Efficient and robust algorithms for adversarial linear contextual bandits

Gergely Neu GERGELY.NEU@GMAIL.COM
Universitat Pompeu Fabra, Barcelona, Spain

Julia Olkhovskaya JULIA.OLKHOVSKAYA@GMAIL.COM
Universitat Pompeu Fabra, Barcelona, Spain

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We consider an adversarial variant of the classic K-armed linear contextual bandit problem where the sequence of loss functions associated with each arm are allowed to change without restriction over time. Under the assumption that the d-dimensional contexts are generated i.i.d. at random from a known distribution, we develop computationally efficient algorithms based on the classic EXP3 algorithm. Our first algorithm, REALLINEXP3, is shown to achieve a regret guarantee of $\widetilde{O}(\sqrt{KdT})$ over T rounds, which matches the best known lower bound for this problem. Our second algorithm, ROBUSTLINEXP3, is shown to be robust to misspecification, in that it achieves a regret bound of $\widetilde{O}((Kd)^{1/3}T^{2/3}) + \varepsilon\sqrt{d}T$ if the true reward function is linear up to an additive nonlinear error uniformly bounded in absolute value by ε. To our knowledge, our performance guarantees constitute the very first results on this problem setting.

Keywords: Contextual bandits, adversarial bandits, linear contextual bandits

1. Introduction

The contextual bandit problem is one of the most important sequential decision-making problems studied in the machine learning literature. Due to its ability to account for contextual information, the applicability of contextual bandit algorithms is far superior to that of standard multi-armed bandit methods: the framework of contextual bandits can be used to address a broad range of important and challenging real-world decision-making problems such as sequential treatment allocation (Tewari and Murphy, 2017) and online advertising (Li et al., 2010). On the other hand, the framework is far less complex than that of general reinforcement learning, which allows for proving formal performance guarantees under relatively mild assumptions. As a result, there has been significant interest in this problem within the learning-theory community, resulting in a wide variety of algorithms with performance guarantees proven under a number of different assumptions. In the present paper, we fill a gap in this literature and design computationally efficient algorithms with strong performance guarantees for an adversarial version of the *linear contextual bandit* problem.

Perhaps the most well-studied variant of the contextual bandit problem is that of *stochastic linear contextual bandits* (Auer, 2002; Rusmevichientong and Tsitsiklis, 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011; Lattimore and Szepesvári, 2017). First proposed by Abe and Long (1999), this version supposes that the loss of each action is a fixed linear function of the vector-valued context, up to some zero-mean noise. Most algorithms designed for this setting are based on some variation of the "optimism in the face of uncertainty" principle championed by Auer (2002); Auer

et al. (2002a), or more generally by an appropriate exploitation of the concentration-of-measure phenomenon (Boucheron et al., 2013). By now, this problem setting is very well-understood in many respects: there exist several computationally efficient, easy-to-implement algorithms achieving near-optimal worst-case performance guarantees (Abbasi-Yadkori et al., 2011; Agrawal and Goyal, 2013). These algorithms can be even adapted to more involved loss models like generalized linear models, Gaussian processes, or very large structured model classes while retaining their performance guarantees (Filippi et al., 2010; Srinivas et al., 2010; Calandriello et al., 2019; Foster et al., 2019). That said, most algorithms for stochastic linear contextual bandits suffer from the limitation that they are sensitive to *model misspecification*: their performance guarantees become void as soon as the true loss functions deviate from the postulated linear model to the slightest degree. This issue has very recently attracted quite some attention due to the work of Du et al. (2019), seemingly implying that learning an ε-optimal policy in a contextual bandit problem has an extremely large sample complexity when assuming that the linear model is ε-inaccurate (defined formally later in our paper). This claim was quickly countered by Van Roy and Dong (2019) and Lattimore et al. (2020), who both showed that learning a (somewhat worse) $\varepsilon\sqrt{d}$-optimal policy is feasible with the very same sample complexity as learning a near-optimal policy in a well-specified linear model. Yet, since algorithms that are currently known to enjoy these favorable guarantees are quite complex, there is much work left to be done in designing practical algorithms with strong guarantees under model misspecification. This is one of the main issues we address in this paper.

Another limitation of virtually all known algorithms for linear contextual bandits is that they crucially rely on assuming that the loss function is *fixed during the learning procedure* [1]. This is in stark contrast with the literature on multi-armed (non-contextual) bandits, where there is a rich literature on both stochastic bandit models assuming i.i.d. rewards and adversarial bandit models making no assumptions on the sequence of loss functions—see Bubeck and Cesa-Bianchi (2012) and Lattimore and Szepesvári (2019) for an excellent overview of both lines of work. Our main contribution in the present paper is addressing this gap by designing and analyzing algorithms that are guaranteed to work for arbitrary sequences of loss functions. While it is tempting to think that the our bandit problem can be directly addressed by a minor adaptation of algorithms developed for adversarial linear bandits, this is unfortunately not the case: all algorithms developed for such problems require a *fixed decision set*, whereas reducing the linear contextual bandit problem to a linear bandit problem requires the use *decision sets that change as a function of the contexts* (Lattimore and Szepesvári, 2019, Section 18). As a crucial step in our analysis, we will assume that the contexts are generated in an i.i.d. fashion and that the loss function in each round is statistically independent from the context in the same round. This assumption will allow us to relate the contextual bandit problem to a set of auxiliary bandit problems with a *fixed action sets*, and reduce the scope of the analysis to these auxiliary problems.

Our main results are the following. We consider a K-armed linear contextual bandit problem with d-dimensional contexts where in each round, a loss function mapping contexts and actions to real numbers is chosen by an adversary in a sequence of T rounds. The aim of the learner is to minimize its regret, defined as the gap between the total incurred by the learner and that of the best decision-making policy π^* fixed in full knowledge of the loss sequence. We consider two different assumptions on the loss function. Assuming that the loss functions selected by the adversary are all linear, we propose an algorithm achieving a regret bound of order $\sqrt{KdT}$, which is known to be

1. Or make other stringent assumptions about the losses, such as supposing that their total variation is bounded—see, e.g., Cheung et al. (2019); Russac et al. (2019); Kim and Tewari (2019).

minimax optimal even in the simpler case of i.i.d. losses (cf. Chu et al., 2011). Second, we consider loss functions that are "nearly linear" up to an additive nonlinear function uniformly bounded by ε. For this case, we design an algorithm that guarantees regret bounded by $(Kd)^{1/3}T^{2/3} + \varepsilon\sqrt{d}T$. Notably, these latter bounds hold against *any* class of policies and the $\varepsilon\sqrt{d}T$ overhead paid for nonlinearity is optimal when K is large (Lattimore et al., 2020). Both algorithms are computationally efficient, but require some prior knowledge to the distribution of the contexts.

There exist numerous other approaches for contextual bandit problems that do not rely on modeling the loss functions, but rather make use of a class of *policies* that map contexts to actions. Instead of trying to fit the loss functions, these approaches aim to identify the best policy in the class. A typical assumption in this line of work is that one has access to a computational oracle that can perform various optimization problems over the policy class (such as returning an optimal policy given a joint distribution of context-loss pairs for each action). Given access to such an oracle, there exist algorithms achieving near-optimal performance guarantees when the loss function is fixed (Dudík et al., 2011; Agarwal et al., 2014). More relevant to our present work are the works of Rakhlin and Sridharan (2016) and Syrgkanis et al. (2016a,b) who propose efficient algorithms with guaranteed performance for adversarial loss sequences and i.i.d. contexts. Unlike the algorithms we present in this paper, these methods fail to guarantee optimal performance guarantees of order $\sqrt{T}$. Yet another line of work considers optimizing surrogate losses, where achieving regret of order $\sqrt{T}$ is indeed possible, with the caveat that the bounds only hold for the surrogate loss (Kakade et al., 2008; Beygelzimer et al., 2017; Foster and Krishnamurthy, 2018).

The rest of the paper is organized as follows. After defining some basic notation, Section 2 presents our problem formulation and states our assumptions. We present our algorithms and main results in Section 3 and provide the proofs in Section 4. Section 5 concludes the paper by discussing some implications of our results and posing some open questions for future study.

Notation. We use $\langle \cdot, \cdot \rangle$ to denote inner products in Euclidean space and by $\|\cdot\|_2$ we denote the Euclidean norm. For a symmetric positive semidefinite matrix A, we use $\lambda_{\min}(A)$ to denote its smallest eigenvalue. We use $\|A\|_{\text{op}}$ to denote the operator norm of A and we write $\text{tr}\,(A)$ for the trace of a matrix A. Finally, we use $A \succcurlyeq 0$ to denote that an operator A is positive semi-definite, and we use $A \succcurlyeq B$ to denote $A - B \succcurlyeq 0$.

2. Preliminaries

We consider a sequential interaction scheme between a *learner* and its *environment*, where the following steps are repeated in a sequence of rounds $t = 1, 2, \ldots, T$:

1. For each action $a = 1, 2, \ldots, K$, the environment chooses a loss vector $\theta_{t,a} \in \mathbb{R}^d$,
2. independently of the choice of loss vectors, the environment draws the context vector $X_t \in \mathbb{R}^d$ from the context distribution $\mathcal{D}$, and reveals it to the learner,
3. based on X_t and possibly some randomness, the learner chooses action $A_t \in [K]$,
4. the learner incurs and observes loss $\ell_t(X_t, A_t) = \langle X_t, \theta_{t,A_t} \rangle$.

The goal of the learner is to pick its actions in a way that its total loss is as small as possible. Since we make no statistical assumptions about the sequence of losses (and in fact we allow them to depend on all the past interaction history), the learner cannot actually hope to incur as little loss as the best

sequence of actions. A more reasonable aim is to match the performance of the *best fixed policy* that maps contexts to actions in a static way. Formally, the learner will consider the set Π of all policies $\pi : \mathbb{R}^d \to [K]$, and aim to minimize its *total expected regret* (or, simply, *regret*) defined as

$$R_T = \max_{\pi\in\Pi} \mathbb{E}\left[\sum_{t=1}^{T}\big(\ell_t(X_t, A_t) - \ell_t(X_t, \pi(X_t))\big)\right] = \max_{\pi\in\Pi} \mathbb{E}\left[\sum_{t=1}^{T}\left\langle X_t, \theta_{t,A_t} - \theta_{t,\pi(X_t)}\right\rangle\right],$$

where the expectation is taken over the randomness injected by the learner, as well as the sequence of random contexts. For stating many of our technical results, it will be useful to define the filtration $\mathcal{F}_t = \sigma(X_s, A_s, \forall s \le t)$ and the notations $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_{t-1}]$ and $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot|\mathcal{F}_{t-1}]$. We will also often make use of a *ghost sample* $X_0 \sim \mathcal{D}$ drawn independently from the entire interaction history $\mathcal{F}_T$ for the sake of analysis. For instance, we can immediately show using this technique that for any policy π, we have

$$\mathbb{E}\left[\left\langle X_t, \theta_{t,\pi(X_t)}\right\rangle\right] = \mathbb{E}\left[\mathbb{E}_t\left[\left\langle X_t, \theta_{t,\pi(X_t)}\right\rangle\right]\right] = \mathbb{E}\left[\mathbb{E}_t\left[\left\langle X_0, \theta_{t,\pi(X_0)}\right\rangle\right]\right] = \mathbb{E}\left[\left\langle X_0, \mathbb{E}\left[\theta_{t,\pi(X_0)}\right]\right\rangle\right],$$

where the last expectation emphasizes that the loss vector $\theta_{t,a}$ may depend on the past random contexts and actions. This in turn can be used to show

$$\mathbb{E}\left[\sum_{t=1}^{T}\left\langle X_t, \theta_{t,\pi(X_t)}\right\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T}\left\langle X_0, \mathbb{E}\left[\theta_{t,\pi(X_0)}\right]\right\rangle\right] \ge \mathbb{E}\left[\min_a \sum_{t=1}^{T}\left\langle X_0, \mathbb{E}\left[\theta_{t,a}\right]\right\rangle\right],$$

so the optimal policy π_T^* that the learner compares itself to is the one defined through the rule

$$\pi_T^*(x) = \arg\min_a \sum_{t=1}^{T}\langle x, \mathbb{E}[\theta_{t,a}]\rangle \qquad (\forall x \in \mathbb{R}^d). \tag{1}$$

We will refer to policies of the above form as *linear-classifier policies* and are defined through the rule $\pi_\theta(x) = \arg\min_a \langle x, \theta_a\rangle$ for any collection of parameter vectors $\theta \in \mathbb{R}^{K\times d}$. We will also rely on the notion of *stochastic policies* that assign probability distributions over the action set to each state, and use $\pi(a|x)$ to denote the probability that the stochastic policy π takes action a in state x.

Our analysis will rely on the following assumptions. We will suppose the context distribution is supported on the bounded set $\mathcal{X}$ with each $x \in \mathcal{X}$ satisfying $\|x\|_2 \le \sigma$ for some $\sigma > 0$, and also that $\|\theta_{t,a}\|_2 \le R$ for some positive R for all t, a. Additionally, we suppose that the loss function is bounded by one in absolute value: $\left|\ell_t(x, a)\right| \le 1$ for all t, a and all $x \in \mathcal{X}$. We will finally assume that the covariance matrix of the contexts $\Sigma = \mathbb{E}[X_t X_t^\intercal]$ is positive definite with its smallest eigenvalue being $\lambda_{\min} > 0$.

3. Algorithms and main results

Our main algorithmic contribution is a natural adaptation of the classic EXP3 algorithm of Auer et al. (2002a) to the linear contextual bandit setting. The key idea underlying our method is to design a suitable estimator of the loss vectors and use these estimators to define a policy for the learner as follows: letting $\widehat{\theta}_{t,a}$ be an estimator of the true loss vector $\theta_{t,a}$ and their cumulative sum $\widehat{\Theta}_{t,a} = \sum_{k=1}^{t}\widehat{\theta}_{k,a}$, our algorithm will base its decisions on the values $\langle X_t, \widehat{\Theta}_{t-1,a}\rangle$ serving as estimators of the cumulative losses $\langle X_t, \Theta_{t-1,a}\rangle = \sum_{k=1}^{t-1}\langle X_t, \theta_{k,a}\rangle$. The algorithm then uses these

values in an exponential-weights-style algorithm and plays action a with probability proportional to $\exp\left(-\eta\langle X_t, \widehat{\Theta}_{t-1,a}\rangle\right)$, where $\eta > 0$ is a *learning-rate* parameter. We present a general version of this method as Algorithm 1. As a tribute to the LINUCB algorithm, a natural extension of the classic UCB algorithm to linear contextual bandits, we refer to our algorithm as LINEXP3.

Algorithm 1 LINEXP3

Parameters: Learning rate $\eta > 0$, exploration parameter $\gamma \in (0, 1)$, Σ
Initialization: Set $\theta_{0,i} = 0$ for all $i \in [K]$.
For $t = 1, \ldots, T$, **repeat:**

1. Observe X_t and, for all a, set

$$w_t(X_t, a) = \exp\left(-\eta \sum_{s=0}^{t-1} \langle X_t, \widehat{\theta}_{s,a}\rangle\right),$$

2. draw A_t from the policy defined as

$$\pi_t\left(a|X_t\right) = (1-\gamma)\frac{w_t(X_t, a)}{\sum_{a'} w_t(X_t, a')} + \frac{\gamma}{K},$$

3. observe the loss $\ell_t(X_t, A_t)$ and compute $\widehat{\theta}_{t,a}$ for all a.

As presented above, LINEXP3 is more of a template than an actual algorithm since it does not specify the loss estimators $\widehat{\theta}_{t,a}$. Ideally, one may want to use *unbiased* estimators that satisfy $\mathbb{E}\left[\widehat{\theta}_{t,a}\right] = \theta_{t,a}$ for all t, a. Our key contribution is designing two different (nearly) unbiased estimators that will allow us to prove performance guarantees of two distinct flavors. Both estimators are efficiently computable, but require some prior knowledge the context distribution $\mathcal{D}$. In what follows, we describe the two variants of LINEXP3 based on the two estimators and state the corresponding performance guarantees, and relegate the proof sketches to Section 4. We also present two simple variants of our algorithms that work with various degrees of full-information feedback in Appendix C.

3.1. Algorithm for nearly-linear losses: ROBUSTLINEXP3

We begin by describing the simpler one of our two algorithms, which will be seen to be robust to misspecification of the linear loss model. We will accordingly refer to this algorithm as ROBUSTLINEXP3. Specifically, we suppose in this section that $\ell_t(x, a) = \langle x, \theta_{t,a}\rangle + \varepsilon_t(x, a)$, where $\varepsilon_t(x, a) : \mathbb{R}^d \times K \to \mathbb{R}$ is an arbitrary nonlinear function satisfying $|\varepsilon_t(x, a)| \leq \varepsilon$ for all t, x and a and some $\varepsilon > 0$. Also supposing that we have perfect knowledge of the covariance matrix Σ, we define the loss estimator used by ROBUSTLINEXP3 for all actions a as

$$\widehat{\theta}_{t,a} = \frac{\mathbb{I}_{\{A_t=a\}}}{\pi_t(a|X_t)} \Sigma^{-1} X_t \ell_t(X_t, A_t). \tag{2}$$

In case the loss is truly linear, it is easy to see that the above is an unbiased estimate since

$$\mathbb{E}_t\left[\widehat{\theta}_{t,a}\right] = \mathbb{E}_t\left[\mathbb{E}_t\left[\frac{\mathbb{I}_{\{A_t=a\}}}{\pi_t(a|X_t)} \Sigma^{-1} X_t \left\langle X_t, \theta_{t,a}\right\rangle \middle| X_t\right]\right] = \mathbb{E}_t\left[\mathbb{E}_t\left[\frac{\mathbb{I}_{\{A_t=a\}}}{\pi_t(a|X_t)} \middle| X_t\right] \Sigma^{-1} X_t X_t^{\mathsf{T}} \theta_{t,a}\right]$$

$$= \mathbb{E}_t \left[\Sigma^{-1} X_t X_t^\intercal \theta_{t,a} \right] = \theta_{t,a},$$

where we used the definition of Σ and the independence of $\theta_{t,a}$ from X_t in the last step. A key result in our analysis will be that, for nonlinear losses, the estimate above satisfies

$$\left| \mathbb{E}_t \left[\left\langle X_t, \widehat{\theta}_{t,a} \right\rangle - \ell_t(X_t, a) \right] \right| \le \varepsilon \sqrt{d}.$$

Our main result regarding the performance of ROBUSTLINEXP3 is the following:

Theorem 1 *For any positive $\eta \le \frac{\gamma \lambda_{\min}}{K\sigma^2}$ and for any $\gamma \in (0,1)$ the expected regret of* ROBUSTLINEXP3 *satisfies*

$$R_T \le 2\sqrt{d}\varepsilon T + 2\gamma T + \frac{2\eta K d T}{\gamma} + \frac{\log K}{\eta}.$$

Furthermore, letting $\eta = T^{-2/3} (Kd)^{-1/3} (\log K)^{2/3}$, $\gamma = T^{-1/3} (Kd \log K)^{1/3}$ and supposing that T is large enough so that $\eta \le \frac{\gamma \lambda_{\min}}{K\sigma^2}$ holds, the expected regret of ROBUSTLINEXP3 *satisfies*

$$R_T \le 5 T^{2/3} (Kd \log K)^{1/3} + 2\varepsilon \sqrt{d} T.$$

3.2. Algorithm for linear losses: REALLINEXP3

Our second algorithm uses a more sophisticated estimator based on the covariance matrix

$$\Sigma_{t,a} = \mathbb{E}_t \left[\mathbb{I}_{\{A_t = a\}} X_t X_t^\intercal \right],$$

which is used to define the estimate

$$\widetilde{\theta}^*_{t,a} = \mathbb{I}_{\{A_t = a\}} \Sigma_{t,a}^{-1} X_t \left\langle X_t, \theta_{t,a} \right\rangle.$$

This can be easily shown to be unbiased as

$$\mathbb{E}_t \left[\widetilde{\theta}^*_{t,a} \right] = \mathbb{E}_t \left[\mathbb{I}_{\{A_t = a\}} \Sigma_{t,a}^{-1} X_t \left\langle X_t, \theta_{t,a} \right\rangle \right] = \mathbb{E}_t \left[\Sigma_{t,a}^{-1} \mathbb{I}_{\{A_t = a\}} X_t X_t^\intercal \theta_{t,a} \right] = \theta_{t,a},$$

where we used the conditional independence of $\theta_{t,a}$ and X_t once again. Unfortunately, unlike the estimator used by ROBUSTLINEXP3, the bias of this estimator cannot be bounded when the losses are misspecified. However, its variance turns out to be much smaller for well-specified linear losses, which will enable us to prove tighter regret bounds for this case.

One downside of the estimator defined above is that it is very difficult to compute: the matrix $\Sigma_{t,a}$ depends on the joint distribution of the context X_t and the action A_t, which has a very complicated structure. While it is trivially easy to design an unbiased estimator of $\Sigma_{t,a}$, it is very difficult to compute a reliable-enough estimator of its inverse. To address this issue, we design an alternative estimator based on a matrix generalization of the Geometric Resampling method of Neu and Bartók (2013, 2016). The method that we hereby dub *Matrix Geometric Resampling* (MGR) has two parameters $\beta > 0$ and $M \in \mathbb{Z}_+$, and constructs an estimator of $\Sigma_{t,a}^{-1}$ through the following procedure:

Matrix Geometric Resampling

Input: data distribution $\mathcal{D}$, policy π_t, action a.

For $k = 1, \ldots, M$**, repeat**:

1. Draw $X(k) \sim \mathcal{D}$ and $A(k) \sim \pi_t(\cdot|X(k))$,
2. compute $B_{k,a} = \mathbb{I}_{\{A(k)=a\}} X(k)X(k)^\intercal$,
3. compute $A_{k,a} = \prod_{j=1}^{k}(I - \beta B_{k,a})$.

Return $\widehat{\Sigma}^{+}_{t,a} = \beta I + \beta \sum_{k=1}^{M} A_{k,a}$.

Clearly, implementing the MGR procedure requires sampling access to the distribution $\mathcal{D}$. The rationale behind the estimator $\widehat{\Sigma}^{+}_{t,a}$ is the following. Assuming that $M = \infty$ and $\beta \leq \frac{1}{\sigma^2}$, we can observe that $\mathbb{E}_t\left[B_{k,a}\right] = \Sigma_{t,a}$ and, due to independence of the contexts $X(k)$ from each other,

$$\mathbb{E}_t\left[A_{k,a}\right] = \mathbb{E}_t\left[\prod_{j=1}^{k}(I - \beta B_{k,a})\right] = \left(I - \beta\Sigma_{t,a}\right)^k,$$

we can see that $\widehat{\Sigma}^{+}_{t,a}$ is a good estimator of $\Sigma^{-1}_{t,a}$ on expectation:

$$\mathbb{E}_t\left[\widehat{\Sigma}^{+}_{t,a}\right] = \beta I + \beta\sum_{k=1}^{\infty}\left(I - \beta\Sigma_{t,a}\right)^k = \beta\sum_{k=0}^{\infty}\left(I - \beta\Sigma_{t,a}\right)^k = \beta\left(\beta\Sigma_{t,a}\right)^{-1} = \Sigma^{-1}_{t,a}. \tag{3}$$

As we will see later in the analysis, the bias introduced by setting a finite M can be controlled relatively easily.

Based on the above procedure, we finally define our loss estimator used in this section as

$$\widetilde{\theta}_{t,a} = \widehat{\Sigma}^{+}_{t,a} X_t \ell(X_t, A_t)\mathbb{I}_{\{A_t=a\}}. \tag{4}$$

Via a careful incremental implementation, the estimator can be computed in $O(MKd)$ time and M calls to the oracle generating samples from the context distribution $\mathcal{D}$. We present the details of this efficient computation procedure in Appendix D. We will refer to the version of LINEXP3 using the estimates above as REALLINEXP3, alluding to its favorable guarantees obtained for realizable linear losses. Our main result in this section is the following guarantee regarding the performance of REALLINEXP3:

Theorem 2 *For $\gamma \in (0, 1)$, $M \geq 0$, any positive $\eta \leq \frac{2}{M+1}$ and any positive $\beta \leq \frac{1}{2\sigma^2}$, the expected regret of* REALLINEXP3 *satisfies*

$$R_T \leq 2T\sigma R \cdot \exp\left(-\frac{\gamma\beta\lambda_{\min}M}{K}\right) + 2\gamma T + 3\eta KdT + \frac{\log K}{\eta}.$$

Furthermore, letting $\beta = \frac{1}{2\sigma^2}$, $M = \left\lceil \frac{K\sigma^2\log(T\sigma^2R^2)}{\gamma\lambda_{\min}} \right\rceil$, $\gamma = \sqrt{\frac{\log(T\sigma^2R^2)}{T}}$, and $\eta = \sqrt{\frac{\log K}{dKT\log(T\sigma^2R^2)}}$ and supposing that T is large enough so that the above constraints are satisfied, we also have

$$R_T \leq 4\sqrt{T} + \sqrt{dKT\log K}\left(3 + \sqrt{\log(T\sigma^2R^2)}\right).$$

4. Analysis

This section is dedicated to proving our main results, Theorems 1 and 2. We present the analysis in a modular fashion, first proving some general facts about the algorithm template LINEXP3, and then treat the two variants separately in Sections 4.1 and 4.2 that differ in their choice of loss estimator.

The main challenge in the contextual bandit setting is that the comparator term in the regret definition features actions that depend on the observed contexts, which is to be contrasted with the classical multi-armed bandit setting where the comparator strategy always plays a fixed action. The most distinctive element of our analysis is the following lemma that tackles this difficulty by essentially reducing the contextual bandit problem to a set of auxiliary online learning problems defined separately for each context x:

Lemma 3 *Let π^* be any fixed stochastic policy and let X_0 be sample from the context distribution $\mathcal{D}$ independent from $\mathcal{F}_T$. Suppose that $\pi_t \in \mathcal{F}_{t-1}$ and that $\mathbb{E}_t\left[\widehat{\theta}_{t,a}\right] = \theta_{t,a}$ for all t, a. Then,*

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\big(\pi_t(a|X_t) - \pi^*(a|X_t)\big)\left\langle X_t, \theta_{t,a}\right\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\big(\pi_t(a|X_0) - \pi^*(a|X_0)\big)\left\langle X_0, \widehat{\theta}_{t,a}\right\rangle\right]. \tag{5}$$

Proof Fix any t and a. Then, we have

$$\begin{aligned}
&\mathbb{E}_t\left[\big(\pi_t(a|X_0) - \pi^*(a|X_0)\big)\left\langle X_0, \widehat{\theta}_{t,a}\right\rangle\right] = \mathbb{E}_t\left[\mathbb{E}_t\left[\big(\pi_t(a|X_0) - \pi^*(a|X_0)\big)\left\langle X_0, \widehat{\theta}_{t,a}\right\rangle \middle| X_0\right]\right] \\
&\quad = \mathbb{E}_t\left[\mathbb{E}_t\left[\big(\pi_t(a|X_0) - \pi^*(a|X_0)\big)\left\langle X_0, \theta_{t,a}\right\rangle \middle| X_0\right]\right] = \mathbb{E}_t\left[\big(\pi_t(a|X_0) - \pi^*(a|X_0)\big)\left\langle X_0, \theta_{t,a}\right\rangle\right] \\
&\quad = \mathbb{E}_t\left[\big(\pi_t(a|X_t) - \pi^*(a|X_t)\big)\left\langle X_t, \theta_{t,a}\right\rangle\right],
\end{aligned}$$

where the first step uses the tower rule of expectation, the second that $\mathbb{E}_t\left[\widehat{\theta}_{t,a}\middle|X_0\right] = \theta_{t,a}$ that holds due to the independence of $\widehat{\theta}_t$ and θ_t on X_0, the third step is the tower rule again, and the last step uses that X_0 and X_t have the same distribution and both are conditionally independent on θ_t. Summing up for all actions concludes the proof. ■

Notably, the lemma above is not specific to our algorithm LINEXP3 and only uses the properties of the loss estimator. Applying the lemma to the policies π_t produced by LINEXP3 and using *any* comparator π^*, we can notice that the term on the right hand side is the regret R_T of the algorithm. We stress here that the above result is in fact very powerful since it does not assume *anything* (except measurability) about π^*, even allowing it to be non-smooth—we provide a more detailed discussion of this issue in Section 5. In order to interpret the term on the right-hand side of Equation (5), let us consider an auxiliary online learning problem for a fixed x with K actions and losses $\widehat{\ell}_t(x,a) = \left\langle x, \widehat{\theta}_{t,a}\right\rangle$ for each t, a, and consider running a copy of the classic exponential-weights algorithm[2] of Littlestone and Warmuth (1994) fed with these losses. The probability distribution played by this algorithm over the actions a is given as $\pi_t(a|x) \propto \exp\left(-\eta\sum_{s=1}^{t-1}\widehat{\ell}_s(x,a)\right)$, which implies that the regret in the auxiliary game against comparator π^* at x can be written as

$$\widehat{R}_T(x) = \sum_{t=1}^{T}\sum_{a}\big(\pi_t(a|x) - \pi^*(a|x)\big)\left\langle x, \widehat{\theta}_{t,a}\right\rangle.$$

2. For the sake of clarity, we omit the step of mixing in the uniform distribution in this expository discussion.

This brings us to the key observation that the term on the right-hand side of the equality in Lemma 3 is exactly $\mathbb{E}\left[R_T(X_0)\right]$. Thus, our proof strategy will be to prove an almost-sure regret bound for the auxiliary games defined at each x and take expectation of the resulting bounds with respect to the law of X_0, thus achieving a bound on the regret R_T. The following lemma provides the desired bounds for the auxiliary games:

Lemma 4 *Fix any $x \in \mathcal{X}$ and suppose that $\widehat{\theta}_{t,a}$ is such that $\left|\eta\langle x, \widehat{\theta}_{t,a}\rangle\right| < 1$. Then, the regret of* LINEXP3 *in the auxiliary game at x satisfies*

$$\widehat{R}_T(x) \le \frac{\log K}{\eta} + 2\gamma U_T(x) + \eta \sum_{t=1}^{T}\sum_{a=1}^{K} \pi_t(a|x)\langle x, \widehat{\theta}_{t,a}\rangle^2,$$

where $U_T(x) = \sum_{t=1}^{T}\left(\frac{1}{K}\sum_a \langle x, \widehat{\theta}_{t,a}\rangle - \langle x, \widehat{\theta}_{t,\pi^*(x)}\rangle\right)$.

In the above bound, $U_T(x)$ is the regret of the uniform policy, which can be bounded by T for all algorithms on expectation. The proof is a straightforward application of standard ideas from the classical EXP3 analysis due to Auer et al. (2002b), and we include it in Appendix A for completeness.

The lemmas above suggest that all we need to do is to bound the expectation of the second-order terms on the right-hand side, $\mathbb{E}_t\left[\sum_{a=1}^{K}\pi_t(a|X_0)\langle X_0, \widehat{\theta}_{t,a}\rangle^2\right]$. This, however, is not the only challenge due to the fact that the estimators our algorithms use are not necessarily all unbiased. Specifically, supposing that our estimator can be written as $\widehat{\theta}_{t,a} = \widehat{\theta}^*_{t,a} + b_{t,a}$, where $\widehat{\theta}^*_{t,a}$ is such that $\mathbb{E}_t\left[\widehat{\theta}^*_{t,a}\right] = \theta_{t,a}$ and $b_{t,a}$ is a bias term, we can directly deduce the following bound from Lemma 3:

$$R_T \le \mathbb{E}\left[\widehat{R}_T(X_0)\right] + 2\sum_{t=1}^{T}\max_a \left|\mathbb{E}\left[\langle X_t, b_{t,a}\rangle\right]\right|. \tag{6}$$

The rest of the section is dedicated to finding the upper bounds on the bias term above and on the expectation of the second-order term discussed right before for both estimators (2) and (4), therefore completing the proofs of our main results, Theorems 1 and 2.

4.1. Proof of Theorem 1

We first consider ROBUSTLINEXP3 which uses the estimator $\widehat{\theta}_{t,a}$ defined in Equation (2). While we have already shown in Section 3.1 that the estimator is unbiased, we now consider the case where the true loss function may be nonlinear and can be written as $\ell_t(x,a) = \langle x, \theta_{t,a}\rangle + \varepsilon_t(x,a)$ for some nonlinear function ε_t uniformly bounded on $\mathcal{X}$ by ε. Then, we can see that our estimator satisfies

$$\begin{aligned}\mathbb{E}_t\left[\langle X_0, \widehat{\theta}_{t,a}\rangle\right] &= \mathbb{E}_t\left[\frac{\mathbb{I}_{\{A_t=a\}}}{\pi(a|X_t)} X_0^\intercal \Sigma^{-1} X_t\big(\langle X_t, \theta_{t,a}\rangle + \varepsilon_t(X_t,a)\big)\right] \\ &= \mathbb{E}_t\left[\langle X_0, \theta_{t,a}\rangle\right] + \mathbb{E}_t\left[X_0^\intercal \Sigma^{-1} X_t \varepsilon_t(X_t,a)\right],\end{aligned}$$

and thus the bias can be bounded using the Cauchy–Schwarz inequality as

$$\left|\mathbb{E}_t\left[X_0^\intercal \Sigma^{-1} X_t \varepsilon_t(X_t,a)\right]\right| \le \sqrt{\mathbb{E}_t\left[\operatorname{tr}\left(X_0 X_0^\intercal \Sigma^{-1} X_t X_t^\intercal \Sigma^{-1}\right)\right]} \cdot \sqrt{\mathbb{E}_t\left[\left(\varepsilon_t(X_t,a)\right)^2\right]} \le \sqrt{d}\varepsilon. \tag{7}$$

Here, we used $\mathbb{E}_t\left[X_0 X_0^\intercal X_t X_t^\intercal\right] = \Sigma^2$, which follows from the conditional independence of X_0 and X_t and the definition of Σ, and the boundedness of ε_t in the last step. The other key component of the proof is the following bound:

$$\begin{aligned}\mathbb{E}_t\left[\sum_{a=1}^{K}\pi_t(a|X_0)\left\langle X_0,\widehat{\theta}_{t,a}\right\rangle^2\right] &= \mathbb{E}_t\left[\sum_{a=1}^{K}\pi_t(a|X_0)\frac{\mathbb{I}_{\{A_t=a\}}\ell_t(X_t,a)^2}{\pi_t^2(a|X_t)}X_0^\intercal\Sigma^{-1}X_tX_t^\intercal\Sigma^{-1}X_0\right] \\ &\le \mathbb{E}_t\left[\sum_{a=1}^{K}\pi_t(a|X_0)\cdot\frac{K}{\gamma}\cdot\frac{\mathbb{I}_{\{A_t=a\}}}{\pi_t(a|X_t)}\cdot\operatorname{tr}\left(\Sigma^{-1}X_tX_t^T\Sigma^{-1}X_0X_0^\intercal\right)\right] \le \frac{Kd}{\gamma}\end{aligned} \tag{8}$$

where we used $\pi_t(a|X_t) \ge \frac{\gamma}{K}$ in the first inequality and the conditional independence of X_t and X_0 in the last step. The problem we are left with is to prove that $\eta\left|\left\langle X_0,\widehat{\theta}_{t,a}\right\rangle\right| \le 1$:

$$\left|\left\langle X_0,\widehat{\theta}_{t,a}\right\rangle\right| = \frac{\mathbb{I}_{\{A_t=a\}}}{\pi_t(a|X_t)}\left|X_0^\intercal\Sigma^{-1}X_t\right|\ell_t(X_t,A_t) \le \frac{K\sigma^2}{\gamma\lambda_{\min}},$$

where we used the conditions $\pi_t(a|X_t) \ge \frac{\gamma}{K}$ and $|\ell_t(x,a)| \le 1$ and the Cauchy–Schwarz inequality to show $\left|X_0^\intercal\Sigma^{-1}X_t\right| \le \frac{\sigma^2}{\lambda_{\min}}$. Having satisfied its condition, we may now appeal to Lemma 4, and the proof is concluded by combining and Equations (6), (7), and (8).

4.2. Proof of Theorem 2

We now turn to analyzing REALLINEXP3 which uses the slightly more complicated loss estimator $\widetilde{\theta}_{t,a}$ defined to the MGR procedure. Although we have already seen in Section 3.2 that MGR could result in an unbiased estimate if we could set $M = \infty$. However, in order to keep computation at bay, we need to set M to be a finite (and hopefully relatively small) value. Following the same steps as in Equation (3), we can show

$$\mathbb{E}_t\left[\widehat{\Sigma}_{t,a}^{+}\right] = \beta\sum_{k=0}^{M}(I-\beta\Sigma_{t,a})^k = \Sigma_{t,a}^{-1} - (I-\beta\Sigma_{t,a})^M\Sigma_{t,a}^{-1}.$$

Combining this insight with the definition of $\widetilde{\theta}_{t,a}$ and using some properties of our algorithm, we can prove the following useful bound on the bias of the estimator:

Lemma 5 *Suppose that* $M \ge \frac{K\sigma^2\log T}{\gamma\lambda_{\min}}$, $\beta = \frac{1}{2\sigma^2}$. *Then,* $\left|\mathbb{E}_t\left[\left\langle X_t,\theta_{t,a}-\widetilde{\theta}_{t,a}\right\rangle\right]\right| \le \frac{\sigma R}{\sqrt{T}}$.

Proof We first observe that the bias of $\widetilde{\theta}_{t,a}$ can be easily expressed as

$$\begin{aligned}\mathbb{E}_t\left[\widetilde{\theta}_{t,a}\right] &= \mathbb{E}_t\left[\widehat{\Sigma}_{t,a}^{+}X_tX_t^\intercal\theta_{t,a}\mathbb{I}_{\{A_t=a\}}\right] = \mathbb{E}_t\left[\widehat{\Sigma}_{t,a}^{+}\right]\mathbb{E}_t\left[X_tX_t^\intercal\mathbb{I}_{\{A_t=a\}}\right]\theta_{t,a} = \mathbb{E}_t\left[\widehat{\Sigma}_{t,a}^{+}\right]\Sigma_{t,a}\theta_{t,a} \\ &= \theta_{t,a} - (I-\beta\Sigma_{t,a})^M\theta_{t,a},\end{aligned}$$

where we have used our expression for $\mathbb{E}_t\left[\widehat{\Sigma}_{t,a}^{+}\right]$ derived above. Thus, the bias is bounded as

$$\left|\mathbb{E}_t\left[X_t^\intercal(I-\beta\Sigma_{t,a})^M\theta_{t,a}\right]\right| \le \|X_t\|_2\cdot\|\theta_{t,a}\|_2\left\|(I-\beta\Sigma_{t,a})^M\right\|_{\text{op}}.$$

In order to bound the last factor above, observe that $\Sigma_{t,a} \succcurlyeq \frac{\gamma}{K}\Sigma$ due to the uniform exploration used by LINEXP3, which implies that

$$\left\|(I - \beta\Sigma_{t,a})^M\right\|_{\text{op}} \leq \left(1 - \frac{\gamma\beta\lambda_{\min}}{K}\right)^M \leq \exp\left(-\frac{\gamma\beta}{K}\lambda_{\min}M\right) \leq \frac{1}{\sqrt{T}},$$

where the second inequality uses $1 - z \leq e^{-z}$ that holds for all z, and the last step uses our condition on M. This concludes the proof. ■

The other key term in the regret bound is bounded in the following lemma:

Lemma 6 *Suppose that X_t is satisfying $\|X_t\|_2 \leq \sigma$, $0 < \beta \leq \frac{1}{2\sigma^2}$ and $M > 0$. Then for each t,* REALLINEXP3 *guarantees*

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

Unfortunately, the proof of this statement is rather tedious, so we have to relegate it to Appendix B. As a final step, we need to verify that the condition of Lemma 4 is satisfied, that is, that $\eta\left|\left\langle X_0, \widetilde{\theta}_{t,a}\right\rangle\right| < 1$ holds. To this end, notice that

$$\begin{aligned}\eta \cdot \left|\left\langle X_0, \widetilde{\theta}_{t,a}\right\rangle\right| &= \eta \cdot \left|X_0^{\intercal}\widehat{\Sigma}_{t,a}^{+}X_t \left\langle X_t, \theta_{t,a}\right\rangle \mathbb{I}_{\{A_t=a\}}\right| \leq \eta \cdot \left|X_0^{\intercal}\widehat{\Sigma}_{t,a}^{+}X_t\right| \\ &\leq \eta\sigma^2\left\|\widehat{\Sigma}_{t,a}^{+}\right\|_{\text{op}} \leq \eta\sigma^2\beta\left(1 + \sum_{k=1}^{M}\|A_{k,a}\|_{\text{op}}\right) \leq \eta(M+1)/2,\end{aligned}$$

where we used the fact that our choice of β ensures that $\|A_{k,a}\|_{\text{op}} = \left\|\prod_{j=0}^{k}(I - \beta B_{j,a})\right\|_{\text{op}} \leq 1$. Thus, the condition $\eta \leq 2/(M+1)$ allows us to use Lemma 4, so we can conclude the proof of Theorem 2 by applying Lemma 5, Lemma 6 and the bound of Equation (6).

5. Discussion

Our work is the first to address the natural adversarial variant of the widely popular framework of linear contextual bandits, thus filling an important gap in the literature. Our algorithm REALLINEXP3 achieves the optimal regret bound of of order $\sqrt{KdT}$ and runs in time polynomial in the relevant problem parameters. To our knowledge, REALLINEXP3 is the first computationally efficient algorithm to achieve near-optimal regret bounds in an adversarial contextual bandit problem, and is among the first ones to achieve any regret guarantees at all for an infinite set of policies (besides results on learning with surrogate losses, cf. Foster and Krishnamurthy, 2018). In the case of misspecified loss functions, our algorithm ROBUSTLINEXP3 achieves a regret guarantee of order $(Kd)^{1/3}T^{2/3} + \varepsilon\sqrt{d}T$. Whether or not the overhead of $\varepsilon\sqrt{d}T$ can be improved is presently unclear: while Lattimore et al. (2020) proved that the dependence on $\sqrt{d}$ is inevitable even in the stochastic linear bandit setting when K is large (say, order of T), the very recent work of Foster and Rakhlin (2020) shows that the overhead can be reduced to $\varepsilon\sqrt{K}T$ in the same setting. These results together suggest that the regret bound $\sqrt{KdT} + \varepsilon\sqrt{\min\{K,d\}}T$ is achievable in for stochastic linear contextual bandits. Whether such guarantees can be achieved in the more challenging adversarial setting we considered in this paper remains an interesting open problem.

The reader may be curious if it is possible to remove the i.i.d. assumption that we make about the contexts. Unfortunately, it can be easily shown that no learning algorithm can achieve sublinear regret if the contexts and losses are both allowed to be chosen by an adversary. To see this, we observe that one can embed the problem of online learning of one-dimensional threshold classifiers into our setting, which is known to be impossible to learn with sublinear regret (Ben-David et al., 2009; Syrgkanis et al., 2016a). While one can conceive other assumptions on the contexts that make the problem tractable, such as assuming that the entire sequence of contexts is known ahead of time (the so-called *transductive setting* studied by Syrgkanis et al., 2016a), such assumptions may end up being a lot more artificial than our natural i.i.d. condition. In addition, it is unclear what the best achievable performance bounds in such alternative frameworks actually are. In contrast, the regret bounds we prove for REALLINEXP3 are essentially minimax optimal.

Our algorithm design and analysis introduces a couple of new techniques that could be of more general interest. First, a key element in our analysis is introducing a set of auxiliary bandit problems for each context x and relating the regrets in these problems to the expected regret in the contextual bandit problem (Lemma 3). While this lemma is stated in terms of linear losses, it can be easily seen to hold for general losses as long as one can construct unbiased estimates of the entire loss function. In this view, our algorithms can be seen as the first instances of a new family of contextual bandit methods that are based on estimating the loss functions rather than working with a policy class. An immediate extension of our approach is to assume that the loss functions belong to a reproducing kernel Hilbert space and define suitable kernel-based estimators analogously to our estimators—a widely considered setting in the literature on stochastic contextual bandits (Srinivas et al., 2010; Bubeck et al., 2017; Calandriello et al., 2019). We also remark that our technique used to prove Lemma 3 is similar in nature to the reduction of stochastic sleeping bandit problems to static bandit problems used by Kanade et al. (2009); Neu and Valko (2014).

A second potentially interesting algorithmic trick we introduce is the Matrix Geometric Resampling for estimating inverse covariance matrices. While such matrices are broadly used for loss estimation in the literature on adversarial linear bandits (McMahan and Blum, 2004; Awerbuch and Kleinberg, 2004; Dani et al., 2008; Audibert et al., 2014), the complexity of computing them never seems to be discussed in the literature. Our MGR method provides a viable option for tackling this problem. For the curious reader, we remark that the relation between the iterations defining MGR and the dynamics of gradient descent for linear least-squares estimation is well-known in the stochastic optimization literature, where SGD is known to implement a spectral filter function approximating the inverse covariance matrix (Robbins and Monro, 1951; Györfi and Walk, 1996; Bach and Moulines, 2013; Neu and Rosasco, 2018).

Besides the most important question of whether or not our guarantees for the misspecified setting can be improved, we leave a few more questions open for further investigation. One limitation of our methods is that they require prior knowledge of the context distribution $\mathcal{D}$. We conjecture that it may be possible to overcome this limitation by designing slightly more sophisticated algorithms that estimate this distribution from data. Second, it appears to be an interesting challenge to prove versions of our performance guarantees that hold with high probability by using optimistically estimators as done by Beygelzimer et al. (2011); Neu (2015b), or if data-dependent bounds depending on the total loss of the best expert rather than T can be achieved in our setting (Agarwal et al., 2017; Allen-Zhu et al., 2018; Neu, 2015a). We find it likely that such improvements are possible at the expense of a significantly more involved analysis.

Acknowledgments

We thank the three anonymous reviewers for their valuable feedback that helped us improve the paper. G. Neu was supported by "la Caixa" Banking Foundation through the Junior Leader Postdoctoral Fellowship Programme, a Google Faculty Research Award, and a Bosch AI Young Researcher Award.

References

Y. Abbasi-Yadkori, D. Pál, and Cs. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320. 2011.

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

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646, 2014.

A. Agarwal, A. Krishnamurthy, J. Langford, H. Luo, and S. R. E. Open problem: First-order regret bounds for contextual bandits. In *Proceedings of the 30th Conference on Learning Theory*, pages 4–7, 2017.

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

Z. Allen-Zhu, S. Bubeck, and Y. Li. Make the minority great again: First-order regret bound for contextual bandits. In *Proceedings of the 35th International Conference on Machine Learning*, pages 186–194, 2018.

J.-Y. Audibert, S. Bubeck, and G. Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39:31–45, 2014.

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

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning Journal*, 47(2-3):235–256, 2002a.

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

B. Awerbuch and R. D. Kleinberg. Adaptive routing with end-to-end feedback: distributed learning and geometric approaches. In *STOC 2004*, pages 45–53, 2004. ISBN 1-58113-852-0.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems 26*, pages 773–781, 2013.

S. Ben-David, D. Pál, and S. Shalev-Shwartz. Agnostic online learning. In *COLT*, 2009.

A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *AISTATS 2011*, pages 19–26, 2011.

A. Beygelzimer, F. Orabona, and C. Zhang. Efficient online bandit multiclass learning with $\widetilde{O}(\sqrt{T})$ regret. In *International Conference on Machine Learning*, pages 488–497, 2017.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities:A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

S. Bubeck and N. Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Now Publishers Inc, 2012.

S. Bubeck, Y. T. Lee, and R. Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 72–85, 2017.

D. Calandriello, L. Carratino, A. Lazaric, M. Valko, and L. Rosasco. Gaussian process optimization with adaptive sketching: Scalable and no regret. In *Conference on Learning Theory*, pages 533–557, 2019.

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

W. C. Cheung, D. Simchi-Levi, and R. Zhu. Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1079–1087, 2019.

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

V. Dani, T. Hayes, and S. Kakade. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, volume 20, pages 345–352, 2008.

S. S. Du, S. M. Kakade, R. Wang, and L. F. Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

M. Dudík, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 169–178, 2011.

S. Filippi, O. Cappé, A. Garivier, and Cs. Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

D. J. Foster and A. Krishnamurthy. Contextual bandits with surrogate losses: Margin bounds and efficient algorithms. In *Advances in Neural Information Processing Systems*, pages 2621–2632, 2018.

D. J. Foster and A. Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, 2020.

D. J. Foster, A. Krishnamurthy, and H. Luo. Model selection for contextual bandits. In *Advances in Neural Information Processing Systems*, pages 14714–14725, 2019.

L. Györfi and H. Walk. On the averaged stochastic approximation for linear regression. *SIAM Journal on Control and Optimization*, 34(1):31–61, 1996.

S. M. Kakade, S. Shalev-Shwartz, and A. Tewari. Efficient bandit algorithms for online multiclass prediction. In *International Conference on Machine Learning*, pages 440–447. ACM, 2008.

V. Kanade, H. B. McMahan, and B. Bryan. Sleeping experts and bandits with stochastic action availability and adversarial rewards. In *AISTATS 2009*, pages 272–279, 2009.

B. Kim and A. Tewari. Near-optimal oracle-efficient algorithms for stationary and non-stationary stochastic linear bandits. *arXiv preprint arXiv:1912.05695*, 2019.

T. Lattimore and Cs. Szepesvári. The end of optimism? An asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737, 2017.

T. Lattimore and Cs. Szepesvári. Bandit algorithms. *book draft*, 2019.

T. Lattimore, Cs. Szepesvári, and G. Weisz. Learning with good feature representations in bandits and in RL with a generative model. In *International Conference on Machine Learning*, 2020.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

N. Littlestone and M. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

H. B. McMahan and A. Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *COLT 2004*, pages 109–123, 2004.

G. Neu. First-order regret bounds for combinatorial semi-bandits. In *Conference on Learning Theory*, pages 1360–1375, 2015a.

G. Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 3150–3158, 2015b.

G. Neu and G. Bartók. An efficient algorithm for learning with semi-bandit feedback. In *Algorithmic Learning Theory*, pages 234–248, 2013.

G. Neu and G. Bartók. Importance weighting without importance weights: An efficient algorithm for combinatorial semi-bandits. *Journal of Machine Learning Research*, 17:1–21, 2016.

G. Neu and L. Rosasco. Iterate averaging as regularization for stochastic gradient descent. In *Proceedings of the 31st Conference On Learning Theory*, pages 3222–3242, 2018.

G. Neu and M. Valko. Online combinatorial optimization with stochastic decision sets and adversarial losses. In *Advances in Neural Information Processing Systems*, pages 2780–2788, 2014.

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

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

P. Rusmevichientong and J. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35:395–411, 2010.

Y. Russac, C. Vernade, and O. Cappé. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, pages 12017–12026, 2019.

N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, pages 1015–1022, 2010.

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

V. Syrgkanis, H. Luo, A. Krishnamurthy, and R. E. Schapire. Improved regret bounds for oracle-based adversarial contextual bandits. In *Advances in Neural Information Processing Systems*, pages 3135–3143, 2016b.

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

B. Van Roy and S. Dong. Comments on the Du-Kakade-Wang-Yang lower bounds. *arXiv preprint arXiv:1911.07910*, 2019.

Appendix A. Proof of Lemma 4

The proof follows the standard analysis of EXP3 originally due to Auer et al. (2002b). We begin by recalling the notation $w_t(x,a) = \exp\bigl(-\eta \sum_{s=1}^{t-1} \langle x, \widehat{\theta}_{t,a}\rangle\bigr)$ and introducing $W_t(x) = \sum_{a=1}^{K} w_t(x,a)$. The proof is based on analyzing $\log W_{T+1}(x)$, which can be thought of as a potential function in terms of the cumulative losses. We first observe that $\log W_{T+1}(x)$ can be lower-bounded in terms of the cumulative loss:

$$\log\left(\frac{W_{T+1}(x)}{W_1(x)}\right) \geq \log\left(\frac{w_{T+1}(x,\pi^*(x))}{W_1(x)}\right) = -\eta \sum_{t=1}^{T} x^\intercal \widehat{\theta}_{t,\pi^*(x)} - \log K.$$

On the other hand, for any t, we can prove the upper bound

$$\begin{aligned}\log \frac{W_{t+1}(x)}{W_t(x)} &= \log\left(\sum_{a=1}^{K} \frac{w_{t+1}(x,a)}{W_t(x)}\right) = \log\left(\sum_{a=1}^{K} \frac{w_t(x,a) e^{-\eta\langle x, \widehat{\theta}_{t,a}\rangle}}{W_t(x)}\right) \\ &= \log\left(\sum_{i=1}^{K} \frac{\pi_t(a|x) - \gamma/K}{1-\gamma} \cdot e^{-\eta\langle x, \widehat{\theta}_{t,a}\rangle}\right)\end{aligned}$$

$$\overset{(a)}{\leq} \log\left(\sum_{i=1}^{K}\frac{\pi_t(a|x)-\gamma/K}{1-\gamma}\left(1-\eta\langle x,\widehat{\theta}_{t,a}\rangle+\left(\eta\langle x,\widehat{\theta}_{t,a}\rangle\right)^2\right)\right)$$
$$\overset{(b)}{\leq} \sum_{a=1}^{K}\frac{\pi_t(a|x)}{1-\gamma}\left(-\eta\langle x,\widehat{\theta}_{t,a}\rangle+\left(\eta\langle x,\widehat{\theta}_{t,a}\rangle\right)^2\right)+\frac{\eta\gamma}{K(1-\gamma)}\sum_{a}\langle x,\widehat{\theta}_{t,a}\rangle,$$

where in step (a) we used the inequality $e^{-z}\leq 1-z+z^2$, which holds for $z\geq -1$, and in step (b) we used the inequality $\log(1+z)\leq z$ that holds for any z. Noticing that $\sum_{t=1}^{T}\log\frac{W_{t+1}}{W_t}=\log\frac{W_{T+1}}{W_1}$, we can sum both sides of the above inequality for all $t=1,\ldots,T$ and compare with the lower bound to get

$$-\eta\sum_{t=1}^{T}x^{\intercal}\widehat{\theta}_{t,\pi^*(x)}-\ln K\leq\sum_{t=1}^{T}\sum_{a=1}^{K}\frac{\pi_t(a|x)}{1-\gamma}\left(-\eta\langle x,\widehat{\theta}_{t,a}\rangle+\left(\eta\langle x,\widehat{\theta}_{t,a}\rangle\right)^2\right)+\frac{\eta\gamma\sum_a\langle x,\widehat{\theta}_{t,a}\rangle}{K(1-\gamma)}.$$

Reordering and multiplying both sides by $\frac{1-\gamma}{\eta}$ gives

$$\sum_{t=1}^{T}\left(\sum_{a=1}^{K}\pi_t(a|x)\langle x,\widehat{\theta}_{t,a}\rangle-\langle x,\widehat{\theta}_{t,\pi^*(x)}\rangle\right)$$
$$\leq\frac{(1-\gamma)\log K}{\eta}+\eta\sum_{t=1}^{T}\sum_{a=1}^{K}\left(\langle x,\widehat{\theta}_{t,a}\rangle\right)^2+\gamma\sum_{t=1}^{T}\left(\frac{1}{K}\sum_{a}\langle x,\widehat{\theta}_{t,a}\rangle-\langle x,\widehat{\theta}_{t,\pi^*(x)}\rangle\right).$$

This concludes the proof. ■

Appendix B. Proof of Lemma 6

The proof relies on a series of matrix operations, and makes repeated use of the following identity that holds for any symmetric positive definite matrix S:

$$\sum_{k=0}^{M}(I-S)^k=S^{-1}-(I-S)^MS^{-1}.$$

We start by plugging in the definition of $\widetilde{\theta}_{t,a}$ and writing

$$\mathbb{E}_t\left[\sum_{a=1}^{K}\pi_t(a|X_0)\langle X_0,\widetilde{\theta}_{t,a}\rangle^2\right]=\mathbb{E}_t\left[\sum_{a=1}^{K}\pi_t(a|X_0)\left(X_0^{\intercal}\Sigma_{t,a}^{+}X_tX_t^{\intercal}\theta_{t,a}\mathbb{I}_{\{A_t=a\}}\right)^2\right]$$
$$\leq\mathbb{E}_t\left[\mathbb{E}\left[\sum_{a=1}^{K}\mathrm{tr}\left(\pi_t(a|X_0)X_0X_0^{\intercal}\Sigma_{t,a}^{+}X_tX_t^{\intercal}\Sigma_{t,a}^{+}\mathbb{I}_{\{A_t=a\}}\right)\middle|X_0\right]\right]$$
$$=\sum_{a=1}^{K}\mathbb{E}_t\left[\mathrm{tr}\left(\Sigma_{t,a}\Sigma_{t,a}^{+}\Sigma_{t,a}\Sigma_{t,a}^{+}\right)\right],$$

where we used $\left\langle X_0, \theta_{t,a} \right\rangle \leq 1$ in the inequality and observed that $\Sigma_{t,a} = \mathbb{E}_t\left[\pi_t(a|X_0) X_0 X_0^\intercal\right]$. In what follows, we suppress the t, a indexes to enhance readability. Using the definition of Σ^+ and elementary manipulations, we can get

$$\begin{aligned}\mathbb{E}\left[\mathrm{tr}\left(\Sigma\Sigma^+\Sigma\Sigma^+\right)\right] &= \mathbb{E}\left[\beta^2 \cdot \mathrm{tr}\left(\Sigma\left(\sum_{k=0}^{M} A_k\right)\Sigma\left(\sum_{j=0}^{M} A_j\right)\right)\right] \\ &= \beta^2 \sum_{k=0}^{M}\sum_{j=0}^{M} \mathrm{tr}\left(\mathbb{E}\left[\Sigma A_k \Sigma A_j\right]\right) = \beta^2 \sum_{k=0}^{M} \mathrm{tr}\left(\mathbb{E}\left[\Sigma A_k \Sigma A_k\right]\right) + 2\beta^2 \sum_{k=0}^{M}\sum_{j=k+1}^{M} \mathrm{tr}\left(\mathbb{E}\left[\Sigma A_k \Sigma A_j\right]\right).\end{aligned}$$

Let us first address the first term on the right hand side. To this end, consider any symmetric positive definite matrix H that commutes with Σ and observe that

$$\begin{aligned}\mathbb{E}\left[(I-\beta B_k)H(I-\beta B_k)\right] &= \mathbb{E}\left[(I-\beta X(k)X(k)^\intercal)H(I-\beta X(k)X(k)^\intercal)\right] \\ &= H - \beta\mathbb{E}\left[X(k)X(k)^\intercal H\right] - \beta\mathbb{E}\left[HX(k)X(k)^\intercal\right] + \beta^2\mathbb{E}\left[X(k)X(k)^\intercal H X(k)X(k)^\intercal\right] \\ &\preccurlyeq H - 2\beta H\Sigma + \beta^2\sigma^2 H\Sigma = H\left(I - \beta(2-\beta\sigma^2)\Sigma\right),\end{aligned}$$

where we used our assumption that $\|X(k)\| \leq \sigma$ which implies $\mathbb{E}\left[\|X(k)\|_2^2 X(k)X(k)^\intercal\right] \preccurlyeq \sigma^2\Sigma$. Now, recalling the definition $A_k = \prod_{j=1}^{k} B_j$ and using the above relation repeatedly, we can obtain

$$\begin{aligned}\mathrm{tr}\left(\mathbb{E}\left[\Sigma A_k \Sigma A_k\right]\right) &= \mathrm{tr}\left(\mathbb{E}\left[\Sigma A_{k-1}\mathbb{E}\left[(I-\beta B_k)\Sigma(I-\beta B_k)\right]A_{k-1}\right]\right) \\ &\leq \mathrm{tr}\left(\mathbb{E}\left[\Sigma A_{k-1}\Sigma\left(I-\beta(2-\beta\sigma^2)\Sigma\right)A_{k-1}\right]\right) \\ &\leq \ldots \leq \mathrm{tr}\left(\Sigma^2(I-\beta(2-\beta\sigma^2)\Sigma)^k\right).\end{aligned} \tag{9}$$

Thus, we can see that

$$\begin{aligned}\beta^2\sum_{k=0}^{M}\mathrm{tr}\left(\mathbb{E}\left[\Sigma A_k\Sigma A_k\right]\right) &= \beta^2\sum_{k=0}^{M}\mathrm{tr}\left(\Sigma^2(I-\beta(2-\beta\sigma^2)\Sigma)^k\right) \\ &= \frac{\beta^2}{\beta(2-\beta\sigma^2)}\mathrm{tr}\left(\Sigma^2\Sigma^{-1}\left(I-(I-\beta(2-\beta\sigma^2)\Sigma)^M\right)\right) \leq \frac{\beta\mathrm{tr}\left(\Sigma\right)}{2-\beta\sigma^2} \leq \frac{2\beta\mathrm{tr}\left(\Sigma\right)}{3},\end{aligned}$$

where we used the condition $\beta \leq \frac{1}{2\sigma^2}$ and the fact that $(I-\beta(2-\beta\sigma^2)\Sigma)^M \succcurlyeq 0$ by the same condition. We can finally observe that our assumption on the contexts implies $\mathrm{tr}\left(\Sigma\right) \leq \mathrm{tr}\left(\sigma^2 I\right) = \sigma^2 d$, so again by our condition on β we have $\beta\mathrm{tr}\left(\Sigma\right) \leq \frac{d}{2}$, and the first term is bounded by $\frac{d}{3}$.

Moving on to the second term, we first note that for any $j > k$, the conditional expectation of B_j given $B_{\leq k} = (B_1, B_2, \ldots B_k)$ satisfies $\mathbb{E}\left[A_j | B_{\leq k}\right] = A_k(I-\beta\Sigma)^{j-k}$ due to conditional independence of all B_i given B_k, for $i > k$. We make use of this equality by writing

$$\begin{aligned}\beta^2\sum_{k=0}^{M}\sum_{j=k+1}^{M}\mathbb{E}\left[\mathrm{tr}\left(\Sigma A_k\Sigma A_j\right)\right] &= \beta^2\sum_{k=0}^{M}\mathbb{E}\left[\mathbb{E}\left[\sum_{j=k+1}^{M}\mathrm{tr}\left(\Sigma A_k\Sigma A_j\right)\middle| B_{\leq k}\right]\right] \\ &= \beta^2\sum_{k=0}^{M}\mathbb{E}\left[\mathbb{E}\left[\sum_{j=k+1}^{M}\mathrm{tr}\left(\Sigma A_k\Sigma A_k(I-\beta\Sigma)^{j-k}\right)\middle| B_{\leq k}\right]\right]\end{aligned}$$

$$
\begin{aligned}
&= \beta \sum_{k=0}^{M} \mathbb{E}\left[\mathbb{E}\left[\operatorname{tr}\left(\Sigma A_k \Sigma A_k \Sigma^{-1}\left(I - (I-\beta\Sigma)^{M-k}\right)\right)\middle| B_{\le k}\right]\right] \\
&\le \beta \sum_{k=0}^{M} \mathbb{E}\left[\mathbb{E}\left[\operatorname{tr}\left(\Sigma A_k \Sigma A_k \Sigma^{-1}\right)\middle| B_{\le k}\right]\right] \\
&\qquad\qquad \text{(due to } (I-\beta\Sigma)^{M-k} \succcurlyeq 0\text{)} \\
&\le \beta \sum_{k=0}^{M} \operatorname{tr}\left(\Sigma^2 (I-\beta(2-\beta\sigma^2)\Sigma)^k \Sigma^{-1}\right) \\
&\qquad\qquad \text{(by the same argument as in Equation (9))} \\
&\le \frac{1}{(2-\beta\sigma^2)} \operatorname{tr}\left(\Sigma^2 \Sigma^{-1}\left(I - (I-\beta(2-\beta\sigma^2)\Sigma)^M \Sigma^{-1}\right)\right) \\
&\le \operatorname{tr}\left(\Sigma^2 \Sigma^{-1} \Sigma^{-1}\right) \le d,
\end{aligned}
$$

where the last line again used the condition $\beta \le \frac{1}{2\sigma^2}$ and $(I-\beta(2-\beta\sigma^2)\Sigma)^M \succcurlyeq 0$. The proof of the theorem is concluded by putting everything together. ■

Appendix C. Algorithms for contextual learning with full information

Clearly, our algorithm LINEXP3 can be simply adapted to simpler settings where the learner gets more feedback about the loss functions ℓ_t chosen by the adversary. In this section, we show results for two such natural settings: one where the learner observes the *entire* loss function ℓ_t, and one where the learner observes the losses $\ell_t(X_t, a)$ for each action a. We refer to the first of these observation models as *counterfactual feedback* and call the second one *full-information feedback*. We describe two variants of our algorithm for these settings and give their performance guarantees below. Both results will hold for general nonlinear losses taking values in $[0, 1]$.

In case of counterfactual feedback, we can modify our algorithm so that, in each round t, it computes the weights $w_{t,a}(X_t) = \exp\left(-\eta \sum_{k=1}^{t-1} \ell_k(X_t, a)\right)$ for each action, and then plays action $A_t = a$ with probability proportional to the obtained weight. Using our general analytic tools, this algorithm can be easily shown to achieve the following guarantee:

Proposition 7 *For any $\eta > 0$, the regret of the algorithm described above for counterfactual feedback satisfies*

$$R_T \le \frac{\log K}{\eta} + \frac{\eta T}{8}.$$

Setting $\eta = \sqrt{\frac{8 \log K}{T}}$, the regret also satisfies $R_T \le \sqrt{(T/2)\log K}$.

Notably, this bound does not depend at all on the dimension of the context space, the complexity of the policy class, or any property of the loss function, and only shows dependence on the number of actions K. The caveat is of course that the counterfactual model provides the learner with a level of feedback that is entirely unrealistic in any practical setting: it requires the ability to evaluate all past loss functions at *any* context-action pair.

The full-information setting is arguably much more realistic in that it only requires evaluating the losses corresponding to the observed context X_t, which which is typically the case in online

classification problems. For this setting, we use our LINEXP3 algorithm with the loss estimator defined for each action a as

$$\widehat{\ell}_{t,a} = \Sigma^{-1} X_t \ell_t(X_t, a).$$

Using our analysis, we can show that the bias of this estimator is uniformly bounded by $\varepsilon\sqrt{d}$ (cf. Equation 7). The following bound is then easy to prove by following the same steps as in Section 4.1:

Proposition 8 *For any positive $\eta \leq \frac{\lambda_{\min}}{\sigma^2}$, the regret of the algorithm described above for full-information feedback*

$$R_T \leq \frac{\log K}{\eta} + \eta d T + \varepsilon\sqrt{d}T.$$

Setting $\eta = \sqrt{\frac{d \log K}{T}}$, the regret also satisfies $R_T \leq 2\sqrt{dT \log K} + \varepsilon\sqrt{d}T$ for large enough T.

As expected, this bound scales with the dimension as $\sqrt{d}$ due to the fact that the algorithm has to "estimate" d, parameters, as opposed to the Kd parameters that need to be learned in the contextual bandit problem we consider in the main text. We also note that this online learning setting is closely related to that of prediction with expert advice, with the set of experts being the class of linear-classifier policies (Cesa-Bianchi and Lugosi, 2006). As a result, it is possible to make use of this framework by running any online prediction algorithm on a finely discretized set of policies, resulting in a regret bound of order $\sqrt{dT \log(KT)}$. Our result above improves on this by a logarithmic factor of T, while being efficient to implement.

Appendix D. Efficient implementation of MGR

The naïve implementation of the MGR procedure presented in the main text requires $O(MKd + Kd^2)$ time due to the matrix-matrix multiplications involved. In this section we explain how to compute $\widehat{\ell}_t(x, a) = \langle x, \widetilde{\theta}_{t,a} \rangle$ in $O(MKd)$ time, exploiting the fact that the matrices $\widehat{\Sigma}_{t,a}$ never actually need to be computed, since the algorithm only works with products of the form $\widehat{\Sigma}_{t,a} X_t$ for a fixed vector X_t. This motivates the following procedure:

Fast Matrix Geometric Resampling

Input: context vector x, data distribution $\mathcal{D}$, policy π_t.
Initialization: Compute $Y_{0,a} = Ix$.
For $k = 1, \ldots, M$**, repeat**:

1. Draw $X(k) \sim \mathcal{D}$ and $A(k) \sim \pi_t(\cdot|X(k))$,
2. if $a = A(k)$, set
   $Y_{k,a} = Y_{k-1,a} - \beta \langle Y_{k-1,a}, X(k) \rangle X(k)$,
3. otherwise, set $Y_{k,a} = Y_{k-1,a}$.

Return $q_{t,a} = \beta Y_{0,a} + \beta \sum_{k=1}^{M} Y_{k,a}$.

It is easy to see from the above procedure that each iteration k can be computed using $(K+1)d$ vector-vector multiplications: sampling each action $A(k)$ takes Kd time due to having to compute the products $\langle X(k), \widehat{\Theta}_{t,a} \rangle$ for each action a, and updating $Y_{k,a}$ can be done by computing the product $\langle Y_{k-1,a}, X(k) \rangle$. Overall, this results in a total runtime of order MKd as promised above.