

---

# Adversarial Contextual Bandits Go Kernelized

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

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

## 14 1 Introduction

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

21 One of the main challenges of the contextual bandit problem is that the partial observations made  
22 about the losses handed out by the environment must be generalized efficiently to a possibly infinite  
23 set of contexts that are yet to be encountered in future decision-making rounds. One possible way  
24 to address this challenge is by making suitable assumptions about the structure of the losses. One  
25 particularly well-studied model is that of *linear contextual bandit*, where the losses are assumed to be  
26 linear in some known low-dimensional representation of the contexts. In the most broadly considered  
27 version of this setup, the sequence of contexts is completely arbitrary and the losses are determined  
28 by fixed linear functions. Advancements in this model have been made in a range of works, including  
29 [3, 4, 5, 6].

30 This model has been successfully generalized to deal with non-linear loss functions that belong to  
31 reproducing kernel Hilbert spaces. This assumption is broadly applicable, as the RKHS associated  
32 with commonly used kernels has the capacity to approximate nearly all continuous functions on  
33 compact subsets of  $\mathbb{R}^d$  [7, 8]. Viewed through the lens of kernel maps, this setting represents an  
34 extreme extension of the parametric linear bandit setting mentioned above, where the contexts can  
35 be represented in infinite-dimensional vector spaces. Works like [9, 10, 11] have provided efficient  
36 algorithms with strong performance guarantees for contextual bandits with such nonlinear loss  
37 functions that remain fixed throughout the online learning process.

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

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

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

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

71 **Notation.** We let  $\ell_2$  denote the space of square-summable sequences. For any two elements  
 72  $v, w \in \ell_2$ , we use  $\langle w, v \rangle$  to denote the standard  $\ell_2$  inner product  $\sum_{i=1}^{\infty} w_i u_i$ , and we define the  $\ell_2$   
 73 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  
 74 the operator that acts on elements  $u$  of  $\ell_2$  as  $(v \otimes w)u = v \langle w, u \rangle$ . For a positive definite operator  $B$   
 75 on  $\ell_2$ , we define  $\|v\|_B = \sqrt{\langle v, Bv \rangle}$ , and its trace as  $\text{tr}(B) = \sum_i \|e_i\|_B^2$ , where  $e_i$  is the  $i$ th canonical  
 76 basis vector in  $\ell_2$ . In the context of sequential-decision making problems, we will use  $\mathcal{F}_t$  to denote  
 77 the interaction history between the learner and the environment, and use the shorthand notations  
 78  $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_{t-1}]$  and  $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot | \mathcal{F}_{t-1}]$ .

## 79 2 Preliminaries

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

### 81 2.1 Adversarial contextual bandits

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

- 84 1. The environment draws the context vector  $X_t \in \mathbb{R}^d$  from the context distribution  $\mathcal{D}$ , and  
 85 reveals it to the learner;
- 86 2. Independently of the context  $X_t$ , the environment chooses a loss function  $\ell_t : \mathcal{X} \times \mathcal{A} \rightarrow$   
 87  $[0, 1]$ ;

- 88 3. Based on  $X_t$  and possibly some randomness, the learner chooses action  $A_t \in [K]$ ;  
 89 4. The learner incurs and observes loss  $\ell_t(X_t, A_t)$ .

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

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

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

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

## 100 2.2 RKHS loss functions

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

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

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

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

111 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  
 112 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  
 113 context  $x$  in  $\ell_2$  induced by Mercer's theorem. An important implication of Mercer's theorem that we  
 114 will repeatedly use is that  $\langle \varphi(x), \varphi(x') \rangle = \kappa(x, x')$  holds for all  $x, x' \in \mathcal{X}$ .

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

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

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

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

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

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

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

### 137 3 Algorithm and main result

138 We now present our algorithm which is based on a regret-decomposition approach first proposed by  
139 [13] for finite-dimensional linear contextual bandits. The core idea of this method is to instantiate  
140 an online learning algorithm in every context  $x \in \mathcal{X}$  and feed it with an appropriately designed  
141 estimator of the loss function that allows generalization across different contexts. Concretely, we will  
142 run an instance of the standard Follow-the-Regularized-Leader (FTRL) algorithm with log-barrier  
143 regularization (as popularized in online learning by 16) as the online learning method, and derive a  
144 new loss estimator based on the Matrix Geometric Resampling procedure proposed by [13] along  
145 with an optimistic exploration idea that is novel within this context. While the algorithm formally  
146 needs to calculate its policies and loss estimates that are valid on the whole context-action space,  
147 we will show that it can be implemented efficiently by querying the policy and the estimates only in  
148 the contexts encountered in runtime. To preserve readability in this section, we present a relatively  
149 abstract version of our algorithm first without worrying about implementability, and defer a fully  
150 detailed operational description to Appendix C.

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

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

155 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  
156 only has to compute the distribution  $\pi_t(\cdot | X_t)$  locally at  $X_t$  which can be done efficiently as long as  
157  $\hat{L}_{t-1}(X_t, a)$  can be efficiently computed for all actions  $a$  [16]. We will show later that this condition  
158 holds true for our loss estimators. We present this method as Algorithm 1 below.

---

#### Algorithm 1 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  $a$ , set

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

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  $\hat{\ell}_t$ .
- 

159 To describe our loss estimator, we first introduce the following operator on  $\ell_2$ :

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

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

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

162 which can be easily demonstrated to be unbiased:

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

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

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

#### Kernel Geometric Resampling

**Input:** Context  $x$ ,  $X_t$ , policy  $\pi_t$ , data distribution  $\mathcal{D}$ , parameters  $\beta, M$ .

**For**  $k = 1, \dots, M$ , **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

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

179 Notice that all operations performed by Kernel Geometric Resampling can be implemented by

180 applying simple rank-one operators to elements of  $\ell_2$ , so  $\widehat{\ell}_t(x, a)$  can be computed without having  
 181 to hold in memory  $C_{k,a}$  and  $B_{k,a}$ , which can both be infinite-dimensional objects. In Appendix C,  
 182 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  
 183 evaluations, and describe all the implementation details of KERNELFTRL. The overall idea is  
 184 that using Mercer's theorem (Theorem 2) shows that the kernel function can be represented as  
 185  $\kappa(x, x') = \langle \varphi(x), \varphi(x') \rangle$ , and the algorithm only needs to evaluate the kernel function for various  
 186 values of  $x, x'$ .

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

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

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

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

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

## 195 4 Analysis

196 In this section we provide the main arguments forming the proof of Theorem 1, relegating the proofs  
 197 of some technical lemmas to Appendix B. First, we introduce some important notations that will  
 198 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  
 199 through the KGR subroutine for the  $t, a$  pair in question), so that we can write the estimate of  $f_{t,a}$  as

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

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

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

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

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

$$\begin{aligned} 213 \quad & \widetilde{R}_T = \mathbb{E} \left[ \sum_{t=1}^T \sum_{a \in \mathcal{A}} (\pi_t(a|X_0) - \pi^*(a|X_0)) \widehat{\ell}_t(X_0, a) \right], \\ 214 \quad & B_T^* = \mathbb{E} \left[ \sum_{t=1}^T \sum_{a \in \mathcal{A}} \pi^*(a|X_0) \left( \widehat{\ell}_t(X_0, a) - \ell_t(X_0, a) \right) \right], \\ 215 \quad & B_T = \mathbb{E} \left[ \sum_{t=1}^T \sum_{a \in \mathcal{A}} \pi_t(a|X_0) \left( \ell_t(X_0, a) - \widehat{\ell}_t(X_0, a) \right) \right]. \end{aligned}$$

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

$$R_T = \widetilde{R}_T + B_T^* + B_T. \quad (6)$$

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

219 The terms in the decomposition can be interpreted as follows. First,  $B_T^*$  is the *overestimation bias*  
 220 of the total loss of the comparator policy  $\pi^*$ , measuring the extent to which the expectation of the  
 221 estimated loss of  $\pi^*$  exceeds the actual loss of the same policy. Similarly,  $B_T$  is the *underestimation*  
 222 *bias* of the total loss incurred by the learner. As we will show, the overestimation bias can be  
 223 uniformly upper bounded for all comparator policies thanks to the optimistic adjustment term  $b_t(x, a)$   
 224 added to the loss function. Furthermore, we will show that the price of this adjustment is a term of  
 225 the order  $\mathbb{E}_t [b_t(X_t, A_t)] = \beta \mathbb{E}_t [\text{tr}(\Sigma_{t,a} \Sigma_{t,a}^+)]$ , which can be controlled in terms of the effective  
 226 dimension of the kernel.

227 To provide an interpretation for the term  $\widetilde{R}_T$ , let's consider an auxiliary online learning problem  
 228 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  
 229 then execute a copy of FTRL with the log-barrier regularizer on this sequence of losses, resulting in



230 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  
 231 regret in the auxiliary game against the comparator  $\pi^*$  at  $x$  can be expressed as

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

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

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

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

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

#### 239 4.1 The bias of the loss estimator

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

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

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

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

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

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

253 *Proof.* We appeal to Lemma 2 to show that

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

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

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

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

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

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

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

## 265 4.2 Bounding the auxiliary regret

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

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

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

269 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  $y \in \Delta(\mathcal{A})$ ,

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

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

274 Before we apply the above result to bounding  $\tilde{R}_T$ , we state the following useful technical result  
 275 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  $\varepsilon > 0$ :*

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

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

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

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

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

## 280 4.3 The proof of Theorem 1

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

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

282 It remains to derive the concrete rates claimed in the theorem for the two separate eigendecay regimes  
 283 considered therein. First, consider the polynomial decay rate  $\mu_i \leq gi^{-c}$  and recall from Section 2.2  
 284 that we have  $m(\varepsilon) = \mathcal{O}(((c-1)\varepsilon/g)^{1/(c-1)})$  in this case. Thus, we can set  $\varepsilon = \frac{g}{c-1} T^{\frac{1-c}{c}}$  which



285 yields  $m = \mathcal{O}(T^{1/c})$ . Taking  $M = T$ ,  $\eta = \beta = T^{-\frac{1}{2}(1+\frac{1}{c})} \sqrt{\frac{(c-1)\ln T}{g}}$  and plugging into the  
 286 bound above, the expected regret of KERNELFTRL can be seen to satisfy

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

287 proving the first claim.

288 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

289 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  
 290 values into the previously derived bound, we can observe that

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

291 which proves the second claim. □

## 292 5 Discussion

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

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

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

313 Finally, let us comment on the computational complexity of our method. In Appendix D, we show  
 314 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  
 315 total computational complexity of  $\mathcal{O}(KT^5)$  over  $T$  rounds due to our choice of  $M = T$ . While  
 316 polynomial in  $T$ , this rate is obviously not the most practical that one can wish for, and thus it is a  
 317 natural question to ask if a faster method can be devised without compromising the regret bounds. A  
 318 potential idea to consider is to use sketching methods such as the ones used by [35, 36] or [32] to  
 319 reduce the computational burden. We note that it is not obvious at all if such methods can achieve the  
 320 desired goal, as none of these sketching-based methods are able to attain the near-optimal rates of  
 321 inefficient algorithms like that of [10].

## 322 References

- 323 [1] Ambuj Tewari and Susan A. Murphy. From ads to interventions: Contextual bandits in mobile  
 324 health. In *Mobile Health - Sensors, Analytic Methods, and Applications*, 2017.
- 325 [2] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual  
 326 bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth  
 327 International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop  
 328 and Conference Proceedings, 2011.

- 329 [3] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff  
330 functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence  
331 and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.
- 332 [4] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear  
333 stochastic bandits. *Advances in neural information processing systems*, 24, 2011.
- 334 [5] Yingkai Li, Yining Wang, and Yuan Zhou. Nearly minimax-optimal regret for linearly parame-  
335 terized bandits. In *Conference on Learning Theory*, pages 2173–2174. PMLR, 2019.
- 336 [6] Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to mis-  
337 specification in contextual bandits. *Advances in Neural Information Processing Systems*,  
338 33:11478–11489, 2020.
- 339 [7] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines,  
340 regularization, optimization, and beyond*. MIT press, 2002.
- 341 [8] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*.  
342 MIT press Cambridge, MA, 2006.
- 343 [9] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian pro-  
344 cess optimization in the bandit setting: No regret and experimental design. *arXiv preprint  
345 arXiv:0912.3995*, 2009.
- 346 [10] Michal Valko, Nathan Korda, Rémi Munos, Ilias Flaounas, and Nello Cristianini. Finite-time  
347 analysis of kernelised contextual bandits. In *Uncertainty in Artificial Intelligence*, 2013.
- 348 [11] Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International  
349 Conference on Machine Learning*, pages 844–853. PMLR, 2017.
- 350 [12] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic  
351 multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.
- 352 [13] Gergely Neu and Julia Olkhovskaya. Efficient and robust algorithms for adversarial linear  
353 contextual bandits. In *Conference on Learning Theory*, pages 3049–3068. PMLR, 2020.
- 354 [14] Haolin Liu, Chen-Yu Wei, and Julian Zimmert. Bypassing the simulator: Near-optimal adver-  
355 sarial linear contextual bandits. *arXiv preprint arXiv:2309.00814*, 2023.
- 356 [15] Niladri Chatterji, Aldo Pacchiano, and Peter Bartlett. Online learning with kernel losses. In  
357 Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International  
358 Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*,  
359 pages 971–980. PMLR, 09–15 Jun 2019.
- 360 [16] Dylan J Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Eva Tardos. Learning in  
361 games: Robustness of fast convergence. *Advances in Neural Information Processing Systems*,  
362 29, 2016.
- 363 [17] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-  
364 theoretic metric learning. In *Proceedings of the 24th international conference on Machine  
365 learning*, pages 209–216. ACM, 2007.
- 366 [18] Prateek Jain, Brian Kulis, Inderjit S Dhillon, and Kristen Grauman. Online metric learning and  
367 fast similarity search. In *Advances in neural information processing systems*, pages 761–768,  
368 2009.
- 369 [19] Brian Kulis and Peter L Bartlett. Implicit online learning. In *Proceedings of the 27th Interna-  
370 tional Conference on Machine Learning (ICML-10)*, pages 575–582, 2010.
- 371 [20] Pranjal Awasthi, Moses Charikar, Kevin A. Lai, and Andrej Risteski. Label optimal regret  
372 bounds for online local learning. In *Proceedings of The 28th Conference on Learning Theory  
373 (COLT)*, pages 150–166, 2015.
- 374 [21] Paul Christiano. Provably manipulation-resistant reputation systems. In *Proceedings of the 29th  
375 Annual Conference on Learning Theory (COLT)*, pages 670–697, 2016.

- 376 [22] Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corraling a band  
377 of bandit algorithms. In *Conference on Learning Theory*, pages 12–38, 2017.
- 378 [23] Sébastien Bubeck, Michael Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-  
379 armed bandits. In *Algorithmic Learning Theory*, pages 111–127, 2018.
- 380 [24] Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference  
381 On Learning Theory*, pages 1263–1291. PMLR, 2018.
- 382 [25] Haipeng Luo, Chen-Yu Wei, and Kai Zheng. Efficient online portfolio with logarithmic regret.  
383 *Advances in neural information processing systems*, 31, 2018.
- 384 [26] Matthias W. Seeger, Sham M. Kakade, and Dean P. Foster. Information consistency of nonpara-  
385 metric gaussian process methods. *IEEE Transactions on Information Theory*, 54(5):2376–2382,  
386 2008.
- 387 [27] Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I Jordan. On function  
388 approximation in reinforcement learning: Optimism in the face of large state spaces. *Advances  
389 in Neural Information Processing Systems*, 2020, 2020.
- 390 [28] Peter Bartlett, Varsha Dani, Thomas Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj  
391 Tewari. High-probability regret bounds for bandit online linear optimization. In *Proceedings of  
392 the 21st Annual Conference on Learning Theory-COLT 2008*, pages 335–342, 2008.
- 393 [29] Julian Zimmert and Tor Lattimore. Return of the bias: Almost minimax optimal high probability  
394 bounds for adversarial linear bandits. In *Proceedings of Thirty Fifth Conference on Learning  
395 Theory*, pages 3285–3312, 2022.
- 396 [30] Yan Dai, Haipeng Luo, Chen-Yu Wei, and Julian Zimmert. Refined regret for adversarial mdps  
397 with linear function approximation. *arXiv preprint arXiv:2301.12942*, 2023.
- 398 [31] Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. Lower bounds on regret for noisy  
399 gaussian process bandit optimization. In *Conference on Learning Theory*, pages 1723–1742,  
400 2017.
- 401 [32] Houssam Zenati, Alberto Bietti, Eustache Diemert, Julien Mairal, Matthieu Martin, and Pierre  
402 Gaillard. Efficient kernelized ucb for contextual bandits. In *International Conference on  
403 Artificial Intelligence and Statistics*, pages 5689–5720, 2022.
- 404 [33] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent  
405 learning. *Constructive Approximation*, 26:289–315, 2007.
- 406 [34] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric  
407 regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*,  
408 15(1):335–366, 2014.
- 409 [35] Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco.  
410 Gaussian process optimization with adaptive sketching: Scalable and no regret. In *Conference  
411 on Learning Theory*, pages 533–557, 2019.
- 412 [36] Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco.  
413 Near-linear time gaussian process optimization with adaptive batching and resparsification. In  
414 *International Conference on Machine Learning*, pages 1295–1305, 2020.
- 415 [37] J. Mercer. Functions of positive and negative type, and their connection with the theory  
416 of integral equations. *Philosophical Transactions of the Royal Society of London. Series A,  
417 Containing Papers of a Mathematical or Physical Character*, 209:415–446, 1909.
- 418 [38] Andreas Christmann and Ingo Steinwart. *Support Vector Machines*. Springer New York, NY,  
419 2008.

## 420 A Mercer's theorem

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

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

426 **Theorem 2.** (Mercer's Theorem) Let  $\mathcal{X}$  be a compact metric space and  $\nu$  be a finite Borel measure on  
 427  $\mathcal{X}$ . Let  $\kappa$  be a continuous and square-integrable kernel, inducing an integral operator  $T_\kappa : L_\nu^2(\mathcal{X}) \rightarrow$   
 428  $L_\nu^2(\mathcal{X})$  defined by

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

429 where  $f \in L_\nu^2(\mathcal{X})$ . Then, there exists a sequence of eigenvalue-eigenfunction pairs  $\{(\mu_i, \psi_i)\}_{i=1}^\infty$   
 430 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

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

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

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

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

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

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

## 438 B Omitted proofs

### 439 B.1 Proof of the regret decomposition (6)

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

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

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

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

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

446 and

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

447 The proof is concluded by collecting all terms together.  $\square$

## 448 B.2 Proof of Lemma 1

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

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

453 Let  $\{e_1, e_2, \dots\}$  be the canonical basis in  $\ell_2$  and recall that  $\text{tr}(U) = \sum_{i=1}^{\infty} \langle e_i, U e_i \rangle$ . Also introduc-  
454 ing the notation  $\text{tr}_n(U) = \sum_{i=n}^{\infty} \langle e_i, U e_i \rangle$  for  $n \in \mathbb{N}$ , we observe that the following holds for each  
455  $t, a$ :

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

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

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

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

### 462 B.3 Proof of Lemma 2

463 *Proof.* We start by writing the bias as

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

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

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

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

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

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

### 470 B.4 Proof of Lemma 6

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

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



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

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

$$\text{tr}(\Sigma \Sigma^+ \Sigma \Sigma^+) \leq \text{tr}(\Sigma^2 (\Sigma^+)^2).$$

475 Define  $G_k = (I - B_k)$ . Using the definition of  $\Sigma^+$  and elementary manipulations, we can get

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

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

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

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

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

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

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

482 and write

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

483 This then gives

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

484 where the last step follows from using Lemma 1. □

## 485 B.5 Proof of Lemma 4

486 By applying Lemma 2, we get that

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

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

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

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

489 where the last step follows from an application of Lemma 1. The proof is concluded by summing up  
490 for all  $t$ .  $\square$

## 491 B.6 Proof of Lemma 7

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

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

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

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

494 To proceed, we appeal to Lemma 2 to show that

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

495 and also observe that the exploration bonus can be bounded as

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

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

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

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

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

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:

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

499 Moving on to the second term, we have

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

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

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

501 thus concluding the proof.  $\square$

502 **C Implementing KERNELFTRL**

503 We now present a fully operational definition of KERNELFTRL that does away with all the ab-  
 504 stractions used in the main text. In particular, we here provide a version that fully unpacks the  
 505 computation of the cumulative loss estimates  $\widehat{L}_t(X_t, a)$  needed by the FTRL subroutine to calculate  
 506 the policy  $\pi_t(\cdot|X_t)$ . As some pondering of the abstract description reveals, this computation requires  
 507 rerunning the entire KGR subroutine for the whole sequence of past observations, including reusing  
 508 the context-action pairs generated by KGR in each time step preceding  $t$ . In order to accommodate  
 509 this sample reuse, in Algorithm 2, we present a version of KERNELFTRL that uses the subroutine  
 510 KGRLOSSESTIMATE to compute the cumulative loss estimates and a *data buffer*  $\mathcal{B}_t$  that stores all  
 511 relevant data for computing said estimates. This subroutine is presented as Algorithm 4, and it makes  
 512 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

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

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

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


---

---

**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  $q_{k,a}(x) = \langle \varphi(x), C_{k,a} \varphi(x) \rangle$ , 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  $\widehat{L}_i(x, a) = \widehat{L}_{i-1}(x, a) + \widehat{\ell}_i(x, a)$ .
- 

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

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

## 521 D Computational analysis of KGR

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

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

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

525 For step  $m + 1$ , we have

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

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

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

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

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

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

528 Thus, we get that for  $i \leq k - 1$ ,  $p_{a,k,i} = p_{a,k-1,i}$  and  $p_{a,k,k} =$   
529  $-\mathbb{I}_{\{a(k)=a\}} \sum_{i=1}^{k-1} p_{a,k-1,i} \kappa(X(i), X(k))$ , which means that computing  $p_{a,k,i}$  takes  $k - 1$   
530 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  
531 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  
532 line of computations apply to computing  $b_{t,a}(x)$ . Thus, given that the time of computing kernel is  $K$ ,  
533 to compute  $q_{t,a}(x)$  and  $b_{t,a}(x)$ , we need  $\frac{M(M+1)}{2} K$  steps.  $\square$