Adversarial Contextual Bandits Go Kernelized

Gergely Neu
Universitat Pompeu Fabra
gergely.neu@gmail.com

Julia Olkhovskaya
TU Delft
julia.olkhovskaya@gmail.com

Sattar Vakili
MediaTek Research
sattar.vakili@mtkresearch.com

Abstract

We study a generalization of the problem of online learning in adversarial linear contextual bandits by incorporating loss functions that belong to a reproducing kernel Hilbert space, which allows for a more flexible modeling of complex decision-making scenarios. We propose a computationally efficient algorithm that makes use of a new optimistically biased estimator for the loss functions and achieves near-optimal regret guarantees under a variety of eigenvalue decay assumptions made on the underlying kernel. Specifically, under the assumption of polynomial eigendecay with exponent $c > 1$, the regret is $\widetilde{\mathcal{O}}(KT^{\frac{1}{2}\left(1+\frac{1}{c}\right)})$, where T denotes the number of rounds and K the number of actions. Furthermore, when the eigendecay follows an exponential pattern, we achieve an even tighter regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$. These rates match the lower bounds in all special cases where lower bounds are known at all, and match the best known upper bounds available for the more well-studied stochastic counterpart of our problem.

1 Introduction

In the domain of sequential decision-making, the framework of contextual bandits has emerged as an important tool for modeling interactions between a learner and environment in a sequence of rounds. Within each such round, the learner observes a context and subsequently selects an action and incurs a loss. The objective of the learner in this iterative process is to minimize her cumulative losses over a sequence of rounds. This model has been employed in a large variety of applications, including medical treatments [1], the domain of personalized recommendations [2], and online advertising [3].

One of the main challenges of the contextual bandit problem is that the partial observations made about the losses handed out by the environment must be generalized efficiently to a possibly infinite set of contexts that are yet to be encountered in future decision-making rounds. One possible way to address this challenge is by making suitable assumptions about the structure of the losses. One particularly well-studied model is that of *linear contextual bandit*, where the losses are assumed to be linear in some known low-dimensional representation of the contexts. In the most broadly considered version of this setup, the sequence of contexts is completely arbitrary and the losses are determined by fixed linear functions. Advancements in this model have been made in a range of works, including [3, 4, 5, 6].

This model has been successfully generalized to deal with non-linear loss functions that belong to reproducing kernel Hilbert spaces. This assumption is broadly applicable, as the RKHS associated with commonly used kernels has the capacity to approximate nearly all continuous functions on compact subsets of $\mathbb{R}^d$ [7, 8]. Viewed through the lens of kernel maps, this setting represents an extreme extension of the parametric linear bandit setting mentioned above, where the contexts can

17th European Workshop on Reinforcement Learning (ewrl 2024).

be represented in infinite-dimensional vector spaces. Works like [9, 10, 11] have provided efficient algorithms with strong performance guarantees for contextual bandits with such nonlinear loss functions that remain fixed throughout the online learning process.

The primary focus of this paper lies in a distinct model known as the adversarial contextual bandit. In this setup, we assume that the context is drawn from a fixed distribution, and losses are chosen by a potentially adaptive adversary. For this setting, the simplest approach is to make use of a finite class of policies that map contexts to actions, as done by the classic EXP4 algorithm of [12]. An alternative to this line of work takes inspiration from the stochastic linear contextual bandit literature, and models the losses as linear functions of some known finite-dimensional feature map [13, 14].

Our principal contribution is extending the understanding of the adversarial linear contextual bandit model to work with a large class of nonlinear loss functions. To enhance model flexibility, we consider the setting where the sequence of loss functions drawn by the adversary belong to a fixed and known RKHS. Within this framework, we establish a regret bound of $\widetilde{\mathcal{O}}(KT^{1/2(1+1/c)})$ for loss functions characterized by polynomial eigendecay ($\mu_i = \mathcal{O}(i^{-c})$) and a $\widetilde{\mathcal{O}}(K\sqrt{T})$ bound for those exhibiting exponential eigendecay ($\mu_i = \mathcal{O}(e^{-ci})$). These conditions are well-studied in the broader literature on learning with kernels, and in particular our results align with the lower bounds established for kernelized bandits with adversarial losses by [15], and match the best known upper bounds in the stochastic version of our problem by [10].

At a high level, our approach is based on the regret decomposition idea of [13] originally proposed for finite-dimensional linear bandits: we place a suitably chosen online learning algorithm in each context x, and feed each algorithm with a suitably chosen estimator for the loss functions that allows generalization across different contexts. Our key technical contribution is the construction of an optimistically biased loss estimator that can be effectively computed via a kernelized version of the Matrix Geometric Resampling estimator proposed for finite-dimensional linear losses by [13]. The optimistic bias is achieved by adding a context-dependent exploration bonus to the standard estimator, in order to offset its potentially large positive bias that could otherwise be problematic to handle for a standard analysis. Another key component of our algorithm design is the now-classic log-barrier regularization function popularized in the online learning literature by [16]—see also the earlier works of [17, 18, 19, 20, 21] and follow-ups by [22, 23, 24, 25] that made use of the same regularizer. In our case, we use the special property of the log-barrier that it can appropriately handle loss functions in an FTRL scheme that are potentially unbounded (as will be the case with our estimators).

The remainder of this paper is structured as follows. In the next section, we introduce the essential notation and definitions. Section 3 presents our algorithm and provides its performance guarantees. Detailed proofs supporting our analysis can be found in Section 4. We draw our conclusions in Section 5, where we also delve into the implications of our results.

Notation. We let ℓ_2 denote the space of square-summable sequences. For any two elements $v, w \in \ell_2$, we use $\langle w, v\rangle$ to denote the standard ℓ_2 inner product $\sum_{i=1}^{\infty} w_i u_i$, and we define the ℓ_2 norm of v as $\|v\|_2 = \sqrt{\langle v, v\rangle}$. The tensor product of v and w is denoted by $v \otimes w$, and is defined as the operator that acts on elements u of ℓ_2 as $(v \otimes w)\, u = v \langle w, u\rangle$. For a positive definite operator B on ℓ_2, we define $\|v\|_B = \sqrt{\langle v, Bv\rangle}$, and its trace as $\mathrm{tr}\,(B) = \sum_i \|e_i\|_B^2$, where e_i is the ith canonical basis vector in ℓ_2. In the context of sequential-decision making problems, we will use $\mathcal{F}_t$ to denote the interaction history between the learner and the environment, and use the shorthand notations $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_{t-1}]$ and $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot|\mathcal{F}_{t-1}]$.

2 Preliminaries

We now introduce our learning setting and the assumptions that we make about the loss functions.

2.1 Adversarial contextual bandits

We investigate a sequential interaction scheme between a learner and its environment, where the subsequent steps are iteratively executed over a fixed number of rounds $t = 1, 2, \ldots, T$:

1. The environment draws the context vector $X_t \in \mathbb{R}^d$ from the context distribution $\mathcal{D}$, and reveals it to the learner;
2. Independently of the context X_t, the environment chooses a loss function $\ell_t : \mathcal{X} \times \mathcal{A} \to [0, 1]$;
3. Based on X_t and possibly some randomness, the learner chooses action $A_t \in [K]$;
4. The learner incurs and observes loss $\ell_t(X_t, A_t)$.

The primary objective of the learner is to strategically choose actions to minimize its cumulative loss. It is important to note that we refrain from making any statistical assumptions about the sequence of losses. In fact, we allow these losses to depend on the entire historical interaction, making it impractical for the learner to aim for a loss level as low as that of the best sequence of actions. A more realistic goal is to strive to match the performance of the best fixed policy that maps contexts to actions. To formalize this objective, the learner considers the set Π, which contains all policies $\pi : \mathbb{R}^d \to [K]$, and seeks to minimize its total expected regret, which is formally defined as

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

Here, the expectation is taken over the randomness injected by the learner, as well as the sequence of random contexts. It is easy to show that the optimal policy π_T^*, which serves as the benchmark for the learner's performance, is defined by the following rule:

$$\pi_T^*(x) = \arg\min_{a} \sum_{t=1}^{T} \ell_t(x, a), \qquad \forall x \in \mathbb{R}^d. \tag{1}$$

2.2 RKHS loss functions

Throughout the paper, we will make the assumption that the loss functions $\ell_t(\cdot, a)$ belong to a known reproducing kernel Hilbert space (RKHS) for each t, a. Specifically, we will suppose that the space of contexts $\mathcal{X} \subseteq \mathbb{R}^d$, and we are given a positive definite kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We let $\mathcal{H}_\kappa \subseteq \mathbb{R}^{\mathcal{X}}$ be the RKHS induced by κ. Without loss of generality, we assume $\kappa(x, x) \leq 1$ for all $x \in \mathcal{X}$. The inner product and norm of $\mathcal{H}_\kappa$ are represented by $\langle \cdot, \cdot \rangle_{\mathcal{H}_\kappa} : \mathcal{H}_\kappa \times \mathcal{H}_\kappa \to \mathbb{R}$ and $\| \cdot \|_{\mathcal{H}_\kappa} : \mathcal{H}_\kappa \to \mathbb{R}$, respectively. Mercer's theorem implies that, under certain mild conditions, κ can be represented using an infinite-dimensional feature map:

$$\kappa(x, x') = \sum_{j=1}^{\infty} \mu_j \psi_j(x) \psi_j(x'), \tag{2}$$

where $\mu_j \in \mathbb{R}_+$ are the Mercer eigenvalues and $\psi_j \in \mathcal{H}_\kappa$ are the corresponding eigenfunctions, and $\sqrt{\mu_j}\psi_j$ form an orthonormal basis of $\mathcal{H}_\kappa$. Using this basis, any $h \in \mathcal{H}_\kappa$ can be represented through the real-valued square summable sequence $(w_j)_{j=1}^{\infty} \in \ell_2$ as

$$h = \sum_{j=1}^{\infty} w_j \sqrt{\mu_j} \psi_j,$$

where $\|h\|_{\mathcal{H}_\kappa}^2 = \sum_{j=1}^{\infty} w_j^2$. A formal statement and the details can be found in Appendix A. We will use the notation $\varphi_i(x) = \sqrt{\mu_i}\psi_i(x)$ and use $\varphi(x) = (\varphi_i(x))_{i=1}^{\infty} \in \ell_2$ to denote the representation of context x in ℓ_2 induced by Mercer's theorem. An important implication of Mercer's theorem that we will repeatedly use is that $\langle \varphi(x), \varphi(x') \rangle = \kappa(x, x')$ holds for all $x, x' \in \mathcal{X}$.

Attempting to obtain a sublinear regret bound without any assumptions regarding the regularity of the loss function would be an arduous task. In this paper, we will impose such regularity conditions by making assumptions about the Mercer eigenvalues of the kernel κ.

Assumption 1. *We assume that the Mercer eigenvalues $\{\mu_j\}_{j \geq 1}$ of the kernel κ over $\mathcal{X}$ are ordered as $\mu_1 \geq \mu_2 \geq \ldots$, and are such that they meet one of the following two eigenvalue decay profiles for some constants $g > 0, c > 0$:*

- *(g, c)-exponential decay: for all $j \in \mathbb{N}$, we have $\mu_j \leq g e^{-cj}$.*

- (g, c)-*polynomial decay with* $c > 1$*: for all* $j \in \mathbb{N}$*, we have* $\mu_j \leq g j^{-c}$.

As an alternative way to measure the decay rate of the kernel κ, we will also define the following quantity for each $\varepsilon > 0$:

$$m(\varepsilon) = \min \left\{ m \in \mathbb{N} : \textstyle\sum_{j=m+1}^{\infty} \mu_j \leq \varepsilon \right\}.$$

It is easy to see that for kernels that satisfy the exponential decay condition, $m(\varepsilon) = \mathcal{O}\left(\log(g/(c\varepsilon))/c\right)$ and for kernels that satisfy the polynomial decay condition, $m(\varepsilon) = \mathcal{O}\left(((c-1)\varepsilon/g)^{1/(1-c)}\right)$. Many practically used kernels are consistent with Assumption 1. For instance, the squared exponential kernel satisfies the exponential decay condition with $c = 1/d$, and the Matérn kernel with smoothness parameter $v > 2$ satisfies the polynomial decay condition with $c = 1 + 2v/d$. We refer to [26] and the discussion in [27] for proofs of these facts and further examples.

Now we can precisely state our assumptions on the loss functions and the contexts. We will suppose the context distribution is supported on the bounded compact set $\mathcal{X} \subset \mathbb{R}^d$ with each $x \in \mathcal{X}$ satisfying $\|\varphi(x)\|_2 \leq 1$. Furthermore, we will suppose that the loss function satisfies $\ell_t(\cdot, a) \in \mathcal{H}_\kappa$ and in particular that it can be written as $\ell_t(x, a) = \langle f_{t,a}, \varphi(x) \rangle$ for some $f_{t,a} \in \ell_2$ that satisfies $\|f_{t,a}\|_2 \leq 1$ for all t, a.

3 Algorithm and main result

We now present our algorithm which is based on a regret-decomposition approach first proposed by [13] for finite-dimensional linear contextual bandits. The core idea of this method is to instantiate an online learning algorithm in every context $x \in \mathcal{X}$ and feed it with an appropriately designed estimator of the loss function that allows generalization across different contexts. Concretely, we will run an instance of the standard Follow-the-Regularized-Leader (FTRL) algorithm with log-barrier regularization (as popularized in online learning by 16) as the online learning method, and derive a new loss estimator based on the Matrix Geometric Resampling procedure proposed by [13] along with an optimistic exploration idea that is novel within this context. While the algorithm formally needs to calculate its policies and loss estimates that are valid on the whole context-action space, we will show that it can be implemented efficiently by querying the policy and the estimates only in the contexts encountered in runtime. To preserve readability in this section, we present a relatively abstract version of our algorithm first without worrying about implementability, and defer a fully detailed operational description to Appendix C.

We start by describing the algorithm (that we call KERNELFTRL) for a generic choice of loss estimators $\widehat{\ell}_t : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ whose specifics will be described shortly. Letting $\widehat{L}_t(x, a) = \sum_{\tau=1}^{t} \widehat{\ell}_\tau(x, a)$ denote the cumulative sum of the estimated losses, our algorithm calculates its policy $\pi_t : \mathcal{X} \to \Delta_{\mathcal{A}}$ by solving the following optimization problem in each round t:

$$\pi_t(\cdot|X_t) = \operatorname*{arg\,min}_{p \in \Delta(\mathcal{A})} \left(\Psi(p) + \eta \sum_a p_a \widehat{L}_{t-1}(X_t, a) \right).$$

Here, $\eta > 0$ serves as a learning-rate parameter and $\Psi(p) = \sum_{a \in \mathcal{A}} \ln\left(\frac{1}{p_a}\right)$. Note that the algorithm only has to compute the distribution $\pi_t(\cdot|X_t)$ *locally at* X_t which can be done efficiently as long as $\widehat{L}_{t-1}(X_t, a)$ can be efficiently computed for all actions a [16]. We will show later that this condition holds true for our loss estimators. We present this method as Algorithm 1 below.

To describe our loss estimator, we first introduce the following operator on ℓ_2:

$$\Sigma_{t,a} = \mathbb{E}_t \left[\mathbb{I}_{\{A_t = a\}} \left(\varphi(X_t) \otimes \varphi(X_t) \right) \right]. \tag{4}$$

Then, supposing for the sake of argument that $\Sigma_{t,a}$ is invertible (which will not be necessary for our actual estimator), we can define the estimate

$$\widehat{f}_{t,a} = \Sigma_{t,a}^{-1} \varphi(X_t) \ell_t(X_t, a) \mathbb{I}_{\{A_t = a\}},$$

which can be easily demonstrated to be unbiased:

$$\mathbb{E}_t \left[\widehat{f}_{t,a} \right] = \mathbb{E}_t \left[\Sigma_{t,a}^{-1} \varphi(X_t) \ell_t(X_t, a) \mathbb{I}_{\{A_t = a\}} \right] = \mathbb{E}_t \left[\Sigma_{t,a}^{-1} \mathbb{I}_{\{A_t = a\}} \varphi(X_t) \langle \varphi(X_t), f_{t,a} \rangle \right] = f_{t,a}.$$

Algorithm 1 KERNELFTRL

Parameters: Learning rate $\eta > 0$.
Initialization: Set $\mathcal{B}_t = \emptyset$.
For $t = 1, \ldots, T$, **repeat:**

1. Observe X_t and, for all a, set

$$\pi_t(\cdot|X_t) = \operatorname*{arg\,min}_{p \in \Delta(\mathcal{A})} \left(\Psi(p) + \eta \sum_a p_a \widehat{L}_{t-1}(X_t, a) \right), \tag{3}$$

2. draw A_t from the policy $\pi_t(\cdot|X_t)$,
3. observe the loss $\ell_t(X_t, A_t)$ and call Kernel Geometric Resampling to produce $\widehat{\ell}_t$.

Here, we used that $\ell_t(X_t, a) = \langle \varphi(X_t), f_{t,a} \rangle$, and that $\varphi(X_t)\langle \varphi(X_t), f_{t,a} \rangle = (\varphi(X_t) \otimes \varphi(X_t))\, f_{t,a}$ holds by definition of the tensor product. Note that the dimension of the $\mathcal{H}_\kappa$ could be infinite (for example when κ is the Gaussian kernel), so neither $\Sigma_{t,a}$ or $\widehat{f}_{t,a}$ can be computed explicitly. Another challenge is that, even in the case of a fixed dimension, the operator $\Sigma_{t,a}$ relies on the joint distribution of both the context X_t and the action A_t, which exhibits a highly intricate structure. As a final note, the eigenvalues of $\Sigma_{t,a}$ can be arbitrary small, which may result in a loss estimator of unbounded norm $\left\| \widehat{f}_{t,a} \right\|_2$ even in the unlikely case that $\Sigma_{t,a}$ is invertible.

To deal with the difficulties stated above, we propose an estimator derived by adapting the idea of Matrix Geometric Resampling (MGR) from [13] to the kernel setting. The estimator is efficiently computable, but requires sampling access to the context distribution $\mathcal{D}$. To deal with the bias of the standard MGR estimator, we also introduce a new element in our algorithm design: an *optimistic exploration bonus* whose purpose is to make sure that the estimates are negatively biased which we will see to be beneficial to the analysis. The bonus for context-action pair (x, a) added in round t will be denoted by $b_t(x, a)$, and will be computed within the same procedure as the base loss estimates themselves. The procedure (which we call Kernel Geometric Resampling or KGR) is presented below:

Kernel Geometric Resampling

Input: Context x, X_t, policy π_t, data distribution $\mathcal{D}$, parameters β, M.
For $k = 1, \ldots, M$**, repeat**:

1. Draw $X(k) \sim \mathcal{D}$ and $A(k) \sim \pi_t(\cdot|X(k))$,
2. compute $q_{t,k,a}(x) = \langle \varphi(x), C_{k,a}\varphi(X_t) \rangle$ and $b_{t,k,a}(x) = \beta \langle \varphi(x), C_{k,a}\varphi(x) \rangle$,
   where $C_{k,a} = \prod_{j=1}^{k} (I - B_{j,a})$
   and $B_{k,a} = \mathbb{I}_{\{A(k)=a\}} \varphi(X(k)) \otimes \varphi(X(k))$.

Return $q_t(x, a) = \kappa(x, X_t) + \sum_{k=1}^{M} q_{t,k,a}(x)$,
$b_t(x, a) = \kappa(x, x) + \sum_{k=1}^{M} b_{t,k,a}(x)$.

Then, the estimator of $\ell_t(x, a)$ can be written as

$$\widehat{\ell}_t(x, a) = q_t(x, a)\ell_t(X_t, a)\mathbb{I}_{\{A_t=a\}} - b_t(x, a).$$

Notice that all operations performed by Kernel Geometric Resampling can be implemented by applying simple rank-one operators to elements of ℓ_2, so $\widehat{\ell}_t(x, a)$ can be computed without having to hold in memory $C_{k,a}$ and $B_{k,a}$, which can both be infinite-dimensional objects. In Appendix C, we show that $q_t(x, a)$ and $b_t(x, a)$ can both be computed for any given x using $\mathcal{O}(tM^3)$ kernel evaluations, and describe all the implementation details of KERNELFTRL. The overall idea is that using Mercer's theorem (Theorem 2) shows that the kernel function can be represented as $\kappa(x, x') = \langle \varphi(x), \varphi(x) \rangle$, and the algorithm only needs to evaluate the kernel function for various values of x, x'.

Our main result regarding the performance of KERNELFTRL for the two different eigenvalue decay conditions is the following:

Theorem 1. *Suppose that the kernel κ satisfies Assumption 1 with the polynomial eigenvalues decay rate $\mu_i \leq gi^{-c}$. Then, setting the parameters as $M = T$, $\eta = \beta = T^{-\frac{1}{2}\left(1+\frac{1}{c}\right)}\sqrt{\frac{(c-1)\ln T}{g}}$ the expected regret of* KERNELFTRL *satisfies*

$$R_T = \mathcal{O}\left(KT^{\frac{1}{2}\left(1+\frac{1}{c}\right)}\sqrt{(g/(c-1))\ln T}\right).$$

Furthermore, suppose that the kernel κ satisfies Assumption 1 with the exponential decay rate $\mu_i \leq ge^{-ci}$. Then, setting the parameters as $M = T$, $\eta = \beta = \sqrt{\frac{c\ln T}{gT}}$, the expected regret of KERNELFTRL *satisfies*

$$R_T = \mathcal{O}\left(K\sqrt{(g/c)T\left(\ln T\right)^3}\right).$$

4 Analysis

In this section we provide the main arguments forming the proof of Theorem 1, relegating the proofs of some technical lemmas to Appendix B. First, we introduce some important notations that will be useful throughout the proof. We first define the operator $\widehat{\Sigma}^{+}_{t,a} = I + \sum_{k=1}^{M} C_k$ (with C_k defined through the KGR subroutine for the t, a pair in question), so that we can write the estimate of $f_{t,a}$ as

$$\widetilde{f}_{t,a} = \widehat{\Sigma}^{+}_{t,a}\varphi(X_t)\ell_t(X_t, A_t)\mathbb{I}_{\{A_t=a\}}. \tag{5}$$

Similarly, the exploration bonus $b_t(x, a)$ can be written using this notation as

$$b_t(x, a) = \beta \left\|\varphi(x)\right\|^2_{\Sigma^{+}_{t,a}}.$$

Using this notation, we denote $\widehat{\ell}_t(x, a) = \left\langle \varphi(x), \widetilde{f}_{t,a}\right\rangle - b_t(x, a)$. When written in this form, it becomes readily apparent that our bonus is closely related to the adjustment proposed by [28] for proving high-probability bounds in linear bandits (see also 29). That said, the purpose of our adjustment is quite different in that it mainly serves to remove a potentially harmful bias from the KGR estimators. As for computing the estimates and bonuses defined above, note that the full functions $\widetilde{f}_{t,a}$ and b_t are never computed by the algorithm, and are only evaluated at the contexts $X_{t+1}, X_{t+2}, \ldots, X_T$ encountered in runtime. As explained in Appendix D, each such evaluation has a cost of $\mathcal{O}(tM^2)$.

Our analysis will use ideas from [13] and a number of new techniques that are necessary for dealing with the infinite-dimensional loss functions $f_{t,a}$. For the sake of analysis, we define X_0 as a sample from the context distribution $\mathcal{D}$ drawn independently from the history of interactions $\mathcal{F}_T$. We introduce the following notations:

- $\widetilde{R}_T = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}(\pi_t(a|X_0) - \pi^*(a|X_0))\widehat{\ell}_t(X_0, a)\right]$,
- $B^*_T = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\pi^*(a|X_0)\left(\widehat{\ell}_t(X_0, a) - \ell_t(X_0, a)\right)\right]$,
- $B_T = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\pi_t(a|X_0)\left(\ell_t(X_0, a) - \widehat{\ell}_t(X_0, a)\right)\right]$.

The first step in our proof is then to rewrite the regret as the sum of these three terms:

$$R_T = \widetilde{R}_T + B^*_T + B_T. \tag{6}$$

The proof of this claim is a straightforward extension of the regret decomposition of Lemma 3 in [13] and can be found in Appendix B.1.

The terms in the decomposition can be interpreted as follows. First, B^*_T is the *overestimation bias* of the total loss of the comparator policy π^*, measuring the extent to which the expectation of the estimated loss of π^* exceeds the actual loss of the same policy. Similarly, B_T is the *underestimation bias* of the total loss incurred by the learner. As we will show, the overestimation bias can be uniformly upper bounded for all comparator policies thanks to the optimistic adjustment term $b_t(x, a)$ added to the loss function. Furthermore, we will show that the price of this adjustment is a term of

the order $\mathbb{E}_t\left[b_t(X_t, A_t)\right] = \beta\mathbb{E}_t\left[\mathrm{tr}\left(\Sigma_{t,a}\Sigma_{t,a}^{+}\right)\right]$, which can be controlled in terms of the effective dimension of the kernel.

To provide an interpretation for the term $\widetilde{R}_T$, let's consider an auxiliary online learning problem where x is fixed, there are K actions, and the losses are defined as $c_{t,a} = \widehat{\ell}_t(x,a)$ for each t, a. We then execute a copy of FTRL with the log-barrier regularizer on this sequence of losses, resulting in the sequence of action distributions $\pi_t = \arg\min_{p\in\Delta(\mathcal{A})}\left(\Psi(p) + \eta\sum_a\sum_{\tau=1}^{t}\widehat{\ell}_\tau(x,a)\right)$. Thus, the regret in the auxiliary game against the comparator π^* at x can be expressed as

$$\widehat{R}_T(x) = \sum_{t=1}^{T}\sum_a\big(\pi_t(a|x) - \pi^*(a|x)\big)\widehat{\ell}_t(x,a). \tag{7}$$

Now it is easy to notice that $\widetilde{R}_T$ can be expressed in terms of the regret in these auxiliary games as $\widetilde{R}_T = \mathbb{E}\left[\widehat{R}_T(X_0)\right]$. Our proof strategy will be to prove an almost-sure regret bound for the auxiliary games defined at each x and take the expectation of the resulting bounds with respect to the law of X_0, thus achieving a bound on the regret.

Before we jump into the analysis of each term discussed above, we state a technical result that will be used repeatedly in nearly all proofs. The simple proof is provided in Appendix B.2.

Lemma 1. *For all t, a and $\varepsilon > 0$, we have*

$$tr\left(\mathbb{E}_t\left[\Sigma_{t,a}^{+}\Sigma_{t,a}\right]\right) = tr\left((I - (I-\Sigma_{t,a})^M)\right) \le m(\varepsilon) + M\varepsilon.$$

4.1 The bias of the loss estimator

The most important ingredient in our analysis is establishing a bound on the bias of our loss estimators $\widehat{\ell}_t$. The following lemma is our key tool that we use to this end.

Lemma 2. *For any $x \in \mathcal{X}$, for any $\beta > 0, \gamma > 0, \lambda > 0$, we have*

$$\left|\mathbb{E}_t\left[\left\langle\varphi(x), f_{t,a} - \widetilde{f}_{t,a}\right\rangle\right]\right| \le \beta\mathbb{E}_t\left[\|\varphi(x)\|^2_{\widehat{\Sigma}_{t,a}^{+}}\right] + \frac{1}{\beta(M+1)}. \tag{8}$$

The proof follows from a more or less straightforward calculation regarding the bias arising from the truncated geometric series we use to approximate the "inverse" of $\Sigma_{t,a}$. While the building blocks are standard, the result itself is new and and valuable in the sense that it gives a tighter control on the bias of the geometric resampling estimator than previous works (e.g., 13). This tighter bound is enabled by our use of the log-barrier policy that allows us to set M significantly larger than what the previous analysis of [13] could have tolerated, which in turns enables meaningful control of the additional bias term $\frac{1}{\beta(M+1)}$ appearing in the above bound. We relegate the proof of this result to Appendix B.3.

We are now well-equipped to tackle the bias terms B_T^* and B_T. We first show a bound on the overestimation bias:

Lemma 3. *The overestimation bias can be bounded as $B_T^* \le \frac{T}{\beta(M+1)}$.*

Proof. We appeal to Lemma 2 to show that

$$\begin{aligned}\mathbb{E}_t\left[\widehat{\ell}_t(x,a)\right] - \ell_t(x,a) &= \left\langle\varphi(x), \mathbb{E}_t\left[\widetilde{f}_{t,a}\right] - f_{t,a}\right\rangle - \mathbb{E}_t\left[b_t(x,a)\right] \\ &\le \beta\mathbb{E}_t\left[\|\varphi(x)\|^2_{\widehat{\Sigma}_{t,a}^{+}}\right] + \frac{1}{\beta(M+1)} - \mathbb{E}_t\left[b_t(x,a)\right] = \frac{1}{\beta(M+1)},\end{aligned}$$

where we recalled the definition of $b_t(x,a)$ in the last step. The claim then follows from averaging both sides with the joint distribution of X_0 and $A^* \sim \pi^*(X_0)$, and summing up for all t. □

Notice that without the optimistic adjustment $b_t(x,a)$, the overestimation bias would scale with $\mathbb{E}\left[\sum_a \pi^*(a|X_0)\,\|\varphi(X_0)\|^2_{\widehat{\Sigma}_{t,a}^{+}}\right]$, which cannot be meaningfully bounded in general. The second lemma takes care of the underestimation bias, and also establishes the price of adding the exploration bonus $b_t(x,a)$ to the loss estimator.

Lemma 4. *The underestimation bias can be bounded for any $\varepsilon > 0$ as*

$$B_T \leq 2KT\beta(m(\varepsilon) + M\varepsilon) + \frac{T}{\beta(M+1)}.$$

The proof of this lemma can be found in Appendix B.5.

In words, the effect of the optimistic bias is a factor of 2 multiplying the term $\mathbb{E}_t\left[\pi_t(a|X_0)\,\|\varphi(X_0)\|^2_{\widetilde{\Sigma}^+_{t,a}}\right]$, which itself can be bounded effectively in terms of the effective dimension.

4.2 Bounding the auxiliary regret

The first major step in our proof is to bound the regret in the auxiliary games, which is done in the following standard lemma concerning the bound of FTRL with log-barrier regularization:

Lemma 5. *Let $p_1, \ldots, p_T \in \Delta(\mathcal{A})$ be defined as*

$$p_t = \operatorname*{arg\,min}_{p \in \Delta(\mathcal{A})} \left\{ \eta \sum_a p_a \sum_{\tau \leq t} c_{\tau,a} + \Psi(p) \right\}, \forall t = 1, \ldots, T,$$

where $c_t \in \mathbb{R}^{\mathcal{A}}$ is an arbitrary loss vector and $\Psi(p) = \sum_{a \in \mathcal{A}} \ln \frac{1}{p_a}$. Then, for any $y \in \Delta(\mathcal{A})$,

$$\sum_{t=1}^{T} \sum_a (p_{t,a} - y_a)\, c_{t,a} \leq \frac{\Psi(y) - \Psi(p_1)}{\eta} + \eta \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} p_{t,a} c_{t,a}^2.$$

The proof of this result is standard and can be found in a number of references—we point the reader to Lemma 3.1 from [30] for concreteness. Notably, the second term in this bound has the the same qualitative form as the standard bound for FTRL with negative entropy, with the key advantage that it does not require any assumptions regarding the range of losses c_t.

Before we apply the above result to bounding $\widetilde{R}_T$, we state the following useful technical result regarding the second moment of the KGR estimator:

Lemma 6. *Suppose that X_t satisfies $\|\varphi(X_t)\| \leq 1$ for each t. Then for each t, the following inequality holds for any $\varepsilon > 0$:*

$$\mathbb{E}_t\left[\sum_{a=1}^{K} \pi_t(a|X_0) \big\langle \varphi(X_0), \widetilde{f}_{t,a} \big\rangle^2\right] \leq 2K\left(1 + (m(\varepsilon) + M\varepsilon)\right).$$

The proof of this lemma follows from a rather tedious calculation that can be found in Appendix B.4. With this result at hand, we are ready to state and prove the last remaining part of our regret bound.

Lemma 7. *For any positive $\eta, \beta, M, \varepsilon,$* KernelFTRL *guarantees*

$$\widetilde{R}_T \leq \frac{K \ln T}{\eta} + 2 + 2\beta(M+1) + \frac{2}{\beta(M+1)} + 2\eta KT\left(2 + (m(\varepsilon) + M\varepsilon)\right) \cdot \left(2 + \beta^2 M\right).$$

The proof of this lemma can be found in Appendix B.6.

4.3 The proof of Theorem 1

The proof now follows from putting together the results of Lemmas 3, 4, and 7, yielding

$$\begin{aligned} R_T &= \widetilde{R}_T + B_T^* + B_T \\ &\leq \frac{K \ln T}{\eta} + 2 + 2\beta(M+1) + \frac{2}{\beta(M+1)} + 2\eta KT\left(2 + (m(\varepsilon) + M\varepsilon)\right) \cdot \left(2 + \beta^2 M\right) \\ &\quad + \frac{2T}{\beta(M+1)} + 2\beta K(m(\varepsilon) + M\varepsilon)T. \end{aligned}$$

It remains to derive the concrete rates claimed in the theorem for the two separate eigendecay regimes considered therein. First, consider the polynomial decay rate $\mu_i \leq gi^{-c}$ and recall from Section 2.2 that we have $m(\varepsilon) = \mathcal{O}\left(((c-1)\varepsilon/g)^{1/(c-1)}\right)$ in this case. Thus, we can set $\varepsilon = \frac{g}{c-1}T^{\frac{1-c}{c}}$ which yields $m = \mathcal{O}\left(T^{1/c}\right)$. Taking $M = T$, $\eta = \beta = T^{-\frac{1}{2}\left(1+\frac{1}{c}\right)}\sqrt{\frac{(c-1)\ln T}{g}}$ and plugging into the bound above, the expected regret of KERNELFTRL can be seen to satisfy

$$R_T = \mathcal{O}\left(K\sqrt{(g/(c-1))\ln(T)}T^{\frac{1}{2}\left(1+\frac{1}{c}\right)}\right),$$

proving the first claim.

As for the exponential decay $\mu_i \leq ge^{-ci}$, recall from Section 2.2 that $m(\varepsilon) = \mathcal{O}\left(\frac{\ln(g/(c\varepsilon))}{c}\right)$, so that we can set $\varepsilon = \frac{g}{cT}$ to get $m(\varepsilon) = \frac{\ln T}{c}$. Letting $M = T$, $\eta = \beta = \sqrt{\frac{c\ln T}{gT}}$, and substituting these values into the previously derived bound, we can observe that

$$R_T = \mathcal{O}\left(K\sqrt{(g/c)T\ln(T)}\ln(T)\right),$$

which proves the second claim. □

5 Discussion

We now turn to discussing our results in some more detail, focusing on comparison with related work and the possibility to improve certain aspects of our algorithm and its theoretical guarantees.

The first question one may ask is if our results match the best achievable regret bounds in this context. While we cannot provide a fully affirmative answer to this question, there are definitely reasons to believe that at least the dependence of our bounds on T is optimal. In the special cases of Matérn and Gaussian kernels, our upper bounds match the lower bounds proved by [31] and also the lower bounds of [15] that were proved for more general kernels but in a slightly different setting. In the general case, our bounds can also be shown to match the best known rates for the stochastic version of our problem claimed by [10]—see the discussion in Appendix D of [32] that relates the various notions of "effective dimension" used in these works. A comparison with these results is made possible by noticing that $\operatorname{tr}\left(I - (I - \Sigma_{t,a})^M\right)$ can be also seen as an effective dimension that closely matches the other dimensions proposed in the previously mentioned papers [33, 34]. In light of these observations, we conjecture that our bounds are optimal in terms of T under the set of assumptions we make.

One remarkable downside of our bounds is their linear scaling with the number of actions K. This obviously suboptimal scaling is due to the use of the log-barrier regularizer in our algorithm. We conjecture that this factor can be improved by a more sophisticated algorithm design. One potential idea that we believe could work would be to adapt the very recently proposed "magnitude-reduced" loss estimators of [30] in tandem with a standard entropy regularizer, but we can see many potential failure modes for this approach and as such we leave its exploration for future work.

Finally, let us comment on the computational complexity of our method. In Appendix D, we show that the computation of $\widehat{L}_{t-1,a}$ for all actions takes $\mathcal{O}(K(t-1)M^3)$ steps, which makes for a total computational complexity of $\mathcal{O}(KT^5)$ over T rounds due to our choice of $M = T$. While polynomial in T, this rate is obviously not the most practical that one can wish for, and thus it is a natural question to ask if a faster method can be devised without compromising the regret bounds. A potential idea to consider is to use sketching methods such as the ones used by [35, 36] or [32] to reduce the computational burden. We note that it is not obvious at all if such methods can achieve the desired goal, as none of these sketching-based methods are able to attain the near-optimal rates of inefficient algorithms like that of [10].

References

[1] Ambuj Tewari and Susan A. Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health - Sensors, Analytic Methods, and Applications*, 2017.

[2] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.

[3] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

[4] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[5] Yingkai Li, Yining Wang, and Yuan Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory*, pages 2173–2174. PMLR, 2019.

[6] Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33:11478–11489, 2020.

[7] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[8] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

[9] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

[10] Michal Valko, Nathan Korda, Rémi Munos, Ilias Flaounas, and Nello Cristianini. Finite-time analysis of kernelised contextual bandits. In *Uncertainty in Artificial Intelligence*, 2013.

[11] Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2017.

[12] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[13] Gergely Neu and Julia Olkhovskaya. Efficient and robust algorithms for adversarial linear contextual bandits. In *Conference on Learning Theory*, pages 3049–3068. PMLR, 2020.

[14] Haolin Liu, Chen-Yu Wei, and Julian Zimmert. Bypassing the simulator: Near-optimal adversarial linear contextual bandits. *arXiv preprint arXiv:2309.00814*, 2023.

[15] Niladri Chatterji, Aldo Pacchiano, and Peter Bartlett. Online learning with kernel losses. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 971–980. PMLR, 09–15 Jun 2019.

[16] Dylan J Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Eva Tardos. Learning in games: Robustness of fast convergence. *Advances in Neural Information Processing Systems*, 29, 2016.

[17] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007.

[18] Prateek Jain, Brian Kulis, Inderjit S Dhillon, and Kristen Grauman. Online metric learning and fast similarity search. In *Advances in neural information processing systems*, pages 761–768, 2009.

[19] Brian Kulis and Peter L Bartlett. Implicit online learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 575–582, 2010.

[20] Pranjal Awasthi, Moses Charikar, Kevin A. Lai, and Andrej Risteski. Label optimal regret bounds for online local learning. In *Proceedings of The 28th Conference on Learning Theory (COLT)*, pages 150–166, 2015.

[21] Paul Christiano. Provably manipulation-resistant reputation systems. In *Proceedings of the 29th Annual Conference on Learning Theory (COLT)*, pages 670–697, 2016.

[22] Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38, 2017.

[23] Sébastien Bubeck, Michael Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. In *Algorithmic Learning Theory*, pages 111–127, 2018.

[24] Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291. PMLR, 2018.

[25] Haipeng Luo, Chen-Yu Wei, and Kai Zheng. Efficient online portfolio with logarithmic regret. *Advances in neural information processing systems*, 31, 2018.

[26] Matthias W. Seeger, Sham M. Kakade, and Dean P. Foster. Information consistency of nonparametric gaussian process methods. *IEEE Transactions on Information Theory*, 54(5):2376–2382, 2008.

[27] Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I Jordan. On function approximation in reinforcement learning: Optimism in the face of large state spaces. *Advances in Neural Information Processing Systems*, 2020, 2020.

[28] Peter Bartlett, Varsha Dani, Thomas Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory-COLT 2008*, pages 335–342, 2008.

[29] Julian Zimmert and Tor Lattimore. Return of the bias: Almost minimax optimal high probability bounds for adversarial linear bandits. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 3285–3312, 2022.

[30] Yan Dai, Haipeng Luo, Chen-Yu Wei, and Julian Zimmert. Refined regret for adversarial mdps with linear function approximation. *arXiv preprint arXiv:2301.12942*, 2023.

[31] Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. Lower bounds on regret for noisy gaussian process bandit optimization. In *Conference on Learning Theory*, pages 1723–1742, 2017.

[32] Houssam Zenati, Alberto Bietti, Eustache Diemert, Julien Mairal, Matthieu Martin, and Pierre Gaillard. Efficient kernelized ucb for contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 5689–5720, 2022.

[33] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26:289–315, 2007.

[34] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

[35] Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco. Gaussian process optimization with adaptive sketching: Scalable and no regret. In *Conference on Learning Theory*, pages 533–557, 2019.

[36] Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco. Near-linear time gaussian process optimization with adaptive batching and resparsification. In *International Conference on Machine Learning*, pages 1295–1305, 2020.

[37] J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209:415–446, 1909.

[38] Andreas Christmann and Ingo Steinwart. *Support Vector Machines*. Springer New York, NY, 2008.

A Mercer's theorem

Mercer's theorem [37] provides a representation of a positive-definite kernel κ in terms of an infinite dimensional feature map (see, e.g. [38], Theorem 4.49). Let $\mathcal{X}$ be a compact metric space and ν be a finite Borel measure on $\mathcal{X}$ (we consider Lebesgue measure in a Euclidean space). Let $L^2_\nu(\mathcal{X})$ be the set of square-integrable functions on $\mathcal{X}$ with respect to ν. We further say that the kernel κ square-integrable if

$$\int_{\mathcal{X}} \int_{\mathcal{X}} \kappa(x, x')^2 \, d\nu(x) d\nu(x') < \infty. \tag{9}$$

Theorem 2. *(Mercer's Theorem) Let $\mathcal{X}$ be a compact metric space and ν be a finite Borel measure on $\mathcal{X}$. Let κ be a continuous and square-integrable kernel, inducing an integral operator $T_k : L^2_\nu(\mathcal{X}) \to L^2_\nu(\mathcal{X})$ defined by*

$$\left(T_\kappa f\right)(\cdot) = \int_{\mathcal{X}} \kappa(\cdot, x') f(x') \, d\nu(x') \,, \tag{10}$$

where $f \in L^2_\nu(\mathcal{X})$. Then, there exists a sequence of eigenvalue-eigenfunction pairs $\{(\mu_i, \psi_i)\}_{i=1}^\infty$ such that $\mu_i > 0$, and $T_\kappa \psi_i = \mu_i \psi_i$, for $m \geq 1$. Moreover, the kernel function can be represented as

$$\kappa\left(x, x'\right) = \sum_{i=1}^{\infty} \mu_i \psi_i(x) \psi_i\left(x'\right), \tag{11}$$

where the convergence of the series holds uniformly on $\mathcal{X} \times \mathcal{X}$.

Additionally, the Mercer representation theorem (see, e.g., [38], Theorem 4.51) states that the RKHS induced by κ can consequently be represented in terms of $\{(\mu_i, \psi_i)\}_{i=1}^\infty$.

Theorem 3. *(Mercer Representation Theorem) Let $\{(\mu_i, \psi_i)\}_{i=1}^\infty$ be the Mercer eigenvalue-eigenfunction pairs. Then, the RKHS associated with κ is given by*

$$\mathcal{H}_\kappa = \left\{ f(\cdot) = \sum_{i=1}^{\infty} w_i \mu_i^{\frac{1}{2}} \psi_i(\cdot) : w_i \in \mathbb{R}, \|f\|^2_{\mathcal{H}_\kappa} := \sum_{i=1}^{\infty} w_i^2 < \infty \right\} \tag{12}$$

In particular, the Mercer representation theorem indicates that the scaled eigenfunctions $\{\sqrt{\mu_i}\psi_i\}_{i=1}^\infty$ form an orthonormal basis for $\mathcal{H}_\kappa$.

B Omitted proofs

B.1 Proof of the regret decomposition (6)

Let us rewrite the estimate $\widehat{\ell}_{t,a}(x) = \left\langle \varphi(x), \widetilde{f}^*_{t,a} \right\rangle + \langle \varphi(x), \delta_{t,a} \rangle + b_{t,a}(x)$, where $\widetilde{f}^*_{t,a}$ and $\delta_{t,a}$ are such that $\mathbb{E}_t \left[\widetilde{f}^*_{t,a} \right] = f_{t,a}$ and $\widetilde{f}_{t,a} = \widetilde{f}^*_{t,a} + \delta_{t,a}$, so $\delta_{t,a}$ is the bias of $\widetilde{f}_{t,a}$. Also, let X_0 be a sample from the context distribution $\mathcal{D}$, drawn independently from $\mathcal{F}_T$. We will consider each term separately on the right-hand side of Equation (6). First, for $\widetilde{R}_T$, we have:

$$\begin{aligned}
\widetilde{R}_T &= \mathbb{E}_t \left[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} (\pi_t(a|X_0) - \pi^*(a|X_0)) \widehat{\ell}_t(X_0, a) \right] \\
&= \mathbb{E}_t \left[\mathbb{E}_t \left[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} (\pi_t(a|X_0) - \pi^*(a|X_0)) \widehat{\ell}_t(X_0, a) \middle| X_0 \right] \right] \\
&= \mathbb{E}_t \left[\mathbb{E}_t \left[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} (\pi_t(a|X_0) - \pi^*(a|X_0)) (\left\langle \varphi(X_0), \widetilde{f}^*_{t,a} \right\rangle + \langle \varphi(X_0), \delta_{t,a} \rangle - b_{t,a}(X_0)) \middle| X_0 \right] \right] \\
&= \mathbb{E}_t \left[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} (\pi_t(a|X_0) - \pi^*(a|X_0)) (\langle \varphi(X_0), f_{t,a} \rangle + \langle \varphi(X_0), \mathbb{E}\left[\delta_{t,a}\right] \rangle - b_{t,a}(X_0)) \right] \\
&= \mathbb{E}_t \left[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} (\pi_t(a|X_t) - \pi^*(a|X_t)) (\langle \varphi(X_t), f_{t,a} \rangle + \langle \varphi(X_t), \mathbb{E}\left[\delta_{t,a}\right] \rangle - b_{t,a}(X_t)) \right]
\end{aligned}$$

$$= \mathbb{E}_t \left[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} (\pi_t(a|X_t) - \pi^*(a|X_t))(\ell_{t,a}(X_t) + \langle \varphi(X_t), \mathbb{E}\left[\delta_{t,a}\right] \rangle - b_{t,a}(X_t)) \right],$$

where we used the independence of X_t and $\widehat{\Sigma}_{t,a}$. Applying the same sequence of equations to B_T^* and B_T, we get

$$B_T^* = \mathbb{E}_t \left[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \pi^*(a|X_t)(\langle \varphi(X_t), \mathbb{E}\left[\delta_{t,a}\right] \rangle - b_{t,a}(X_t)) \right]$$

and

$$B_T = \mathbb{E}_t \left[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \pi_t(a|X_t)(\langle \varphi(X_0), b_{t,a}(X_t) - \mathbb{E}\left[\delta_{t,a}\right] \rangle) \right].$$

The proof is concluded by collecting all terms together. □

B.2 Proof of Lemma 1

For the first part, recall the construction of $\widehat{\Sigma}_{t,a}^{+}$ defined through the KGR procedure. Note that $\mathbb{E}_t\left[B_{k,a}\right] = \Sigma_{t,a}$, and $\{X(k)\}_{k=1}^{M}$ are independent, and recall the identity $\left(I + \sum_{i=1}^{M} U_i\right)(I - U) = \left(I - U^{M+1}\right)$ that holds for any Hermitian operator $U : \ell_2 \to \ell_2$. Applying this identity with $U = I - \Sigma_{t,a}$, we get

$$\mathbb{E}\left[\widehat{\Sigma}_{t,a}^{+}\right] \Sigma_{t,a} = \left(I + \sum_{i=1}^{M} \left(I - \Sigma_{t,a}\right)^i \right) \Sigma_{t,a} = \left(I - \left(I - \Sigma_{t,a}\right)^{M+1} \right). \tag{13}$$

Let $\{e_1, e_2, \dots\}$ be the canonical basis in ℓ_2 and recall that $\mathrm{tr}\left(U\right) = \sum_{i=1}^{\infty} \langle e_i, U e_i \rangle$. Also introducing the notation $\mathrm{tr}_n\left(U\right) = \sum_{i=n}^{\infty} \langle e_i, U e_i \rangle$ for $n \in \mathbb{N}$, we observe that the following holds for each t, a:

$$\begin{aligned}
\mathrm{tr}_n\left(\Sigma_{t,a}\right) &= \mathrm{tr}_n\left(\mathbb{E}\left[\pi_t(a|X_t)\varphi(X_t) \otimes \varphi(X_t)\right]\right) \\
&= \mathbb{E}\left[\pi_t(a|X_t)\mathrm{tr}_n\left(\varphi(X_t) \otimes \varphi(X_t)\right)\right] \\
&\qquad \text{(by Fubini's theorem)} \\
&= \mathbb{E}\left[\pi_t(a|X_t) \sum_{i=n}^{\infty} \langle e_i, \varphi(X_t) \otimes \varphi(X_t) e_i \rangle\right] \\
&\leq \mathbb{E}\left[\sum_{i=n}^{\infty} \langle e_i, \varphi(X_t) \otimes \varphi(X_t) e_i \rangle\right] \\
&= \mathbb{E}\left[\sum_{i=n}^{\infty} \left(\langle e_i, \varphi(X_t) \rangle\right)^2\right] \\
&\leq \mathbb{E}\left[\sum_{i=n}^{\infty} \left\|\varphi(X_t)\right\|_2^2 \left\|e_i\right\|_2^2\right] \\
&\qquad \text{(Cauchy–Schwarz inequality)} \\
&= \mathbb{E}\left[\left(\sum_{i=n}^{\infty} \varphi_i^2(X_t)\right)\right] \\
&\qquad \text{(using } \left\|e_i\right\|_2^2 = 1\text{)} \\
&= \mathbb{E}\left[\left(\sum_{i=n}^{\infty} \mu_i \psi_i^2(X_t)\right)\right] \leq \sum_{i=n}^{\infty} \mu_i,
\end{aligned} \tag{14}$$

where we used $|\psi_i(x)| \leq 1$ and $\mu_i \geq 0$ in the last step. Moving on to bounding the trace of $\Sigma_{t,a}\widehat{\Sigma}^+_{t,a}$, we observe

$$\begin{aligned}
\mathrm{tr}\left(\mathbb{E}\left[\widehat{\Sigma}^+_{t,a}\right]\Sigma_{t,a}\right) &= \mathrm{tr}\left((I-(I-\Sigma_{t,a})^M)\right) = \sum_{i=1}^{\infty}\left\langle e_i, \left(I-(I-\Sigma_{t,a})^M\right)\cdot e_i\right\rangle \\
&\leq m + \sum_{i=m+1}^{\infty}\left\langle e_i, \left(I-(I-\Sigma_{t,a})^M\right)\cdot e_i\right\rangle \\
&= m + \sum_{i=m+1}^{\infty}\left(1-(1-\langle e_i, \Sigma_{t,a}\cdot e_i\rangle)^M\right) \\
&\leq m + M\sum_{i=m+1}^{\infty}\langle e_i, \Sigma_{t,a}\cdot e_i\rangle = m + M\mathrm{tr}_m(\Sigma_{t,a}) \\
&\leq m + M\sum_{i=m+1}^{\infty}\mu_i,
\end{aligned}$$

where we have split the sum at some arbitrary m, used $\left\langle e_i, \left(I-(I-\Sigma_{t,a})^M\right)\cdot e_i\right\rangle \leq \|e_i\|_2^2 \leq 1$ for the first m terms, and used the inequality $1-M\lambda \leq (1-\lambda)^M$ that holds for $\lambda \in [0,1]$ and $M > 1$ for the rest of the terms. Finally, we used the inequality (14) in the last step. The statement then follows from the taking $m = m(\varepsilon)$. □

B.3 Proof of Lemma 2

Proof. We start by writing the bias as

$$\begin{aligned}
\mathbb{E}_t\left[\langle\varphi(x), f_{t,a}\rangle - \left\langle\varphi(x), \widetilde{f}_{t,a}\right\rangle\right] &= \langle\varphi(x), f_{t,a}\rangle - \mathbb{E}_t\left[\left\langle\varphi(x), \widehat{\Sigma}^+_{t,a}\varphi(X_t)\right\rangle\langle\varphi(X_t), f_{t,a}\rangle\,\mathbb{I}_{\{A_t=a\}}\right] \\
&= \langle\varphi(x), f_{t,a}\rangle - \mathbb{E}_t\left[\left\langle\varphi(x), \widehat{\Sigma}^+_{t,a}\Sigma_{t,a}f_{t,a}\right\rangle\right] \\
&= \mathbb{E}_t\left[\left\langle\varphi(x), \left(I-\widehat{\Sigma}^+_{t,a}\Sigma_{t,a}\right)f_{t,a}\right\rangle\right] \\
&= \left\langle\varphi(x), \left((I-\Sigma_{t,a})^{M+1}\right)f_{t,a}\right\rangle \\
&\quad \text{(by (13))} \\
&\leq \alpha\left\|\varphi(x)\right\|^2_{(I-\Sigma_{t,a})^{M+1}} + \frac{1}{\alpha}\left\|f_{t,a}\right\|^2_{(I-\Sigma_{t,a})^{M+1}},
\end{aligned}$$

for an arbitrary $\alpha > 0$, where we have used the Cauchy–Schwarz inequality in the last step. The last term can be conveniently upper bounded by $\frac{1}{\alpha}\left\|f_{t,a}\right\|^2 \leq \frac{1}{\alpha}$, as $\|f_{t,a}\| \leq 1$. To bound the first term, notice that

$$\alpha\left\|\varphi(x)\right\|^2_{(I-\Sigma_{t,a})^{M+1}} \leq \frac{\alpha}{M+1}\sum_{i=0}^{M}\left\|\varphi(x)\right\|^2_{(I-\Sigma_{t,a})^i} = \frac{\alpha}{(M+1)}\mathbb{E}_t\left[\left\|\varphi(x)\right\|^2_{\widehat{\Sigma}^+_{t,a}}\right],$$

where we have used that $(I-\Sigma_{t,a})^{M+1} \preccurlyeq (I-\Sigma_{t,a})^i$ holds for all $i \leq M+1$ due to $\left\|\Sigma_{t,a}\right\|_{\mathrm{op}} \leq 1$, and the definition of $\widehat{\Sigma}^+_{t,a}$. Putting the above statements together, we obtain

$$\mathbb{E}_t\left[\langle\varphi(x), f_{t,a}\rangle - \left\langle\varphi(x), \widetilde{f}_{t,a}\right\rangle\right] \leq \frac{\alpha}{(M+1)}\mathbb{E}_t\left[\left\|\varphi(x)\right\|^2_{\widehat{\Sigma}^+_{t,a}}\right] + \frac{1}{\alpha}.$$

The statement is then proved by taking $\alpha = \beta\,(M+1)$. □

B.4 Proof of Lemma 6

The proof follows the steps of the proof of Lemma 6 of [13]. We start by plugging in the definition of $\widetilde{f}_{t,a}$ and writing

$$\mathbb{E}_t\left[\sum_{a=1}^{K}\pi_t(a|X_0)\left\langle\varphi(X_0), \widetilde{f}_{t,a}\right\rangle^2\right] = \mathbb{E}_t\left[\sum_{a=1}^{K}\pi_t(a|X_0)\left(\left\langle\varphi(X_0), \widehat{\Sigma}^+_{t,a}\varphi(X_t)\right\rangle\cdot\langle\varphi(X_t), f_{t,a}\rangle\,\mathbb{I}_{\{A_t=a\}}\right)^2\right]$$

$$\leq \mathbb{E}_t\left[\sum_{a=1}^{K}\pi_t(a|X_0)\left(\left\langle\varphi(X_0),\widehat{\Sigma}^{+}_{t,a}\varphi(X_t)\right\rangle\mathbb{I}_{\{A_t=a\}}\right)^2\right]$$
$$= \mathbb{E}_t\left[\sum_{a=1}^{K}\mathrm{tr}\left(\Sigma_{t,a}\widehat{\Sigma}^{+}_{t,a}\Sigma_{t,a}\widehat{\Sigma}^{+}_{t,a}\right)\right],$$

where we used $\left\langle\varphi(X_0), f_{t,a}\right\rangle \leq 1$ in the inequality. We omit t, a indexes in the following text. Using the Araki–Lieb–Thirring inequality, we get

$$\mathrm{tr}\left(\Sigma\Sigma^{+}\Sigma\Sigma^{+}\right) \leq \mathrm{tr}\left(\Sigma^2\left(\Sigma^{+}\right)^2\right).$$

Define $G_k = (I - B_k)$. Using the definition of Σ^{+} and elementary manipulations, we can get

$$\left(\Sigma^{+}\right)^2 = \left(I + \sum_{k=1}^{M}\prod_{j=1}^{k}G_i\right)^2 = I + 2\sum_{k=1}^{M}\prod_{j=1}^{k}G_j + \sum_{k,k'=1}^{M}\left(\prod_{j=1}^{k}G_j\right)\left(\prod_{j=0}^{k'}G_j\right)$$
$$= I + 2\sum_{k=1}^{M}\prod_{j=1}^{k}G_j + 2\sum_{k=1}^{M}\sum_{k'=k}^{M}\prod_{j=1}^{k}G_j^2\prod_{j=k+1}^{k'}G_j - \sum_{k=1}^{M}\prod_{j=1}^{k}G_j^2$$
$$\preccurlyeq 2I + 2\sum_{k=1}^{M}\prod_{j=1}^{k}G_j + 2\sum_{k=1}^{M}\sum_{k'=k}^{M}\prod_{j=1}^{k}G_j^2\prod_{j=k+1}^{k'}G_j,$$

where in the second line we reordered the sum $\sum_{k,k'=1}^{M}a_k a_{k'} = 2\sum_{k=1}^{M}\sum_{k'=k}^{M}a_k a_{k'} - \sum_{k=1}^{M}a_k^2$, while in the third line we dropped the last term and added I. Denote $D = \mathbb{E}_t\left[G_j\right]$ and $E = \mathbb{E}_t\left[G_j^2\right]$. Using independence of G_j's we get:

$$\mathbb{E}_t\left[\left(\Sigma^{+}\right)^2\right] \preccurlyeq 2\sum_{k=0}^{M}D^k + 2\sum_{k=1}^{M}E^k\sum_{k'=0}^{M-k}D^{k'}.$$

Using the fact that $D = I - \Sigma$, we have $\sum_{k=0}^{M}D^k\Sigma = I - D^M$ and thus

$$\mathbb{E}_t\left[\left(\Sigma^{+}\right)^2\right]\Sigma^2 \preccurlyeq 2\left(I - D^M\right)\Sigma + 2\sum_{k=1}^{M}E^k\left(I - D^{M-k}\right)\Sigma \preccurlyeq 2\Sigma + 2\sum_{k=1}^{M}E^k\Sigma,$$

where we have also used the fact that if $A \preccurlyeq B$, then for any positive semi-definite operator C holds the inequality $\mathrm{tr}\,(CA) \leq \mathrm{tr}\,(CB)$. Furthermore, since we have $B \preccurlyeq I$, we can also simplify

$$E = \mathbb{E}_t\left[(I - B)^2\right] \preccurlyeq \mathbb{E}_t\left[(I - B)\right] = D,$$

and write

$$\sum_{k=1}^{M}E^k\Sigma \preccurlyeq \sum_{k=1}^{M}D^k\Sigma = (I - (I - \Sigma)^M).$$

This then gives

$$\mathrm{tr}\left(\Sigma^2\mathbb{E}_t\left[\left(\Sigma^{+}\right)^2\right]\right) \leq 2\mathrm{tr}\left(\Sigma\right) + 2\sum_{k=1}^{M}\mathrm{tr}\left(D^k\Sigma\right) = 2\mathrm{tr}\left(\Sigma\right) + 2\mathrm{tr}\left((I - (I - \Sigma)^M)\right)$$
$$\leq 2 + m(\varepsilon) + M\varepsilon,$$

where the last step follows from using Lemma 1. □

B.5 Proof of Lemma 4

By applying Lemma 2, we get that

$$\ell_t(x,a) - \mathbb{E}_t\left[\widehat{\ell}_t(x,a)\right] = \left\langle\varphi(x), f_{t,a} - \mathbb{E}_t\left[\widetilde{f}_{t,a}\right]\right\rangle + \mathbb{E}_t\left[b_t(x,a)\right]$$

$$\leq \beta \mathbb{E}_t \left[\|\varphi(x)\|^2_{\widehat{\Sigma}^+_{t,a}} \right] + \frac{1}{\beta(M+1)} + \mathbb{E}_t \left[b_t(x,a) \right] = 2\mathbb{E}_t \left[b_t(x,a) \right] + \frac{1}{\beta(M+1)},$$

where we recalled the definition of $b_t(x,a)$ in the last step. Taking expectations and averaging with respect to $\pi_t(\cdot|X_0)$ and the distribution of X_0, we get

$$\mathbb{E}_t \left[\pi_t(a|X_0) \|\varphi(X_0)\|^2_{\widehat{\Sigma}^+_{t,a}} \right] = \mathbb{E}_t \left[\mathrm{tr}\left(\Sigma_{t,a} \widehat{\Sigma}^+_{t,a} \right) \right] \leq m(\varepsilon) + M\varepsilon,$$

where the last step follows from an application of Lemma 1. The proof is concluded by summing up for all t. □

B.6 Proof of Lemma 7

Let us fix $x \in \mathcal{X}$ and apply Lemma 5 to obtain the following:

$$\begin{aligned} \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} (\pi_t(a|x) - \pi^*(a|x)) \left(\hat{\ell}_k(a,x) \right) &\leq \frac{\Psi(\tilde{\pi}^*(\cdot|x)) - \Psi(\pi_1(\cdot|x))}{\eta} \\ &+ \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} (\tilde{\pi}^*(a|x) - \pi^*(a|x)) \left(\widehat{\ell}_t(x,a) \right) + \eta \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \pi_t(a|x) \left(\widehat{\ell}_t(x,a) \right)^2. \end{aligned} \tag{15}$$

By picking $\tilde{\pi}^*(a|x) = \left(1 - \frac{K}{T}\right) \pi^*(a|x) + \frac{1}{T}$, the first term is bounded by

$$\frac{\Psi(\tilde{\pi}^*(\cdot|x)) - \Psi(\pi_1(\cdot|x))}{\eta} \leq \frac{K \ln T}{\eta}.$$

To proceed, we appeal to Lemma 2 to show that

$$\left| \mathbb{E}_t \left[\widehat{\ell}_t(x,a) \right] - \ell_t(x,a) \right| \leq 2\beta \mathbb{E}_t \left[\|\varphi(x)\|^2_{\widehat{\Sigma}^+_{t,a}} \right] + \frac{1}{\beta(M+1)} = 2\mathbb{E}_t \left[b_t(x,a) \right] + \frac{1}{\beta(M+1)},$$

and also observe that the exploration bonus can be bounded as

$$b_t(x,a) = \beta \|\varphi(x)\|^2_{\widehat{\Sigma}^+_{t,a}} \leq \beta \left\| \widehat{\Sigma}^+_{t,a} \right\|_2 \leq \beta(M+1).$$

Altogether, these observations can be used to simply bound the second term on the right-hand side of Equation (15) as

$$\mathbb{E} \left[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} (\tilde{\pi}^*(a|x) - \pi^*(a|x)) \widehat{\ell}_t(x,a) \right] \leq 2 \left(1 + \beta(M+1) + \frac{1}{\beta(M+1)} \right).$$

It remains to bound the last term on the right-hand side of Equation (15), which we start by writing

$$\eta \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \pi_t(a|x) \left(\widehat{\ell}_t(x,a) \right)^2 \leq 2\eta \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \pi_t(a|x) \left(\left(\left\langle \varphi(x), \widetilde{f}_{t,a} \right\rangle \right)^2 + \left(b_t(x,a) \right)^2 \right).$$

We will bound these terms on expectation with respect to the random context X_0. The first term in the resulting expression can be upper bounded by using Lemma 6:

$$\mathbb{E}_t \left[\sum_{t=1}^{T} \sum_{a=1}^{A} \pi_t(a|X_0) \left\langle \varphi(X_0), \widetilde{f}_{t,a} \right\rangle^2 \right] \leq 2K \left(1 + (m(\varepsilon) + M\varepsilon) \right).$$

Moving on to the second term, we have

$$\begin{aligned} \mathbb{E}_t \left[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \pi_t(a|X_0) \left(b_t(X_0,a) \right)^2 \right] &\leq \beta^2 M \mathbb{E}_t \left[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \pi_t(a|X_0) \|\varphi(X_0)\|^2_{\widehat{\Sigma}^+_{t,a}} \right] \\ &= \beta^2 M \mathbb{E}_t \left[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \mathrm{tr}\left(\Sigma_{t,a} \widehat{\Sigma}^+_{t,a} \right) \right] \leq \beta^2 M \left(m(\varepsilon) + M\varepsilon \right) KT, \end{aligned}$$

where the last inequality uses Lemma 1. Collecting all terms together, we get

$$\widetilde{R}_T \leq \frac{K \ln T}{\eta} + 2 + 2\beta(M+1) + \frac{2}{\beta(M+1)} + 2\eta KT \left(2 + (m(\varepsilon) + M\varepsilon) \right) \cdot \left(2 + \beta^2 M \right),$$

thus concluding the proof. □

C Implementing KERNELFTRL

We now present a fully operational definition of KERNELFTRL that does away with all the abstractions used in the main text. In particular, we here provide a version that fully unpacks the computation of the cumulative loss estimates $\widehat{L}_t(X_t, a)$ needed by the FTRL subroutine to calculate the policy $\pi_t(\cdot|X_t)$. As some pondering of the abstract description reveals, this computation requires rerunning the entire KGR subroutine for the whole sequence of past observations, including reusing the context-action pairs generated by KGR in each time step preceding t. In order to accommodate this sample reuse, in Algorithm 2, we present a version of KERNELFTRL that uses the subroutine KGRLOSSESTIMATE to compute the cumulative loss estimates and a *data buffer* $\mathcal{B}_t$ that stores all relevant data for computing said estimates. This subroutine is presented as Algorithm 4, and it makes use of the KGR subroutine presented as Algorithm 3.

Algorithm 2 KERNELFTRL

Parameters: Learning rate $\eta > 0$.
Initialization: Set $\mathcal{B}_t = \emptyset$.
For $t = 1, \dots, T$, **repeat:**

1. Observe X_t and, for all a, compute $L_t(X_t, a) = \text{KGRLOSSESTIMATE}(X_t, a, \mathcal{B}_t)$.
2. Observe X_t and, for all a, set

$$\pi_t(\cdot|X_t) = \operatorname*{arg\,min}_{p \in \Delta(\mathcal{A})} \left(\Psi(p) + \eta \sum_a p_a \widehat{L}_{t-1}(X_t, a) \right), \tag{16}$$

3. draw A_t from the policy $\pi_t(\cdot|X_t)$,
4. observe the loss $\ell_t(X_t, A_t)$,
5. for $k = 1, 2, \dots, M$, draw $X(k) \sim \mathcal{D}$ and $A(k) \sim \pi_t(\cdot|X(k))$ using the procedure above,
6. update buffer with the tuple

$$\mathcal{B}_{t+1} = \mathcal{B}_t \cup \left(X_t, A_t, \ell_t(X_t, A_t), \{X_t(k), A_t(k)\}_{k=1}^M\right).$$

Algorithm 3 Kernel Geometric Resampling (KGR)

Parameters: $\beta \geq 0$.
Input: Context-action pairs (x, a), (x', a'), and $\{x(k), a(k)\}_{k=1}^M$.
For $k = 1, \dots, M$**, repeat**:

1. set $B_{k,a} = \mathbb{I}_{\{a(k)=a\}} \varphi(x(k)) \otimes \varphi(x(k))$,
2. set $C_{k,a} = \prod_{j=1}^k (I - B_{j,a})$,
3. compute $q_{k,a}(x) = \langle \varphi(x), C_{k,a} \varphi(X_t) \rangle$, and
4. compute $b_{k,a}(x) = \beta \langle \varphi(x), C_{k,a} \varphi(x) \rangle$.

Return $q(x, a) = \mathbb{I}_{\{a=a'\}} \left(\kappa(x, x') + \sum_{k=1}^M q_{k,a}(x) \right)$,
$b(x, a) = \kappa(x, x) + \sum_{k=1}^M b_{k,a}(x)$.

Algorithm 4 KGRLOSSESTIMATE

Input: context x, action a, a set of tuples $\left(X_i, A_i, \ell_i(X_i, A_i), \{X_i(k), A_i(k)\}_{k=1}^M\right)_{i=1}^n$.
Initialize: $L_0(x, a) = 0$ for all a.
For $i = 1, 2, \dots, n$**, repeat:**

1. let $(q_i(x, a), b_i(x, a)) = \text{KGR}(x, a, X_i, A_i, \{X_i(k), A_i(k)\}_{k=1}^M)$,
2. let $\widehat{\ell}_i(x, a) = q_i(x, a) \ell_i(X_i, a) \mathbb{I}_{\{A_i=a\}} - b_i(x, a)$,
3. update $\widehat{L}_i(x, a) = \widehat{L}_{i-1}(x, a) + \widehat{\ell}_i(x, a)$.

As we show in Appendix D, the KGR procedure with a given set of inputs runs in $\mathcal{O}(M^2)$ time. In round t, the KGR subroutine is called by KGRLOSSESTIMATE t times, which costs a total of

$\mathcal{O}(tM^2)$ time. Finally, KGRLOSSESTIMATE is called by the main algorithm KERNELFTRL $M+1$ times when generating the action A_t and the independent copies $\{A_t(k)\}_{k=1}^{M}$, which altogether makes for a time complexity of $\mathcal{O}(tM^3)$ per round. Thus, the total time complexity of implementing KERNELFTRL is $\mathcal{O}(T^2M^3)$. As for memory complexity, the main bottleneck is having to store the data buffer $\mathcal{B}_t$, which consists of $t(M+1)$ context-action pairs and t observed losses. Overall, this means that KERNELFTRL requires to store a total of $\mathcal{O}(TM)$ context-action pairs in memory.

D Computational analysis of KGR

Lemma 8. *Kernel Geometric Resampling requires $\mathcal{O}(M^2)$ elementary operations and requires $\mathcal{O}(M)$ memory.*

Proof. The proof goes by induction. For $M=1$, we get

$$\begin{aligned} q_t(x,a) &= k(x,X_t) + \left\langle \varphi(x), \left(I - \mathbb{I}_{\{a(1)=a\}}\varphi(X(1))\otimes\varphi(X(1))\right)\varphi(x)\right\rangle \\ &= 2k(x,X_t) - \mathbb{I}_{\{A(1)=a\}}\kappa(x,X(1))k(X(1),X_t). \end{aligned}$$

For step $m+1$, we have

$$\begin{aligned} q_{t,k,a}(x) &= \langle \varphi(x), \prod_{j=1}^{k}\left(I - B_{j,a}\right)\varphi(X_t)\rangle \\ &= \langle \varphi(x), \prod_{j=1}^{k-1}\left(I - B_{j,a}\right)\cdot\left(I - \mathbb{I}_{\{a(k)=a\}}\varphi(X(k))\otimes\varphi(X(k))\right)\varphi(X_t)\rangle \\ &= q_{t,k-1,a} + \mathbb{I}_{\{a(k)=a\}}\langle \varphi(x), \prod_{j=1}^{k-1}\left(I - B_{j,a}\right)\cdot\varphi(X(k))\otimes\varphi(X(k))\varphi(X_t)\rangle. \end{aligned}$$

Note that there exists a set of coefficients $\{p_{a,k-1,i}\}_{i=1}^{k-1}$ such that

$$\langle \varphi(x), \prod_{j=1}^{k-1}\left(I - B_{j,a}\right)\cdot\varphi(X_t)\rangle = \sum_{i=1}^{k-1} p_{a,k-1,i}\left\langle \varphi(X(i)), \varphi(X_t)\right\rangle.$$

We can compute $\{p_{a,k,i}\}_{i=1}^{k}$ in k steps, as:

$$\begin{aligned} &\langle \varphi(x), \prod_{j=1}^{k}\left(I - B_{j,a}\right)\cdot\varphi(X_t)\rangle \\ &= \sum_{i=1}^{k} p_{a,k-1,i}\left\langle \varphi(X(i)), \left(I - B_{k,a}\right)\cdot\varphi(X_t)\right\rangle \\ &= \sum_{i=1}^{k} p_{a,k-1,i}\left\langle \varphi(X(i)), \varphi(X_t)\right\rangle \\ &\qquad - \mathbb{I}_{\{a(k)=a\}}\sum_{i=1}^{k} p_{a,k-1,i}\left\langle \varphi(X(i)), \varphi(X(k+1))\otimes\varphi(X(k+1))\varphi(X_t)\right\rangle \\ &= \sum_{i=1}^{k} p_{a,k-1,i}\left\langle \varphi(X(i)), \varphi(X_t)\right\rangle \\ &\qquad - \mathbb{I}_{\{a(k)=a\}}\sum_{i=1}^{k} p_{a,k-1,i}\left\langle \varphi(X(i)), \varphi(X(k+1))\right\rangle\left\langle \varphi(X(k+1)), \varphi(X_t)\right\rangle \\ &= \sum_{i=1}^{k} p_{a,k-1,i}\left\langle \varphi(X(i)), \varphi(X_t)\right\rangle \end{aligned}$$

$$- \mathbb{I}_{\{a(k)=a\}} \sum_{i=1}^{k} p_{a,k-1,i} \kappa(X(i), X(k+1)) \langle \varphi(X(k+1)), \varphi(X_t) \rangle .$$

Thus, we get that for $i \leq k-1$, $p_{a,k,i} = p_{a,k-1,i}$ and $p_{a,k,k} = -\mathbb{I}_{\{a(k)=a\}} \sum_{i=1}^{k-1} p_{a,k-1,i} \kappa((X(i), X(k))$, which means that computing $p_{a,k,i}$ takes $k-1$ operations, which results in $\frac{M(M+1)}{2}$ operations to compute $\{p_{a,k-1,i}\}_{i=1}^{M}$. In order to do this we need to store in memory an array of size M with coefficients $\{p_{a,k-1,i}\}_{i=1}^{k-1}$. Notice, that the same line of computations apply to computing $b_{t,a}(x)$. Thus, given that the time of computing kernel is K, to compute $q_{t,a}(x)$ and $b_{t,a}(x)$, we need $\frac{M(M+1)}{2} K$ steps. □