbol{\Sigma})}{2K}\cdot M_t\right)&\le\frac{1}{t^2},\end{aligned}$$

where we used $\|X_t\| \leq 1$ and $\|\boldsymbol{\theta}_{t,a}\|_2 \leq 1$ in the second inequality, we used the fact that the policy $\pi(\cdot|X_t)$ employs the uniform exploration with mixing rate γ_t in the third inequality, and the last step follows by $M_t = \left\lceil \frac{4K}{\gamma_t \lambda_{\min}(\boldsymbol{\Sigma})} \ln t \right\rceil$.

Next we consider the second statement of (ii), which can be shown via our careful tuning of learning parameters. For the output of MGR procedure in Algorithm 7 with $\rho = \frac{1}{2}$ and any $\boldsymbol{x} \in \mathcal{X}$, $|\eta_t \langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a} \rangle|$ for each $a \in [K]$ is bounded as follows:

$$
\begin{aligned}
|\eta_t \langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a} \rangle| &= \eta_t \left| \langle \boldsymbol{x}, \widehat{\boldsymbol{\Sigma}}^{+}_{t,a} X_t \ell_t(X_t, A_t) \mathbb{1}[A_t = 1] \rangle \right| \leq \eta_t \left| \boldsymbol{x}^\top (\widehat{\boldsymbol{\Sigma}}^{+}_{t,a} X_t) \right| \leq \eta_t \|\widehat{\boldsymbol{\Sigma}}^{+}_{t,a}\|_{\mathrm{op}} \\
&\leq \eta_t \left(\left\| \rho I + \rho \sum_{k=1}^{M_t} \mathbf{A}_{k,a} \right\|_{\mathrm{op}} \right) \leq \frac{\eta_t}{2} \left(1 + \sum_{k=1}^{M_t} \left\| \Pi_{j=1}^{k} \left(I - \frac{1}{2} \mathbf{B}_{k,a} \right) \right\|_{\mathrm{op}} \right) \leq \frac{\eta_t (M_t + 1)}{2},
\end{aligned}
\tag{48}
$$

where the first equality follows from the definition of $\widetilde{\boldsymbol{\theta}}_{t,a}$, the first inequality follows from $\ell_t(X_t, A_t) \leq 1$, and the second inequality follows from $\max_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{x}\|_2 \leq 1$. Setting $M_t = \left\lceil \frac{4K}{\gamma_t \lambda_{\min}(\boldsymbol{\Sigma})} \ln t \right\rceil$ gives

$$
\frac{1}{\eta_t} = \frac{2}{\eta_t} - \frac{\alpha_t}{\alpha_t \eta_t} = \frac{2}{\eta_t} - \frac{\alpha_t}{\gamma_t} \leq \frac{2}{\eta_t} - (M_t - 1),
$$

where we used the definition of $\gamma_t = \alpha_t \eta_t$ for $\alpha_t = \frac{4K \ln t}{\lambda_{\min}(\boldsymbol{\Sigma})}$. Therefore, from the definition of $\eta_t \leq \frac{1}{2}$, we have $2 \leq \frac{2}{\eta_t} - (M_t - 1) \Leftrightarrow \eta_t \leq \frac{2}{M_t + 1}$. Combining it with (48) guarantees that $|\eta_t \langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a} \rangle| \leq 1$, as desired. □

G.5 Proof of Lemma 3

Proof of Lemma 3. By Lemma 2 and the definitions of $\beta_t, \eta_t, \gamma_t$ and M_t, we can see that $|\eta_t \langle X_0, \widetilde{\boldsymbol{\theta}}_{t,a} \rangle| \leq 1$ holds, which allow us to use Lemma 1 for fixed X_0. Then we have

$$
\mathbb{E}[\widetilde{R}_T(X_0)] \leq \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} (\beta_{t+1} - \beta_t) H(p_{t+1}(\cdot|X_0)) \right]}_{\text{term } A} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \eta_t \sum_{a \in [K]} \pi_t(a|X_0) \langle X_0, \widetilde{\boldsymbol{\theta}}_{t,a} \rangle^2 \right]}_{\text{term } B} + \mathbb{E}\left[U(X_0) \right] + \beta_1 \ln K.
\tag{49}
$$

Using the definition of $\beta_t = \max\{2, c_2' \ln T, \beta_t'\}$, we have

$$
\beta_1 \ln K \leq c_2' \ln K \ln T.
\tag{50}
$$

Next, we will evaluate term B and $\mathbb{E}[U(X_0)]$. From the definition of β_t' in (13), we see that

$$
\beta_t' = c_1' + \sum_{s=1}^{t-1} \frac{c_1'}{\sqrt{1 + (\ln K)^{-1} \sum_{u=1}^{s-1} H(p_u(\cdot|X_u))}} \geq \frac{c_1' t}{\sqrt{1 + (\ln K)^{-1} \sum_{s=1}^{t} H(p_s(\cdot|X_s))}},
$$

and thus

$$
\begin{aligned}
\sum_{t=1}^{T} \eta_t &\leq \sum_{t=1}^{T} \frac{1}{\beta_t'} \leq \sum_{t=1}^{T} \frac{\sqrt{1 + (\ln K)^{-1} \sum_{s=1}^{t} H(p_s(\cdot|X_s))}}{c_1' t} \\
&\leq \frac{1 + \ln T}{c_1'} \sqrt{1 + (\ln K)^{-1} \sum_{s=1}^{T} H(p_s(\cdot|X_s))} = \mathcal{O}\left(\frac{\ln T}{c_1' \sqrt{\ln K}} \sqrt{\sum_{t=1}^{T} H(p_t(\cdot|X_t))} \right),
\end{aligned}
\tag{51}
$$

where we used $H(p_1(\cdot|X_1)) = \ln K$.

By Lemma 8 and (51), we obtain

$$\text{term } B = \mathbb{E}\left[\sum_{t=1}^{T}\eta_t \sum_{a\in[K]}\pi_t(a|X_0)\langle X_0, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle^2\right] = \mathcal{O}\left(\mathbb{E}\left[\frac{3Kd\cdot\ln T}{c_1'\sqrt{\ln K}}\sqrt{\sum_{t=1}^{T}H(p_t(\cdot|X_t))}\right]\right)$$
$$= \mathcal{O}\left(\frac{3Kd\cdot\ln T}{c_1'\sqrt{\ln K}}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}H(p_t(\cdot|X_0))\right]}\right), \tag{52}$$

where we used the fact that $\mathbb{E}_{X_0\sim\mathcal{D}}[p_t(\cdot|X_0)|\widetilde{\boldsymbol{\theta}}_t] = \mathbb{E}_{X_t\sim\mathcal{D}}[p_t(\cdot|X_t))|\widetilde{\boldsymbol{\theta}}_t]$.

For $\mathbb{E}[U(X_0)]$, from Lemma 2, we have $|\mathbb{E}[\langle X_t, \widetilde{\boldsymbol{\theta}}_{t,a} - \widehat{\boldsymbol{\theta}}_{t,a}\rangle|\mathcal{F}_{t-1}]| \leq \frac{1}{t^2} \leq 1$, and thus

$$\mathbb{E}[U(X_0)] = \mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\sum_{a\in[K]}\left(\frac{1}{K} - \pi^*(a|X_0)\right)\langle \boldsymbol{x}, \widetilde{\boldsymbol{\theta}}_{t,a}\rangle\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\max_{a\in[K]}\langle X_0, \widetilde{\boldsymbol{\theta}}_{t,a} - \widehat{\boldsymbol{\theta}}_{t,a} + \widehat{\boldsymbol{\theta}}_{t,a}\ \rangle\right]$$
$$\leq \mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\left(\max_{a\in[K]}\langle X_0, \widetilde{\boldsymbol{\theta}}_{t,a} - \widehat{\boldsymbol{\theta}}_{t,a}\rangle + \ell_t(X_0,a)\right)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\left(\max_{a\in[K]}\left|\langle X_0, \widetilde{\boldsymbol{\theta}}_{t,a} - \widehat{\boldsymbol{\theta}}_{t,a}\rangle\right| + 1\right)\right]$$
$$\leq 2\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\right]. \tag{53}$$

where we used $\mathbb{E}[\widehat{\boldsymbol{\theta}}_{t,a}] = \boldsymbol{\theta}_{t,a}$ and $\mathbb{E}[\ell_t(X_0,a)] \leq 1$ in the second and third inequality. From the definition of γ_t and (51), we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\right] = \mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\eta_t\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\frac{4K\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})}\cdot\eta_t\right] = \mathcal{O}\left(\frac{K\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})}\cdot\frac{\ln T}{c_1'\sqrt{\ln K}}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}H(p_t(\cdot|X_0))\right]}\right). \tag{54}$$

Thus from (53) and (54), we obtain

$$\mathbb{E}[U(X_0)] \leq 2\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\right] = \mathcal{O}\left(\frac{K\ln^2 T}{c_1'\lambda_{\min}(\boldsymbol{\Sigma})\sqrt{\ln K}}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}H(p_t(\cdot|X_0))\right]}\right). \tag{55}$$

Finally, we will evaluate term A. Let t_0 be the first round in which β_t' becomes larger than the constant $F := \max\{2, c_2'\ln T\}$, i.e., $t_0 = \min\{t\in[T] : \beta_t' \geq F\}$. Then, by the definition of β_t, we have that

$$\mathbb{E}\left[\sum_{t=1}^{T}(\beta_{t+1}-\beta_t)\,H(p_{t+1}(\cdot|X_0))\right]$$
$$= \mathbb{E}\left[\sum_{t=1}^{t_0-2}(\beta_{t+1}-\beta_t)\,H(p_{t+1}(\cdot|X_0)) + (\beta_{t_0}-\beta_{t_0-1})\,H(p_{t+1}(\cdot|X_0)) + \sum_{t=t_0}^{T}(\beta_{t+1}-\beta_t)\,H(p_{t+1}(\cdot|X_0))\right]$$
$$\leq \mathbb{E}\left[0 + \left(\beta_{t_0}' - \beta_{t_0-1}'\right)H(p_{t+1}(\cdot|X_0)) + \sum_{t=t_0}^{T}\left(\beta_{t+1}'-\beta_t'\right)H(p_{t+1}(\cdot|X_0))\right]$$
$$\leq \mathbb{E}\left[\sum_{t=1}^{T}\left(\beta_{t+1}'-\beta_t'\right)H(p_{t+1}(\cdot|X_0))\right] = \mathcal{O}\left(c_1'\sqrt{\ln K}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[H(p_t(\cdot|X_0))\right]}\right), \tag{56}$$

where the first inequality is due to the fact that β_t is the constant while $t\in[t_0-1]$, $\beta_t' \leq \beta_t$ for any t, and $\beta_t' = \beta_t$ for $t \geq t_0$. The last step follows by Lemma 18. Hence using (56) and the fact that X_0 and X_t follows the same distribution $\mathcal{D}$, we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T}(\beta_{t+1}-\beta_t)\,H(p_{t+1}(\cdot|X_0))\right] = \mathcal{O}\left(c_1'\sqrt{\ln K}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}H(p_t(\cdot|X_0))\right]}\right). \tag{57}$$

Combining (50), (52), (55), (57) with (49), we obtain

$$\mathbb{E}[\widetilde{R}_T(X_0)] = \mathcal{O}\left(\left(c_1'\sqrt{\ln K} + \frac{\left(3Kd + \frac{2K\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)\ln T}{c_1'\sqrt{\ln K}}\right)\sqrt{\mathbb{E}\left[\sum_{t=1}^{T} H(p_t(\cdot|X_0))\right]} + c_2'\ln K\ln T\right),$$

and plugging $c_2' = \frac{8K}{\lambda_{\min}(\boldsymbol{\Sigma})}$ to this bound concludes the proof.

□

G.6 Proof of Theorem 2

Proof of Theorem 2. Using Lemmas 7 and 3, we have

$$\begin{aligned}
R_T &\le \mathbb{E}[\widetilde{R}_T(X_0)] + 2\sum_{t=1}^{T}\max_{a\in[K]}|\mathbb{E}[\langle X_t, \boldsymbol{b}_{t,a}\rangle]| \\
&= \mathcal{O}\left(\left(c_1'\sqrt{\ln K} + \frac{\left(3Kd + \frac{K\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)\ln T}{c_1'\sqrt{\ln K}}\right)\sqrt{\mathbb{E}\left[\sum_{t=1}^{T} H(p_t(\cdot|X_0))\right]} + c_2'\ln K\ln T + 4\right)
\end{aligned}$$

where in the second step we used Lemma 2 with $M_t = \left\lceil \frac{4K}{\gamma_t\lambda_{\min}(\boldsymbol{\Sigma})}\ln t\right\rceil$ to have

$$\sum_{t=1}^{T}\max_{a\in[K]}|\mathbb{E}[\langle X_t, \boldsymbol{b}_{t,a}\rangle]| \le \sum_{t=1}^{T}\frac{1}{t^2} \le 2. \tag{58}$$

Setting

$$c_1' = \sqrt{\left(3Kd + \frac{2K\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)\frac{\ln T}{\ln K}}$$

gives

$$\begin{aligned}
R_T &= \mathcal{O}\left(\left(c_1'\sqrt{\ln K} + \frac{\left(Kd + \frac{K\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)\ln T}{c_1'\sqrt{\ln K}}\right)\sqrt{\mathbb{E}\left[\sum_{t=1}^{T} H(p_t(\cdot|X_0))\right]} + c_2'\ln K\ln T\right) \\
&= \mathcal{O}\left(\sqrt{\left(Kd + \frac{K\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)\ln T}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T} H(p_t(\cdot|X_0))\right]} + c_2'\ln K\ln T\right) \\
&= \mathcal{O}\left(\sqrt{\left(d + \frac{\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)K\ln T\cdot\mathbb{E}\left[\sum_{t=1}^{T} H(p_t(\cdot|X_0))\right]} + \frac{K}{\lambda_{\min}(\boldsymbol{\Sigma})}\ln K\ln T\right)
\end{aligned} \tag{59}$$

where we used $c_2' = \frac{8K}{\lambda_{\min}(\boldsymbol{\Sigma})}$ in the third equality.

For the adversarial regime, due to (59) and the fact that $\sum_{t=1}^{T} H(p_t(\cdot|X_0)) \le T\ln K$, it holds that

$$R_T = O\left(\sqrt{T\left(d + \frac{\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)K\ln(T)\ln(K)} + \frac{K}{\lambda_{\min}(\boldsymbol{\Sigma})}\ln K\ln T\right),$$

as desired.

Applying self-bounding techniques. Now, we will apply self-bounding techniques [Zimmert and Seldin, 2021, Wei and Luo, 2018] to proceed with further analysis.

Lemma 19. *For any corrupted stochastic regime, the regret is bounded from below by*

$$R_T \geq \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{X_t}(A_t)\right] - 2C.$$

Proof. Recall that $\Delta_x(a)$ is defined as $\Delta_x(a) := \langle \boldsymbol{x}, \boldsymbol{\theta}_a - \boldsymbol{\theta}_{\pi^*(\boldsymbol{x})}\rangle$ for $\boldsymbol{x} \in \mathcal{X}$ and each action $a \in [K]$.

We have

$$\begin{aligned}
R_T &= \mathbb{E}\left[\sum_{t=1}^{T}(\ell_t(X_t, A_t) - \ell_t(X_t, \pi^*(X_t)))\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\langle X_t, \boldsymbol{\theta}_{t,A_t} - \boldsymbol{\theta}_{t,\pi^*(X_t)}\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T}\langle X_t, \boldsymbol{\theta}_{A_t} - \boldsymbol{\theta}_{A_t}\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T}\langle X_t, \boldsymbol{\theta}_{\pi^*(X_t)} - \boldsymbol{\theta}_{\pi^*(X_t)}\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\langle X_t, \boldsymbol{\theta}_{A_t} - \boldsymbol{\theta}_{\pi^*(X_t)}\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T}\langle X_t, \boldsymbol{\theta}_{t,A_t} - \boldsymbol{\theta}_{A_t}\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T}\langle X_t, \boldsymbol{\theta}_{\pi^*(X_t)} - \boldsymbol{\theta}_{t,\pi^*(X_t)}\rangle\right] \\
&\geq \mathbb{E}\left[\sum_{t=1}^{T}\langle X_t, \boldsymbol{\theta}_{A_t} - \boldsymbol{\theta}_{\pi^*(X_t)}\rangle\right] - \mathbb{E}\left[\sum_{t=1}^{T}\left|\langle X_t, \boldsymbol{\theta}_{t,A_t} - \boldsymbol{\theta}_{A_t}\rangle\right|\right] - \mathbb{E}\left[\sum_{t=1}^{T}\left|\langle X_t, \boldsymbol{\theta}_{\pi^*(X_t)} - \boldsymbol{\theta}_{t,\pi^*(X_t)}\rangle\right|\right] \\
&\geq \mathbb{E}\left[\sum_{t=1}^{T}\langle X_t, \boldsymbol{\theta}_{A_t} - \boldsymbol{\theta}_{\pi^*(X_t)}\rangle\right] - 2\mathbb{E}\left[\sum_{t=1}^{T}\max_{a\in[K]}\|X_t\|_2\|\boldsymbol{\theta}_{t,a} - \boldsymbol{\theta}_a\|_2\right] \\
&\geq \sum_{t=1}^{T}\Delta_{X_t}(A_t) - 2\mathbb{E}\left[\sum_{t=1}^{T}\max_{a\in[K]}\|\boldsymbol{\theta}_{t,a} - \boldsymbol{\theta}_a\|_2\right], \\
&\geq \sum_{t=1}^{T}\Delta_{X_t}(A_t) - 2C,
\end{aligned}$$

where we used the definition of $\Delta_{X_t}(A_t)$ in the third inequality, and we used the definition of the corruption level $C \geq 0$ in the last inequality.

□

We further show the regret upper bound based on the following notation. For the optimal policy $\pi^* \in \Pi$,

$$\varrho_0(\pi^*) := \sum_{t=1}^{T}(1 - p_t(\pi^*(X_0)|X_0)), \quad \varrho_{(X_t)_{t=1}^T}(\pi^*) := \sum_{t=1}^{T}(1 - p_t(\pi^*(X_t)|X_t)), \quad \overline{\varrho}_X(\pi^*) := \mathbb{E}[\varrho_{(X_t)_{t=1}^T}(\pi^*)]. \tag{60}$$

Note that it holds that $0 \leq \overline{\varrho}_X(\pi^*) \leq T$. We also confirm the property on them in the following lemma.

Lemma 20. *Let π^* be the optimal policy defined in (1). Then we have $\overline{\varrho}_X(\pi^*) = \mathbb{E}[\varrho_0(\pi^*)]$.*

Proof of Lemma 20. Notice that since the optimal policy $\pi^* \in \Pi$ is the deterministic policy, it holds that $\mathbb{E}_{X_0\sim\mathcal{D}}[\pi^*(X_0)] = \mathbb{E}_{X_t\sim\mathcal{D}}[\pi^*(X_t)]$. Let $\widetilde{\boldsymbol{\theta}}_t = (\widetilde{\boldsymbol{\theta}}_{t,1}, \ldots, \widetilde{\boldsymbol{\theta}}_{t,K})$. Then we have

$$\mathbb{E}_t[p_t(\pi^*(X_0)|X_0)] = \mathbb{E}_{X_0\sim\mathcal{D}}[p_t(\pi^*(X_0)|X_0)|\widetilde{\boldsymbol{\theta}}_t] = \mathbb{E}_{X_t\sim\mathcal{D}}[p_t(\pi^*(X_t)|X_t)|\widetilde{\boldsymbol{\theta}}_t] = \mathbb{E}_t[p_t(\pi^*(X_t)|X_t)].$$

Hence, we have

$$\sum_{t=1}^{T}\mathbb{E}_t[(p_t(\pi^*(X_0)|X_0)] = \sum_{t=1}^{T}\mathbb{E}_t[(p_t(\pi^*(X_t)|X_t)],$$

which concludes the proof.

□

We next show that the regret is bounded in terms of $\overline{\varrho}_X(\pi^*)$.

Lemma 21. *In the corrupted stochastic setting, the regret is bounded from below as*

$$R_T \geq \frac{\Delta_{\min}}{2}\overline{\varrho}_X(\pi^*) - 2C. \tag{61}$$

Proof of Lemma 21. Recall that $\Delta_x(a) := \boldsymbol{x}^\top(\theta_a - \theta_{\pi^*(\boldsymbol{x})})$ for $a \in [K] \setminus \{\pi^*(\boldsymbol{x})\}$, where π^* is the unique optimal policy given by (1). Also recall that $\Delta_{\min}(\boldsymbol{x}) := \min_{a \neq \pi^*(\boldsymbol{x})} \Delta_x(a)$ and $\Delta_{\min} := \min_{\boldsymbol{x}\in\mathcal{X}} \Delta_{\min}(\boldsymbol{x})$. Then, using these gap definitions and Lemma 19, the regret is bounded from below as

$$\begin{aligned}
R_T &\geq \mathbb{E}\left[\sum_{t=1}^T \Delta_{X_t}(A_t) - 2C\right] = \mathbb{E}\left[\sum_{t=1}^T \sum_{a\in[K]\setminus\{\pi^*(X_t)\}} \pi_t(a|X_t)\Delta_{X_t}(a)\right] - 2C \\
&\geq \mathbb{E}\left[\sum_{t=1}^T \sum_{a\in[K]\setminus\{\pi^*(X_t)\}} (1-\gamma_t)p_t(a|X_t)\Delta_{X_t}(a)\right] - 2C \\
&\geq \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^T \sum_{a\in[K]\setminus\{\pi^*(X_t)\}} p_t(a|X_t)\Delta_{X_t}(a)\right] - 2C \\
&\geq \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^T \sum_{a\in[K]\setminus\{\pi^*(X_t)\}} p_t(a|X_t) \min_{a\in[K]\setminus\{\pi^*(X_t)\}} \Delta_{X_t}(a)\right] - 2C \\
&= \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^T \sum_{a\in[K]\setminus\{\pi^*(X_t)\}} p_t(a|X_t)\Delta_{\min}(X_t)\right] - 2C \\
&\geq \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^T \sum_{a\in[K]\setminus\{\pi^*(X_t)\}} p_t(a|X_t) \min_{\boldsymbol{x}\in\mathcal{X}} \Delta_{\min}(\boldsymbol{x})\right] - 2C \\
&= \frac{\Delta_{\min}}{2}\mathbb{E}\left[\sum_{t=1}^T \sum_{a\in[K]\setminus\{\pi^*(X_t)\}} p_t(a|X_t)\right] - 2C \\
&= \frac{\Delta_{\min}}{2}\mathbb{E}\left[\sum_{t=1}^T (1 - p_t(\pi^*(X_t)|X_t))\right] - 2C \\
&= \frac{\Delta_{\min}}{2}\mathbb{E}\left[\varrho_{(X_t)_{t=1}^T}(\pi^*)\right] - 2C = \frac{\Delta_{\min}}{2}\overline{\varrho}_X(\pi^*) - 2C,
\end{aligned}$$

where the second inequality follows by (11), the third inequality follows by $\gamma_t \leq \frac{1}{2}$, and the last steps follows by the definitions of $\varrho_{(X_t)_{t=1}^T}(\pi^*) := \sum_{t=1}^T (1 - p_t(\pi^*(X_t)|X_t))$ and $\overline{\varrho}_X(\pi^*) := \mathbb{E}[\varrho_{(X_t)_{t=1}^T}(\pi^*)]$. ☐

The following lemma that bounds the sum of entropy in terms of $\varrho_0(\pi^*)$ follows by a similar argument as Lemma 4 of Ito et al. [2022].

Lemma 22. *For any $\pi \in \Pi$ and for a fixed ghost sample X_0, we have*

$$\sum_{t=1}^T H(p_t(\cdot|X_0)) \leq \varrho_0(\pi^*) \ln \frac{\mathrm{e}KT}{\varrho_0(\pi)},$$

where $\varrho_0(\pi) = \sum_{t=1}^T (1 - p_t(\pi(X_0)|X_0))$.

Proof of Lemma 22. By the similar calculation of (30) in Ito et al. [2022], we see that for any distribution $p \in \Delta([K])$, and for any $i^* \in [K]$, it holds that

$$H(p) \leq (1 - p_{i^*})\left(\ln \frac{K-1}{1-p_{i^*}} + 1\right).$$

Using this inequality, for a fixed X_0, it holds that

$$\sum_{t=1}^{T} H(p_t(\cdot|X_0)) \le \sum_{t=1}^{T}(1 - p_t(\pi^*(X_0)|X_0)) \left(\ln \frac{K-1}{1 - p_t(\pi^*(X_0)|X_0)} + 1\right)$$
$$\le \varrho_0(\pi^*) \left(\ln \frac{(K-1)T}{\varrho_0(\pi^*)} + 1\right) \le \varrho_0(\pi^*) \ln \frac{\mathrm{e}KT}{\varrho_0(\pi^*)},$$

where the second inequality follows from Jensen's inequality. □

Using Lemma 22, we have $\sum_{t=1}^{T} H(p_t(\cdot|X_0)) \le \mathrm{e}\ln(\mathrm{e}KT) + \mathrm{e}^{-1}$ in the case of $\varrho_0(\pi^*) < \mathrm{e}$, which gives us the desired bound. Next, we consider the case of $\varrho_0(\pi^*) \ge \mathrm{e}$. In this case, we have $\sum_{t=1}^{T} H(p_t(\cdot|X_0)) \le \varrho_0(\pi^*)\ln(KT)$. Hence, for $\pi^* \in \Pi$ we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} H(p_t(\cdot|X_0))\right] \le \mathbb{E}\left[\varrho_0(\pi^*)\ln(KT)\right] = \mathbb{E}\left[\varrho_{(X_t)_{t=1}^T}(\pi^*)\ln(KT)\right] = \overline{\varrho}_X(\pi^*)\ln(KT), \tag{62}$$

where we used Lemma 20 in the first equality.

Let $c_4 = \frac{K}{\lambda_{\min}(\boldsymbol{\Sigma})}\ln(K)\ln(T) + 4$. Therefore, by Lemma 21, (59), and (62) for any $\lambda > 0$, it holds that

$$\begin{aligned}
R_T &= (1+\lambda)R_T - \lambda R_T \\
&\le \mathbb{E}\left[(1+\lambda)\sqrt{\left(d + \frac{\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})}\right) K \ln T \ln(KT) \cdot \overline{\varrho}_X(\pi^*)} - \lambda \frac{\Delta_{\min}}{2}\overline{\varrho}_X(\pi^*)\right] + \lambda \cdot 2C + (1+\lambda)c_4 \\
&= \mathcal{O}\left(\frac{(1+\lambda)^2(d + \frac{\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})})K\ln T\ln(KT)}{\lambda\Delta_{\min}} + \lambda C + \lambda c_4\right) \\
&= \mathcal{O}\left(\frac{(d + \frac{\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})})K\ln T\ln(KT)}{\Delta_{\min}} + \lambda\left(\frac{(d + \frac{\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})})K\ln T\ln(KT)}{\Delta_{\min}} + C\right)\right. \\
&\qquad \left. + \frac{(d + \frac{\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})})K\ln T\ln(KT)}{\Delta_{\min}\cdot\lambda} + \lambda c_4\right),
\end{aligned}$$

where we used $a\sqrt{x} - \frac{bx}{2} \le \frac{a^2}{2b}$ for any $a, b, x \ge 0$ in the first equality.

By letting $0 \le \lambda \le 1$ to be

$$\lambda = \sqrt{\frac{(d + \frac{\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})})K\ln T\ln(KT)\Delta_{\min}^{-1}}{(d + \frac{\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})})K\ln T\ln(KT)\Delta_{\min}^{-1} + C + c_4}},$$

we have

$$\begin{aligned}
R_T &= \mathcal{O}\left(\frac{(d + \frac{\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})})K\ln T\ln(KT)}{\Delta_{\min}} + \sqrt{\frac{(d + \frac{\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})})K\ln T\ln(KT)}{\Delta_{\min}}\cdot C}\right. \\
&\qquad \left. + \sqrt{\frac{c_4(d + \frac{\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})})K\ln T\ln(KT)}{\Delta_{\min}}}\right) \\
&= \mathcal{O}\left(\frac{(d + \frac{\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})})K\ln T\ln(KT)}{\Delta_{\min}} + \sqrt{\frac{(d + \frac{\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})})K\ln T\ln(KT)}{\Delta_{\min}}\cdot C}\right. \\
&\qquad \left. + K\ln T\sqrt{\frac{\frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})}\left(d + \frac{\ln T}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)\ln(K)\ln(KT)}{\Delta_{\min}}}\right),
\end{aligned}$$

which concludes the proof of the theorem. □