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

Yuko Kuroki
CENTAI Institute
Turin, Italy
Alberto Rumi
Università degli Studi di Milano
and CENTAI Institute
Milan, Italy
Taira Tsuchiya
The University of Tokyo
Tokyo, Japan
Fabio Vitale
CENTAI Institute
Turin, Italy
Nicolò Cesa-Bianchi
Università degli Studi di Milano
and Politecnico di Milano
Milan, Italy

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 $\frac{dK}{\Delta_{\min}^{\frac{5}{2}}} \log(dKT)$, where $\Delta_{\min}$ is the minimum suboptimality gap over the $d$-dimensional context space. In the adversarial regime, we obtain either the first-order $\tilde{O}(dK\sqrt{L^*})$ bound, or the second-order $\tilde{O}(dK\sqrt{\Lambda^*})$ bound, where $L^*$ is the cumulative loss of the best action and $\Lambda^*$ 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 $\tilde{O}(dK\sqrt{T})$ 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 (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 $\Omega(\sqrt{dKT})$ 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 $\Lambda^*$ is a notion of cumulative second moment for the losses incurred by the algorithm.

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

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 Take-mura, 2023a,b), 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 $\frac{dK \ln(T)}{\Delta_{\text{min}}}$ in order of $\frac{dK}{\Delta_{\text{min}}}$ for the stochastic setting, ignoring logarithmic factors in $K$ and $d$, where $\Delta_{\text{min}}$ is the minimum sub-optimality gap over the context space. In the adversarial regime, we obtain either a first-order $\tilde{O}(dK\sqrt{T})$ bound, or a second order $\tilde{O}(dK\sqrt{K}T)$ 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 $\tilde{O}(dK\sqrt{T})$ regret in the adversarial regime (Theorem 2), without prior knowledge of the inverse of the contextual covariance matrix $\Sigma$. 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 $(TK^{-2}d^{-1})^{Kd}$ (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 $O(\sqrt{T})$ — 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 $\Sigma^{-1}$ 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 $\Sigma^{-1}$ computed using Matrix Geometric Resampling (MGR) of Neu and Bartók (2013); Neu and Bartók (2016), thus avoiding the advance knowledge of $\Sigma^{-1}$. 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 Ziehrn 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 $O(\sqrt{T})$ 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 $\Sigma^{-1}$.

Table 1 summarizes our results in the context of the previous literature. The upper bound of Ziehrn et al. (2023) is for a combinatorial contextual setting where the action space satisfies $A \subseteq \{0,1\}^K$ and we assume $\max_{a \in A} ||a||_1 \leq 1$. The best known lower
Table 1: A comparison of regret bounds for linear contextual bandits. \( \tilde{O} \) 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.

<table>
<thead>
<tr>
<th>reference</th>
<th>stochastic</th>
<th>adversarial</th>
<th>( \sqrt{C} )</th>
<th>( \Sigma^{-1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neu and Olkhovskaya (2020)</td>
<td>–</td>
<td>( \tilde{O}\left(\sqrt{TK \max \left{ d, \frac{\ln T}{\lambda_{\min}(\Sigma)} \right} \ln(K)\right) )</td>
<td>–</td>
<td>Unknown</td>
</tr>
<tr>
<td>Olkhovskaya et al. (2023)</td>
<td>–</td>
<td>( \tilde{O}\left(K\sqrt{dK^*}\right) )</td>
<td>Known</td>
<td></td>
</tr>
<tr>
<td>Olkhovskaya et al. (2023)</td>
<td>–</td>
<td>( \tilde{O}\left(K\sqrt{dT}\right) )</td>
<td>Known</td>
<td></td>
</tr>
<tr>
<td>Zierahn et al. (2023)</td>
<td>–</td>
<td>( \tilde{O}\left(\sqrt{TK \max \left{ d, \frac{\ln T}{\lambda_{\min}(\Sigma)} \right} \ln(K)\right) )</td>
<td>–</td>
<td>Unknown</td>
</tr>
<tr>
<td>Proposition 8</td>
<td>( \tilde{O}\left(\frac{K^2}{\lambda_{\min}} \left( d + \frac{1}{\lambda_{\min}(\Sigma)} \right)^2 \ln(K) \ln(T) \right) )</td>
<td>( \tilde{O}\left(\sqrt{TK^2 \left( d + \frac{1}{\lambda_{\min}(\Sigma)} \right)^2 \ln(K) \ln(T) \ln(K)\right) )</td>
<td>✔</td>
<td>Known</td>
</tr>
<tr>
<td>Theorem 1</td>
<td>( \tilde{O}\left(\frac{(dK)^2}{\lambda_{\min}} \ln^2(dKT) \ln^3(T) \right) )</td>
<td>( \tilde{O}\left(dK\sqrt{\lambda_{\min}(\Sigma)}\right) )</td>
<td>✔</td>
<td>Known</td>
</tr>
<tr>
<td>Corollary 1</td>
<td>( \tilde{O}\left(\frac{(dK)^2}{\lambda_{\min}} \ln^2(dKT) \ln^3(T) \right) )</td>
<td>( \tilde{O}\left(dK\sqrt{\min{L^*,\Sigma}}\right) )</td>
<td>✔</td>
<td>Known</td>
</tr>
<tr>
<td>Theorem 2</td>
<td>( \tilde{O}\left(\frac{K}{\lambda_{\min}} \left( d + \frac{\ln T}{\lambda_{\min}(\Sigma)} \right) \ln(KT) \ln(T) \right) )</td>
<td>( \tilde{O}\left(\sqrt{TK \left( d + \frac{\ln T}{\lambda_{\min}(\Sigma)} \right) \ln(T) \ln(K)\right) )</td>
<td>✔</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

bound for the adversarial or distribution-free setting is \( \Omega(\sqrt{dKT}) \) 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; Liu et al., 2023), 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 \( \theta_{t,a} \in \mathbb{R}^d \).
- Independently of the choice of loss vectors \( \theta_{t,a} \) 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; Dunn et al., 2023), we make the following assumptions:

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

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

Adversarial regime. The loss function is defined by \( \ell_t(X_t, \theta_{t,a}) := (X_t, \theta_{t,a}) \), where \( \theta_{t,a} \) is chosen by an oblivious adversary for all \( a \) and \( t \).
Stochastic regime. The loss function is defined by
\[ \ell_t(X_t, a) := \langle X_t, \theta_a \rangle + \varepsilon_t(X_t, a), \]
where \( \theta_a \) 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
\[ \ell_t(X_t, a) := \langle X_t, \theta_{t,a} \rangle + \varepsilon_t(X_t, a), \]
where \( \theta_{t,a} \) are independent and bounded zero-mean noise and the vectors \( \theta_{t,a} \) are such that there exist fixed and unknown vectors \( \theta_1, \ldots, \theta_K \) and an unknown constant \( C > 0 \) for which \( \sum_{i=1}^{K} \max_{a \in [K]} \| \theta_{i,a} - \theta_a \| \leq C \) 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 \( \Pi \) 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:
\[ \pi^*(x) := \arg \min_{a \in [K]} \mathbb{E} \left[ \sum_{t=1}^{T} \ell_t(x, a) \right] \quad \forall x \in \mathcal{X}, \quad (1) \]
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 \( \pi^* \):
\[ R_T = \mathbb{E} \left[ \sum_{t=1}^{T} \left( \ell_t(X_t, A_t) - \ell_t(X_t, \pi^*(X_t)) \right) \right], \quad (2) \]
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 \( \theta_1, \ldots, \theta_K \), let \( \Delta_{\min}(x) := \min_{a \neq \pi^*(x)} \langle x, \theta_a - \theta_{\pi^*(x)} \rangle \) for all \( x \in \mathcal{X} \). Then, we define the minimum sub-optimality gap by \( \Delta_{\min} := \min_{x \in \mathcal{X}} \Delta_{\min}(x) > 0 \).

We denote the cumulative loss incurred by the optimal policy by \( L^* := \mathbb{E} \left[ \sum_{t=1}^{T} \ell_t(X_t, \pi^*(X_t)) \right] \) and the cumulative variance of a policy choosing actions \( A_1, A_2, \ldots \) with respect to a predictable loss sequence \( m_{t,a} \in \mathbb{R}^d \) for action \( a \) by \( \Lambda^* := \mathbb{E} \left[ \sum_{t=1}^{T} (\ell_t(X_t, A_t) - \langle X_t, m_{t,a} \rangle)^2 \right] \). We use \( \Delta := \mathbb{E} \left[ \sum_{t=1}^{T} \left( \ell_t(X_t, A_t) - \langle X_t, \theta_a \rangle \right)^2 \right] \) with \( \theta := \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sum_{s=1}^{K} \theta_{t,a} \).

Additional notation. We denote by \( \mathbb{E}[\cdot] \) the expectation over a random variable (r.v.) \( X \). We denote by \( \mathbb{E} [X | Y] \) the expectation over \( X \) conditioned on \( Y \). When we write \( \mathbb{E}[X] = \mathbb{E}[X|Y] \), we take the expectation conditioned on \( Y \) with respect to all sources of randomness in \( X \). We denote by \( \mathcal{F}_t = \sigma(X_s, A_s, \forall s \leq t) \) the filtration generated by all the random variables \( X_s \) and the set of actions \( A_s \), for each \( s \leq t \). Then we write \( \mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_{t-1}] \). For any semi-definite matrix \( B \in \mathbb{R}^{d \times d} \), we use \( \lambda_{\min}(B) \) to denote the smallest eigenvalue of \( B \), and write \( \| u \|_B = \sqrt{u^T Bu} \) for \( u \in \mathbb{R}^d \). We also define the probabilistic policy mapping each context \( x \) to a probability distribution \( \pi(\cdot | x) \) over \([K]\) (i.e., an element of the simplex \( \Delta([K]) \)). For the analysis of data-dependent bounds, we use the notation \( \xi_{t,a} := \langle \ell_t(X_t, a) - \langle X_t, m_{t,a} \rangle \rangle \in \mathbb{R} \) with a loss predictor \( m_{t,a}^* \) for \( t \in [T] \) and \( a \in [K] \). We write \( 1[\cdot] \) 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
\[ p_t(\cdot | X_t) = \arg \min_{r \in \Delta([K])} \left\{ \sum_{s=1}^{t-1} \langle r, \hat{\ell}_s(X_t) \rangle + \psi_t(r) \right\}, \]
where \( \hat{\ell}_s(X_t) := (\langle X_t, \hat{\theta}_{s,1} \rangle, \ldots, \langle X_t, \hat{\theta}_{s,K} \rangle)^T \in \mathbb{R}^K \), and \( \hat{\theta}_{t,a} \) is an estimator of the linear loss \( \theta_{t,a} \in \mathbb{R}^d \). 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
\[ p_t(a | X_t) = \frac{\exp(-\eta_t \sum_{s=1}^{t-1} \langle r, \hat{\ell}_s(X_t) \rangle) \exp(-\eta_t \sum_{s=1}^{t-1} \langle r, \hat{\theta}_{s,a} \rangle)}{\sum_{b \in [K]} \exp(-\eta_t \sum_{s=1}^{t-1} \langle r, \hat{\theta}_{s,b} \rangle)}. \]
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 \( m_{t,a} \in \mathbb{R}^d \) for each action \( a \) is available to the learner at the beginning of each round \( t \). OFTRL can be viewed as adding \( m_{t,a} \) to the objective as a guess for the next loss vector. The OFTRL prediction \( p_t(\cdot | X_t) \) is then defined as
\[ \arg \min_{r \in \Delta([K])} \left\{ \sum_{s=1}^{t-1} \langle r, \hat{\ell}_s(X_t) \rangle + \langle r, m_t(X_t) \rangle + \psi_t(r) \right\}, \]
where \( m_t(X_t) := (\langle X_t, m_{t,1} \rangle, \ldots, \langle X_t, m_{t,K} \rangle) \in \mathbb{R}^K \).

In the following sections, we apply OFTRL in Theorem 1 exploiting the predicted loss \( m_t(X_t) \) 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 $m_{t,a}$ (Theorem 1). By choosing the appropriate loss predictor $m_{t,a}$, we also show how to simultaneously achieve first- and second-order bounds (Corollary 1).

MWU-LC (Algorithm 1) is an instance of OFTL 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 $h : \mathbb{R}^d \to \mathbb{R}_{>0}$ 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 $\Sigma^{-1}$, the inverse of the covariance matrix of contexts. However, these assumptions will be dropped in Section 5. We assume that loss predictors satisfy $(X_t, m_{t,a}) \in [-1, 1]$ 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 cost predictor $m_{t,a} \in \mathbb{R}^d$ 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 $w_t(\cdot|X_t)$:

$$w_t(r|X_t) := \exp \left( -\eta_t \left( \sum_{k=1}^{t-1} \langle r, \tilde{\ell}_k(X_t) \rangle + \langle r, m_t(X_t) \rangle \right) \right),$$

$$p_t(r|X_t) := \frac{w_t(r|X_t)}{\int_{\Delta([K])} w_t(y|X_t) \, dy},$$

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

For the rejection sampling step in Algorithm 1-1, we use the following covariance matrix $\Sigma_{t,a} \in \mathbb{R}^{d \times d}$:

$$\Sigma_{t,a} := \mathbb{E}_{X_t \sim p_t(\cdot|X)} \left[ y_t(a)^2 X X^\top | F_{t-1} \right].$$

The number of steps required for the rejection sampling is $O(1)$, 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 $\tilde{p}_t(\cdot|X_t)$ of $p_t(\cdot|X_t)$ is defined as:

$$\tilde{p}_t(r|X_t) := \frac{p_t(r|X_t) \mathbb{1} \left[ \sum_{a=1}^K y_{t,a}^2 \|X_t\|_2^2 \sum_{a=1}^K \sum_{a=1}^K y_{t,a}^2 \|X_t\|_2^2 \leq dK^2 \gamma_t^2 \right]}{\mathbb{P}_{y \sim p_t(\cdot|X)} \left[ \sum_{a=1}^K y_{t,a}^2 \|X_t\|_2^2 \sum_{a=1}^K \sum_{a=1}^K y_{t,a}^2 \|X_t\|_2^2 \leq dK^2 \gamma_t^2 \right]}$$

for $r \in \Delta([K])$, where $\gamma_t > 1$ is the truncation level to be specified soon. Thus, $Q_t \in \Delta([K])$ is sampled from the truncated distribution $\tilde{p}_t(\cdot|X_t)$ 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 $\text{upd}_t$ is set to 1, otherwise $\text{upd}_t$ is set to 0. Then MWU-LC constructs an unbiased estimator $\tilde{\theta}_{t,a}$ of $\theta_{t,a}$ for each $a \in [K]$ as follows:

$$\tilde{\theta}_{t,a} := m_{t,a} + \frac{\text{upd}_t}{q_t} Q_t(a) \sum_{t,a} \xi_{t,a} X_t \mathbb{1} \left[ A_t = a \right],$$

where $\xi_{t,a} = (\ell_t(X_t, a) - (X_t, m_{t,a}))$ and $\Sigma_{t,a} \in \mathbb{R}^{d \times d}$ is given by:

$$\Sigma_{t,a} := \mathbb{E}_X \left[ Q_t(a)^2 X X^\top | F_{t-1} \right].$$

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

$$\eta_t := \left( \frac{800dK_\gamma^2}{\min_{j \leq t} \beta_j} + \sum_{j=1}^{t-1} \beta_j \right)^{-\frac{1}{2}},$$

where we set $\beta_j := 16\gamma_j^2 \xi_{t,A_t}^2$ and $\gamma_t := 4\ln(10dKt)$ 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 $\Sigma_{t,a}$ in (5) and $\Sigma_{t,a}$ in (7) are known to the learner at each round $t$ and action $a \in [K]$. 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 $\kappa(d, K, \tau) = \mathbb{E} \left[ \sum_{t=1}^\tau \ell_t(X_t, A_t) - \ell_t(X_t, a^*) \right]$
Algorithm 1: Continuous MWU (MWU-LC)

**Input**: Set of $K$ arms

Receive update probability $q_t$;

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

| Observe $X_t$;
| Draw $Q = p_t(\cdot|X_t)$ defined in (4)
| while $\sum_{a=1}^{K} Q(a)^2 \|X_t\|_{F_{t,a}}^2 \leq dK^2 \tau$;

$Q_t := Q \in \Delta([K])$;

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 $\theta_{t,a}$ for $a \in [K]$ as in (6);

Update $p_t(\cdot|X_t)$ as in (4);

Update $\eta_t$ as in (8);

| End do

32$Kd\ln(10dK\tau)\ln(\tau)$. Then, for any $\tau$ and $a^*$, the regret $R(\tau, a^*)$ of MWU-LC is bounded by

$$
\kappa(d, K, \tau) \left( \mathbb{E} \left[ \sum_{t=1}^{\tau} \frac{\ell_t p_{t,a}^2}{q_t} \right] + \mathbb{E} \left[ \sqrt{\frac{50dK}{\min_{j \leq \tau} q_j}} \right] \right).
$$

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 $\Sigma_{t,a}$ in (5) and $\Sigma_{t,a}$ in (7) are known to the learner at each round $t$ and action $a$. Let $\kappa_1(d, K, T) = K^2d^2\ln^2(dKT)\ln^2(T)$ 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

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

in the adversarial regime and

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

in the corrupted stochastic regime.

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

$$
\|m\|_S^2 + \sum_{j=1}^{T} I[A_j = a] (\langle \theta_{j,a} - m, X_j \rangle)^2
$$

(9)

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

**Corollary 1.** Let $m_{t,a}$ 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 $m^* \in M$, $R_T$ is bounded by

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

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

**Remark 1.** Although the first-order bound is obtained by just setting $m_{t,a} = 0$ (see Corollary 2 in Appendix F), computing the minimizer of (9) as $m_{t,a}$ 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 $D$. However, removing the prior knowledge of $\Sigma^{-1}$ seems challenging, as computation of (5) and (7) involves expectation depending on both $D$ and a learner’s policy. Moreover, the continuous exponential weights incur a high (yet polynomial) sampling cost, resulting in $O((K^5 + \ln T) g_{\Sigma})$ per round running time, where $g_{\Sigma}$ 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 UNKNOWN $\Sigma^{-1}$ CASE**

We present a computationally efficient algorithm, called FTRL-LC, based on FTRL with negative Shannon entropy. This algorithm does not require knowledge of $\Sigma^{-1}$, and only needs access to context distribution $D$ and minimum eigenvalue $\lambda_{\min}(\Sigma)$.
While this estimate is unbiased, \( T \) o 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 \( \hat{\Sigma}_{t,a}^+ \) as the estimate of \( \Sigma_{t,a}^{-1} \). MGR can be implemented in \( \mathcal{O}(M_tKd + Kd^2) \) time (Neu and Olkhovskaya, 2020). Using \( \hat{\Sigma}_{t,a}^+ \), we can define the estimator of \( \theta_{t,a} \) by

\[
\hat{\theta}_{t,a} := \hat{\Sigma}_{t,a}^+X_t\ell_t(X_t, A_t)1 \quad \forall a \in [K].
\]

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

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

\[
\beta_{t+1}^r = \beta_t^r + c_t^r \left/ \sqrt{1 + (\ln T)^{-1} \sum_{s=1}^t H(p_s(\cdot|X_s))} \right., \quad \beta_t^r = \max \left\{ 2, c_2^r \ln T, \beta_t^r \right\},
\]

\[
\eta_t = \frac{1}{\beta_t^r}, \quad \gamma_t = \alpha_t \cdot \eta_t, \quad \alpha_t = 4K \ln(t) / \lambda_{\min}(\Sigma),
\]

where \( c_t^r = \sqrt{3Kd + 2K \ln T \ln K} \), \( c_2^r = 8K / \lambda_{\min}(\Sigma) \), and we set \( \beta_0^r = c_1^r = 1 \). 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 \( t \) to 1 to

\[
M_t = \left\lceil \frac{4K}{\gamma_t \lambda_{\min}(\Sigma) \ln(t)} \right\rceil \quad (\geq 1).
\]

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

\[
R_T = \mathcal{O} \left( \sqrt{T \left( d + \frac{\ln T}{\lambda_{\min}(\Sigma)} \right) K \ln(K) \ln(T) + c_4} \right).
\]
For the stochastic regime, the regret is bounded by
\[ R_T = O\left( \frac{K}{\Delta_{\min}} \left( d + \frac{\ln T}{\lambda_{\min}(\Sigma)} \right) \ln(KT) \ln T \right) =: R^{\text{sto}}_T, \]
and for the corrupted stochastic regime, the regret is bounded by
\[ R_T = O\left( R^{\text{sto}}_T + \sqrt{C R^{\text{sto}}_T} \right). \]

Our bound achieves \( \tilde{O}\left( \sqrt{TK \max\left\{ d, \frac{1}{\lambda_{\min}(\Sigma)} \right\}} \right) \) recovering the best-known result in the adversarial regime (Neu and Olkhovskaya, 2020; Ziebart et al., 2023) up to log-factors when \( T \geq \frac{k^2}{\lambda_{\min}(\Sigma)} \), 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 \( x \in \mathcal{X} \) does not change over time, and for each trial \( t \in [T] \) the incurred loss is obtained replacing \( \theta_{t,a} \) by a (possibly biased) loss vector estimator \( \tilde{\theta}_{t,a} \) as follows. Let \( \tilde{\theta}_{t,a} \in \mathbb{R}^d \) be an estimator of the loss vector \( \theta_{t,a} \) with bias \( b_{t,a} \in \mathbb{R}^d \) and \( a \in [K] \). Suppose that the learner’s action \( A_t \) is selected by a probabilistic policy \( \pi_t(\cdot|x) \in \Delta([K]) \). Then, the regret in the auxiliary game against the comparator \( \pi^*(x) \) defined in (1) for the estimated loss is defined as
\[ \widetilde{R}_T(x) := E\left[ \sum_{t=1}^{T} \langle x, \tilde{\theta}_{t,a} \rangle - \langle x, \tilde{\theta}_{t,a} \rangle \right]. \]

As in Neu and Olkhovskaya (2020); Olkhovskaya et al. (2023); Ziebart 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 \( X_1, \ldots, X_T \). We introduce the notation that \( R_T \) is bounded as follows (see Eq.(6) in Neu and Olkhovskaya (2020) and Lemma 7 in Appendix D):
\[ R_T \leq E[\widetilde{R}_T(X_0)] + 2 \sum_{t=1}^{T} \max_{a \in [K]} \left| E[\langle X_t, b_{t,a} \rangle] \right|. \]

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 \( \max_{x \in \mathcal{X}} |\eta_t(x, \tilde{\theta}_{t,a})| \leq 1 \) holds, and \( A_t \) is chosen by \( \pi_t(\cdot|x) \) defined by (11) for \( x \in \mathcal{X} \). Then, we have
\[ \widetilde{R}_T(x) \leq \sum_{t=1}^{T} \left( \beta_{t+1} - \beta_t \right) H(p_{t+1}(\cdot|x)) + \beta_1 \ln K \]
\[ + \sum_{t=1}^{T} \eta_t \sum_{a=1}^{K} \pi_t(a|x) \langle x, \tilde{\theta}_{t,a} \rangle^2 + U(x), \]
where \( U(x) = \sum_{t=1}^{T} \gamma_t \sum_{a \neq \pi^*(x)} \frac{1}{K} \langle x, \tilde{\theta}_{t,a} \rangle \) 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) \( \mathbb{E}[|\langle X_t, \tilde{\theta}_{t,a} - \tilde{\theta}_{t,a} \rangle|] \leq \exp\left( -\frac{\gamma_t \lambda_{\min}(\Sigma) M_t}{2K} \right) \leq 1/t^2 \)
and (ii) \( |\eta_t(x, \tilde{\theta}_{t,a})| \leq 1, \forall x \in \mathcal{X} \).

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 \( \theta_{t,a} \), i.e., \( \mathbb{E}[U(X_0)] \) and \( \sum_{t=1}^{T} \max_{a \in [K]} \left| \mathbb{E}[\langle X_t, b_{t,a} \rangle] \right| \).

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 \( \kappa = c_1' \sqrt{\ln K} + \frac{3Kd + 2\ln T}{c_1' \sqrt{\ln K}} \ln T \) and \( A_t \) 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
\[ \mathcal{O}\left( \kappa \left[ \sum_{t=1}^{T} H(p_t(\cdot|X_0)) \right] + \frac{K \ln K}{\lambda_{\min}(\Sigma)} \ln T \right), \]

We introduce the following notation for the further analysis: Let \( \varphi_t(\pi^*) := \sum_{t=1}^{T} (1 - p_t(\pi^*(X_0)|X_0)) \) and \( \varphi_t(X_t) := \sum_{t=1}^{T} (1 - p_t(X_t)|X_t) \). 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 \( R_T \geq \frac{2}{\kappa} \mathbb{E}[\varphi_t(X_t)^2] - 2C \) 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 $\sum_{t=1}^{T} H(p_t(\cdot|X_0)) \leq \varrho_0(\pi^*) \ln \frac{e^{K T}}{\varrho_0(\pi^*)}$ (Lemma 22 in Appendix G). When $\varrho_0(\pi^*) < e$, then we have the desired bound trivially from this inequality. In the case of $\varrho_0(\pi^*) \geq e$, 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 $R_T = \tilde{O}(\text{poly}(\ln T) \cdot \sqrt{} \mathbb{E}[^{\varrho_0(\pi^*)}] + c_4)$, 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 $\mathbb{E}[\varrho_0(X_{1:t}^{T})^\pi(\cdot)] = \mathbb{E}[\varrho_0(\pi^*)]$ (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.

\section{Conclusions}

We proposed the first 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 bound 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 $\Sigma^{-1}$ at each round $t$ for action $a$, and achieves the worst-case regret in the adversarial regime without sacrificing the polylogarithmic regret in the (corrupted) stochastic regime.

It is important to develop a computationally efficient algorithm that can achieve data-dependent bounds without relying on knowledge of $\Sigma^{-1}$. 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.

Acknowledgements

YK was partially supported by JST, ACT-X Grant Number JPMJAX200E. TT was supported by JST, ACT-X Grant Number JPMJAX210E, Japan. NCB acknowledges the financial support from the MUR PRIN grant 2022EKNE5K (Learning in Markets and Society), the NextGenerationEU program within the PNRR-PE-AI scheme (project FAIR), the EU Horizon CL4-2022-HUMAN-02 research and innovation action under grant agreement 101120237 (project ELIAS).

References


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


Checklist

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

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

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

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

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]
In this appendix, we provide Table A summarizing the most important notations used in the paper.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$[K] := {1, 2, \ldots, K}$</td>
<td>Finite action set</td>
</tr>
<tr>
<td>$d \in \mathbb{N}$</td>
<td>Dimension of loss vectors and contexts</td>
</tr>
<tr>
<td>$\mathcal{X} \subseteq \mathbb{R}^d$</td>
<td>A context space of a full-dimensional compact set</td>
</tr>
<tr>
<td>$\mathcal{D} \in \Delta(\mathcal{X})$</td>
<td>Context distribution over $\mathcal{X}$</td>
</tr>
<tr>
<td>$\Sigma \in \mathbb{R}^{d \times d}$</td>
<td>Covariance matrix of contexts, $\mathbb{E}_{X \sim \mathcal{P}}[XX^\top]$</td>
</tr>
<tr>
<td>$\theta_{t,a} \in \mathbb{R}^d$</td>
<td>Loss vector of action $a \in [K]$ at round $t \in [T]$</td>
</tr>
<tr>
<td>$\theta_a \in \mathbb{R}^d$</td>
<td>Fixed and unknown vectors of action $a \in [K]$ at round $t \in [T]$ (corrupted and stochastic regime)</td>
</tr>
<tr>
<td>$C \in [0,T]$</td>
<td>Corruption level, upper bound of $\sum_{t=1}^T \max_{a \in [K]} |\theta_{t,a} - \theta_a|^2$</td>
</tr>
<tr>
<td>$\pi(\cdot</td>
<td>x) \in \Delta([K])$</td>
</tr>
<tr>
<td>$\Pi$</td>
<td>Set of all deterministic policies $\pi : \mathcal{X} \to [K]$</td>
</tr>
<tr>
<td>$\pi^* \in \Pi$</td>
<td>Optimal policy</td>
</tr>
<tr>
<td>$\Delta_{\min} &gt; 0$</td>
<td>Minimum sub-optimal gap over a context space, $\min_{x \in \mathcal{X}} \min_{a \in \pi^<em>(x)} \langle x, \theta_a - \theta_{\pi^</em>(x)} \rangle$</td>
</tr>
<tr>
<td>$m_{t,a} \in \mathbb{R}^d$</td>
<td>Loss predictor for action $a \in [K]$ and $t \in [T]$</td>
</tr>
<tr>
<td>$L^*$</td>
<td>$\mathbb{E} \left[ \sum_{t=1}^T \ell_t(X_t, \pi^*(X_t)) \right]$</td>
</tr>
<tr>
<td>$\Lambda^*$</td>
<td>$\mathbb{E} \left[ \sum_{t=1}^T (\ell_t(X_t, A_t) - \langle X_t, m_{a_t} \rangle)^2 \right]$</td>
</tr>
<tr>
<td>$\tilde{\ell}_t(X_t) \in \mathbb{R}^K$</td>
<td>Estimated loss vector for $X_t$ at round $t \in [T]$, $(X_t, \tilde{\theta}<em>{a_1}, \ldots, X_t, \tilde{\theta}</em>{a_K})$</td>
</tr>
<tr>
<td>$\tilde{m}_t(X_t) \in \mathbb{R}^K$</td>
<td>Predicted loss vector for $X_t$ at round $t \in [T]$, $(X_t, m_{a_1}, \ldots, X_t, m_{a_K})$</td>
</tr>
<tr>
<td>$\bar{R}_{\mathcal{P}}(x)$</td>
<td>Regret of auxiliary game for context $x$</td>
</tr>
<tr>
<td>$\tilde{\theta}_{t,a} \in \mathbb{R}^d$</td>
<td>Biased estimator for $\theta_{t,a}$ for $a \in [K]$ and $t \in [T]$</td>
</tr>
<tr>
<td>$\tilde{\ell}_t(X_t) \in \mathbb{R}^K$</td>
<td>Estimated loss vector for $X_t$ at round $t \in [T]$, $(X_t, \tilde{\theta}<em>{a_1}, \ldots, X_t, \tilde{\theta}</em>{a_K})$</td>
</tr>
<tr>
<td>$\bar{R}_{\mathcal{P}}(x)$</td>
<td>Regret of auxiliary game for context $x$ and loss estimator $\tilde{\theta}<em>{t,a}$ at round $t$, $\mathbb{E} \left[ \sum</em>{t=1}^T \langle x, \tilde{\theta}<em>{t,a} \rangle - \langle x, \tilde{\theta}</em>{t,\pi^*(x)} \rangle \right]$</td>
</tr>
</tbody>
</table>
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 $\pi : \mathcal{X} \rightarrow [K]$, 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.
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 \( m_{t,a} = 0 \) are orthogonal, as formally stated in the following lemmas.

**Lemma 4.** Suppose that \( \mathcal{X} \) consists of only basis vectors i.e., \( \mathcal{X} = \{e_1, \ldots, e_K\} \), and pick some \( t \in [T] \). Let \( X_{t'} \neq X_t \) and let \( a \in [K] \). Then, \( \langle X_{t'}, \tilde{\theta}_{t,a} \rangle = 0 \) holds for the biased estimator \( \tilde{\theta}_{t,a} \) in (12) of FTRL-LC, and \( \langle X_{t'}, \hat{\theta}_{t,a} \rangle = 0 \) holds for the unbiased estimator \( \hat{\theta}_{t,a} \) with \( m_{t,a} = 0 \) in (6) of MWU-LC.

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

\[
\tilde{\Sigma}_{t,a}^+ = \rho \sum_{k=0}^M \prod_{j=1}^k (I - \rho B_{k,a}).
\]

Notice that \( B_{k,a} = 1[A(k) = a]X(k)X(k)^\top \) is diagonal since \( D \) has the support of \( \mathcal{X} = \{e_1, \ldots, e_K\} \). So as \( \tilde{\Sigma}_{t,a}^+ \) for all \( a \in [K] \). Let \( X_{t'} = e_i \) and \( X_t = e_j \) and pick \( a \in [K] \). Then we see that

\[
\langle X_{t'}, \tilde{\theta}_{t,a} \rangle = e_i^\top \tilde{\theta}_{t,a} = e_i^\top \tilde{\Sigma}_{t,a}^+ X_t \ell_t(X_t, A_t) 1[A_t = a] = e_i^\top \tilde{\Sigma}_{t,a}^+ e_j \ell_t(X_t, A_t) 1[A_t = a] = (\tilde{\Sigma}_{t,a}^+)_{i,j} \ell_t(X_t, A_t) 1[A_t = a],
\]

concluding that \( \langle X_{t'}, \tilde{\theta}_{t,a} \rangle = 0 \) if \( i \neq j \).

Next, we consider \( \hat{\theta}_{t,a} \) in (6), where \( \hat{\Sigma}_{t,a}^{-1} \) is given by (7) and \( \xi_t,a = \langle \ell_t(X_t, a) - \langle X_t, m_{t,a} \rangle \rangle \). By a similar discussion, we have

\[
\langle X_{t'}, \hat{\theta}_{t,a} \rangle = e_i^\top \hat{\theta}_{t,a} = e_i^\top (m_{t,a} + \frac{\text{upd}}{q_t} Q_t(a) \hat{\Sigma}_{t,a}^{-1} X_t \xi_t,a 1[A_t = a]) = e_i^\top (m_{t,a} + \frac{\text{upd}}{q_t} Q_t(a) \hat{\Sigma}_{t,a}^{-1} X_t \xi_t,a 1[A_t = a]) = m_{t,a}(i) + (\hat{\Sigma}_{t,a}^{-1})_{i,j} \frac{\text{upd}}{q_t} Q_t(a) \xi_t,a 1[A_t = a].
\]

Therefore, we conclude that \( \langle X_{t'}, \hat{\theta}_{t,a} \rangle = 0 \) if \( i \neq j \) and \( m_{t,a} = 0 \), since \( \hat{\Sigma}_{t,a}^+ \) is diagonal in this case.

**Lemma 5.** Suppose that \( \mathcal{X} \) consists of only basis vectors i.e., \( \mathcal{X} = \{e_1, \ldots, e_K\} \). 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 \( m_{t,a} = 0 \) 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 $X_t = e_i$, then $p_t(\cdot|X_t)$ in (11) in FTRL-LC can be written as
\[ p_t(\cdot|e_i) = \arg\min_{r \in \Delta([K])} \left\{ \sum_{s<t; X_s=e_i} \langle r, \ell_s(e_i) \rangle + \psi_t(r) \right\}, \]
where $\ell_s(e_i) := (\langle e_i, \tilde{\theta}_{s,1} \rangle, \ldots, \langle e_i, \tilde{\theta}_{s,K} \rangle)^T \in \mathbb{R}^K$. Also, we can write $w_t(r|X_t)$ for $r \in \Delta([K])$ in (4) of MWU-LC as
\[ w_t(r|e_i) = \exp\left( -\eta_t \sum_{a \in [K]} r_a \left( \psi_t(e_i, \sum_{s<t; X_s=e_i} \tilde{\theta}_{s,a} + m_{t,a}) \right) \right), \]
where $m_{t,a} = 0$. These equations mean that both algorithms do not use the information at round $s < t$ wherein $X_t \neq X_s$. \hfill \Box

\section{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.

\subsection{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 $x \in \mathcal{X}$. 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 $p_t(\cdot|x)$ be a FTRL prediction with loss estimators $\tilde{\theta}_{t,a}$ for each $a \in [K]$, which is given by (10) with any convex regularizer $\psi_t(\cdot)$. Suppose that $A_t$ is chosen by $\pi_t(\cdot|x) := (1 - \gamma_t)p_t(\cdot|x) + \gamma_t \frac{1}{K}$, where $\gamma_t \in [0,1]$ is the mixture rate. Then, for any context $x \in \mathcal{X}$, we have
\[
\mathbb{E}_{A_t} \left[ \sum_{t=1}^T \left( \langle x, \tilde{\theta}_{t,A_t} \rangle - \langle x, \tilde{\theta}_{t,\pi^*(x)} \rangle \right) \right] \leq \sum_{t=1}^T \left( \psi_t(p_{t+1}(\cdot|x)) - \psi_{t+1}(p_{t+1}(\cdot|x)) + \psi_{t+1}(\pi^*(\cdot|x)) - \psi_1(p_1(\cdot|x)) \right) + \sum_{t=1}^T (1 - \gamma_t) \left( \langle p_t(\cdot|x) - p_{t+1}(\cdot|x), \ell_t(x) \rangle - D_t(p_{t+1}(\cdot|x), p_t(\cdot|x)) \right) + U(x),
\]
where $U(x) = \sum_{t=1}^T \gamma_t \left( \frac{1}{K} \mathbf{1} - \pi^*(\cdot|x), \ell_t(x) \right)$, and $\pi^*(a|x) = 1$ if $a = \pi^*(x)$ otherwise 0.

Proof of Lemma 6. From the definition of the auxiliary game and the design of the algorithm, for any $x \in \mathcal{X}$, we have
\[
\mathbb{E}_{A_t} \left[ \sum_{t=1}^T \left( \langle x, \tilde{\theta}_{t,A_t} \rangle - \langle x, \tilde{\theta}_{t,\pi^*(x)} \rangle \right) \right] = \sum_{t=1}^T \sum_{a \in [K]} (\pi_t(a|x) - \pi^*(a|x)) \langle x, \tilde{\theta}_{t,a} \rangle = \sum_{a \in [K]} (1 - \gamma_t) \sum_{t=1}^T (p_t(a|x) - \pi^*(a|x)) \langle x, \tilde{\theta}_{t,a} \rangle + \sum_{t=1}^T \gamma_t \left( \frac{1}{K} \mathbf{1} - \pi^*(\cdot|x), \ell_t(x) \right) = \sum_{a \in [K]} (1 - \gamma_t) \sum_{t=1}^T (p_t(a|x) - \pi^*(a|x)) \langle x, \tilde{\theta}_{t,a} \rangle + U(x).
\]
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
\[
\sum_{t=1}^{T} (1 - \gamma_t) \sum_{a \in [K]} (p_t(a|x) - \pi^*(a|x)) \langle x, \tilde{\theta}_{t,a} \rangle \\
\leq \sum_{t=1}^{T} (1 - \gamma_t) \left( (p_t(\cdot|x) - p_{t+1}(\cdot|x), \tilde{\ell}_t(x)) - D_t(p_{t+1}(\cdot|x), p_t(\cdot|x)) \right) \\
+ \sum_{t=1}^{T} (\psi_t(p_{t+1}(\cdot|x)) - \psi_{t+1}(p_{t+1}(\cdot|x))) + \psi_{t+1}(\pi^*(\cdot|x)) - \psi_1(p_1(\cdot|x)).
\]

Combining the above arguments completes the proof. \qed

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 \( X_0 \sim D \) be a ghost sample drawn independently from the entire interaction history. Then we have
\[
R_T \leq \mathbb{E}[\tilde{R}_T(X_0)] + 2 \sum_{t=1}^{T} \max_{a \in [K]} |\mathbb{E}[\langle X_t, b_{t,a} \rangle]|.
\]

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 \( X_0 \sim D \) 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 \( \tilde{\theta}_{t,a} \), we have
\[
\mathbb{E}_t \left[ \sum_{a=1}^{K} \pi_t(a|x_0) \langle X_0, \tilde{\theta}_{t,a} \rangle^2 \right] \leq 3Kd. \tag{17}
\]

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 \( \tilde{\theta}_{t,a} = \sum_{t,a}^{-1} X_t \ell_t(X_t, A_t) 1[A_t = a] \) for all \( a \in [K] \), and let \( \tilde{\theta}_{t,a} = \sum_{t,a}^{+} X_t \ell_t(X_t, A_t) 1[A_t = a] \) for all \( a \in [K] \) where \( \sum_{t,a}^{+} \) is obtained by MGR with \( \rho = \frac{1}{2} \) of Algorithm 7. Then, we have
\[
|\mathbb{E}[\langle X_t, \tilde{\theta}_{t,a} - \hat{\theta}_{t,a} \rangle |F_{t-1}]| \leq \exp \left( -\frac{\gamma_t \lambda_{\min}^{\Sigma}(\Sigma) M_t}{2K} \right).
\]

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 \( \mathbb{E} \left[ \sum_{t=1}^{2} \ell_t(X_t, A_t) - \ell_t(X_t, a^*) \right] \) 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 \( \alpha \)-local-self-bounding condition (LSB).
Algorithm 3: BoBW via local-self-bounding (LSB) algorithm, Adaption of Algorithm 1 in Dann et al. (2023)

Input : LSB algorithm \( L \)

\[
T_1 \leftarrow 0; \quad T_0 \leftarrow -c_2 \ln T; \\
\hat{A}_1 \sim \text{unif}(\{K\}); \quad t \leftarrow 1;
\]

for \( k = 1, 2, \ldots \) do

- Initialize \( L \) with candidate action \( \hat{A}_k \);
- Set the number of pulls \( N_k(a) \) for all \( a \in [K] \);

for \( t = T_k + 1, T_k + 2, \ldots \) do

- Observe \( X_t \);
- Choose action \( A_t \) according to \( L \), and advance \( L \) 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 \( \exists a \in [K] \setminus \{ \hat{A}_k \} \) such that \( N_k(a) \geq \frac{t - T_k}{2} \) then

\[
\hat{A}_{k+1} \leftarrow a; \\
T_{k+1} \leftarrow t;
\]

break

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

Input : candidate action \( \hat{a} \in [K], \frac{1}{2} \)-iw-stable algorithm \( B \) over \( [K] \setminus \{ \hat{a} \} \) 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 \( t = 1, 2, \ldots \) do

- Observe \( X_t \);
- Compute

\[
\bar{q}_t \leftarrow \arg \min_{q \in \Delta(2)} \left\{ q, \sum_{\tau=1}^{t-1} z_{\tau} - \left[ 0 \right. B_{t-1} + \psi_t(q) \right\}, q_t \leftarrow \left( 1 - \frac{1}{2t^2} \right) \bar{q}_t + \frac{1}{4t^2} \frac{1}{ \sqrt{q_{t-1}} }, \beta = \frac{1}{c_2};
\]

- Sample \( i_t \sim q_t \);
- if \( i_t = 1 \) then

\[
\text{Choose } A_t = \hat{a} \text{ and observe } \ell_t(X_t, A_t);
\]

else

\[
\text{Choose } A_t \text{ according to base algorithm } B \text{ and observe } \ell_t(X_t, A_t);
\]

Define \( z_{t,i} \leftarrow \frac{\ell_t(X_t, A_{t,i})}{q_{t,i}} - 1 \) and \( B_t \leftarrow \sqrt{c_1} \sum_{\tau=1}^{t} \frac{1}{q_{\tau,2}} + \min_{\tau \leq t} q_{\tau,2} ;
\]
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 $\hat{a} \in [K]$ as input and has the following pseudo-regret guarantee for any stopping time $\tau \in [1, T]$ and for any $a^* \in [K]$:

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

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 $O(c_0^{1-\alpha}T^\alpha + c_2 \ln^2(T))$ in the adversarial regime and by $O(c_1 \ln(T)\Delta_{\min}^{-\frac{\alpha}{2}} + (c_1 \ln T)^{1-\alpha}(C\Delta_{\min}^{-1})^{\alpha} + c_2 \ln(T) \ln(C\Delta_{\min}^{-1}))$ in the corrupted stochastic regime.

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 $\tilde{a}$ and an importance-weighting stable algorithm $\mathcal{B}$ over the action set $[K] \setminus \{\tilde{a}\}$.

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 $a^* \in [K]$:

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

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 $\overline{c}_0 = \overline{c}_1 = O(c_1)$ and $\overline{c}_2 = O(c_2)$.

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 $\tilde{a} \in \mathcal{A}$ as input and satisfies the following pseudo-regret guarantee for any stopping time $\tau \in [1, T]$ and action $a^* \in [K]$,

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

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 $O(\sqrt{c_1 E \sum_{t=1}^{T} \xi_{t,a}^2} \ln^2 T + c_2 \ln^2(T))$ in the adversarial regime and by $O\left( c_1 \ln(T) \Delta_{\min}^{-\frac{\alpha}{2}} + \sqrt{c_1 \ln T \Delta_{\min}^{-1}} + c_2 \ln(T) \ln(C\Delta_{\min}^{-1}) \right)$ 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 \( \hat{a} \) and the other is an algorithm with the data-dependent-importance-weighting-stable (dd-iw-stable) condition over the action set of \( A \setminus \{ \hat{a} \} \), 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 \( a^* \in [K] \):

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

where \( \text{upd}_t = 1 \) if feedback is observed in round \( t \) and \( \text{upd}_t = 0 \) otherwise.

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

**Algorithm 5:** dd-LSB via Corral, Adaption of Algorithm 6 in Dann et al. (2023)

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

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

for \( t = 1, 2, \ldots \) do

Observe \( X_t \);

Let \( B \) output an action \( \bar{A}_t \);

Receive predictors \( m_{t,a} \) for all \( a \in [K] \), and set \( y_{t,1} = (X_t, m_{\hat{a},a}) \) and \( y_{t,2} = (X_t, m_{\bar{a},a}) \);

Compute

\[
\bar{q}_t = \arg \min_{q \in \Delta_2} \left\{ \frac{1}{q_1} \left[ \left( \frac{1}{t} \sum_{\tau=1}^{t-1} z_\tau + y_t - 0 \right) + \frac{1}{q_t} \psi(q) \right] \right\}, \quad q_t = \left( 1 - \frac{1}{2t^2} \right) \bar{q}_t + \frac{1}{4t^2} 1,
\]

where \( \eta_t = \frac{1}{4}(\ln T)^\frac{1}{2} \left( \sum_{\tau=1}^{t-1} (1 - q_{t,1}^2) + (c_1 + c_2^2) \ln T \right)^{-\frac{1}{2}} \);

Sample \( i_t \sim q_t \);

if \( i_t = 1 \) then

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

else

Choose \( A_t = \bar{A}_t \) and observe \( \ell_t(X_t, A_t) \);

Define \( z_{t,i} = \left( \frac{\ell_t(X_t, A_t) - y_{t,i}}{q_t, i} \right) \mathbb{1}_{[i = 1]} + y_{t,i} \) and \( B_t = \sqrt{c_1 \sum_{\tau=1}^{t} \frac{\ell_{\tau, A_t}^2 \mathbb{1}_{[\tau = 2]}}{q_{\tau, 2}^2}} + \frac{c_2}{\min_{\tau \leq t} q_{\tau, 2}} \).

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 \( \frac{\ln T}{\Delta_{\text{min}}} \) 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 \emph{RealLinExp3} in Neu and Olkhovskaya (2020), but we assume that \( \Sigma^{-1} \) is known to the learner.

**Proposition 7** (iw-stable condition of ADAPTIVE-REALLINEXP3 as a base algorithm). Assume that \( \Sigma^{-1} \) 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 \( c_1 = \mathcal{O} \left( \ln(K)K^2 \left( d + \frac{1}{\lambda_{\text{min}}(\Sigma)} \right)^2 \right) \) and \( c_2 = \frac{K \ln K}{\lambda_{\text{min}}(\Sigma) \Delta_{\text{min}}} \).
Algorithm 6: RealLinExp3 with adaptive learning rate (Adaptive-RealLinExp3)

**Input**: Arms $[K]$

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

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

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

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

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

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

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 $\text{upd}_t = 1$, otherwise set $\text{upd}_t = 0$);

Compute $\hat{\theta}_{t,a} = \frac{\text{upd}_t}{q_t} \Sigma_{t,a}^{-1} X_t \ell_t(X_t, A_t) I [A_t = a]$ for all $a \in [K]$;

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 Adaptive-RealLinExp3). Assume that $\Sigma^{-1}$ is known to the learner. Combining Algorithms 3, 4 and 6 results in the following the regret bound: for the adversarial regime,

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

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

Proposition 8 implies that we obtain desired BoBW bounds if the learner access to $\Sigma_{t,a}^{-1} := \mathbb{E}_t [I [A_t = a] X_t X_t^\top]$ for computing the unbiased estimator $\hat{\theta}_{t,a}$ 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 $\Sigma_{t,a}^{-1}$ 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 $\pi^*(a|x) = 1$ if $a = \pi^*(x)$ otherwise 0 for $a \in [K]$, and $x \in X$. Let $X_0 \sim \mathcal{D}$ be a ghost sample chosen independently from the entire history. Then, we have

$$ \mathbb{E}_t [(X_t, \theta_t, \pi(X_t))] = \mathbb{E}_t [(X_0, \theta_t, \pi(X_0))]. $$

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

$$ \hat{R}_T(x) := \mathbb{E}_t \left[ \sum_{t=1}^{T} \langle x, \hat{\theta}_{t,A_t} \rangle - \langle x, \hat{\theta}_{t,\pi^*(x)} \rangle \right]. \tag{20} $$
Using this property and unbiased estimator $\hat{\theta}_{t,a}$, as also analyzed in Lemma 3 in Olkhovskaya et al. (2023), we have

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

$$= \mathbb{E} \left[ \sum_{t=1}^{\tau} \left( \ell_t(X_0, A_t) - \ell_t(X_0, \pi^*(X_0)) \right) \right]$$

$$= \mathbb{E} \left[ \sum_{t=1}^{\tau} \left( \langle X_0, \hat{\theta}_{t,A_t} \rangle - \langle X_0, \hat{\theta}_{t,\pi^*(X_0)} \rangle \right) \right].$$

(21)

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

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

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

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

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

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

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

we see that

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

$$\leq \sum_{t=1}^{\tau} \mathbb{E}_t \left[ (1 - \gamma_t) \sum_{a \in [K]} (p_t(a | x) - \pi^*(a | x)) \langle x, \hat{\theta}_{t,a} \rangle \right] + \sum_{t=1}^{\tau} \mathbb{E}_t \left[ \frac{\gamma_t}{K} \sum_{a \in [K]} \langle x, \hat{\theta}_{t,a} \rangle - \langle x, \hat{\theta}_{t,\pi^*(x)} \rangle \right].$$

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

$$p_t(\cdot | x) \in \arg \min_{p \in \Delta([K])} \left\{ \sum_{s=1}^{t-1} (p, \hat{\ell}_s(x)) + \psi_t(p) \right\},$$

(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

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

$$\leq \sum_{t=1}^{\tau} \mathbb{E}_t \left[ \eta_t \sum_{a=1}^{K} \pi_t(a | x) \langle x, \hat{\theta}_{t,a} \rangle^2 \right] + \mathbb{E} \left[ \frac{\ln K}{\eta_t} \right].$$
Since
\[
\sum_{t=1}^{\tau} \mathbb{E}_t \left[ \frac{\gamma_t}{K} \sum_{a \in [K]} \langle (x, \hat{\theta}_{t,a}) - \langle x, \hat{\theta}_{t,\pi^*(x)} \rangle \rangle \right] = \sum_{t=1}^{\tau} \mathbb{E}_t \left[ \frac{\gamma_t}{K} \sum_{a \in [K]} \langle (x, \theta_{t,a}) - \langle x, \theta_{t,\pi^*(x)} \rangle \rangle \right] \leq 2 \sum_{t=1}^{\tau} \mathbb{E}_t [\gamma_t]
\]
by \(|\langle x, \theta_{t,a} \rangle| \leq 1\), combining above equalities gives the desired result.

We next introduce the following lemma, which is implied by Lemma 8, for known \(\Sigma_{t,a}^{-1}\) and unbiased estimator \(\hat{\theta}_{t,a}\) in Algorithm 6 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
\[
E_t \left[ \sum_{a=1}^{K} \pi_t(a|X_0) \langle X_0, \hat{\theta}_{t,a} \rangle^2 \right] \leq \sum_{a=1}^{K} E_t [\text{tr}(\Sigma_{t,a}^{-1} \Sigma_{t,a}^{-1} \Sigma_{t,a}^{-1})] \leq 3Kd \ell_t/q_t.
\]

Then, we are ready to prove Proposition 7. We first show \(|\eta_t \langle x, \hat{\theta}_{t,a} \rangle| \leq 1\).

\[
|\eta_t \langle x, \hat{\theta}_{t,a} \rangle| = \eta_t |\langle x, \hat{\theta}_{t,a} \rangle| = \eta_t \left| x^\top \text{upd}_t \Sigma_{t,a}^{-1} X_t \ell_t(X_t, a) 1 [A_t = a] \right| \leq \frac{\eta_t}{q_t} \left| x^\top \Sigma_{t,a}^{-1} X_t \right|
\]
\[
\leq \frac{\eta_t}{q_t} \|\Sigma_{t,a}^{-1}\|_{\text{op}} \cdot \max_{x \in \mathbb{R}} \|x\|^2 \leq \frac{\eta_t}{q_t} \frac{1}{\lambda_{\min}(\Sigma_{t,a})} \leq \frac{\eta_t}{q_t} K \gamma_t \leq 1,
\]
where we used \(\ell_t(X_t, a) \leq 1\) in the first inequality, \(\lambda_{\min}(\Sigma_{t,a}) \geq \frac{\gamma_t \lambda_{\max}(\Sigma)}{K}\) in the forth inequality, and the definition of \(\gamma_t = \frac{\eta_t K}{q_t \lambda_{\min}(\Sigma)}\) in the last inequality.

Next, we will give the bound of \(\sum_{t=1}^{\tau} \frac{\eta_t}{q_t}\). Since we have
\[
\sum_{t=1}^{\tau} \frac{1}{q_t} \leq 2 \sum_{t=1}^{\tau} \frac{1}{\sqrt{\sum_{s=1}^{t} 1/q_s}} = 2 \sum_{t=1}^{\tau} \left( \sum_{s=1}^{t} \frac{1}{q_s} - \sum_{s=1}^{t-1} \frac{1}{q_s} \right) = 2 \sqrt{\sum_{s=1}^{t} 1/q_s}
\]
and using the definition of \(\eta_t\), we obtain
\[
\sum_{t=1}^{\tau} \frac{\eta_t}{q_t} \leq \sqrt{\ln K} \sum_{t=1}^{\tau} \frac{1}{q_t} \sqrt{\sum_{s=1}^{t} \frac{1}{q_s}} \leq \sqrt{4 \ln K} \sum_{t=1}^{\tau} \frac{1}{q_t}.
\]

Furthermore, by the definition of \(\eta_t\), it is easy to see that
\[
\frac{1}{\eta_t} \leq \sqrt{\frac{\sum_{t=1}^{\tau} 1/q_t}{\ln K}} + \frac{c}{\min_{t \leq \tau} q_t}.
\]

Therefore, by combining the above inequalities, we have for any \(a^* \in [K]\) and \(\tau \in [T]\),
\[
R(\tau, a^*) = E \left[ \sum_{t=1}^{\tau} (\ell_t(X_t, A_t) - \ell_t(X_t, a^*)) \right] = E \left[ \sum_{t=1}^{\tau} (\ell_t(X_0, A_t) - \ell_t(X_0, a^*)) \right] \\
\leq E \left[ \sum_{t=1}^{\tau} (\ell_t(X_0, A_t) - \ell_t(X_0, \pi^*(X_0))) \right] \\
= E \left[ \sum_{t=1}^{\tau} (X_0, \hat{\theta}_{t,A_t}) - \langle X_0, \hat{\theta}_{t,\pi^*(X_0)} \rangle \right] \\
= E \left[ \hat{R}_T(X_0) \right]
\]
We prove Corollary F.2 Proof of Corollary X-linear optimization oracle over \( M \) can be done in polynomial time, making \( O(d^2 \ln d) \)-call for linear optimization oracle over \( \mathcal{X} \).

where the first and second equalities follow from the property of \( X_0 \) and the fact that \( \hat{\theta}_{t,a} \) 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 \( \gamma_t \) and Lemma 11, the fourth inequality follows from (25) and the definition of \( \eta_t \). Lastly, we have the statement plugging in the definition of \( c = \frac{K}{\lambda_{\min}(\Sigma)} \).

\[ \square \]

F APPENDIX FOR DATA-DEPENDENT BOUNDS

In this section, we describe how to find a positive semidefinite matrix \( S \in \mathbb{R}^{d \times d} \) to compute a loss predictor \( m_{t,a} \) 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) for linear bandits, if we have the prior knowledge of the support of \( \mathcal{D} \), i.e., context space \( \mathcal{X} \), we can find an appropriate matrix \( S \) such that \( \|m^*\|_S^2 = O(d) \) for any vector \( m^* \in \mathcal{M} \), and \( \max_{x \in \mathcal{X}} \|x\|_{S^{-1}}^2 = O(d) \) in our case. \( \mathcal{X}_{\text{span}} = \{x_1, \ldots, x_d\} \subseteq \mathcal{X} \) is said to be 2-barycentric spanner for \( \mathcal{X} \) if each \( x \in \mathcal{X} \) can be expressed as linear combination of elements in \( \mathcal{X}_{\text{span}} \) with coefficients in \([-2, 2]\). Define \( S \in \mathbb{R}^{d \times d} \) as

\[ M = (x_1 x_2 \cdots x_d), \quad S = MM^\top = \sum_{i=1}^d x_i x_i^\top. \] (26)

Then, for \( m \in \mathcal{M} \), we can easily confirm \( \|m\|_S^2 = m^\top \left( \sum_{i=1}^d x_i x_i^\top \right) m \leq d \) and \( \|x\|_{S^{-1}}^2 = x^\top (M^{-1})^\top M^{-1} x = u^\top u \leq 4d \) using some \( u \in [-2, 2]^d \) such that \( x = Mu \). 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 \( m_{t,a} \) by (9).

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

\[
m_{t,a} \in \arg\min_{m \in \mathcal{M}} \left\{ \|m\|_S^2 + \sum_{j=1}^{t-1} 1 \left[ A_j = a \right] \left( \langle \theta_{j,a} - m, X_j \rangle \right)^2 \right\}.
\]
Then, for any $m^* \in \mathcal{M}$, it holds that

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

where $\xi_{t,A_t} = (\langle \theta_{t,A_t} - m^*, X_t \rangle)$.

**Proof of Lemma 12.** The proof can be shown in a manner 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 $m_{t,a}$.

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

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

From this, we have that

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

Therefore, we obtain

$$
E \left[ \sum_{t=1}^{T} (\langle \theta_{t,A_t} - m_{t,A_t}, X_t \rangle)^2 \right] \leq E \sum_{t=1}^{T} E_{A_t \sim Q_t} [ (\langle \theta_{t,A_t} - m_{t,A_t}, X_t \rangle)^2 ]
$$

$$
= E \sum_{t=1}^{T} \sum_{a \in [K]} \mathbb{1}_{\{A_t = a\}} (\langle \theta_{t,a} - m_{t,a}, X_t \rangle)^2
$$

$$
\leq E \sum_{t=1}^{T} \sum_{a \in [K]} \mathbb{1}_{\{A_t = a\}} (\langle \theta_{t,a} - m^*, X_t \rangle)^2 + K \|m^*\|_S^2 + 8T \sum_{t=1}^{T} \|X_t\|_{G^{-1}}^2
$$

$$
\leq E \sum_{t=1}^{T} (\langle \theta_{t,A_t} - m^*, X_t \rangle)^2 + K \|m^*\|_S^2 + 8K E \sum_{t=1}^{T} \|X_t\|_{G^{-1}}^2.
$$

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:

$$
f_0(m) = \frac{1}{2} \|m\|_S^2,
$$

$$
f_t(m) = \frac{1}{2} \mathbb{1}_{\{A_t = a\}} (\langle \theta_{t,a} - m, X_t \rangle)^2 \quad (t \in [T]),
$$

$$
F_t(m) = \sum_{j=0}^{t} f_j(m) \quad (t \in \{0, 1, \ldots, T\}).
$$

Then, the definition of $m_{t,a}$ in (9) can be rewritten as:

$$
m_{t,a} \in \arg\min_{m \in \mathcal{M}} F_{t-1}(m).
$$

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

$$
F_T(m^*) \geq F_T(m_{T+1,a}) = F_{T-1}(m_{T+1,a}) + f_T(m_{T+1,a}) \geq F_{T-1}(m_{t,a}) + f_T(m_{T+1,a})
$$

$$
= f_{T-2}(m_{t,a}) + f_{T-1}(m_{t,a}) + f_T(m_{T+1,a}) \geq \cdots \geq f_0(m_{1,a}) + \sum_{t=1}^{T} f_t(m_{T+1,a})
$$
We next show

\[ f_t(m_t,a) - f_t(m_{T+1,a}) \leq 4 \|X_t\|_{G_t^{-1}}^2, \]

where we define positive semi-definite matrices \( G_t \in \mathbb{R}^{d \times d} \) for \( t = 0, 1, \ldots, T \) by

\[ G_t = S + \sum_{j=1}^{t} X_j X_j^\top. \]

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

\[ F_t(m') \geq F_t(m) + \langle \nabla F_t(m), m' - m \rangle + \|m' - m\|_{G_t}^2. \tag{30} \]

Further, (29) implies that

\[ \langle \nabla F_{t-1}(m_{t,a}), m - m_{t,a} \rangle \geq 0 \tag{31} \]

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

\[
\begin{align*}
    f_t(m_{t,a}) - f_t(m_{T+1,a}) \\
    = F_t(m_{t,a}) - F_t(m_{T+1,a}) + F_t(m_{T+1,a}) - F_t(m_{t,a}) \\
    \leq \langle \nabla F_t(m_{t,a}), m_{t,a} - m_{T+1,a} \rangle + \langle \nabla F_t(m_{T+1,a}), m_{T+1,a} - m_{t,a} \rangle \\
    \leq \langle \nabla F_t(m_{t,a}), m_{t,a} - m_{T+1,a} \rangle + \langle \nabla F_t(m_{T+1,a}), m_{T+1,a} - m_{t,a} \rangle \\
    = \langle \nabla f_t(m_{t,a}), m_{t,a} - m_{T+1,a} \rangle + \langle \nabla f_t(m_{T+1,a}), m_{T+1,a} - m_{t,a} \rangle \\
    \leq \|\nabla f_t(m_{t,a})\|_{G_t^{-1}} \|m_{t,a} - m_{T+1,a}\|_{G_t} + \|\nabla f_t(m_{T+1,a})\|_{G_t^{-1}} \|m_{T+1,a} - m_{t,a}\|_{G_t} \\
    \leq \frac{1}{4} \|\nabla f_t(m_{t,a})\|_{G_t^{-1}}^2 + \frac{1}{4} \|\nabla f_t(m_{T+1,a})\|_{G_t^{-1}}^2 \\
    \leq 4 \|X_t\|_{G_t^{-1}}^2.
\end{align*}
\]

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

\[ \sum_{t=1}^{T} 1[A_t = a] \langle \theta_{t,a} - m_{t,a}, X_t \rangle^2 - \sum_{t=1}^{T} 1[A_t = a] \langle \theta_{t,a} - m^*, X_t \rangle^2 \leq \|m^*\|_S^2 + 8 \sum_{t=1}^{T} \|X_t\|_{G_t^{-1}}^2, \]
which is (27). We next show

\[ \sum_{t=1}^{T} \|X_t\|^2_{G_t^{-1}} \leq d \ln \left( 1 + \frac{T}{d} \max_{x \in \mathcal{X}} \|x\|^2_{S^{-1}} \right). \] (32)

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

\[
\ln \det G_t - \ln \det G_{t-1} = - (\ln \det (G_t - X_tX_t^\top) - \ln \det G_t) \\
= - \ln \det \left( G_t^{-\frac{1}{2}} (G_t - X_tX_t^\top) G_t^{-\frac{1}{2}} \right) \\
= - \ln \left( 1 - \left\| G_t^{-\frac{1}{2}} X_t \right\|_2^2 \right) \geq \left\| G_t^{-\frac{1}{2}} X_t \right\|_2^2 = \|X_t\|^2_{G_t^{-1}},
\]

where the forth equality holds since the matrix \( I - G_t^{-\frac{1}{2}} X_tX_t^\top G_t^{-\frac{1}{2}} \) has eigenvalues \( \lambda'_1 = 1 - \left\| G_t^{-\frac{1}{2}} X_t \right\|_2^2 \) 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

\[ X_t^\top G_t^{-1} X_t \leq \ln \frac{\det G_t}{\det G_{t-1}}. \]

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

\[
\sum_{t=1}^{T} \|X_t\|^2_{G_t^{-1}} \leq \ln \det G_t - \ln \det G_0 = \ln \det \left( I + \sum_{t=1}^{T} S^{-\frac{1}{2}} X_tX_t^\top S^{-\frac{1}{2}} \right) = \sum_{i=1}^{d} \ln (1 + \lambda_i).
\]

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

\[
E \left[ \sum_{t=1}^{T} \langle (\theta_{t,A_t} - m_{t,A_t}, X_t)^2 \right] \\
\leq E \left[ \sum_{t=1}^{T} \langle (\theta_{t,A_t} - m^*, X_t)^2 \right] + K \|m^*\|^2_S + 8K E \left[ \sum_{t=1}^{T} \|X_t\|^2_{G_t^{-1}} \right] \\
\leq E \left[ \sum_{t=1}^{T} \langle (\theta_{t,A_t} - m^*, X_t)^2 \right] + K \|m^*\|^2_S + 8K d \ln \left( 1 + \frac{T}{d} \max_{x \in \mathcal{X}} \|x\|^2_{S^{-1}} \right),
\]

which concludes the proof.

We are ready to prove Corollary 1.

\textbf{Proof of Corollary 1.} Since we choose \( S \) by (26), it holds that \( \|m^*\|^2_S = O(d) \) and \( \max_{x \in \mathcal{X}} \|x\|^2_{S^{-1}} = O(d) \).
Then, by Lemma 12 and Theorem 1, it holds that for any $m^* \in \mathcal{M}$,

$$R_T = \mathcal{O} \left( \sqrt{\kappa_1(d, K, T) E \left[ \sum_{t=1}^{T} \xi_{t,A_t}^2 \right]} \ln^2 T + \kappa_2(d, K, T) \ln^2(T) \right)$$

$$= \mathcal{O} \left( \sqrt{\kappa_1(d, K, T) \left( E \left[ \sum_{t=1}^{T} (\langle \theta_{t,A_t} - m^*, X_t \rangle)^2 \right] + K \|m^*\|_S^2 + K d \ln \left( 1 + \frac{T}{d \max_{x \in A} \|x\|_{S-1}} \right) \right]) \ln^2 T \right.$$  

$$+ \kappa_2(d, K, T) \ln^2(T))$$

$$= \mathcal{O} \left( \sqrt{\kappa_1(d, K, T) \left( E \left[ \sum_{t=1}^{T} (\langle \theta_{t,A_t} - m^*, X_t \rangle)^2 \right] + K d \ln (1 + T) \right) \ln^2 T + \kappa_2(d, K, T) \ln^2(T)} \right)$$

$$= \mathcal{O} \left( K d \ln(dKT) \ln^2(T) \sqrt{E \left[ \sum_{t=1}^{T} (\langle \theta_{t,A_t} - m^*, X_t \rangle)^2 \right] + (dK)^{3/2} \ln(dKT) \ln^3(T)} \right)$$

$$= \tilde{\mathcal{O}} \left( K d \sqrt{E \left[ \sum_{t=1}^{T} (\langle \theta_{t,A_t} - m^*, X_t \rangle)^2 \right] + (dK)^{3/2}} \right),$$

in the adversarial regime. Therefore, we obtain

$$R_T = \tilde{\mathcal{O}} \left( K d \sqrt{E \left[ \sum_{t=1}^{T} (\langle \theta_{t,A_t} - m^*, X_t \rangle)^2 \right] + (dK)^{3/2}} \right) = \tilde{\mathcal{O}} \left( K d \sqrt{\mathbb{A}} + (dK)^{3/2} \right). \quad (33)$$

On the other hand, for $m^* = 0$, we also have that

$$\frac{R_T}{\mathcal{C}} \leq \sqrt{\kappa_1(d, K, T) \left( E \left[ \sum_{t=1}^{T} (X_t, \theta_{t,A_t})^2 \right] \ln^2 T + \kappa_2(d, K, T) \ln^2(T) \right)}$$

$$\leq \sqrt{E \left[ \sum_{t=1}^{T} (X_t, \theta_{t,A_t}) \right] + K d \ln (1 + T) \sqrt{\kappa_1(d, K, T) \ln^2 T + \kappa_2(d, K, T)}}$$

$$\leq \sqrt{\kappa_1(d, K, T) \ln^2 T} \sqrt{E \left[ \sum_{t=1}^{T} (X_t, \theta_{t,A_t}) \right] + \sqrt{\kappa_1(d, K, T) \ln^2 T \sqrt{K d \ln (1 + T) + \kappa_2(d, K, T) \ln^2(T)}}},$$

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

$$\sqrt{E \left[ \sum_{t=1}^{T} (X_t, \theta_{t,A_t}) \right]} \leq \sqrt{\kappa_1(d, K, T) \ln^2 T} + \sqrt{\kappa_1(d, K, T) \ln^2 T + 4 \left( L^* + \sqrt{\kappa_1(d, K, T) \ln^2 T \sqrt{K d \ln (1 + T) + \kappa_2(d, K, T) \ln^2(T)}} \right)}$$

$$\leq \frac{\sqrt{\kappa_1(d, K, T) \ln^2 T + \kappa_1(d, K, T) \ln^2 T + 4 \left( L^* + \sqrt{\kappa_1(d, K, T) \ln^2 T \sqrt{K d \ln (1 + T) + \kappa_2(d, K, T) \ln^2(T)}} \right)}}{2}.$$
This indicates that
\[
\sqrt{\mathbb{E} \left[ \sum_{t=1}^{T} (X_t, \theta_{t,A_t}) \right]} = \mathcal{O} \left( \sqrt{L^* + K^2d^2 \ln^2(dKT) \ln^4(T)} \right)
\]
\[
= \tilde{\mathcal{O}} \left( \sqrt{L^* + K^2d^2} \right).
\]

Therefore, in this case, we obtain
\[
R_T = \mathcal{O} \left( \sqrt{\kappa_1(d,K,T) \ln^2 T \cdot L^* + K^2d^2 \ln^2(dKT) \ln^4(T)} \right)
\]
\[
+ \sqrt{\kappa_1(d,K,T) \ln^2 T \sqrt{Kd \ln (1 + T) + \kappa_2(d,K,T) \ln^2(T)}}
\]
\[
= \mathcal{O} \left( Kd \ln(dKT) \ln^2(T) \sqrt{L^*} + K^2d^2 \ln^2(dKT) \ln^4(T) \right)
\]
\[
= \tilde{\mathcal{O}} \left( Kd \sqrt{L^*} + K^2d^2 \right).
\]

From (33) and this, we conclude that
\[
R_T = \tilde{\mathcal{O}} \left( Kd \sqrt{\min\{L^*,1\} + K^2d^2} \right).
\]

We note that instead computing \( m_{t,a} \) 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 \( m_{t,a} = 0 \in \mathbb{R}^d \).

**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, 2, and 5 results in the following 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 \( m_{t,a} = 0 \) in Theorem 1, for a universal constant \( \tilde{c} > 0 \), we have
\[
\frac{R_T}{\tilde{c}} \leq \sqrt{\kappa_1(d,K,T) \ln^2 T \cdot \mathbb{E} \left[ \sum_{t=1}^{T} (X_t, \theta_{t,A_t}) \right]^2} + \kappa_2(d,K,T) \ln^2(T)
\]
\[
\leq \sqrt{\kappa_1(d,K,T) \ln^2 T \cdot \mathbb{E} \left[ \sum_{t=1}^{T} X_t, \theta_{t,A_t} \right]} + \kappa_2(d,K,T) \ln^2(T),
\]
where the second inequality follows from \( 0 \leq \ell_t(X_t, A_t) \leq 1 \). By the definition of \( R_T = \mathbb{E} \left[ \sum_{t=1}^{T} (X_t, \theta_{t,A_t}) \right] - L^* \), and solving the quadratic inequality for \( \mathbb{E} \left[ \sum_{t=1}^{T} (X_t, \theta_{t,A_t}) \right] \), we obtain \( \mathbb{E} \left[ \sum_{t=1}^{T} (X_t, \theta_{t,A_t}) \right] = \mathcal{O}(L^* + \ln T \sqrt{\kappa_1(d,K,T)}) \). 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.

\section*{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
\[
\Pr[\|y\|^2_2 \geq da^2] \leq d \exp(1 - \alpha)
\]
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 \( x \in \mathcal{X} \), we denote the \( dK \)-dimensional vector \( z(r, x) := (r_1 \cdot x^\top, \ldots, r_K \cdot x^\top)^\top \in \mathbb{R}^{dK} \). We define the \( dK \times dK \) matrix \( \Sigma_0(t) := \text{diag}_{a \in [K]}(\Sigma_{t,a}) \in \mathbb{R}^{dK \times dK} \) as a block diagonal arrangement of the covariance matrices per arm, where \( \Sigma_{t,a} \) is given in (5). Similarly, we also define \( \tilde{\Sigma}_0(t) := \text{diag}_{a \in [K]}(\tilde{\Sigma}_{t,a}) \in \mathbb{R}^{dK \times dK} \), where \( \tilde{\Sigma}_{t,a} \) is given in (7). Using these notation, we can rewrite the \( \bar{p}_t(r|x) \) for \( r \in \Delta([K]) \) and a context \( x \in \mathcal{X} \) as follows:

\[
\bar{p}_t(r|x) = \frac{p_t(r|x) \mathbb{I}\left[\sum_{a=1}^{K} r_a^2 \|x\|^2_{\Sigma^{-1}_{t,a}} \leq dK \gamma_t^2\right]}{\mathbb{P}_{y \sim p_t(|x)} \left[\sum_{a=1}^{K} y_a^2 \|x\|^2_{\Sigma^{-1}_{t,a}} \leq dK \gamma_t^2\right]} = \frac{p_t(r|x) \mathbb{I}\left[\|z(r, x)\|^2_{\Sigma^{-1}_0(t)} \leq dK \gamma_t^2\right]}{\mathbb{P}_{y \sim p_t(|x)} \left[\|z(y, x)\|^2_{\tilde{\Sigma}^{-1}_0(t)} \leq dK \gamma_t^2\right]}.
\]

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

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

\[
Z(x) := z(Q(x), x) = (Q_{t,1}(x) \cdot x^\top, \ldots, Q_{t,K}(x) \cdot x^\top)^\top \in \mathbb{R}^{dK}.
\]

And, we define \( \tilde{Z}_t(x) \) for context \( x \) and \( Q_t(x) \sim \bar{p}_t(|x) \) as:

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

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

\[
Z^*(x) := (0^\top, \ldots, x^\top)^\top \in \mathbb{R}^{dK},
\]

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

\[
\hat{Z}(x) := \left(\frac{1}{K} x^\top, \ldots, \frac{1}{K} x^\top\right)^\top \in \mathbb{R}^{dK}.
\]

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

\[
R_\tau = \mathbb{E} \left[ \sum_{t=1}^{\tau} (\ell_t(X_t, A_t) - \ell_t(X_t, \pi^*(X_t))) \right]
= \mathbb{E} \left[ \sum_{t=1}^{\tau} \left\langle \tilde{Z}_t(x_t) - Z^*(X_t), \theta_t \right\rangle \right]
= \mathbb{E} \left[ \sum_{t=1}^{\tau} \left\langle \tilde{Z}_t(X_t) - Z(X_t), \theta_t \right\rangle \right] + \mathbb{E} \left[ \sum_{t=1}^{\tau} \left\langle Z(X_t) - Z^*(X_t), \theta_t \right\rangle \right].
\]

(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 $\hat{\theta}_t$, and a fixed context $x \in X$, we define

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

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

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

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

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

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

To see that the first term of $E \left[ \sum_{t=1}^{\tau} \left\langle \tilde{Z}_t(X_t) - Z(X_t), \theta_t \right\rangle \right]$ 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 $\tilde{Z}_t(X_t)$ is close to $Z(X_t)$, and also provides a useful relation between covariance matrices $\Sigma_b(t)$ and $\Sigma_b(t)$. 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 $\tilde{\gamma}_t \geq 4 \ln(10dK\tau)$ and $\langle r_1 \cdot x^\top, \ldots, r_K \cdot x^\top \rangle, \theta_t \rangle \in [-1,1]$ for any $t$, a policy $r \in \Delta([K])$ and context $X_t \in X$. Then, we have

$$\left\| E_t \left[ \left\langle \tilde{Z}_t(X_t) - Z(X_t), \theta_t \right\rangle \right] \right\| \leq \frac{1}{2\gamma^2}.$$ 

Further, we have

$$\frac{3}{4} \Sigma_b(t) \preceq \Sigma_b(t) \preceq \frac{4}{3} \Sigma_b(t).$$

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 $\mu_0$ be a uniform distribution over $[K]$ and $\phi(y) = \exp(y) - y - 1$. Then, the regret $\hat{R}_\tau(x)$ for fixed $x \in X$ of Algorithm 1 almost surely satisfies

$$\hat{R}_\tau(x) \leq \frac{1}{\tau} \sum_{t=1}^{\tau} \left\langle Z(x) - Z^*(x), \hat{\theta}_t \right\rangle + \frac{K \ln \tau}{\eta_t} + \sum_{t=1}^{\tau} \frac{1}{\eta_t} E_t \left[ \phi(-\eta_t (Z(X_t), \hat{\theta}_t - m_t)) \right].$$

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 $E[y^2] \leq \frac{1}{100}$, we have

$$E[\phi(y)] \leq E[y^2] + 30 \exp \left( -\frac{1}{\sqrt{E[y^2]}^2} \right) \leq 2E[y^2]$$

where $\phi(x) = \exp(y) - y - 1$. 

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

$$
\hat{\theta}_{t,a} := m_{t,a} + \frac{\text{upd}_t}{q_t} Q_{t,a}(X_t) \Sigma_{t,a}^{-1} X_t \xi_{t,a} 1[a_t = a],
$$

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

$$
\mathbb{E}_{Q(X_0) \sim p_1(X_0), \text{upd} \sim q_t} \left[ (-\eta_t(Z(X_0), \hat{\theta}_t - m_t))^2 \mid \mathcal{F}_{t-1} \right]
$$

$$
= \mathbb{E}_{\text{upd} \sim q_t} \left[ \frac{\eta_t^2}{q_t} \mathbb{E}_t \left[ \xi_{t,A_t}^2 Z(X_t)^\top \Sigma_b^{-1}(t) Z(X_0) Z(X_0)^\top \Sigma_b^{-1}(t) Z(X_t) \right] \right]
$$

$$
= \mathbb{E}_{\text{upd} \sim q_t} \left[ \frac{\eta_t^2}{q_t} \mathbb{E}_t \left[ \xi_{t,A_t}^2 Z(X_t)^\top \Sigma_b^{-1}(t) \Sigma_b(t) \Sigma_b^{-1}(t) Z(X_t) \right] \right]
$$

$$
\leq \frac{4}{3} \eta_t^2 \mathbb{E}_{\text{upd} \sim q_t} \left[ \frac{\eta_t^2}{q_t} \mathbb{E}_t \left[ \xi_{t,A_t}^2 Z(X_t)^\top \Sigma_b^{-1}(t) Z(X_t) \right] \right]
$$

$$
\leq \frac{4}{3} \eta_t^2 \mathbb{E}_{\text{upd} \sim q_t} \left[ \frac{\eta_t^2}{q_t} \mathbb{E}_t \left[ \xi_{t,A_t}^2 Z(X_t) \right] \right]
$$

$$
\leq \frac{4}{3} \eta_t^2 \mathbb{E}_{\text{upd} \sim q_t} \left[ \frac{\eta_t^2}{q_t} \mathbb{E}_t \left[ \xi_{t,A_t}^2 \right] \right]
$$

$$
\leq \frac{1}{100},
$$

where the first and second inequalities follow from Lemma 15, the third inequality follows from Algorithm 1 in Algorithm 1 of $\|Z(X_t)\|^2 \Sigma_b^{-1}(t) \leq dK^2$, and we used $\eta_t \leq \frac{2}{\sqrt{\log K}}$ and the assumptions that $|\ell_t(X_t, A_t)| \leq 1$ and $|\langle X_t, m_{t,A_t} \rangle| \leq 1$ in the last inequality. Then using Lemma 17 for $y = -\eta_t(Z(X_0), \hat{\theta}_t - m_t)$ and (39), we obtain

$$
\frac{1}{\eta_t} \mathbb{E} \left[ \psi(-\eta_t(Z(X_0), \hat{\theta}_t - m_t)) \right] \leq \frac{2}{\eta_t} \mathbb{E} \left[ (-\eta_t(Z(X_0), \hat{\theta}_t - m_t))^2 \right]
$$

$$
\leq \frac{4dK^2 \eta_t^2}{q_t} \mathbb{E}_t \left[ \xi_{t,A_t}^2 \right].
$$

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

$$
\mathbb{E} \left[ \frac{1}{t} \sum_{t=1}^T \langle Z(X_0) - Z^*(X_0), \hat{\theta}_t \rangle \right] = \frac{1}{t} \sum_{t=1}^T \langle Z(X_0) - Z^*(X_0), \theta_t \rangle \leq 2.
$$

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

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

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

$$
= \mathbb{E} \left[ \sum_{t=1}^\tau (Z_t(X_t) - Z(X_t), \theta_t) \right] + \mathbb{E} \left[ \sum_{t=1}^\tau (Z_t(X_t) - Z^*(X_t), \theta_t) \right]
$$
We also define \( \eta_t \). Therefore, using the definition of \( \tilde{\gamma}_t \) in the first inequality and we use (41) in the second inequality and we use (40) in the last inequality.

Recall that \( \beta_t := 16 \tilde{\gamma}_t^2 \xi_t^2 \), and \( \tilde{\gamma}_t = 4 \ln(10 d K t) \). Also, recall that the learning rate \( \eta_t \) is defined as follows:

\[
\eta_t = \frac{1}{\sqrt{\frac{800 d K \tilde{\gamma}_t^2}{\min_j \leq q_j} + \sum_{j=1}^{t-1} \frac{\beta_j}{q_j}}}.
\]

We also define \( \eta'_t \) as follows:

\[
\eta'_t := \frac{1}{\sqrt{\frac{800 d K \tilde{\gamma}_t^2}{\min_j \leq q_j} + \sum_{j=1}^{t-1} \frac{\beta_j}{q_j}}}
\]

Using \( \frac{x}{2 \sqrt{y}} \leq \sqrt{y} - \sqrt{y - x} \) for \( x = \frac{\beta_t}{q_t}, y = \frac{800 d K \tilde{\gamma}_t^2}{\min_j \leq q_j} + \sum_{j=1}^{t-1} \frac{\beta_j}{q_j} \), we have that

\[
\frac{\beta_t}{q_t} \leq 2 \sqrt{\frac{800 d K \tilde{\gamma}_t^2}{\min_j \leq q_j} + \sum_{j=1}^{t-1} \frac{\beta_j}{q_j}} - \sqrt{\frac{800 d K \tilde{\gamma}_t^2}{\min_j \leq q_j} + \sum_{j=1}^{t-1} \frac{\beta_j}{q_j}}
\]

\[
\leq 2 \sqrt{\frac{800 d K \tilde{\gamma}_t^2}{\min_j \leq q_j} + \sum_{j=1}^{t} \frac{\beta_j}{q_j}} - \sqrt{\frac{800 d K \tilde{\gamma}_t^2}{\min_j \leq q_j} + \sum_{j=1}^{t-1} \frac{\beta_j}{q_j}}
\]

\[
= \frac{1}{\eta'_{t+1}} - \frac{1}{\eta'_t}.
\]

where we used \( \frac{\tilde{\gamma}_t}{\min_j \leq q_j} \geq \frac{\tilde{\gamma}_{t-1}}{\min_j \leq q_j} \) in the second inequality. Summing up over \( t = 1, \ldots, \tau \) gives

\[
\sum_{t=1}^{\tau} \left( \frac{1}{\eta'_{t+1}} - \frac{1}{\eta'_t} \right) = \frac{1}{\eta'_{\tau+1}} - \frac{1}{\eta'_1} \leq \frac{1}{\eta'_{\tau+1}} = \sqrt{\frac{800 d K \tilde{\gamma}_t^2}{\min_j \leq q_j} + \sum_{j=1}^{\tau} \frac{\beta_j}{q_j}}.
\]

Therefore, using the definition of \( \eta_t \) and \( \beta_t \), we have that

\[
\mathbb{E} \left[ \sum_{t=1}^{\tau} \frac{4 d K \eta_t \tilde{\gamma}_t^2 \xi_t^2}{q_t} \xi_{t,A_t} \right] \leq \mathbb{E} \left[ \sum_{t=1}^{\tau} \frac{d K \beta_t}{4 q_t} \eta_t \right] = \mathbb{E} \left[ \sum_{t=1}^{\tau} \frac{d K \beta_t}{4 q_t} \sqrt{\frac{800 d K \tilde{\gamma}_t^2}{\min_j \leq q_j} + \sum_{j=1}^{t-1} \frac{\beta_j}{q_j}} \right]
\]

\[
\leq d K \mathbb{E} \left[ \sum_{t=1}^{\tau} \left( \frac{1}{\eta'_{t+1}} - \frac{1}{\eta'_t} \right) \right] \leq d K \mathbb{E} \left[ \sqrt{\frac{800 d K \tilde{\gamma}_t^2}{\min_j \leq q_j} + \sum_{j=1}^{\tau} \frac{\beta_j}{q_j}} \right]
\]
\begin{align*}
&= dK \mathbb{E} \left[ \frac{800dK \gamma^2_	au}{\min_{j \leq \tau} q_j} + \sum_{t=1}^{\tau} \frac{16 \gamma^2_	au \xi_{t, A_t}^2}{q_t} \right] \\
&\leq 4dK \gamma \mathbb{E} \left[ \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) \mathbb{E} \left[ \frac{50dK}{\min_{j \leq \tau} q_j} + \mathbb{E} \left[ \sum_{t=1}^{\tau} \frac{\text{upd}_t \xi_{t, A_t}^2}{q_t^2} \right] \right],
\end{align*}

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

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

\begin{align*}
\mathbb{E} \left[ K \ln \frac{\tau}{\eta} \right] &\leq K \ln \tau \mathbb{E} \left[ \frac{800dK \gamma^2_	au}{\min_{j \leq \tau} q_j} + \sum_{t=1}^{\tau} \frac{16 \gamma^2_	au \xi_{t, A_t}^2}{q_t} \right] \\
&\leq 16K \ln(\tau) \ln(10dK \tau) \cdot \mathbb{E} \left[ \frac{50dK}{\min_{j \leq \tau} q_j} + \mathbb{E} \left[ \sum_{t=1}^{\tau} \frac{\text{upd}_t \xi_{t, A_t}^2}{q_t^2} \right] \right] + 3 \\
&\leq 32Kd \ln(10dK \tau) \ln(\tau) \left( \mathbb{E} \left[ \sum_{t=1}^{\tau} \frac{\text{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).\end{align*}

\hfill \Box

**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 $D$, policy $\pi_t$, action $a \in [K]$, number of iterations $M_t$, and constant $\rho$, and outputs $\hat{\Sigma}_{t, a}^+ = \rho I + \rho \sum_{k=1}^{M_t} A_{k, a}$ as the estimate of the inverse of the covariance matrix $\Sigma_{t, a}^-$. 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 $\beta_t'$ 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).
Lemma 18. Let $\beta_t'$ be updated by (13) for each round $t$. Then for a ghsat sample $X_0$, we have
\[
\mathbb{E} \left[ \sum_{t=1}^{T} (\beta_{t+1}^t - \beta_t^t) H(p_{t+1}(.|X_0)) \right] = O \left( c' \sqrt{\ln K} \sum_{t=1}^{T} \mathbb{E}[H(p_t(.|X_0))] \right).
\]

Proof of Lemma 18. From our definition of $\beta_t'$, we have
\[
\mathbb{E} \left[ \sum_{t=1}^{T} (\beta_{t+1}^t - \beta_t^t) H(p_{t+1}(.|X_0)) \right] = \mathbb{E} \left[ \sum_{t=1}^{T} \frac{c'_t}{\sqrt{1 + (\ln K)^{-1} \sum_{s=1}^{t} H(p_s(.|X_s))}} H(p_{t+1}(.|X_0)) \right]
\]
\[
= 2c'_t \sqrt{\ln K} \mathbb{E} \left[ \sum_{t=1}^{T} \frac{H(p_{t+1}(.|X_0))}{\sqrt{4 \ln K + 4 \sum_{s=1}^{t} H(p_s(.|X_s))}} \right]
\]
\[
= 2c'_t \sqrt{\ln K} \mathbb{E} \left[ \sum_{t=1}^{T} \frac{H(p_{t+1}(.|X_0))}{\sqrt{\ln K + \sum_{s=1}^{t} H(p_s(.|X_s)) + \sqrt{\ln K + \sum_{s=1}^{t} H(p_s(.|X_s))}} \right]
\]
\[
\leq 2c'_t \sqrt{\ln K} \mathbb{E} \left[ \sum_{t=1}^{T} \frac{H(p_{t+1}(.|X_0))}{\sqrt{\sum_{s=1}^{t+1} H(p_s(.|X_s)) + \sqrt{\sum_{s=1}^{t} H(p_s(.|X_s))}}} \right],
\]
where in the last step we used the fact that $H(p_s(.|X_s)) \leq H(p_1(.|X_1)) = \ln K$. Using the property that $\mathbb{E}_{X_{t+1} \sim D}[H(p_{t+1}(.|X_{t+1})),|F_t]] = \mathbb{E}_{X_0 \sim D}[H(p_{t+1}(.|X_0)),|F_t]]$, we have
\[
2c'_t \sqrt{\ln K} \mathbb{E} \left[ \sum_{t=1}^{T} \frac{H(p_{t+1}(.|X_0))}{\sqrt{\sum_{s=1}^{t+1} H(p_s(.|X_s)) + \sqrt{\sum_{s=1}^{t} H(p_s(.|X_s))}}} \right]
\]
\[
= 2c'_t \sqrt{\ln K} \mathbb{E} \left[ \sum_{t=1}^{T} \left( \sqrt{\sum_{s=1}^{t+1} H(p_{s}(.|X_s))} - \sqrt{\sum_{s=1}^{t} H(p_{s}(.|X_s))} \right) \right]
\]
\[
= 2c'_t \sqrt{\ln K} \mathbb{E} \left[ \left( \sum_{s=1}^{t+1} H(p_{s}(.|X_s)) - \sqrt{H(p_{1}(.|X_1))} \right) \right]
\]
\[
\leq 2c'_t \sqrt{\ln K} \mathbb{E} \left[ \sum_{s=1}^{T} H(p_{s}(.|X_s)) \right],
\]
where in the last step we again used the fact that $H(p_s(.|X_s)) \leq H(p_1(.|X_1)) = \ln K$. Hence, again using the
We first bound the stability term

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

which concludes the proof.

\[ \Box \]

### 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 \( x \in \mathcal{X} \), we have

\[
\bar{R}_T(x) = \mathbb{E}_{A_t} \left[ \sum_{t=1}^{T} (\langle x, \bar{\theta}_{t,A_t} \rangle - \langle x, \bar{\theta}_{t,\pi^*(x)} \rangle) \right] \\
\leq \sum_{t=1}^{T} (\psi_t(p_{t+1}(\cdot|X)) - \psi_t(p_{t+1}(\cdot|X)) + \psi_T(\pi^*(\cdot|X)) - \psi_1(p_1(\cdot|X)) \\
+ \sum_{t=1}^{T} (1 - \gamma_t) \left( \langle p_t(\cdot|X) - p_{t+1}(\cdot|X), \bar{\ell}_t(x) \rangle - D_t(p_{t+1}(\cdot|X), p_t(\cdot|X)) \right) + U(x). \tag{44}
\]

We first bound the stability term \( \langle p_t(\cdot|X) - p_{t+1}(\cdot|X), \bar{\ell}_t(x) \rangle - D_t(p_{t+1}(\cdot|X), p_t(\cdot|X)) \). Since the function \( f(q) = \sum_{a \in [K]} (p_t(a|x) - q(\cdot|X)) \langle x, \bar{\theta}_{t,a} \rangle - D_t(q, p_t(\cdot|X)) \) 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

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

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

\[
\sum_{a \in [K]} (p_t(a|x) - p_{t+1}(a|x)) \langle x, \bar{\theta}_{t,a} \rangle - D_t(p_{t+1}(\cdot|X), p_t(\cdot|X)) \\
\leq \sum_{a \in [K]} (p_t(a|x) - q^*(a)) \langle x, \bar{\theta}_{t,a} \rangle - D_t(q^*, p_t(\cdot|X)) \\
= \sum_{a \in [K]} \left( \langle x, \bar{\theta}_{t,a} \rangle (p_t(a|x) - q^*(a)) - \frac{1}{\eta_t} (q^*(a) \ln p_t(a|x) - p_t(a|x) \ln p_t(a|x) - (\ln p_t(a|x) + 1)(q^*(a) - p_t(a|x))) \right) \\
= \sum_{a \in [K]} \left( \langle x, \bar{\theta}_{t,a} \rangle p_t(a|x) + \frac{1}{\eta_t} (q^*(a) - p_t(a|x)) \right) \\
= \frac{1}{\eta_t} \sum_{a \in [K]} p_t(a|x) \left( \exp(-\eta_t \langle x, \bar{\theta}_{t,a} \rangle) + \eta_t \langle x, \bar{\theta}_{t,a} \rangle - 1 \right). \tag{45}
\]

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

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

implying that

\[
(1 - \gamma_t) \sum_{a \in [K]} (p_t(a|x) - p_{t+1}(a|x)) \langle x, \bar{\theta}_{t,a} \rangle - D_t(p_{t+1}(\cdot|X), p_t(\cdot|X)) \leq (1 - \gamma_t) \eta_t \sum_{a \in [K]} p_t(a|x) \langle x, \bar{\theta}_{t,a} \rangle^2.
\]
Since \( p_t(a|x) = \frac{1}{1-\gamma_t}(\pi_t(a|x) - \frac{\gamma_t}{K}) \) from the definition of \( \pi_t(a|x) \), we obtain

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

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

\[
\sum_{t=1}^{T} (\psi_t(p_{t+1}(\cdot|x)) - \psi_t(p_t(\cdot|x))) + \psi_{t+1}(\pi^*(\cdot|x)) - \psi_t(\pi(\cdot|x)) \leq \sum_{t=1}^{T} (\beta_{t+1} - \beta_t) H(p_{t+1}(\cdot|x)) + \beta_1 \ln K. \tag{47}
\]

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

### G.4 Proof of Lemma 2

**Proof of Lemma 2.** Let \( \| \cdot \|_{\text{op}} \) be the operator norm of any positive semi-definite matrix. Recall that definitions of the biased estimator \( \tilde{\theta}_{t,a} = \tilde{\Sigma}_{t,a}^+X_t \xi_t(X_t, A_t) \mathbb{I} [A_t = a] \) and unbiased estimator \( \check{\theta}_{t,a} = \sum_{a} \tilde{\Sigma}_{t,a}^+X_t \xi_t(X_t, A_t) \mathbb{I} [A_t = a] \).

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 \( \tilde{\Sigma}_{t,a}^+ \), the output of MGR procedure in Algorithm 7 with \( \rho = \frac{1}{K} \), we have \( \mathbb{E}_t[A_{t,a}] = \mathbb{E}_t[\Pi_{j=1}^k (I - \rho B_{t,a})] = (I - \frac{1}{2} \Sigma_{t,a})^k \) for each \( a \in [K] \). Then, it gives \( \mathbb{E}_t[\tilde{\Sigma}_{t,a}^+] = \frac{1}{2} \sum_{k=0}^{M_t} (I - \frac{1}{2} \Sigma_{t,a})^k = \Sigma_{t,a}^{-1} - (I - \frac{1}{2} \Sigma_{t,a})^{M_t} \Sigma_{t,a}^{-1} \). Using these expressions, for the biased estimator \( \tilde{\theta}_{t,a} \) of each action \( a \in [K] \), we have that

\[
\mathbb{E}_t[\tilde{\theta}_{t,a}] = \mathbb{E}_t[\tilde{\Sigma}_{t,a}^+X_t \xi_t(X_t, a) \mathbb{I} [A_t = a]]
= \mathbb{E}_t[\tilde{\Sigma}_{t,a}^+] \mathbb{E}_t[X_t \xi_t(X_t, \check{\theta}_{t,a}) \mathbb{I} [A_t = a]]
= \mathbb{E}_t[\tilde{\Sigma}_{t,a}^+] \mathbb{E}_t[X_t X_t^\top \mathbb{I} [A_t = a] \cdot \check{\theta}_{t,a}]
= \mathbb{E}_t[\tilde{\Sigma}_{t,a}^+] \Sigma_{t,a} \check{\theta}_{t,a}
= \left( \Sigma_{t,a}^{-1} - (I - \frac{1}{2} \Sigma_{t,a})^{M_t} \right) \Sigma_{t,a} \check{\theta}_{t,a}
= \theta_{t,a} - (I - \frac{1}{2} \Sigma_{t,a})^{M_t} \check{\theta}_{t,a}.
\]

implying that

\[
\mathbb{E}_t[\tilde{\theta}_{t,a} - \check{\theta}_{t,a}] = - \left( I - \frac{1}{2} \Sigma_{t,a} \right)^{M_t} \check{\theta}_{t,a}.
\]

Therefore, we obtain

\[
\mathbb{E}_t[(X_t, \tilde{\theta}_{t,a} - \check{\theta}_{t,a})] \leq \|X_t\|_2 \|\check{\theta}_{t,a}\|_2 \left\| \left( I - \frac{1}{2} \Sigma_{t,a} \right)^{M_t} \right\|_{\text{op}} \leq \left\| \left( I - \frac{1}{2} \Sigma_{t,a} \right)^{M_t} \right\|_{\text{op}} \leq \left( 1 - \frac{\gamma_t \lambda_{\text{min}}(\Sigma)}{2K} \right) \leq \exp \left( - \frac{\gamma_t \lambda_{\text{min}}(\Sigma)}{2K} \cdot M_t \right) \leq \frac{1}{e^t},
\]

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

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 \( x \in X, \|\eta_t(x, \check{\theta}_{t,a})\| \) for each \( a \in [K] \) is
bounded as follows:

$$
|\eta_t(x, \tilde{\theta}_{t,a})| = \eta_t|\langle x, \tilde{\Sigma}_{t,a}^+ X_t \ell_t(X_t, A_t) 1[A_t = 1]\rangle| \leq \eta_t|x^\top (\tilde{\Sigma}_{t,a}^+ X_t)\rangle \leq \eta_t\|\tilde{\Sigma}_{t,a}^+\|_\text{op}
$$

$$
\leq \eta_t \left(\|pI + \rho \sum_{k=1}^{M_t} A_{k,a}\|_\text{op}\right) \leq \eta_t \left(\frac{1}{2} \left(1 + \sum_{k=1}^{M_t} \|\Pi_{j=1}^k \left(I - \frac{1}{2} B_{k,a}\right)\|_\text{op}\right) \right) \leq \frac{\eta_t(M_t + 1)}{2},
$$

(48)

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

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

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

\section{Proof of Lemma 3.}

By Lemma 2 and the definitions of $\beta_t, \eta_t, \gamma_t$ and $M_t$, we can see that $|\eta_t(x_0, \tilde{\theta}_{t,a})| \leq 1$ holds, which allow us to use Lemma 1 for fixed $X_0$. Then we have

$$
\mathbb{E}[\hat{R}_T(X_0)] \leq \mathbb{E} \left[ \sum_{t=1}^{T} (\beta_{t+1} - \beta_t) H(p_{t+1}(\cdot|X_0)) \right] + \mathbb{E} \left[ \sum_{t=1}^{T} \eta_t \sum_{a \in [K]} \pi_t(a|X_0)(X_0, \tilde{\theta}_{t,a}) \right] + \mathbb{E} [U(X_0)] + \beta_1 \ln K.
$$

(49)

Using the definition of $\beta_t = \max\{2, c'_2 \ln T, \beta_t'\}$, we have

$$
\beta_1 \ln K \leq c'_2 \ln K \ln T.
$$

(50)

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

$$
\beta_t' = c'_1 + \frac{t-1}{\sqrt[4]{1 + \ln K^{-1} \sum_{a=1}^{t-1} H(p_a(\cdot|X_a))}} \geq \frac{c'_1 t}{\sqrt[4]{1 + \ln K^{-1} \sum_{a=1}^{t-1} H(p_a(\cdot|X_a))}}
$$

and thus

$$
\sum_{t=1}^{T} \eta_t \leq \sum_{t=1}^{T} \frac{1}{\beta_t'} \leq \sum_{t=1}^{T} \frac{1}{c'_1} \sqrt[4]{1 + \ln K^{-1} \sum_{a=1}^{t-1} H(p_a(\cdot|X_a))} \leq \frac{1 + \ln T}{c'_1} \sqrt[4]{1 + \ln K^{-1} \sum_{a=1}^{t-1} H(p_a(\cdot|X_a))} = \mathcal{O} \left( \frac{\ln T}{c'_1 \sqrt{\ln K}} \sum_{t=1}^{T} H(p_t(\cdot|X_t)) \right),
$$

(51)

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

By Lemma 8 and (51), we obtain

$$
term B = \mathbb{E} \left[ \sum_{t=1}^{T} \eta_t \sum_{a \in [K]} \pi_t(a|X_0)(X_0, \tilde{\theta}_{t,a}) \right] = \mathcal{O} \left( \frac{3Kd \cdot \ln T}{c'_1 \sqrt{\ln K}} \sum_{t=1}^{T} H(p_t(\cdot|X_t)) \right),
$$

(52)
where we used the fact that $\mathbb{E}_{\gamma \sim \mathcal{D}}[p_t(\cdot|X_0)\tilde{\theta}_t] = \mathbb{E}_{\gamma \sim \mathcal{D}}[p_t(\cdot|X_1)]\tilde{\theta}_t$.

For $\mathbb{E}[U(X_0)]$, from Lemma 2, we have $|\mathbb{E}[\langle X_t, \tilde{\theta}_{t,a} - \tilde{\theta}_{t,a} | \mathcal{F}_{t-1} |]| \leq \frac{1}{T}$, and thus
\[
\mathbb{E}[U(X_0)] = \mathbb{E} \left[ \sum_{t=1}^T \gamma_t \sum_{a \in [K]} \left( \frac{1}{K} - \pi^*(a|X_0) \right) \langle x, \tilde{\theta}_{t,a} \rangle \right] 
\leq \mathbb{E} \left[ \sum_{t=1}^T \gamma_t \max_{a \in [K]} \langle X_0, \tilde{\theta}_{t,a} - \tilde{\theta}_{t,a} + \tilde{\theta}_{t,a} \rangle \right] 
\leq 2\mathbb{E} \left[ \sum_{t=1}^T \gamma_t \right].
\] (53)

where we used $\mathbb{E}[\tilde{\theta}_{t,a}] = \theta_{t,a}$ and $\mathbb{E}[\ell_t(X_0, a)] \leq 1$ in the second and third inequality. From the definition of $\gamma_t$ and (51), we have
\[
\mathbb{E} \left[ \sum_{t=1}^T \gamma_t \right] = \mathbb{E} \left[ \sum_{t=1}^T \alpha_t \eta_t \right] \leq \mathbb{E} \left[ \sum_{t=1}^T \frac{4K\ln T}{\lambda_{\min}(\Sigma)} \cdot \eta_t \right] = O \left( \frac{K\ln T}{\lambda_{\min}(\Sigma)} \cdot \ln T \cdot c_1' / \sqrt{K} \right) \mathbb{E} \left[ \sum_{t=1}^T H(p_t(\cdot|X_0)) \right].
\] (54)

Thus from (53) and (54), we obtain
\[
\mathbb{E}[U(X_0)] \leq 2\mathbb{E} \left[ \sum_{t=1}^T \gamma_t \right] = O \left( \frac{K\ln^2 T}{c_1'\lambda_{\min}(\Sigma)\sqrt{\ln K}} \mathbb{E} \left[ \sum_{t=1}^T H(p_t(\cdot|X_0)) \right] \right).
\] (55)

Finally, we will evaluate term $A$. Let $t_0$ be the first round in which $\beta'_t$ 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 $\beta_t$, we have that
\[
\mathbb{E} \left[ \sum_{t=1}^T (\beta_{t+1} - \beta_t) H(p_{t+1}(\cdot|X_0)) \right]
= \mathbb{E} \left[ \sum_{t=1}^{t_0-2} (\beta_{t+1} - \beta_t) H(p_{t+1}(\cdot|X_0)) + (\beta_{t_0} - \beta_{t_0-1}) H(p_{t+1}(\cdot|X_0)) + \sum_{t=t_0}^T (\beta_{t+1} - \beta_t) H(p_{t+1}(\cdot|X_0)) \right]
\leq \mathbb{E} \left[ 0 + (\beta'_0 - \beta'_{t_0-1}) H(p_{t+1}(\cdot|X_0)) + \sum_{t=t_0}^T (\beta'_t - \beta'_t) H(p_{t+1}(\cdot|X_0)) \right]
\leq \mathbb{E} \left[ \sum_{t=1}^T (\beta'_t - \beta'_t) H(p_{t+1}(\cdot|X_0)) \right] = O \left( c_1' \sqrt{\ln K} \mathbb{E} \left[ \sum_{t=1}^T H(p_t(\cdot|X_0)) \right] \right),
\] (56)

where the first inequality is due to the fact that $\beta_t$ 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
\[
\mathbb{E} \left[ \sum_{t=1}^T (\beta_{t+1} - \beta_t) H(p_{t+1}(\cdot|X_0)) \right] = O \left( c_1' \sqrt{\ln K} \mathbb{E} \left[ \sum_{t=1}^T H(p_t(\cdot|X_0)) \right] \right).
\] (57)

Combining (50), (52), (55), (57) with (49), we obtain
\[
\mathbb{E} \left[ \tilde{R}_T(X_0) \right] = O \left( c_1' \sqrt{\ln K} + \frac{3Kd + \frac{2K\ln T}{\lambda_{\min}(\Sigma)}}{c_1' \sqrt{\ln K}} \ln T \right) \mathbb{E} \left[ \sum_{t=1}^T H(p_t(\cdot|X_0)) \right] + c_2' \ln K \ln T,
\]
and plugging $c_2' = \frac{8K}{\lambda_{\min}(\Sigma)}$ to this bound concludes the proof.
G.6 Proof of Theorem 2

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

\[ R_T \leq \mathbb{E}[\tilde{R}_T(X_0)] + 2 \sum_{t=1}^{T} \max_{a \in [K]} \mathbb{E}[(X_t, b_t, a)] \]

= \( \mathcal{O} \left( \left( c'_1 \sqrt{\ln K} + \frac{K \ln T}{\lambda_{\min}(\Sigma)} \ln T \right) \right) \sqrt{\mathbb{E} \left[ \sum_{t=1}^{T} H(p_t(\cdot|X_0)) \right] + c'_2 \ln K \ln T} \)

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

\[ \sum_{t=1}^{T} \max_{a \in [K]} \mathbb{E}[(X_t, b_t, a)] \leq \sum_{t=1}^{T} \frac{1}{t^2} \leq 2. \quad (58) \]

Setting

\[ c'_1 = \sqrt{\left( 3Kd + \frac{2K \ln T}{\lambda_{\min}(\Sigma)} \right) \ln K} \]

gives

\[ R_T = \mathcal{O} \left( \left( c'_1 \sqrt{\ln K} + \frac{K \ln T}{\lambda_{\min}(\Sigma)} \ln T \right) \right) \sqrt{\mathbb{E} \left[ \sum_{t=1}^{T} H(p_t(\cdot|X_0)) \right] + c'_2 \ln K \ln T} \]

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

(59)

where we used \( c'_2 = \frac{8K}{\lambda_{\min}(\Sigma)} \) in the third equality.

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

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

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 \( \Delta_x(a) := \langle x, \theta_a - \theta^{*}(x) \rangle \) for \( x \in X \) and each action \( a \in [K] \).
We have
\[
R_T = 
\mathbb{E} \left[ \sum_{t=1}^{T} (\ell_t(X_t, A_t) - \ell_t(X_t, \pi^*(X_t))) \right] 
\]
\[
= \mathbb{E} \left[ \sum_{t=1}^{T} (X_t, \theta_{t,A_t} - \theta_{t,\pi^*(X_t)}) \right] + \mathbb{E} \left[ \sum_{t=1}^{T} (X_t, \theta_{A_t} - \theta_{A_t}) \right] + \mathbb{E} \left[ \sum_{t=1}^{T} (X_t, \theta_{\pi^*(X_t)} - \theta_{\pi^*(X_t)}) \right] 
\]
\[
\geq \mathbb{E} \left[ \sum_{t=1}^{T} (X_t, \theta_{A_t} - \theta_{\pi^*(X_t)}) \right] - \mathbb{E} \left[ \sum_{t=1}^{T} (X_t, \theta_{t,A_t} - \theta_{t,A_t}) \right] - \mathbb{E} \left[ \sum_{t=1}^{T} (X_t, \theta_{t,\pi^*(X_t)} - \theta_{t,\pi^*(X_t)}) \right] 
\]
\[
\geq \mathbb{E} \left[ \sum_{t=1}^{T} (X_t, \theta_{A_t} - \theta_{\pi^*(X_t)}) \right] - 2\mathbb{E} \left[ \max_{a \in [K]} \| X_t \|_2 \| \theta_{t,a} - \theta_{a} \|_2 \right] 
\]
\[
\geq \sum_{t=1}^{T} \Delta_{X_t}(A_t) - 2\mathbb{E} \left[ \sum_{t=1}^{T} \max_{a \in [K]} \| \theta_{t,a} - \theta_{a} \|_2 \right], 
\]
\[
\geq \sum_{t=1}^{T} \Delta_{X_t}(A_t) - 2C, 
\]
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 \( \pi^* \in \Pi \),
\[
\varrho_{0}(\pi^*) := \sum_{t=1}^{T} (1 - p_t(\pi^*|X_t)|X_0)) \quad \text{and} \quad \varrho_{(X_t)_{t=1}^{T}}(\pi^*) := \sum_{t=1}^{T} (1 - p_t(\pi^*|X_t)|X_t)) \quad \text{and} \quad \varrho_X(\pi^*) := \mathbb{E}[\varrho_{(X_t)_{t=1}^{T}}(\pi^*)]. 
\]
Note that it holds that \( 0 \leq \varrho_X(\pi^*) \leq T \). We also confirm the property on them in the following lemma.

**Lemma 20.** Let \( \pi^* \) be the optimal policy defined in (1). Then we have \( \varrho_X(\pi^*) = \mathbb{E}[\varrho_{0}(\pi^*)] \).

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

**Lemma 21.** In the corrupted stochastic setting, the regret is bounded from below as
\[
R_T \geq \frac{\Delta_{\min}}{2} \varrho_X(\pi^*) - 2C. 
\]

(61)
Proof of Lemma 21. Recall that $\Delta_x(a) := x^T(\theta_a - \theta_{x^*}(a))$ for $a \in [K] \setminus \{x^*(a)\}$, where $x^*$ is the unique optimal policy given by (1). Also recall that $\Delta_{\min}(x) := \min_{a \neq x^*} \Delta_x(a)$ and $\Delta_{\min} := \min_{x \in X} \Delta_{\min}(x)$. Then, using these gap definitions and Lemma 19, the regret is bounded from below as

$$R_T \geq E \left[ \sum_{t=1}^T \Delta_{x_t}(A_t) - 2C \right] = E \left[ \sum_{t=1}^T \sum_{a \in [K] \setminus \{x^*(X_t)\}} \pi_t(a|X_t)\Delta_{x_t}(a) \right] - 2C$$

$$\geq E \left[ \sum_{t=1}^T \sum_{a \in [K] \setminus \{x^*(X_t)\}} (1 - \gamma_t)p_t(a|X_t)\Delta_{x_t}(a) \right] - 2C$$

$$\geq \frac{1}{2} E \left[ \sum_{t=1}^T \sum_{a \in [K] \setminus \{x^*(X_t)\}} p_t(a|X_t)\Delta_{\min}(X_t) \right] - 2C$$

$$= \frac{1}{2} E \left[ \sum_{t=1}^T \sum_{a \in [K] \setminus \{x^*(X_t)\}} p_t(a|X_t)\min_{x \in X} \Delta_{\min}(x) \right] - 2C$$

$$\geq \frac{1}{2} E \left[ \sum_{t=1}^T \sum_{a \in [K] \setminus \{x^*(X_t)\}} p_t(a|X_t)\min_{x \in X} \Delta_{\min}(x) \right] - 2C$$

$$= \frac{\Delta_{\min}}{2} E \left[ \sum_{t=1}^T \sum_{a \in [K] \setminus \{x^*(X_t)\}} p_t(a|X_t) \right] - 2C$$

$$= \frac{\Delta_{\min}}{2} E \left[ \sum_{t=1}^T (1 - p_t(x^*(X_t)|X_t)) \right] - 2C$$

$$= \frac{\Delta_{\min}}{2} E \left[ \Delta_{\min}(X_t) \right] - 2C = \frac{\Delta_{\min}}{2} \bar{\theta}_X(x^*) - 2C,$$

where the second inequality follows by (11), the third inequality follows by $\gamma_t \leq \frac{1}{t}$, and the last steps follows by the definitions of $\bar{\theta}_X(x^*) := \sum_{t=1}^T (1 - p_t(x^*(X_t)|X_t))$ and $\bar{\theta}_X(x^*) := E[\bar{\theta}_X(x^*)]$. \hfill \Box

The following lemma that bounds the sum of entropy in terms of $q_0(x^*)$ 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

$$\sum_{t=1}^T H(p_t(-|X_0)) \leq q_0(x^*) \ln \frac{eKT}{q_0(\pi)}$$

where $q_0(\pi) = \sum_{t=1}^T (1 - p_t(x^*(X_0)|X_0))$.

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

$$\sum_{t=1}^T H(p_t(-|X_0)) \leq \sum_{t=1}^T (1 - p_t(x^*(X_0)|X_0)) \left( \ln \frac{K - 1}{1 - p_t(x^*(X_0)|X_0)} + 1 \right)$$

$$\leq q_0(x^*) \left( \ln \frac{(K - 1)T}{q_0(x^*)} + 1 \right) \leq q_0(x^*) \ln \frac{eKT}{q_0(x^*)}. $$
where the second inequality follows from Jensen’s inequality.

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

\[
\mathbb{E} \left[ \sum_{t=1}^{T} H(p_t(\cdot|X_0)) \right] \leq \mathbb{E} [\varrho_0(\pi^*) \ln(KT)] = \mathbb{E} [\varrho(X_t)] \ln(\pi^*) \ln(KT)] = \varrho_X(\pi^*) \ln(KT),
\]

where we used Lemma 20 in the first equality.

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

\[
R_T = (1 + \lambda)R_T - \lambda R_T \\
\leq \mathbb{E} \left[ (1 + \lambda) \sqrt{\left( d + \frac{\ln T}{\lambda_{\min}(\Sigma)} \right) K \ln T \ln(KT) \cdot \varrho_X(\pi^*) - \lambda \frac{\Delta_{\min}}{2} \varrho_X(\pi^*)} \right] + \lambda \cdot 2C + (1 + \lambda)c_4 \\
= \mathcal{O} \left( (1 + \lambda)^2 (d + \frac{\ln T}{\lambda_{\min}(\Sigma)}) K \ln T \ln(KT) \cdot \lambda \Delta_{\min} + \lambda C + \lambda c_4 \right) \\
= \mathcal{O} \left( \frac{d + \ln T}{\lambda_{\min}(\Sigma)} K \ln T \ln(KT) \cdot \lambda \Delta_{\min} + \frac{(d + \ln T)}{\lambda_{\min}(\Sigma)} K \ln T \ln(KT) \cdot \Delta_{\min} + C \Delta_{\min} \lambda + \lambda C + \lambda c_4 \right),
\]

where we used \(a\sqrt{x} - \frac{b \sqrt{x}}{x} \leq \frac{a^2}{2x}\) for any \(a, b, x \geq 0\) in the first equality.

By letting \(0 \leq \lambda \leq 1\) to be

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

we have

\[
R_T = \mathcal{O} \left( \frac{(d + \frac{\ln T}{\lambda_{\min}(\Sigma)}) K \ln T \ln(KT)}{\Delta_{\min}} + \sqrt{\frac{(d + \frac{\ln T}{\lambda_{\min}(\Sigma)}) K \ln T \ln(KT)}{\Delta_{\min}}} \cdot C \\
+ \sqrt{\frac{c_4 (d + \frac{\ln T}{\lambda_{\min}(\Sigma)}) K \ln T \ln(KT)}}{\Delta_{\min}} \right) \\
= \mathcal{O} \left( \frac{(d + \frac{\ln T}{\lambda_{\min}(\Sigma)}) K \ln T \ln(KT)}{\Delta_{\min}} + \sqrt{\frac{(d + \frac{\ln T}{\lambda_{\min}(\Sigma)}) K \ln T \ln(KT)}}{\Delta_{\min}} \cdot C \\
+ K \ln T \left( \frac{1}{\lambda_{\min}(\Sigma)} \left( d + \frac{\ln T}{\lambda_{\min}(\Sigma)} \right) \ln(K) \ln(KT) \right),
\]

which concludes the proof of the theorem.