Kernel ϵ-Greedy for Contextual Bandits

Sakshi Arya and Bharath K. Sriperumbudur

Department of Statistics, Pennsylvania State University
University Park, PA 16802, USA.
arya.sakshi44@gmail.com,bks18@psu.edu

Abstract

We consider a kernelized version of the ϵ-greedy strategy for contextual bandits. More precisely, in a setting with finitely many arms, we consider that the mean reward functions lie in a reproducing kernel Hilbert space (RKHS). We propose an online weighted kernel ridge regression estimator for the reward functions. Under some conditions on the exploration probability sequence, $\{\epsilon_t\}_t$, and choice of the regularization parameter, $\{\lambda_t\}_t$, we show that the proposed estimator is consistent. We also show that for any choice of kernel and the corresponding RKHS, we achieve a sub-linear regret rate depending on the intrinsic dimensionality of the RKHS. Furthermore, we achieve the optimal regret rate of $\sqrt{T}$ under a margin condition for finite-dimensional RKHS.

MSC 2010 subject classification: Primary: 62L10; Secondary: 62G05, 68T05.
Keywords and phrases: Contextual bandits, reproducing kernel Hilbert space, covariance operator, ϵ-greedy, cumulative regret, inverse probability weighting, kernel ridge regression

1 Introduction

Sequential decision-making in real time is increasingly becoming important in various applications, such as clinical trials (Bather, 1985; Villar et al., 2015), news article recommendation (Li et al., 2010) and mobile health (Nahum-Shani et al., 2017). In all such problems, the decision-maker is faced with several alternatives, from which they have to make a series of choices (referred to as *arms*) sequentially, based on the information available at any given time. In doing so, the decision-maker takes into account additional information or covariates (characteristics) which help in making informed decisions. This framework is popularly known as the contextual bandit problem (Langford and Zhang, 2007). In a treatment allocation problem, this can be described as follows: given finitely many competing treatments for a disease, the decision-maker (physician) chooses the treatment best suited for individual patients as they arrive, and each allocated treatment results in a *reward* (outcome). While doing so, the decision-maker takes into account the patient's covariates and

information available about previous patients with the same disease, with the eventual goal of maximizing the total reward accumulated over a period of time. The technical challenge in achieving this is two-fold: 1) learning the relationship between the covariates and optimal arms, and, 2) balancing the *exploration-exploitation trade-off*, which arises due to the sequential (or online) nature of the problem. In other words, in a sequential setup, at each time point the physician has to effectively identify the best treatment (exploration) and treat patients as effectively as possible during the trial (exploitation). Since there has been a substantial number of advancements in the contextual bandit problem in recent years, we refer the reader to Lattimore and Szepesvári (2020) for a detailed description of recent developments in this area and Tewari and Murphy (2017) for a comprehensive survey of both parametric and nonparametric methods in contextual bandits.

In this paper, we consider a kernelized contextual bandit framework, where the relationship between the rewards and covariates for each arm is modeled by functions in a reproducing kernel Hilbert space (RKHS), and study a popular heuristic in multi-armed bandit problems known as the annealed ϵ-greedy strategy (Sutton and Barto, 2018). This strategy allocates arms based on a randomized strategy to balance the exploration-exploitation trade-off, with a careful reduction in exploration over time. For example, in a two-armed bandit, the ϵ-greedy strategy chooses the current best-performing arm with probability $1-\epsilon$ and makes a random decision with a small probability ϵ. In the annealed version of the algorithm, ϵ is a non-increasing function of time. This algorithm falls in the broad category of 'algorithms with myopic exploration', which are easy to implement and could result in good empirical performance in some situations with appropriate exploration probability choices (Bietti et al., 2021; Mnih et al., 2015). In practice, they are often selected as the top choices due to their simplicity. However, they have not been studied extensively in the literature. Recently, Dann et al. (2022) studied the ϵ-greedy strategy in the more general reinforcement learning setup and provided theoretical guarantees in terms of what they define as the myopic exploration gap, which is a problem-dependent quantity. Using ϵ-greedy strategy, they achieve the optimal regret rate of $\tilde{O}(T^{2/3})$ for contextual bandits, which matches the rate we obtain for our algorithm when the RKHS is finite-dimensional. We also show that we can get the same rate for an infinite-dimensional RKHS under the *margin condition*. Chen et al. (2021) study the ϵ-greedy strategy for the linear bandit problem (parametric contextual bandit with a linear regression framework) and establish the regret rate of $\tilde{O}(\sqrt{T})$, which our work recovers when the RKHS is finite-dimensional and under the margin condition.

1.1 Contributions

The main contribution of this work is in developing a theoretical understanding of kernelized ϵ-greedy algorithm. More specifically:

- We propose an inverse probability weighted kernel ridge (IPWKR) regression type of online estimator for the mean reward functions in Section 3.2, whose implementation details are provided in Section 3.3. Such estimators have been studied in the context of linear bandits for mitigating the problem of estimation bias in adaptively collected data (Dimakopoulou et al., 2019). Using IPWKR estimator, we propose a kernel ϵ-greedy algorithm for contextual bandits and provide regret bounds. We highlight here

that the inverse probability weights appearing in IPWKR are deterministic known quantities that involve user-determined exploration probabilities.

- In Section 4, we establish upper bounds on the estimation error (see Theorem 1) for the proposed IPWKR estimator, which we specialize to the setting of finite-dimensional RKHS in Theorem 2. These results hold for specific conditions on the exploration probability sequence $\{\epsilon_t\}_t$ and choices of the regularization parameter sequence $\{\lambda_t\}_t$. As a comparison, in the linear bandit framework, Chen et al. (2021) propose an online weighted least squares (WLS) estimator similar to our proposed IPWKR estimator but without regularization. In fact, when the kernel is linear, our proposed estimator can be seen as a dualized representation of their online WLS estimator if $\lambda = 0$. In Section 4 (Theorem 2 and Remark 1), we show that our consistency result is stronger than (Chen et al., 2021, Proposition 4.1) in the sense that we achieve consistency in estimation for large ranges of the exploration probabilities, $\{\epsilon_t\}_t$, i.e., for decaying choices of $\{\epsilon_t\}_t$ faster than those considered in Chen et al. (2021). Interestingly, for these choices of $\{\epsilon_t\}_t$, we obtain the same convergence rate as in Chen et al. (2021)—obtained for linear contextual bandits—even when the true regression functions are non-linear, as long as the RKHS is finite-dimensional. Also, we highlight that, compared to the existing literature (Valko et al., 2013; Zenati et al., 2022), in a finite-dimensional setting, our analysis provides an explicit data-dependent choice for regularization parameter, $\{\lambda_t\}_t$, circumventing the need to tune it when implementing the algorithm.
- In Theorem 3 of Section 5, we establish finite-time regret bounds for kernel ϵ-greedy strategy for contextual bandits with finitely many arms when the mean reward functions are assumed to be in an RKHS. These regret bounds are sub-linear for all choices of bounded, positive definite, and symmetric kernels. Compared to the results in the literature, this is a significant result since the Matérn kernel may have a linear regret for some choices of the kernel parameters (Vakili et al., 2021b; Scarlett et al., 2017). In fact, our finite-time regret bound matches the state-of-the-art upper bound for ϵ-greedy strategy when the RKHS is finite-dimensional (Theorem 4), both with and without the *margin condition*—the margin conditions ensures that there is sufficient gap between the rewards for different arms—, which are $\tilde{O}(T^{1/2})$ and $\tilde{O}(T^{2/3})$, respectively. When the RKHS is infinite-dimensional, the bounds are controlled by the intrinsic dimensionality of the RKHS—in turn, controlled by the decay rate of the eigenvalues of the covariance operator—and a source condition, which captures the smoothness of the target.
- In contrast to the literature, the above results are obtained by only assuming the noise variance to be finite, instead of the commonly assumed sub-Gaussian structure on the noise in (1). Moreover, our analysis does not require any martingale concentration results as in Chen et al. (2021) but uses Chebyshev's inequality, however at the cost of losing exponential concentration for polynomial concentration.

1.2 Related research

Contextual bandits in a parametric framework, especially linear bandits have been extensively studied in the bandit literature (Lattimore and Szepesvári, 2020). In a linear bandit framework, every arm corresponds to a known, finite-dimensional context, and its expected reward is assumed to be an unknown linear function of its context. Upper Confidence Bound (UCB) and Thompson sampling strategies are the most commonly used bandit algorithms. While the UCB uses the *optimism in the face of uncertainty* idea where the exploration-exploitation trade-off is balanced by creating confidence sets on the unknown reward functions, Thompson sampling is a randomized algorithm and a popular heuristic based on Bayesian ideas. Both these types of algorithms are extremely popular, well-studied, and enjoy tight regret guarantees (Dani et al., 2008; Li et al., 2010; Abbasi-Yadkori et al., 2011; Agarwal et al., 2012). On the other hand, there are relatively fewer works on ϵ-greedy strategy except for the epoch-greedy version of Langford and Zhang (2007) and the mostly exploration-free algorithms based on Bastani et al. (2021), which have gained a lot of popularity recently. In a nonparametric framework, ϵ-greedy strategies and modifications of the same have been studied by Yang and Zhu (2002); Qian and Yang (2016); Arya and Yang (2020); Qian et al. (2023).

In the contextual bandit framework, there are also two types of setups that are considered in the literature: (a) continuous action space linear bandits: action (arm) space is the same as the context (covariate) space (Abbasi-Yadkori et al., 2011; Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010) and (b) finite action space linear bandits: finite action (arm) set which is different from the context space (Li et al., 2010; Chen et al., 2021). Under these two steps, infinite-dimensional extensions of contexts and/or arms have been considered in both the frequentist and Bayesian perspectives. In the frequentist perspective, Valko et al. (2013) propose a *KernelUCB* algorithm for finite action space, which is obtained by kernelizing the *LinUCB* and *SupLinUCB* algorithm of Li et al. (2010); Chu et al. (2011), and Auer et al. (2002). They give a bound on the regret in terms of a data-dependent quantity, the effective dimension, $\tilde{d}$. In the Bayesian perspective, Srinivas et al. (2010) propose the Gaussian Process (GP)-UCB for the context-free stochastic bandit problem, which assumes that the reward function is drawn from a GP prior. Krause and Ong (2011) generalize the GP-UCB algorithm by taking context information into account in the decision-making process. Gopalan et al. (2014) study the stochastic multi-armed bandit problem with continuous action space and propose an improved version of GP-UCB, which they call *IGP-UCB*. In this work, they also propose a nonparametric version of Thompson sampling, *GP-Thompson sampling*. The regret bounds in all of the Gaussian process bandits line of work are in terms of the quantity, γ_t, which is the maximum information gain at time t and depends on the choice of kernels that define the underlying RKHS containing the reward functions. Moreover, Valko et al. (2013) show $\tilde{d}$ to be closely related to γ_t. Zhou et al. (2020) propose a neural-net-based algorithm, called NeuralUCB, which uses a neural network-based random feature mapping to construct an upper confidence bound (UCB). More recently, Zenati et al. (2022) propose an efficient contextual UCB type algorithm for computational efficiency. This algorithm relies on incremental Nyström approximations of the joint kernel embedding of contexts and actions. Furthermore, Janz et al. (2020) and Vakili et al. (2021a) provide improved regret bounds for GP bandits with Mátern kernels.

Scarlett et al. (2017) and Cai and Scarlett (2021) provide lower bounds for the Gaussian Process bandit optimization problem with squared exponential and Mátern kernel for the contextual bandit problem with continuous action space. Another approach for kernelized contextual bandits is related to experimental design and also aims at optimal pure exploration in kernel bandits as has been studied by Camilleri et al. (2021) and Zhu et al. (2021). A partial list of some of the other unrelated work in nonparametric contextual bandit problem includes Yang and Zhu (2002); Rigollet and Zeevi (2010); Magureanu et al. (2014); Hu et al. (2022); Kleinberg et al. (2008); Slivkins (2014); Zhou et al. (2020).

While UCB and Thompson sampling algorithms for kernel contextual bandits have received considerable attention in the recent past, to the best of our knowledge, the kernelized version of the ϵ-greedy algorithm has not been studied previously. Our setup is similar to the agnostic setup of Valko et al. (2013) for a contextual bandit framework with finitely many arms, where we propose a kernelized version of the ϵ-greedy algorithm, which can be seen as a nonparametric extension to the linear ϵ-greedy algorithm studied by Chen et al. (2021). Our main contribution is a theoretical analysis of this kernelized ϵ-greedy approach validated by some numerical results. The regret bounds we achieve are different from the previous line of work, as our results do not depend on the data-dependent quantities such as $\tilde{d}$ in Valko et al. (2013) or maximum information gain γ_t as in the GP-bandits line of work. One can quantify this information gain for a specific kernel choice and find the corresponding regret rate, see Scarlett et al. (2017). Therefore, for some choices of kernel parameters for the Mátern kernel, it has been observed that one could obtain a linear cumulative regret. On the other hand, our results depend on the intrinsic dimensionality of the RKHS stemming from the assumption on the rate of the eigenvalue decay for the covariance operator and the source condition for the space in which the true mean reward functions are assumed to belong. As a result, we always obtain a sub-linear regret rate irrespective of the kernel choice. Note that, the ϵ-greedy algorithm that we propose uses an inverse probability-weighted online kernel ridge regression estimator. The inverse probability weighting is reminiscent of the inverse propensity score weighted algorithms considered in parametric frameworks to handle model misspecifications. These are known as balanced bandit algorithms (Dimakopoulou et al., 2019) where each observation is divided by its propensity score to correct for estimation bias. Dimakopoulou et al. (2019) propose the balanced linear UCB and balanced Thompson sampling algorithms, establish their corresponding regret rates, and assess empirical performances under model misspecifications. Bogunovic and Krause (2021) study misspecified Gaussian Process bandit optimization, but for continuous action space and they establish regret bounds in terms of the amount of model misspecification. Chen et al. (2021) consider an 'balanced' (weighted) linear ϵ-greedy strategy for handling model-misspecification with weighted least squares estimator. Our work can be thought of as an extension to their work offering flexible and nonparametric modeling for the relationship between the rewards and covariates, where we study a kernelized version of the weighted linear ϵ-greedy algorithm.

1.3 Organization

The rest of the paper is organized as follows. In Section 2.1, we define the notations used in the rest of the paper. In Section 2.2, we introduce the setting of contextual bandits and the online regression framework along with the definition of regret that is used to assess the performance of the proposed algorithm. In Section 3, we present the kernel ϵ-greedy strategy and the online kernel ridge-regression estimator studied in the following sections. In Section 3.3, we provide an implementable version of the regression estimator that is then employed in the proposed algorithm for empirical evaluation on synthetic data sets in Section 6. We provide convergence rates for the estimation error in Section 4 for both the infinite-dimensional and finite-dimensional RKHS settings. In Section 5, we present the regret rates for both the infinite-dimensional and finite-dimensional RKHSs. Under an additional assumption of the margin condition, improved regret bounds are presented for the kernel ϵ-greedy algorithm in Section 5.1. Finally, all the proofs are provided in Section 8.

2 Background and problem setup

In this section, we introduce the notations followed by the problem setup of the sequential decision-making framework of contextual bandits.

2.1 Notations

For a Hilbert Space $\mathcal{H}$, $\langle f, g\rangle_{\mathcal{H}}$ denotes the inner product of $f, g \in \mathcal{H}$. We denote $\|\cdot\|$ or $\|\cdot\|_{\mathcal{H}}$ to denote the corresponding norm in $\mathcal{H}$. For $h \in \mathcal{H}$, we use $\|h\|_{\mathcal{H}} = \sqrt{\langle h, h\rangle_{\mathcal{H}}}$ to denote the RKHS norm and $\|A\|_\infty$ denotes the operator norm of a bounded operator A. For operators A and B on $\mathcal{H}$, $A \preceq B$ if and only if $B - A$ is a positive definite operator. For two real numbers x and y, $x \lesssim y$ denotes that x is less than or equal to y up to a constant factor, and $\otimes$ denotes the tensor product. The notation $X \perp Y$ for two random variables, X and Y, translates to X is independent of Y. $\tilde{O}(\cdot)$ denotes the order of approximation (big-O) with some additional constant terms or terms of logarithmic order in time.

2.2 Problem Setup

In the contextual bandit problem with finitely many *arms*, the decision-maker has $L \in \mathbb{N}$ competing choices of arms (or actions), say $\mathcal{A} := \{1, \ldots, L\}$, and have to choose arms sequentially over time, $\mathcal{T} = \{1, \ldots, T\}$, while using the contextual information available at each time point, with T being the horizon. The contextual information can be thought of as patient characteristics in a treatment allocation problem or user information in a recommender system application. That is, at each time point $t \in \mathcal{T}$, the decision-maker observes contextual information (covariate), $X_t \in \mathbb{R}^d$, from an underlying probability distribution $\mathcal{P}_X$. Now, based on the available information until time t, the decision-maker then chooses an arm a_t from a finite set of arms, $\mathcal{A} = \{1, \ldots, L\}$. Choosing (pulling) the arm a_t results in a *reward*, $y_t \in \mathbb{R}$. The reward can be thought of as a quantitative outcome of assigning

the arm at that time. For instance, it could mean the amount of benefit caused by assigning a particular treatment to a patient at a given time.

In order to make informed decisions, the decision-maker must understand the relationship between the rewards, $\{y_t\}_{t=1}^T$, and covariates, $\{X_t\}_{t=1}^T$, for each arm in $\{1, \ldots, L\}$. This relationship, usually stochastic in nature, can naturally be formulated as a regression problem. In this work, we assume a nonparametric regression framework to model this relationship. For each arm, $i \in 1, \ldots, L$, we consider the following regression model:

$$y_t = f_i(X_t) + e_t, \tag{1}$$

where the corresponding mean reward function $f_i \in \mathcal{H}$, with $\mathcal{H}$ being a reproducing Kernel Hilbert space (RKHS) with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathcal{H}$, where $\mathcal{X}$ is a separable topological space. We assume that k is bounded and continuous, i.e., there exists a $\kappa > 0$, such that $\sup_{x \in \mathcal{X}} k(x, x) \leq \kappa$. Note that, by the reproducing property, $f_i(x_t) = \langle f_i, k(\cdot, x_t) \rangle_H$. We make the following model assumptions:

($\mathcal{A}_1$). *Errors* $\{e_t\}_t$ *conditioned on an arm, i.e.,* $\{e_t|i\}_t$ *are i.i.d. random variables with mean 0 and finite variance,* $\sigma^2 := \mathbb{E}(e_t^2|i) < \infty$ *for* $i = 1, \ldots, L$.

($\mathcal{A}_2$). $e_t \perp x_t | a_t = i$ *for all* $t = 1, \ldots, T$, *and* $i = 1, \ldots, L$.

Note that the above distributional assumptions on the error are weaker than that made in Chen et al. (2021), where the errors are assumed to be sub-Gaussian in addition to satisfying the assumptions ($\mathcal{A}_1$) and ($\mathcal{A}_2$). Higher variability in the noise would lead to larger regret bounds but would not change the rate of convergence for our proposed allocation strategy.

In Section 3.1, we propose an allocation strategy or an algorithm, $\mathcal{B}$, for choosing the next arm based on the sequence of past information on arms played, the covariates, and the rewards obtained respectively. We will build online estimators for the mean reward functions, $f_i, i = 1, \ldots L$, in Section 3.2. Then, we use these estimators to make optimal decisions about arms in a sequential manner. Next, we formulate the notion of *regret* which is a standard way to assess the performance of contextual bandit algorithms.

For covariate $X_t = x_t$, let,

$$a_t^* := \arg\max_{a \in \{1, \ldots, L\}} f_a(x_t)$$

be the true best arm at time t and $f_{a_t^*}(x_t)$ the corresponding best function value. Then, given the previously observed contexts, arms and rewards $\{(X_s, a_s, y_s)\}_{s=1,\ldots,t-1}$ and the current context $X_t = x_t$, a standard goal in a contextual bandit problem is to choose an action a_t in order to minimize the regret (see Definition 1) after T rounds. Let $\mathcal{F}_t = \sigma\langle a_1, x_1, y_1, a_2, x_2, y_2, \ldots, a_t, x_t, y_t \rangle$ denote the sigma-field generated by all information on the covariates observed, arms pulled, and rewards obtained, respectively, until time t.

Definition 1. *The instantaneous regret at time* t *is* $r_t(\mathcal{B}) := f_{a_t^*}(x_t) - f_{a_t}(x_t)$, *where* a_t^* *is the optimal arm at time* t *and* a_t *is the arm chosen by the bandit algorithm,* $\mathcal{B}$, *at time* t. *The cumulative regret* $R_T(\mathcal{B})$ *with horizon* T *is defined as:*

$$R_T(\mathcal{B}) := \sum_{t=1}^{T} (f_{a_t^*}(X_t) - f_{a_t}(X_t)).$$

Note that, the regret as defined above is a random quantity. Thus, we are interested in finding an upper bound for regret in probability or in expectation. Since our method aims at providing model flexibility by allowing us to discover a non-linear relationship between expected rewards and covariates in an RKHS while trying to achieve contextual bandit designs which are less prone to problems of bias, we also study the estimation error. Let $\hat{f}_a$ denote the proposed estimator for f_a. The estimation error at time t is defined as $\|\hat{f}_{a_t} - f_{a_t}\|_{\mathcal{H}}$, where a_t is the arm chosen by the algorithm.

3 Kernel ϵ-greedy algorithm & IPKWR estimator

A simple policy to make sequential decisions in a contextual bandit framework is to be greedy and choose the arm yielding the highest estimated reward for that covariate. However, this could lead to under-exploring some arms thus adversely affecting the performance of the algorithm. A way around this is to use ϵ-greedy, a randomized version of the greedy algorithm, which chooses the best arm with a large probability, i.e., $(1-\epsilon)$ and explores the remaining arms with a small probability, i.e., ϵ. In the following, we propose a kernel ϵ-greedy algorithm that sequentially makes decisions about which arms to play for the contextual bandits' problem as described in Section 2.2.

3.1 Kernel ϵ-greedy algorithm

The proposed algorithm is a kernelized version of the popular ϵ-greedy algorithm for the contextual bandit problem. Let $\{\epsilon_t\}_t$ be a sequence of non-increasing probabilities, such that $\epsilon_t \to 0$ as $t \to \infty$. We denote $\hat{a}_t$ to be the arm chosen by the proposed algorithm at time t, as it depends on all previous data. Below, we describe the kernel ϵ-greedy strategy.

1. **Initialize.** Randomly select among the L arms up to time t_0 for $t = 1, 2, \ldots, t_0$, such that at least one reward per arm is obtained by time t_0.
2. **Estimate f_i.** At time t_0, construct regression estimators for the L arms and denote them by $\hat{f}_{i,t_0}, i = 1, \ldots, L$.
3. **Most promising arm at time t.** For $t = t_0 + 1$, observe covariate $X_t = x_t$ and define:

   $$A_t = \arg\max_{i \in \mathcal{A}} \hat{f}_{i,t-1}(x_t),$$

   be the arm corresponding to the highest estimated value at the current covariate.
4. **ϵ-greedy step.** For a non-increasing probability sequence $\{\epsilon_t\}_t$, the arm pulled is given by the following randomized scheme:

   $$\hat{a}_t = \begin{cases} A_t & \text{with probability } 1 - \epsilon_t \\ \{1, \ldots, L\} \backslash A_t & \text{with probability } \frac{\epsilon_t}{L-1} \end{cases}. \tag{2}$$

5. **Update the estimators.** Corresponding to the arm pulled at time $t = t_0 + 1$, observe reward Y_t and update $\hat{f}_{\hat{a}_t,t}$. For the remaining arms, $i \neq \hat{a}_t$, $\hat{f}_{i,t} = \hat{f}_{i,t-1}$.

6. Repeat steps 3-5 for $t = t_0 + 2$ and so on up to time T.

Note that step 1 is an initialization step, where we randomly assign the L arms until time t_0, such that at least one reward is observed per arm by then. In step 2, we construct regression estimators for each arm using the information gathered during the initialization phase. Step 2 is presented as a generic step as we do not describe how the regression estimator is constructed. We propose an inverse probability weighted kernel ridge regression estimator in Section 3.2 and study the above algorithm for that specific estimator. In step 3, at time $t = t_0 + 1$, we evaluate estimated mean reward functions at the covariate X_t for each arm using the estimators constructed in step 2 and find the arm A_t that maximizes the estimated mean reward. Note that, at this instant, we face the *exploration-exploitation dilemma*. That is, we can either choose the most promising arm, A_t, based on the data available so far or explore the remaining arms, $a \neq A_t$. We use the ϵ-greedy strategy in step 4 in order to balance this trade-off. This is a randomization scheme where we choose the best promising arm A_t with a larger probability $1 - \epsilon_t$ and explore the other arms $a \neq A_t$ with the remaining probabilities $\epsilon_t/(L-1)$. We also assume that $\epsilon_t \leq (L-1)/L$ for $t > t_0$, so that $1 - \epsilon_t \geq \epsilon_t/(L-1)$. Note that, the exploration probabilities $\{\epsilon_t\}_t$ are chosen to be a decreasing sequence of probabilities converging to 0 as $t \to \infty$, hence exploiting more with time. This is because as we accumulate more data, we gain more confidence in our estimates for the mean rewards for each of the arms. Then, the same process is repeated sequentially until we hit the time horizon T. In Section 3.2, we propose an online kernel regression estimator, which we use in this algorithm. Note that, we have a 'hat' on the proposed arm notation, $\hat{a}_t$, to highlight that the choice of the arm is data dependent.

3.2 Inverse probability weighted kernel ridge regression estimator

In this section, we propose an online version of the kernel ridge regression estimator for the mean reward functions $f_i, i = 1, \ldots, L$. Recall, $\mathcal{F}_t = \sigma\langle \hat{a}_1, x_1, y_1, \hat{a}_2, x_2, y_2, \ldots, \hat{a}_t, x_t, y_t \rangle$ denotes the sigma-field generated by all information on the covariates observed, arms pulled, and rewards obtained, respectively, until time t. In order to build an online kernel ridge regression estimator, we solve the following optimization problem with Tikhonov regularization,

$$\hat{f}_{i,t} = \arg\min_{f_i \in \mathcal{H}} \sum_{s=1}^{t} I\{\hat{a}_s = i\}(Y_s - \langle f_i, k(\cdot, X_s)\rangle_{\mathcal{H}})^2 + \lambda \|f_i\|_{\mathcal{H}}^2,$$

where $\lambda > 0$ is the regularization parameter. Using the same ideas as in kernel ridge regression, it is easy to verify that

$$\hat{f}_{i,t} = \left(\sum_{s=1}^{t} I\{\hat{a}_s = i\} k(\cdot, X_s) \otimes k(\cdot, X_s) + \lambda I \right)^{-1} \sum_{s=1}^{t} I\{\hat{a}_s = i\} k(\cdot, X_s) Y_s, \tag{3}$$

where $I\{\hat{a}_s = i\}$ denotes the indicator function which is 1 if the arm chosen by the algorithm at time s is i and is 0 otherwise. While the online kernel-ridge regression estimator in (3) could be a potential candidate to consider, we work with a modified version of this estimator which provides an unbiased estimator of the covariance operator, i.e., $\mathbb{E}[k(\cdot, X) \otimes k(\cdot, X)]$,

as shown in Lemma 1. Specifically, we consider an online Inverse Probability Weighted Kernel Ridge (IPWKR) regression estimator given by:

$$\hat{f}_{i,t} = \left(\frac{1}{t} \sum_{s=1}^{t} \frac{I\{\hat{a}_s = i\}}{P(\hat{a}_s = i|\mathcal{F}_{s-1}, X_s)} k(\cdot, X_s) \otimes k(\cdot, X_s) + \lambda I \right)^{-1} \frac{1}{t} \sum_{s=1}^{t} \frac{I\{\hat{a}_s = i\}}{P(\hat{a}_s = i|\mathcal{F}_{s-1}, X_s)} k(\cdot, X_s) Y_s, \tag{4}$$

for $i = 1, \ldots, L$. Note that, in the above algorithm, the probability weights $P(\hat{a}_s = i|\mathcal{F}_{s-1}, X_s)$ are known at time s. More specifically, since we know A_s and ϵ_s at time step s, $P(\hat{a}_s = A_s|\mathcal{F}_{s-1}, X_s) = 1 - \epsilon_s$ and $P(\hat{a}_s = a|\mathcal{F}_{s-1}, X_s) = \epsilon_s/(L-1)$ for $a \neq A_s$, therefore $\hat{f}_{i,s}$ for $i = 1, \ldots, L$ and for $s = 1, \ldots, t$ are data-determined estimates and can be used for estimation at the $(t+1)^{\text{th}}$ time. Note that the definitions of the arm pulled (see (2)) and the online IPWKR estimator in (4) depend on each other. Since the data are not independent, the consistency of the estimator does not follow immediately from the classical tools in kernel methods. Therefore, one of our contributions is in analyzing the estimation error associated with the online IPWKR estimator, and establish its consistency and rate of convergence.

From the proposed estimator, it is easy to see that a natural candidate for the estimator of the covariance operator is:

$$\hat{\Sigma}_{i,t} = \frac{1}{t} \sum_{s=1}^{t} \frac{I\{\hat{a}_s = i\}}{P(\hat{a}_s = i|\mathcal{F}_{s-1}, X_s)} k(\cdot, X_s) \otimes k(\cdot, X_s), \tag{5}$$

which can be shown (see Lemma 1) to be an unbiased estimator of the covariance operator $\Sigma := \mathbb{E}(k(\cdot, X_s) \otimes k(\cdot, X_s))$. We highlight that the unbiasedness of the covariance estimator is critical in our analysis. In the following, we re-write the proposed kernel ϵ-greedy algorithm of Section 3.1 in an implementable format.

3.3 Implementation of kernel ϵ-greedy algorithm

In this section, we devise an implementable version of the proposed strategy. To this, we define the following.

- Let $S_{t,X} : \mathcal{H} \to \mathbb{R}^t$ be the sampling operator, such that $S_{t,X} f = \frac{1}{\sqrt{t}}[f(X_1), \ldots, f(X_t)]^\top$.
- The reconstruction operator is given by $S^*_{t,X} : \mathbb{R}^t \to \mathcal{H}$, where $S^*_{t,X} \underline{\alpha} = \frac{1}{\sqrt{t}} \sum_{s=1}^{t} \alpha_s k(\cdot, X_s)$, for $\underline{\alpha} \in \mathbb{R}^t$.
- $S_{t,X} S^*_{t,X} = \frac{K_t}{t} : \mathbb{R}^t \to \mathbb{R}^t$, where K_t is the kernel/Gram matrix.

We express the empirical covariance operator in (5) in terms of these operators. Let $\Lambda_{i,t}$ be a diagonal matrix in $\mathbb{R}^{t \times t}$ with diagonal elements given by,

$$\left\{ w_{s,i} := \frac{I\{\hat{a}_s = i\}}{P(\hat{a}_s = i|\mathcal{F}_{s-1}, X_s)}, s = 1, \ldots, t \right\}.$$

Then note that, $\hat{\Sigma}_{i,t} = S^*_{t,X}\Lambda_{i,t}S_{t,X}$ and,

$$\hat{\Sigma}_{i,t}f = \frac{1}{t}\sum_{s=1}^{t} w_{s,i}f(X_s)k(\cdot, X_s) \text{ for all } i = 1, \ldots, L.$$

Let $Y_t = (y_1, \ldots, y_t)'$. Then, the proposed estimator in (4) can be written as:

$$\begin{aligned}
\hat{f}_{i,t} &= \frac{1}{\sqrt{t}}(S^*_{t,X}\Lambda_{i,t}S_{t,X} + \lambda I)^{-1}\frac{1}{\sqrt{t}}\sum_{s=1}^{t} w_{s,i}k(\cdot, X_s)y_s \\
&= \frac{1}{\sqrt{t}}(S^*_{t,X}\Lambda_{i,t}S_{t,X} + \lambda I)^{-1}S^*_{t,X}\Lambda_{i,t}Y_t \\
&= \frac{1}{\sqrt{t}}S^*_{t,X}(\Lambda_{i,t}S_{t,X}S^*_{t,X} + \lambda I)^{-1}\Lambda_{i,t}Y_t,
\end{aligned}$$

where the last equality follows from the fact that,

$$(S^*_{t,X}\Lambda_{i,t}S_{t,X} + \lambda I)^{-1}S^*_{t,X} = S^*_{t,X}(\Lambda_{i,t}S_{t,X}S^*_{t,X} + \lambda I)^{-1}.$$

Therefore we obtain

$$\hat{f}_{i,t} = \frac{1}{\sqrt{t}}S^*_{t,X}\left(\Lambda_{i,t}\frac{K_t}{t} + \lambda I\right)^{-1}\Lambda_{i,t}Y_t, \text{ for } i = 1, \ldots, L. \tag{6}$$

Then, using the definition of $S^*_{t,X}$, the estimated reward function value at X_{t+1} for arm i is given by:

$$\begin{aligned}
\hat{f}_{i,t}(X_{t+1}) &= \frac{1}{t}\bar{k}^\top_{t+1}\left(\Lambda_{i,t}\frac{K_t}{t} + \lambda_t I\right)^{-1}\Lambda_{i,t}Y_t \\
&= \bar{k}^\top_{t+1}(\Lambda_{i,t}K_t + t\lambda_t I)^{-1}\Lambda_{i,t}Y_t \text{ for } i = 1, \ldots, L,
\end{aligned} \tag{7}$$

where, $\bar{k}_{t+1} = (k(X_1, X_{t+1}), k(X_2, X_{t+1}), \ldots, k(X_t, X_{t+1}))^\top$. Note that (7) involves only inverting a $t \times t$ matrix and therefore this version of the estimator is implementable. In order to facilitate faster computation, we use SVD for finding the inverse in (7).

4 Estimation error: Convergence rates

In this section, we present the theoretical results for the proposed algorithm for which the proofs can be found in Section 8. We make the following assumptions throughout this paper whenever we are working under the assumption that the mean reward functions lie in an RKHS $\mathcal{H}$.

($\mathcal{A}_3$). *$\eta_i(\Sigma) \leq \bar{C}i^{-\alpha}, \alpha > 1$ where $\eta_i(\Sigma)$ denotes the i^{th} eigenvalue of $\Sigma = \mathbb{E}(k(\cdot, X_s) \otimes k(\cdot, X_s))$ and $\bar{C} \in (0, \infty)$.*

($\mathcal{A}_4$). *For all $i = 1, \ldots, L$, $f_i \in \text{Ran}(\Sigma^{\gamma_i})$, $0 < \gamma_i \leq \frac{1}{2}$, i.e., there exists $h \in \mathcal{H}$ such that $f_i = \Sigma^{\gamma_i}h$ for $i = 1, \ldots, L$.*

Algorithm 1 Kernel ϵ-greedy algorithm

1: Randomly select arms $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_{t_0} \in \mathcal{A} = \{1, \ldots, L\}$ until each arm is selected at least once.
2: **for** $t = t_0 + 1, \ldots, T$ **do**
3: Estimate $\hat{f}_{i,t-1}(X_t)$, for each $i = \mathcal{A}$ using (7).
4: Calculate the best-performing arm so far: $A_t = \arg\max_{i \in \mathcal{A}} \hat{f}_{i,t-1}(X_t)$.
5: For a non-increasing exploration probability sequence $\{\epsilon_t, t \geq 1\}$, the arm pulled is given by:

$$\hat{a}_t = \begin{cases} A_t & \text{with probability } 1 - \epsilon_t \\ \{1, \ldots, L\} \backslash A_t & \text{with probability } \frac{\epsilon_t}{L-1}. \end{cases}$$

6: Observe reward Y_t corresponding to $\hat{a}_t$.
7: For $i = \hat{a}_t$, update $\hat{f}_{i,t}$ using (6) and use $\hat{f}_{i,t} = \hat{f}_{i,t-1}$ for $i \in \mathcal{A} \backslash \hat{a}_t$.
8: **end for**

Note that, $(\mathcal{A}_3)$ implies that the effective dimension, $N_{\Sigma,1}(\lambda) := \mathrm{Tr}\left((\Sigma + \lambda I)^{-1}\Sigma\right) \lesssim \lambda^{-1/\alpha}$, which controls the complexity of $\mathcal{H}$ and $(\mathcal{A}_4)$ determines the smoothness of the true mean reward functions. Next, in Theorem 1, which is proved in Section 8.1, we present an upper bound on the estimation error both in probability and in expectation.

Theorem 1. *Suppose* $(\mathcal{A}_1)$*–*$(\mathcal{A}_4)$ *hold and* $\{\epsilon_s\}_{s=1}^t$ *is such that for any* $\delta > 0$ *and* $t \geq 1$*,*

$$\lambda_{i,t} = \left[\frac{1}{\delta t^2}\left(\sum_{s=1}^{t} \frac{1}{\epsilon_s}\right)\right]^{\alpha/(2\gamma_i\alpha+\alpha+1)}, \ i = 1, \ldots, L, \tag{8}$$

satisfies

$$\lambda_{i,t} \geq \left[\frac{4(L-1)\kappa A_1(\bar{C}, \alpha)}{\delta t^2}\left(\sum_{s=1}^{t} \frac{1}{\epsilon_s}\right)\right]^{\alpha/(1+\alpha)}. \tag{9}$$

Then, the following holds with probability at least $1 - 2\delta$*:*

$$\|\hat{f}_{i,t} - f_i\|_{\mathcal{H}} \leq 2\sqrt{2}\max\{C_0, C_i\}\left[\frac{1}{\delta t^2}\left(\sum_{s=1}^{t} \frac{1}{\epsilon_s}\right)\right]^{\gamma_i\alpha/(2\gamma_i\alpha+\alpha+1)}, \ i = 1, \ldots, L. \tag{10}$$

Furthermore, for $0 \leq \zeta < \frac{\gamma_i\alpha+\alpha+1}{\gamma_i\alpha}$*,*

$$\mathbb{E}[\|\hat{f}_{i,t} - f_i\|_{\mathcal{H}}^{1+\zeta}] \leq B(C_0, C_i, \gamma_i, \zeta, \alpha)\left[\frac{1}{t^2}\sum_{s=1}^{t} \frac{1}{\epsilon_s}\right]^{w_i(1+\zeta)}, \ i = 1, \ldots, L, \tag{11}$$

where $C_0 = \sqrt{\sigma^2(L-1)A_1(\bar{C}, \alpha)}$*,* $A_1(\bar{C}, \alpha) = \bar{C}^{-1/\alpha}\int_0^\infty (1 + x^\alpha)^{-1}\,dx$*,* $C_i = \|\Sigma^{-\gamma_i} f_i\|_{\mathcal{H}}$*,* $w_i = \frac{\gamma_i\alpha}{2\gamma_i\alpha+\alpha+1}$*, and* $B(C_0, C_i, \gamma_i, \zeta, \alpha)$ *is a constant depending only on its arguments and not on* t *and* $\{\epsilon_s\}_s$*.*

Note that the upper bound on the estimation error depends on the exploration probability sequence $\{\epsilon_s\}_{s=1}^t$, and for consistent estimation, we require $\sum_{s=1}^t \epsilon_s^{-1} = o(t^2)$ as $t \to \infty$. Under this requirement of $\sum_{s=1}^t \epsilon_s^{-1} = o(t^2)$ as $t \to \infty$, it is easy to verify that (9) holds. Also, we would like to highlight that the restriction on ζ appearing in Theorem 1 is due to polynomial (precisely, quadratic) concentration in (10), which if improved to sub-exponential or sub-Gaussian concentration would remove the upper bound restriction on ζ.

In order to compare Theorem 1 with Proposition 4.1 in Chen et al. (2021), we assume that the underlying RKHS, $\mathcal{H}$, is finite-dimensional, and make the following assumption instead of $(\mathcal{A}_4)$:

$(\mathcal{A}_5)$**.** *The minimum eigenvalue of* Σ*, denoted as* $\eta_{\min}(\Sigma)$ *satisfies* $\eta_{\min}(\Sigma) > \eta$ *for some* $\eta > 0$*.*

Since Chen et al. (2021) study linear contextual bandits which are equivalent to our approach when the kernel is linear, i.e., the corresponding RKHS is finite-dimensional, we would like to specialize Theorem 1 to finite-dimensional RKHS. This assumption of finite-dimensional RKHS is imposed through $(\mathcal{A}_5)$, which is also assumed in Chen et al. (2021).

Theorem 2. *Suppose* $(\mathcal{A}_1)$*,* $(\mathcal{A}_2)$ *and* $(\mathcal{A}_5)$ *hold. For any* $\delta > 0$ *and* $t \geq 1$*, suppose* $\{\epsilon_s\}_{s=1}^t$ *satisfies*

$$\frac{1}{t^2} \sum_{s=1}^{t} \frac{1}{\epsilon_s} \leq \frac{\delta \eta}{4(L-1)d\kappa}, \tag{12}$$

where $d := \dim(\mathcal{H})$*. Then, for any* $\delta > 0$ *and* $t \geq 1$*, with the choice of*

$$\lambda_{i,t} = \left[\frac{1}{\delta t^2} \left(\sum_{s=1}^{t} \frac{1}{\epsilon_s} \right) \right]^{1/2}, \ i = 1, \ldots, L,$$

the following holds with probability at least $1 - 2\delta$*:*

$$\|\hat{f}_{i,t} - f_i\|_{\mathcal{H}} \leq 4 \max\{\tilde{C}_0, \tilde{C}_i\} \left[\frac{1}{\delta t^2} \left(\sum_{s=1}^{t} \frac{1}{\epsilon_s} \right) \right]^{1/2}, \ i = 1, \ldots, L.$$

Moreover, for $0 \leq \zeta < 1$*,*

$$\mathbb{E}[\|\hat{f}_{i,t} - f_i\|_{\mathcal{H}}^{1+\zeta}] \leq B(\tilde{C}_0, \tilde{C}_i, \zeta, \eta) \left[\frac{1}{t^2} \sum_{s=1}^{t} \frac{1}{\epsilon_s} \right]^{\frac{1+\zeta}{2}}, \ i = 1, \ldots, L, \tag{13}$$

where $\tilde{C}_0 := \sqrt{\frac{(L-1)d\sigma^2}{\eta}}$ *and* $\tilde{C}_i := \frac{\|f_i\|_{\mathcal{H}}}{\eta}$*. Here* $B(\tilde{C}_0, \tilde{C}_i, \zeta, \eta)$ *is a constant that depends only on its arguments and not on* t *and* $\{\epsilon_s\}_s$*.*

Note that in the above result, the choice of $\lambda_{i,t}$ is independent of i unlike in the infinite-dimensional case of Theorem 1. Moreover, unlike in Theorem 1, the choice of $\lambda_{i,t}$ in Theorem 2 has an exponent of $\frac{1}{2}$ instead of the term depending on γ_i and α in (10). Clearly, the estimators are consistent if $\sum_{s=1}^t \epsilon_s^{-1} = o(t^2)$ as $t \to \infty$. Under this requirement on $\{\epsilon_s\}_s$, it is easy to verify that (12) holds. The proof of Theorem 2 is provided in Section 8.2.

Remark 1. *Chen et al. (2021) proposed a weighted online least squares estimator similar to the IPWKR estimator for the linear contextual bandit problem with 2 arms ($L = 2$) and a finite-dimensional context space. However, unlike our estimator, they study an unregularized online weighted least squares (WLS) estimator. Under the assumptions of (i) bounded covariates, (ii) reliability of linear approximation, and (iii) minimum eigenvalue are bounded from below, they provide a bound (see Proposition 4.1) on the estimation error, which can be summarized to behave as*

$$\tilde{O}\left(\left[\frac{d^2\log(4d/\delta)}{t\epsilon_t^4}\right]^{1/2}\right) \tag{14}$$

with probability at least $1-\delta$. On the other hand, our method provides a bound of

$$\tilde{O}\left(\left[\frac{d}{\delta}\frac{1}{t^2}\Big(\sum_{s=1}^{t}\frac{1}{\epsilon_s}\Big)\right]^{1/2}\right), \tag{15}$$

which also holds with probability $1-\delta$. By comparing (14) and (15), it can be observed that the bound in Chen et al. (2021) holds with sub-Gaussian concentration while ours holds with polynomial concentration. Another key distinction is the requirement of $\frac{1}{\epsilon_t} = o(t^{1/4})$ in (14) vs. $\sum_{s=1}^{t}\epsilon_s^{-1} = o(t^2)$ in (15) for the estimators to be consistent, as $t \to \infty$. Clearly, a sufficient condition for our estimator to be consistent is $\frac{1}{\epsilon_t} = o(t)$. Therefore, our result is stronger than that of Chen et al. (2021), as the exploration probability can decay at a faster rate while still guaranteeing the consistency of our estimator. This is promising as for situations where the reward functions are relatively easier to learn, more exploitative strategies can significantly reduce the regret. Another key highlight of our work is that these improved results are obtained for any f_i belonging to a finite-dimensional RKHS, that is not necessarily linear.

5 Regret analysis

In this section, we construct upper bounds for the regret defined in Definition 1 for the algorithm proposed in Section 3 that involves using the IPWKR estimator studied in Section 3.2. Recall, $a_t^* = \arg\max_{a\in\mathcal{A}} f_a(x_t)$, where $f_a(x_t) = \langle f_a, k(\cdot, x_t)\rangle_{\mathcal{H}}$. Let $A_t = \arg\max_{a\in\mathcal{A}}\hat{f}_a(x_t)$ and note that by definition of A_t, $\hat{f}_{a_t^*}(X_t) \leq \hat{f}_{A_t}(X_t)$. Then, the cumulative regret for the

proposed algorithm up to some time horizon T is given by,

$$
\begin{aligned}
R_T &= \sum_{t=1}^{T} f_{a_t^*}(X_t) - f_{\hat{a}_t}(X_t) = \sum_{t=1}^{t_0} f_{a_t^*}(X_t) - f_{\hat{a}_t}(X_t) + \sum_{t=t_0+1}^{T} f_{a_t^*}(X_t) - f_{\hat{a}_t}(X_t) \\
&= \sum_{t=1}^{t_0} \langle f_{a_t^*} - f_{\hat{a}_t}, k(\cdot, X_t)\rangle_{\mathcal{H}} + \sum_{t=t_0+1}^{T} f_{a_t^*}(X_t) - f_{\hat{a}_t}(X_t) \\
&\leq \sum_{t=1}^{t_0} \kappa \| f_{a_t^*} - f_{\hat{a}_t}\|_{\mathcal{H}} + \sum_{t=t_0+1}^{T} f_{a_t^*}(X_t) - f_{\hat{a}_t}(X_t) \leq \Lambda t_0 + \sum_{t=t_0+1}^{T} f_{a_t^*}(X_t) - f_{\hat{a}_t}(X_t) \\
&= \Lambda t_0 + \sum_{t=t_0+1}^{T} \Big[f_{a_t^*}(X_t) - \hat{f}_{a_t^*}(X_t) + \hat{f}_{a_t^*}(X_t) - f_{A_t}(X_t) + f_{A_t}(X_t) - f_{\hat{a}_t}(X_t) \Big] \\
&\leq \Lambda t_0 + \sum_{t=t_0+1}^{T} \Big[f_{a_t^*}(X_t) - \hat{f}_{a_t^*}(X_t) + \hat{f}_{A_t}(X_t) - f_{A_t}(X_t) + f_{A_t}(X_t) - f_{\hat{a}_t}(X_t) \Big] \\
&\leq \Lambda t_0 + 2 \underbrace{\sum_{t=t_0+1}^{T} \sup_{a \in \mathcal{A}} |(f_a(X_t) - \hat{f}_{a,t}(X_t))|}_{\text{Cumulative estimation error}} + \underbrace{\sum_{t=t_0+1}^{T} |f_{A_t}(X_t) - f_{\hat{a}_t}(X_t)|}_{\text{Randomization error}}, \qquad (16)
\end{aligned}
$$

where $\Lambda := \sup\{\|f_a - f_{a'}\|_{\mathcal{H}} : a, a' \in \mathcal{A}, a \neq a'\}$, and for simplicity, we did not put the algorithm within parenthesis for R_T (as in Definition 1) though all the presented results are for the proposed kernel ϵ-greedy algorithm. Note that the first term in (16) is the regret incurred due to the random initialization up to time t_0. We call the second term on the right-hand side in (16) as *cumulative estimation error*, as it measures the error in estimating the function accumulated over time, and the third term as *randomization error* since it measures the error incurred due to the randomization scheme (ϵ-greedy in step 4 of the proposed algorithm in Section 3). Note that, the initialization phase regret can be trivially bounded by $O(t_0)$, but the regret incurred during the post-initialization phase dominates over the regret incurred over the initialization phase. Hence, without the loss of generality, we set $t_0 = 0$ in the following results and in the corresponding proofs.

Theorem 3. *Suppose* $(\mathcal{A}_1)$–$(\mathcal{A}_4)$ *hold,* $\sup_{\substack{a,a' \in \mathcal{A} \\ a \neq a'}} \|f_a - f_{a'}\|_{\mathcal{H}} < \infty$, *and* $\{\epsilon_s\}_{s=1}^{T}$ *is such that for any* $\delta > 0$, *and* $T \geq 1$,

$$
\lambda_{i,t} = \left[\frac{L}{\delta t^2} \left(\sum_{s=1}^{t} \frac{1}{\epsilon_s} \right) \right]^{\alpha/(2\gamma_i \alpha + \alpha + 1)}, \; i \in \{1, \ldots, L\}, \; 0 < t \leq T, \qquad (17)
$$

satisfies

$$
\lambda_{i,t} \geq \left[\frac{4L(L-1)\kappa A_1(\bar{C}, \alpha)}{\delta t^2} \left(\sum_{s=1}^{t} \frac{1}{\epsilon_s} \right) \right]^{\alpha/(1+\alpha)} \quad i \in \{1, \ldots, L\}, \; 0 < t \leq T.
$$

Define

$$
\Delta_t = \frac{L}{\delta t^2} \sum_{s=1}^{t} \frac{1}{\epsilon_s}.
$$

Then, for $\delta > 0$, $1 \leq p \leq \infty$ and $T \geq 1$, the following holds with probability at least $1 - 2\delta$:

$$R_T \leq \kappa\Theta \sum_{t=1}^{T} \left(I\{\Delta_t < 1\} \Delta_t^{\frac{(\min_{i\in\mathcal{A}} \gamma_i)\alpha}{2(\min_{i\in\mathcal{A}} \gamma_i)\alpha+\alpha+1}} + I\{\Delta_t \geq 1\} \Delta_t^{\frac{(\max_{i\in\mathcal{A}} \gamma_i)\alpha}{2(\max_{i\in\mathcal{A}} \gamma_i)\alpha+\alpha+1}} \right)$$
$$+ \kappa T^{1-\frac{1}{p}} \left[\sum_{t=1}^{T} \frac{\epsilon_t}{L-1} + \left\{ \frac{1}{\delta} \sum_{t=1}^{T} \frac{\epsilon_t}{L-1} \right\}^{1/2} \right]^{1/p} \sup_{\substack{a,a'\in\mathcal{A} \\ a\neq a'}} \|f_a - f_{a'}\|_{\mathcal{H}}, \tag{18}$$

where $\Theta = 4\sqrt{2}\max\{C_0, \max_{i\in\mathcal{A}} C_i\}$ with C_0, C_i and $A_1(\bar{C}, \alpha)$ being defined in Theorem 1.

The first term in the r.h.s. of (18) corresponds to the estimation error accumulated by time T and the second term corresponds to the randomization error. The cumulative estimation error bound follows from the estimation error analysis in Theorem 1, after applying a union bound over the arms. Note that for a given sequence of exploration probability, the cumulative estimation error and the randomization error behave inversely to each other. Further, note that if $\sum_{s=1}^{t} \epsilon_s^{-1} = o(t^2)$ as $t \to \infty$, then clearly $\Delta_t < 1$ for some large enough t. Let $t_1 = \min_t I\{\Delta_t < 1\}$. Then, we can choose the initialization phase endpoint to be $\tilde{t}_0 = \max\{t_0, t_1\}$, which means only the first term in the accumulated estimation error would contribute to the regret bound. This reflects that the estimation error incurred due to the arm with the lowest smoothness parameter for the respective reward function dominates.

Remark 2. *Suppose $\epsilon_t = t^{-\beta}$ for some $0 < \beta < 1$ and $t \in \mathbb{N}$, which satisfies the requirement that $\sum_{s=1}^{t} \epsilon_s^{-1} \lesssim t^{\beta+1} = o(t^2)$ as $t \to \infty$. Then under the assumptions of Theorem 3, for large enough T,* (18) *reduces to*

$$R_T \lesssim T^{(\beta-1)w+1} + T^{1-\frac{\beta}{p}},$$

for the choice of $\lambda_{i,t}$ as in (17) *and*

$$w = \frac{(\min_{i\in\mathcal{A}} \gamma_i)\alpha}{2(\min_{i\in\mathcal{A}} \gamma_i)\alpha + \alpha + 1}.$$

Since $1 \leq p \leq \infty$ is arbitrary, the best regret is achieved at $p = 1$, yielding

$$R_T \lesssim T^{(\beta-1)w+1} + T^{1-\beta}. \tag{19}$$

In (19)*, the first term corresponds to an upper bound on the cumulative estimation error, which is increasing in β and the second term corresponds to an upper bound on the randomization error, which is decreasing in β. By balancing these two terms, the optimal choice of β is given by*

$$\beta = \frac{w}{w+1} = \frac{(\min_{i\in\mathcal{A}} \gamma_i)\alpha}{3(\min_{i\in\mathcal{A}} \gamma_i)\alpha + \alpha + 1} =: \beta^*.$$

This means the optimal exploration probability sequence depends on the smoothness of the targets and the intrinsic dimensionality of $\mathcal{H}$, which is controlled by α. Clearly, if $\beta > \beta^*$, i.e., exploitation is favored over exploration, the estimation error dominates the randomization error and the converse happens when $\beta < \beta^*$, i.e., exploration is favored over exploitation. Note that for any choice of γ_i and α, we have $\beta^* \leq \frac{1}{5}$, which implies the growth rate of regret is at least of the order $T^{4/5}$.

Next, in order to compare with the regret rate in Section 7.1 of Chen et al. (2021) for the ϵ-greedy strategy, we specialize Theorem 3 to a finite-dimensional setting. While the randomization error bound can be obtained similarly as the second term in (18), we modify the bounding for the cumulative estimation error (see the proof in Section 8.4 for details).

Theorem 4. *Suppose* $(\mathcal{A}_1)$, $(\mathcal{A}_2)$ *and* $(\mathcal{A}_5)$ *hold, and* $\sup_{\substack{a,a'\in\mathcal{A}\\ a\neq a'}} \|f_a - f_{a'}\|_{\mathcal{H}} < \infty$. *For any* $\delta > 0$, *suppose* $\{\epsilon_s\}_{s=1}^T$ *satisfies*

$$\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s} \leq \frac{\delta\eta}{4L(L-1)d\kappa}$$

for all $0 < t \leq T$, *where* $d := \dim(\mathcal{H})$. *Then, for any* $\delta > 0$, $1 \leq p \leq \infty$, $T \geq 1$, *and*

$$\lambda_{i,t} = \left[\frac{L}{\delta t^2}\left(\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)\right]^{1/2},\ i \in \{1,\ldots,L\},\ 0 < t \leq T,$$

the following holds with probability at least $1 - 2\delta$:

$$\begin{aligned} R_T \leq\ & 8\kappa \max\{\tilde{C}_0, \tilde{C}_*\}\sum_{t=1}^{T}\left[\frac{L}{\delta t^2}\Big(\sum_{s=1}^{t}\frac{1}{\epsilon_s}\Big)\right]^{1/2} \\ & + \kappa T^{1-\frac{1}{p}}\left[\sum_{t=1}^{T}\frac{\epsilon_t}{L-1} + \left\{\frac{1}{\delta}\sum_{t=1}^{T}\frac{\epsilon_t}{L-1}\right\}^{1/2}\right]^{1/p}\sup_{\substack{a,a'\in\mathcal{A}\\ a\neq a'}}\|f_a - f_{a'}\|_{\mathcal{H}}, \end{aligned} \tag{20}$$

where $\tilde{C}_0 = \sqrt{\frac{(L-1)d\sigma^2}{\eta}}$ *and* $\tilde{C}_* = \max_{1\leq i\leq L}\frac{\|f_i\|_{\mathcal{H}}}{\eta}$.

Remark 3. *As in Remark 2, the choice of* $\epsilon_t = t^{-\beta}$ *for some* $0 < \beta < 1$ *reduces* (20) *to*

$$R_T \lesssim T^{\frac{\beta+1}{2}} + T^{1-\beta},$$

Balancing these terms yields that R_T *has a growth order of at least* $T^{2/3}$ *with the choice of* $\beta = \frac{1}{3}$. *Note that the estimation error (resp. randomization error) dominates the randomization error (resp. estimation error) if* $\beta > \frac{1}{3}$ *(resp.* $\beta < \frac{1}{3}$*). We would like to highlight that this rate of* $T^{2/3}$ *is optimal for contextual bandits using an* ϵ*-greedy strategy, i.e., this strategy cannot achieve regret rates slower than* $T^{2/3}$ *for contextual bandits (Dann et al., 2022, Theorem 3).*

5.1 Regret analysis with the margin condition

In this section, we make an additional assumption known as the 'margin condition' on the underlying reward functions as is commonly assumed in the contextual bandit's literature (Chen et al., 2021; Goldenshluger and Zeevi, 2013). Under this assumption, we can achieve significant improvement in the expected regret rate as compared to the previous results. Note that, here we construct regret upper bounds in expectation unlike the previous results in Section 5 where we constructed upper bounds on the regret in probability. As in the literature, we focus our analysis on $L = 2$ (two arms) though it can be extended to L arms.

($\mathcal{A}_6$). ***Margin Condition:*** *There exists* $C > 0$ *such that* $P_{X\sim\mathcal{P}_X}\left(0 < |\langle f_1 - f_0, k(\cdot, X)\rangle_{\mathcal{H}}| \le l\right) \le Cl,\ \forall l > 0.$

The assumption is related to the behavior of the distribution of the covariates near the decision boundary $\{x : f_1(x) = f_0(x)\}$. As it is difficult to distinguish between the arms near the boundary, imposing such an assumption helps to control the contribution of incorrect decisions being made near the decision boundary. In the following theorem, we present an upper bound on the expected regret for the kernel ϵ-greedy algorithm when the true mean reward functions lie in an RKHS $\mathcal{H}$.

Theorem 5. *Let* $L = 2$. *Suppose* ($\mathcal{A}_1$)–($\mathcal{A}_4$) *and* ($\mathcal{A}_6$) *hold with* $\gamma_0 = \gamma_1 = \gamma$, $\|f_1 - f_0\|_{\mathcal{H}} < \infty$ *and* $\{\epsilon_s\}_{s=1}^T$ *is such that for* $\delta > 0$, *and* $T \ge 1$,

$$\lambda_t = \left[\frac{1}{\delta t^2}\left(\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)\right]^{\alpha/(2\gamma\alpha+\alpha+1)},\ 0 < t \le T,$$

satisfies

$$\lambda_t \ge \left[\frac{4(L-1)\kappa A_1(\bar{C},\alpha)}{\delta t^2}\left(\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)\right]^{\alpha/(1+\alpha)},\ 0 < t \le T.$$

Then, for $T \ge 1$, $\theta > 0$, *and* $0 \le \zeta < (\gamma\alpha+\alpha+1)/\gamma\alpha$, *the following holds:*

$$\begin{aligned}\mathbb{E}R_T \le{}& \kappa\|f_1 - f_0\|_{\mathcal{H}}\sum_{t=1}^{T}\frac{\epsilon_t}{2}\\ &+ A_0(\zeta,\kappa,C_0,C,C_*,\gamma,\alpha)\left[T^{-\theta}\sum_{t=1}^{T}\left(\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)^{w} + T^{\theta\zeta}\sum_{t=1}^{T}\left(\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)^{w(1+\zeta)}\right],\end{aligned} \tag{21}$$

where $A_0(\zeta,\kappa,C_0,C,C_*,\gamma,\alpha)$ *is a constant that depends only on its arguments and not on* T *and* $\{\epsilon_s\}_s$, $w = \alpha\gamma/(2\alpha\gamma+\alpha+1)$, $C_* = \max_{1\le i\le L}\|\Sigma^{-\gamma_i}f_i\|_{\mathcal{H}}$, C_0 *and* $A_1(\bar{C},\alpha)$ *are defined in Theorem 1, and* C *is defined in* ($\mathcal{A}_6$).

Remark 4. *As in Remark 2, the choice of* $\epsilon_t = t^{-\beta}$ *for some* $0 < \beta < 1$ *and* $0 \le \zeta < (1-w)/w$ *for* $w = \alpha\gamma/(2\alpha\gamma+\alpha+1)$, *reduces* (21) *to*

$$\mathbb{E}R_T \lesssim \underbrace{T^{1-\beta}}_{\text{I}} + \underbrace{T^{-\theta+(\beta-1)w+1}}_{\text{II}} + \underbrace{T^{(\beta-1)(1+\zeta)w+\theta\zeta+1}}_{\text{III}}.$$

To get the final rate, we first balance II *and* III, *and then balance the resulting rate and* I. *Note that* II *is decreasing in* θ *while* III *is increasing in* θ, *and they are balanced when*

$$\theta = \Big(\frac{\zeta}{1+\zeta}\Big)w(1-\beta),$$

resulting in

$$\text{IV} := \text{II} + \text{III} = T^{(\beta-1)w\frac{(1+2\zeta)}{(1+\zeta)}+1}.$$

Clearly, IV *and* I *are increasing and decreasing functions of* β *for* $0 < \beta < 1$*, respectively, which are balanced when*

$$\zeta = \frac{\beta - (1-\beta)w}{2w(1-\beta) - \beta}.$$

For ζ *to satisfy*

$$0 \le \zeta < \frac{1-w}{w},$$

we need the condition

$$\frac{w}{w+1} \le \beta < \frac{2w}{2w+1}.$$

Therefore, ϵ*-greedy algorithm achieves* $\mathbb{E}R_T = O(T^{1-\beta})$ *as* $T \to \infty$ *for* $w/(w+1) \le \beta < 2w/(2w+1)$*. This means the best regret rate that we can achieve is* $T^{\frac{2}{3}+\varepsilon}$ *for any* $\varepsilon > 0$ *when* $\gamma = \frac{1}{2}$ *and* $\alpha \to \infty$ *with the exploration sequence being chosen to satisfy* $\epsilon_t = t^{-\beta}$*,* $\frac{1}{5} \le \beta < \frac{1}{3}$*. In contrast to Remark 2 where the regret rate is at least* $T^{4/5}$*, the margin condition in* $(\mathcal{A}_6)$ *improves the best rate to* $T^{\frac{2}{3}+\varepsilon}$*,* $\varepsilon > 0$*.*

Next, we bound the expected cumulative regret in the finite-dimensional case and show it is almost minimax optimal.

Theorem 6. *Let* $L = 2$*. Suppose* $(\mathcal{A}_1)$*,* $(\mathcal{A}_2)$*,* $(\mathcal{A}_5)$*, and* $(\mathcal{A}_6)$ *hold,* $\|f_1 - f_0\|_{\mathcal{H}} < \infty$*, and for any* $\delta > 0$*,* $T \ge 1$*,* $\{\epsilon_s\}_{s=1}^T$ *satisfies*

$$\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s} \le \frac{\delta\eta}{4(L-1)d\kappa},\ 0 < t \le T,$$

where $d := \dim(\mathcal{H})$*. Then for* $\delta > 0$*,* $T \ge 1$*,* $\theta > 0$*,* $0 \le \zeta < 1$*, and*

$$\lambda_t = \left[\frac{1}{\delta t^2}\left(\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)\right]^{1/2},\ 0 < t \le T, \tag{22}$$

the following holds:

$$\begin{aligned}\mathbb{E}R_T \le{}& \kappa\|f_1 - f_0\|_{\mathcal{H}}\sum_{t=1}^{T}\frac{\epsilon_t}{2} \\ &+ \tilde{A}(\zeta,\kappa,\tilde{C}_0,C,\tilde{C}_*)\left[T^{-\theta}\sum_{t=1}^{T}\left[\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right]^{1/2} + T^{\theta\zeta}\sum_{t=1}^{T}\left(\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)^{(1+\zeta)/2}\right],\end{aligned} \tag{23}$$

where $\tilde{A}(\zeta,\kappa,\tilde{C}_0,C,\tilde{C}_*)$ *is a constant that depends only on its arguments and not on* T *and* $\{\epsilon_s\}_s$*,* $\tilde{C}_0$*, and* $\tilde{C}_*$ *are defined in Theorem 4, and* C *is defined in* $(\mathcal{A}_6)$*.*

Remark 5. *The choice of* $\epsilon_t = t^{-\beta}$ *for some* $0 < \beta < 1$ *and* $0 \leq \zeta < 1$*, reduces* (23) *to*

$$\mathbb{E}R_T \lesssim T^{1-\beta} + T^{-\theta+\frac{\beta+1}{2}} + T^{\frac{(\beta-1)(1+\zeta)}{2}+\theta\zeta+1}. \tag{24}$$

Similar to Remark 4, we first balance the second and the third term in (24)*, and then balance the resulting rate with the first term in* (24)*. As a result we obtain* $\mathbb{E}R_T = O(T^{1-\beta})$ *where* $\frac{3}{7} \leq \beta < \frac{1}{2}$*. Therefore, the best regret we can achieve in the setting of Theorem 6 is* $T^{\frac{1}{2}+\varepsilon}$*,* $\varepsilon > 0$*, which is almost minimax optimal with* $T^{1/2}$ *being the minimax optimal rate for linear contextual bandits (Chen et al., 2021).*

6 Numerical experiments

In this section, we compare the performance of the proposed kernel ϵ-greedy strategy with other contextual bandit algorithms through numerical experiments. We use a Gaussian kernel parameterized by $\gamma > 0$, i.e.,

$$K(x, x') = \exp\left(-\gamma^2 \|x - x'\|^2\right),$$

for the kernel bandit algorithms. We compare the following four strategies (with parameters defined the parentheses to be tuned and selected) based on the cumulative regret incurred until time horizon $T = 1000$:

1. Kernel ϵ-greedy algorithm with Gaussian kernel (regularization parameter λ_t, length-scale parameter γ)
2. (a) Kernel ϵ-greedy algorithm with linear kernel using the choice of λ_t as in Theorem 2,
   (b) Weighted linear ϵ-greedy algorithm of Chen et al. (2021) with ridge regression estimator and regularization parameter λ_t as in Theorem 2,
3. Weighted linear ϵ-greedy algorithm of Chen et al. (2021) (i.e., without regularization),
4. Kernel Upper Confidence Bound (Kernel UCB) algorithm of Valko et al. (2013) with Gaussian kernel (exploration parameter τ, regularization parameter λ_t, length-scale parameter γ).

Note that 2(a) and 2(b) are essentially the same algorithm since our estimator (3.2) with a linear kernel is just a dual representation of the ridge regression version of the weighted linear ϵ-greedy estimator of Chen et al. (2021). This is also reflected in the regret curves in Figures 1(b),(d) and Figure 2(a),(b). For all the ϵ-greedy based algorithms (1-3), we choose the exploration probability sequence $\epsilon_t = \max\{\frac{t^{-1/2}\log(t)}{10}, 0.02\}$, which is the same choice as used by Chen et al. (2021) in their simulation setup. For algorithms 1 and 4, i.e., the kernel ϵ-greedy and kernel UCB with Gaussian kernel, we do cross-validation as described in Section 6.1 to determine the right choice of the parameters in the parentheses.

We consider four simulated data experiments for d-dimensional context space for $d \in \{1, 2, 3\}$ and for $L = 2$ arms. All algorithms are run until the time horizon $T = 1000$ with initial random exploration time, $t_0 = 50$. For the initialization phase until $t_0 = 50$, we randomly assign both arms 25 times each. In Figures 1(b), 1(d) and Figures 2(a), 2(b), we

plot the average cumulative regret (averaged over 25 runs) over time for the four strategies. It can be seen that in all four settings, the kernel ϵ-greedy with Gaussian kernel outperforms the linear strategies by significantly reducing the regret incurred.

Setting 1: We let $d = 1$ and sample $X_t \overset{\text{i.i.d.}}{\sim} \text{Unif}(-1, 1)$ for $t = 1, \ldots, T$. The mean reward functions considered are: $f_1(x) = \sin(\pi x)$ and $f_2(x) = \cos(\pi x)$ for $-1 < x < 1$ as in Figure 1(a). In Figure 1(b), note that kernel UCB performs the poorest amongst all four strategies while kernel ϵ-greedy performs the best.

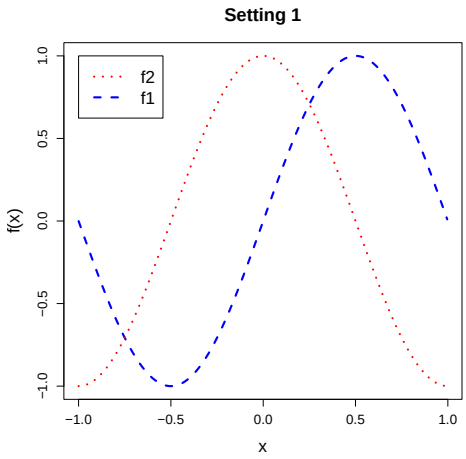
Setting 2: We consider a 'chessboard' like setup (see Figure 1(c)) similar to the experimental setup of Zenati et al. (2022), where $d = 2$, and the mean reward functions are $f_1(x) = 1$ and $f_2(x) = 1$ in the green and red regions, respectively, and 0 elsewhere. Here, we sample each component of the covariates $X_t \in \mathbb{R}^2, t = 1, \ldots, T$ independently from Unif(-1,1) distribution. Note that, in Figure 1(d), kernel UCB and kernel ϵ-greedy perform better than the weighted ϵ-greedy linear algorithms. Both these algorithms give comparable performance with the former being slightly better.

Settings 3 and 4: We consider two arms, $L = 2$, and covariate dimension, $d = 3$. For both these settings, we follow the data generation process of Chen et al. (2021), wherein the covariates X_t are sampled i.i.d. from a truncated normal distribution supported on $[-10, 10]$ with mean zero and scale parameter one. In Setting 3, we use a discretized version of the 'Bump' synthetic environment as in the experimental setup of Zenati et al. (2022). The rewards are generated using the functions, $f_a(x) = \max(0, 1 - \|a - a^*\|_1 - \langle w^*, x - x^* \rangle_2), a = 1, 2$ for some fixed a^*, x^* and w^*. We fix $a^* = 2$ and randomly generate d-dimensional vectors x^* and w^*. In Setting 4, we consider the following function: $f_a(x) = I\{\|x - a + 0.5\|_1 < 4\} + 0.5I\{\|x - (a-1)\|_1 < 4\}$ for $a = 1, 2$. For setting 3, as can be seen in Figure 2(a), both kernel UCB and kernel ϵ-greedy perform at par with each other and result in significantly lower regret than the linear algorithms. For setting 4 as can be seen in Figure 2(b), kernel ϵ-greedy performs better than the kernel UCB algorithm and significantly outperforms the linear algorithms.

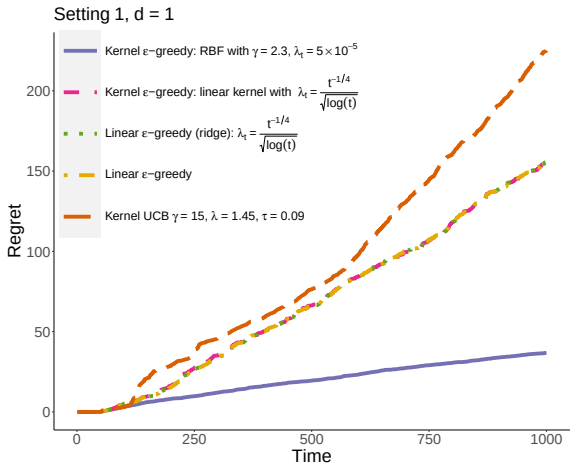
6.1 Choice of kernel parameters

In this section, we describe the methodology we use to tune and select the parameters in the proposed kernel ϵ-greedy algorithm and the kernel UCB algorithm of Valko et al. (2013). Note that for the kernel ϵ-greedy algorithm with Gaussian kernel, we need to tune two parameters, λ_t and γ, while for the kernel UCB algorithm, we need to tune three parameters, λ, γ, and the exploration parameter, τ. Below, we describe the steps for tuning the two parameters in the former, while the same methodology is used to tune the three parameters in the latter.

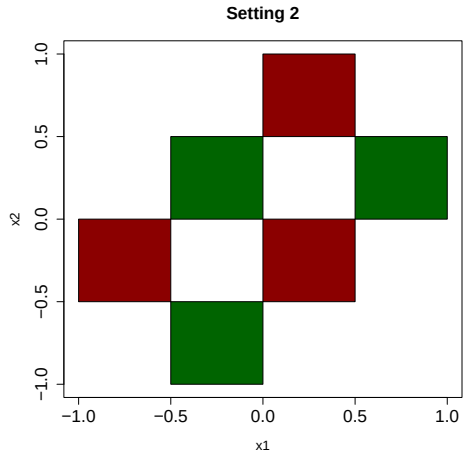
For selecting the two parameters γ and λ in implementing the kernel ϵ-greedy algorithm (corresponding to the purple line in Figures 1(b),(d) and Figure 2), we use the following



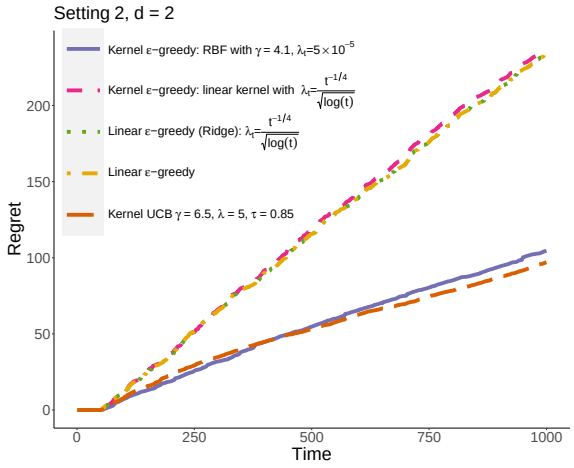
(a) Setting 1: Mean Reward functions



(b) Average cumulative regret over time



(c) Setting 2: Mean reward functions



(d) Average cumulative regret over time.

Figure 1: Left: Mean reward functions for settings 1 and 2. Right: Average cumulative regret over 25 runs for the five strategies over time for $T = 1000$.

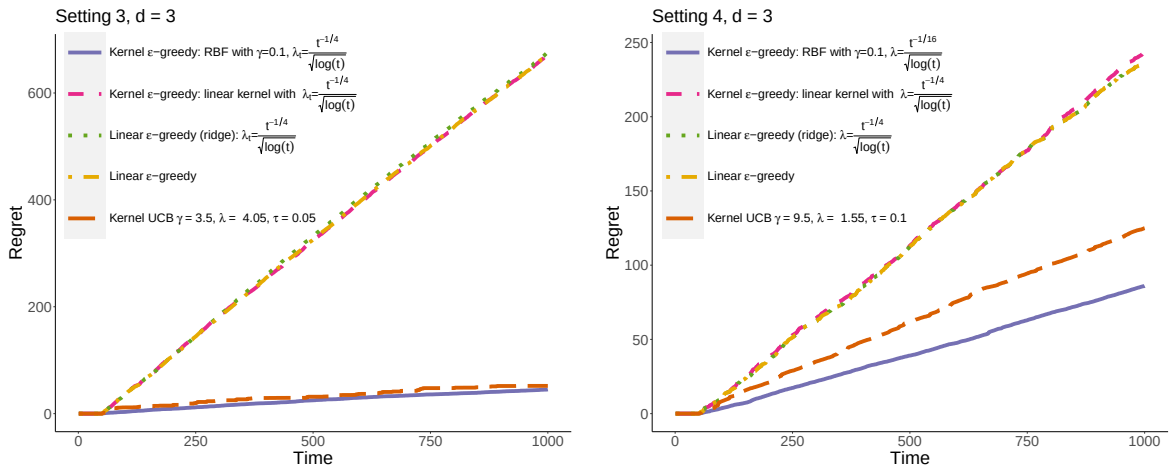
cross-validation approach. We consider the following choices for λ_t and γ, respectively:

$$\lambda_t \in \left\{ \frac{t^{-1/2}}{\sqrt{\log t}}, \frac{t^{-1/4}}{\sqrt{\log t}}, \frac{t^{-1/6}}{\sqrt{\log t}}, \frac{t^{-1/8}}{\sqrt{\log t}}, \frac{t^{-1/16}}{\sqrt{(\log t)}}, 5 \times 10^{-5}, 0.005, 0.5 \right\}, \text{ and}$$

$$\gamma \in \{0.1, 0.3, 0.5, \ldots, 5\}.$$

Then, we use the following k-fold cross-validation approach with $k = 10$:

1. Sample $X \in \mathbb{R}^{T(k+1)\times d}$.
2. Split this data into $(k+1)$ subsets, each of size T. First k subsets are used as training datasets and the $(k+1)$th dataset is used as the test data.
3. For each pair of (λ_t, γ), run the algorithms independently on each of the training sets (first k subsets) and note the regret incurred. Take the average cumulative regret



(a) Setting 3: Average cumulative regret (b) Setting 4: Average cumulative regret

Figure 2: Average cumulative regret over time for kernel ϵ-greedy algorithm with Gaussian kernel, linear kernel (with and without regularization), and kernel UCB with Gaussian kernel.

across all the k subsets.

4. Choose (λ_t^*, γ^*) that minimized the average regret at time T.
5. Run the kernel ϵ-greedy algorithm using the Gaussian kernel with length-scale γ^* and regularization parameter λ_t^* on the test dataset and repeat the experiment on this dataset 25 times. The results reported in Figures 1(b),(d) and 2(a),(b) are the averages of the cumulative regret over these 25 runs.

For Kernel UCB, we follow the same cross-validation approach but with the following grid choices for λ, γ and τ:

$$\lambda \in \{0.05, 0.15, 0.25, \ldots, 5\},\ \gamma \in \{0.5, 1.5, 2.5, \ldots, 15\}, \text{and } \tau \in \{0.05, 0.1, 0.15, \ldots, 0.9\},$$

where it has to be noted that Kernel UCB uses λ that does not vary with time (as suggested in Valko et al., 2013) in contrast to ours which is time dependent.

For cases where a linear approximation would lead to a good classification of arms, our algorithm still performs on par with the linear algorithms. However if one knows that the true model is linear, it might be computationally efficient to opt for the linear ϵ-greedy algorithm. If that is not the case, we can say that using the kernel ϵ- greedy can be beneficial in most settings with the right choice of kernel.

7 Discussion

In this work, we proposes the kernel ϵ-greedy algorithm for contextual bandit problems with finitely many arms. We provided upper bounds on the estimation error for the proposed online regression estimator and provided sub-linear regret rates for the proposed algorithm. While, the kernelized versions of UCB and Thompson sampling have been well-studied, to our knowledge this is the first attempt at studying kernelized ϵ-greedy algorithm. The

theoretical analysis presented is novel, as it utilizes the intrinsic properties of the RKHS and exploits the simplicity of the ϵ-greedy algorithm, resulting in upper bounds that do not depend on quantities like the maximum information gain like previous works. An advantage of the analysis is that we achieve sub-linear regret bounds for wide choices of kernels, along with achieving state-of-the-art regret bounds in a finite-dimensional setting, even when the regressors are not linear. Simple strategies like ϵ-greedy are easy to implement and deeper theoretical understanding helps in supporting their application in real-life sequential decision-making problems. From a practical point of view, addressing computational challenges in implementing the ϵ-greedy algorithm needs further research. One way to address time and computational complexity would be by using incremental Nystr̈om approximations as done by Zenati et al. (2022). While we employ cross-validation to tune for the kernel parameters and regularization parameters, it can be time-consuming to do an exhaustive search. Therefore, new computational techniques need to be devised to help with parameter tuning. Another possible future direction is to study the effect of delayed feedback on the kernel ϵ-greedy strategy, similar to Vakili et al. (2023). On the theoretical front, though the results presented in this paper are non-asymptotic, they are based on Chebyshev's inequality in separable Hilbert spaces and so they only provide quadratic concentration. A future direction is to develop sharp Bernstein-type concentration inequality for operator norm of a self-adjoint Hilbert-Schmidt operator-valued random element defined on a separable Hilbert space, which would provide similar results as in this paper but with sub-Gaussian concentration.

8 Proofs

In this section, we present the proofs of the main results of the paper. Before we present the proofs, we present a result that is used in many of these proofs.

Lemma 1. $\mathbb{E}(\hat{\Sigma}_{i,t}) = \Sigma$, *where* $\hat{\Sigma}_{i,t}$ *is defined in* (5).

Proof. Consider,

$$\begin{aligned}
\mathbb{E}(\hat{\Sigma}_{i,t}) &= \mathbb{E}\left[\frac{1}{t}\sum_{s=1}^{t}\frac{I\{\hat{a}_s = i\}}{P(\hat{a}_s = i|\mathcal{F}_{s-1}, X_s)}k(\cdot, X_s)\otimes k(\cdot, X_s)\right] \\
&= \frac{1}{t}\sum_{s=1}^{t}\mathbb{E}\left[\frac{I\{\hat{a}_s = i\}}{P(\hat{a}_s = i|\mathcal{F}_{s-1}, X_s)}k(\cdot, X_s)\otimes k(\cdot, X_s)\right] \\
&= \frac{1}{t}\sum_{s=1}^{t}\mathbb{E}\left[\mathbb{E}\left(\frac{I\{\hat{a}_s = i\}}{P(\hat{a}_s = i|\mathcal{F}_{s-1}, X_s)}k(\cdot, X_s)\otimes k(\cdot, X_s)\middle|\mathcal{F}_{s-1}, X_s\right)\right] \\
&= \frac{1}{t}\sum_{s=1}^{t}\mathbb{E}\left[\frac{P(\hat{a}_s = i|\mathcal{F}_{s-1}, X_s)}{P(\hat{a}_s = i|\mathcal{F}_{s-1}, X_s)}k(\cdot, X_s)\otimes k(\cdot, X_s)\right] \\
&= \frac{1}{t}\sum_{s=1}^{t}\mathbb{E}(k(\cdot, X_s)\otimes k(\cdot, X_s)) = \Sigma, \text{ for } i = 1, \ldots, L,
\end{aligned}$$

where the third equality follows from the law of iterated expectations. □

8.1 Proof of Theorem 1

Without loss of generality, we will assume $i = 1$. Let $\hat{\pi}_s := P(\hat{a}_s = 1|\mathcal{F}_{s-1}, X_s)$. Then,

$$\begin{aligned}
\hat{f}_{1,t} - f_1 &= \left(\hat{\Sigma}_{1,t} + \lambda I\right)^{-1} \frac{1}{t} \sum_{s=1}^{t} \frac{I\{\hat{a}_s = 1\}}{\hat{\pi}_s} k(\cdot, X_s) y_s - f_1 \\
&= \left(\hat{\Sigma}_{1,t} + \lambda I\right)^{-1} \frac{1}{t} \sum_{s=1}^{t} \frac{I\{\hat{a}_s = 1\}}{\hat{\pi}_s} k(\cdot, X_s)[\langle k(\cdot, X_s), f_1\rangle_{\mathcal{H}} + e_s] - f_1 \\
&= \left(\hat{\Sigma}_{1,t} + \lambda I\right)^{-1} \left[-\lambda f_1 + \frac{1}{t} \sum_{s=1}^{t} \frac{I\{\hat{a}_s = 1\}}{\hat{\pi}_s} k(\cdot, X_s) e_s\right] \\
&= \left(\hat{\Sigma}_{1,t} + \lambda I\right)^{-1/2} \left(\hat{\Sigma}_{1,t} + \lambda I\right)^{-1/2} (\Sigma + \lambda I)^{1/2} (\Sigma + \lambda I)^{-1/2} \\
&\qquad \times \left[-\lambda f_1 + \frac{1}{t} \sum_{s=1}^{t} \frac{I\{\hat{a}_s = 1\}}{\hat{\pi}_s} k(\cdot, X_s) e_s\right].
\end{aligned}$$

This implies that,

$$\begin{aligned}
\|\hat{f}_{1,t} - f_1\|_{\mathcal{H}} &\leq \left\| \left(\hat{\Sigma}_{1,t} + \lambda I\right)^{-1/2} \left(\hat{\Sigma}_{1,t} + \lambda I\right)^{-1/2} (\Sigma + \lambda I)^{1/2} \right\|_{\infty} \\
&\qquad \times \left\| (\Sigma + \lambda I)^{-1/2} \left[-\lambda f_1 + \frac{1}{t} \sum_{s=1}^{t} \frac{I\{\hat{a}_s = 1\}}{\hat{\pi}_s} k(\cdot, X_s) e_s\right] \right\|_{\mathcal{H}} \\
&\leq \|(\hat{\Sigma}_{1,t} + \lambda I)^{-1/2}\|_{\infty} \|(\hat{\Sigma}_{1,t} + \lambda I)^{-1/2} (\Sigma + \lambda I)^{1/2}\|_{\infty} \\
&\qquad \times \left\| (\Sigma + \lambda I)^{-1/2} \left[-\lambda f_1 + \frac{1}{t} \sum_{s=1}^{t} \frac{I\{\hat{a}_s = 1\}}{\hat{\pi}_s} k(\cdot, X_s) e_s\right] \right\|_{\mathcal{H}} \leq \frac{\mathcal{S}_1 \mathcal{S}_2}{\sqrt{\lambda}}, \quad (25)
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{S}_1 &:= \left\| \left(\hat{\Sigma}_{1,t} + \lambda I\right)^{-1/2} (\Sigma + \lambda I)^{1/2} \right\|_{\infty}, \text{ and} \\
\mathcal{S}_2 &:= \left\| (\Sigma + \lambda I)^{-1/2} \left[-\lambda f_1 + \frac{1}{t} \sum_{s=1}^{t} \frac{I\{\hat{a}_s = 1\}}{\hat{\pi}_s} k(\cdot, X_s) e_s\right] \right\|_{\mathcal{H}}.
\end{aligned}$$

We now bound $\mathcal{S}_1$ and $\mathcal{S}_2$. By defining $B_t := (\Sigma + \lambda I)^{-1/2}(\Sigma - \hat{\Sigma}_{1,t})(\Sigma + \lambda I)^{-1/2}$, we have

$$\begin{aligned}
\mathcal{S}_1 = \|(\hat{\Sigma}_{1,t} + \lambda I)^{-1/2} (\Sigma + \lambda I)^{1/2}\|_{\infty} &= \|(\Sigma + \lambda I)^{1/2} (\hat{\Sigma}_{1,t} + \lambda I)^{-1} (\Sigma + \lambda I)^{1/2}\|_{\infty}^{1/2} \\
&= \|(I - B_t)^{-1}\|_{\infty}^{1/2} \\
&\leq (1 - \|B_t\|_{\infty})^{-1/2},
\end{aligned}$$

where the last inequality follows from Lemma 3.6 of (Rudi et al. 2013). Note, $\|B_t\|_\infty \leq \|B_t\|_{HS}$, where $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm. Using Chebyshev's inequality we obtain,

$$P(\|B_t\|_{HS} \geq \epsilon) \leq \frac{\mathbb{E}\|B_t\|_{HS}^2}{\epsilon^2}.$$

It follows from Lemma 1 that

$$\begin{aligned}\mathbb{E}\|B_t\|_{HS}^2 &= \mathbb{E}\|(\Sigma+\lambda I)^{-1/2}(\hat{\Sigma}_{1,t}-\Sigma)(\Sigma+\lambda I)^{-1/2}\|_{HS}^2 \\ &= \mathbb{E}\|(\Sigma+\lambda I)^{-1/2}\hat{\Sigma}_{1,t}(\Sigma+\lambda I)^{-1/2}\|_{HS}^2 - \|(\Sigma+\lambda I)^{-1/2}\Sigma(\Sigma+\lambda I)^{-1/2}\|_{HS}^2. \quad (26)\end{aligned}$$

Define $N_{\Sigma,2}(\lambda) := \|(\Sigma+\lambda I)^{-1/2}\Sigma(\Sigma+\lambda I)^{-1/2}\|_{HS}^2$ and consider,

$$\begin{aligned}&\mathbb{E}\|(\Sigma+\lambda I)^{-1/2}\hat{\Sigma}_{1,t}(\Sigma+\lambda I)^{-1/2}\|_{HS}^2 \\ &\quad= \mathbb{E}\left\langle (\Sigma+\lambda I)^{-1/2}\hat{\Sigma}_{1,t}(\Sigma+\lambda I)^{-1/2}, (\Sigma+\lambda I)^{-1/2}\hat{\Sigma}_{1,t}(\Sigma+\lambda I)^{-1/2}\right\rangle_{HS} \\ &\quad= \mathbb{E}\,\mathrm{Tr}\left[(\Sigma+\lambda I)^{-1/2}\hat{\Sigma}_{1,t}(\Sigma+\lambda I)^{-1}\hat{\Sigma}_{1,t}(\Sigma+\lambda I)^{-1/2}\right] \\ &\quad= \mathbb{E}\,\mathrm{Tr}\left[(\Sigma+\lambda I)^{-1}\hat{\Sigma}_{1,t}(\Sigma+\lambda I)^{-1}\hat{\Sigma}_{1,t}\right].\end{aligned}$$

Now, plugging in the definition of $\hat{\Sigma}_{1,t}$ from (5) and defining $\hat{\tau}_s = I\{\hat{a}_s = 1\}/\hat{\pi}_s$, we obtain

$$\begin{aligned}&\mathbb{E}\|(\Sigma+\lambda I)^{-1/2}\hat{\Sigma}_{1,t}(\Sigma+\lambda I)^{-1/2}\|_{HS}^2 \\ &\quad= \mathbb{E}\left[\frac{1}{t^2}\sum_{s=1}^{t}\sum_{\ell=1}^{t}\hat{\tau}_s\hat{\tau}_\ell\,\mathrm{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)\right].\end{aligned}$$

By breaking the double sum in the above expression into the cases when, (1) $s=\ell$, (2) $s>\ell$, and (3) $s<\ell$, yields

$$\begin{aligned}&\mathbb{E}\|(\Sigma+\lambda I)^{-1/2}\hat{\Sigma}_{1,t}(\Sigma+\lambda I)^{-1/2}\|_{HS}^2 \\ &\quad= \mathbb{E}\left[\frac{1}{t^2}\sum_{s=1}^{t}\hat{\tau}_s^2\,\mathrm{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)\right)\right] \\ &\qquad+ \mathbb{E}\left[\frac{1}{t^2}\sum_{\ell=1}^{t-1}\sum_{s=\ell+1}^{t}\hat{\tau}_s\hat{\tau}_\ell\,\mathrm{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)\right] \\ &\qquad+ \mathbb{E}\left[\frac{1}{t^2}\sum_{\ell=2}^{t}\sum_{s=1}^{\ell-1}\hat{\tau}_s\hat{\tau}_\ell\,\mathrm{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)\right] \\ &\quad= \textcircled{1}+\textcircled{2}+\textcircled{3}, \quad (27)\end{aligned}$$

where we bound ①–③ as follows.

$$
\begin{aligned}
① &= \frac{1}{t^2}\sum_{s=1}^{t}\mathbb{E}[\hat{\tau}_s^2\,\mathrm{Tr}((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s))] \\
&= \frac{1}{t^2}\sum_{s=1}^{t}\mathbb{E}\left[\mathbb{E}[\hat{\tau}_s^2\,\mathrm{Tr}((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s))\big|\mathcal{F}_{s-1},X_s]\right] \\
&= \frac{1}{t^2}\sum_{s=1}^{t}\mathbb{E}\left[\mathbb{E}(\hat{\tau}_s^2|\mathcal{F}_{s-1},X_s)\,\mathrm{Tr}((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s))\right],
\end{aligned}
\tag{28}
$$

where

$$
\mathbb{E}(\hat{\tau}_s^2|\mathcal{F}_{s-1},X_s)=\mathbb{E}\left[\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s^2}\middle|\mathcal{F}_{s-1},X_s\right]=\frac{\hat{\pi}_s}{\hat{\pi}_s^2}=\frac{1}{\hat{\pi}_s}\le\frac{L-1}{\epsilon_s},
$$

since $1-\epsilon_s\ge\frac{\epsilon_s}{L-1}$, resulting in

$$
\begin{aligned}
(28) &\le \frac{1}{t^2}\sum_{s=1}^{t}\mathbb{E}\left[\frac{L-1}{\epsilon_s}\,\mathrm{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)\right)\right] \\
&\le \frac{1}{t^2}\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\mathbb{E}\,\mathrm{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)\right)\sup_{X_s}\|(\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)\|_\infty \\
&\le \frac{1}{t^2}\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\,\mathrm{Tr}\left((\Sigma+\lambda I)^{-1}\Sigma\right)\|(\Sigma+\lambda I)^{-1}\|_\infty\sup_{X_s}\|k(\cdot,X_s)\otimes k(\cdot,X_s)\|_\infty \qquad (29) \\
&\le \frac{1}{t^2}\left(\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\right)N_{\Sigma,1}(\lambda)\frac{\kappa}{\lambda}, \qquad (30)
\end{aligned}
$$

where $N_{\Sigma,1}(\lambda):=\mathrm{Tr}\left((\Sigma+\lambda I)^{-1}\Sigma\right)$.

$$
\begin{aligned}
② &= \frac{1}{t^2}\sum_{\ell=1}^{t-1}\sum_{s=\ell+1}^{t}\mathbb{E}\left[\hat{\tau}_s\hat{\tau}_\ell\,\mathrm{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)\right] \\
&= \frac{1}{t^2}\sum_{\ell=1}^{t-1}\sum_{s=\ell+1}^{t}\mathbb{E}\Big[\mathbb{E}[\hat{\tau}_s\hat{\tau}_\ell\,\mathrm{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}\right. \\
&\qquad\qquad \left.\times k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)|\mathcal{F}_{s-1},X_s]\Big] \\
&\overset{(\dagger)}{=} \frac{1}{t^2}\sum_{\ell=1}^{t-1}\sum_{s=\ell+1}^{t}\mathbb{E}\left[\hat{\tau}_\ell\mathbb{E}(\hat{\tau}_s|\mathcal{F}_{s-1},X_s)\,\mathrm{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}\right.\right. \\
&\qquad\qquad \left.\left.\times k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)\right] \\
&= \frac{1}{t^2}\sum_{\ell=1}^{t-1}\sum_{s=\ell+1}^{t}\mathbb{E}\left[\hat{\tau}_\ell\,\mathrm{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)\right],
\end{aligned}
\tag{31}
$$

where we used

$$\mathbb{E}(\hat{\tau}_s|\mathcal{F}_{s-1}, X_s) = \mathbb{E}\left[\frac{I\{\hat{a}_s = 1\}}{\hat{\pi}_s}\Big|\mathcal{F}_{s-1}, X_s\right] = 1 \tag{32}$$

in (†). Now using the law of iterated expectation, we obtain

$$\begin{aligned}(31) &= \frac{1}{t^2}\sum_{\ell=1}^{t-1}\mathbb{E}\left[\sum_{s=\ell+1}^{t}\mathbb{E}[\hat{\tau}_\ell \operatorname{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)|\mathcal{F}_\ell]\right] \\ &= \frac{1}{t^2}\sum_{\ell=1}^{t-1}\mathbb{E}\left[\hat{\tau}_\ell \operatorname{Tr}\Big((\Sigma+\lambda I)^{-1}\Big(\sum_{s=\ell+1}^{t}\mathbb{E}(k(\cdot,X_s)\otimes k(\cdot,X_s)|\mathcal{F}_\ell)\Big)(\Sigma+\lambda I)^{-1}\right. \\ &\qquad\left.\times k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\Big)\right]\end{aligned}$$

$$\begin{aligned}&= \frac{1}{t^2}\sum_{\ell=1}^{t-1}\mathbb{E}\left[\hat{\tau}_\ell \operatorname{Tr}\left((\Sigma+\lambda I)^{-1}(t-\ell)\Sigma(\Sigma+\lambda I)^{-1}k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)\right] \\ &= \frac{1}{t^2}\sum_{\ell=1}^{t-1}(t-\ell)\mathbb{E}\left[\hat{\tau}_\ell \operatorname{Tr}\left((\Sigma+\lambda I)^{-1}\Sigma(\Sigma+\lambda I)^{-1}k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)\right] \\ &= \frac{1}{t^2}\sum_{\ell=1}^{t-1}(t-\ell)\mathbb{E}\left[\mathbb{E}\left[\hat{\tau}_\ell \operatorname{Tr}\left((\Sigma+\lambda I)^{-1}\Sigma(\Sigma+\lambda I)^{-1}k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)\Big|\mathcal{F}_{\ell-1}, X_\ell\right]\right] \\ &= \frac{1}{t^2}\sum_{\ell=1}^{t-1}(t-\ell)\mathbb{E}\left[\mathbb{E}(\hat{\tau}_\ell|\mathcal{F}_{\ell-1}, X_\ell)\operatorname{Tr}\left((\Sigma+\lambda I)^{-1}\Sigma(\Sigma+\lambda I)^{-1}k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)\right].\end{aligned} \tag{33}$$

Again, using (32), we get

$$\begin{aligned}(33) &= \frac{1}{t^2}\sum_{\ell=1}^{t-1}(t-\ell)\mathbb{E}\left[\operatorname{Tr}\left((\Sigma+\lambda I)^{-1}\Sigma(\Sigma+\lambda I)^{-1}k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)\right] \\ &= \frac{1}{t^2}\sum_{\ell=1}^{t-1}(t-\ell)\left[\operatorname{Tr}\left((\Sigma+\lambda I)^{-1}\Sigma(\Sigma+\lambda I)^{-1}\Sigma\right)\right] \\ &= \frac{1}{t^2}\sum_{\ell=1}^{t-1}(t-\ell)N_{\Sigma,2}(\lambda) = \frac{1}{t^2}\left[t(t-1)-\frac{(t-1)t}{2}\right]N_{\Sigma,2}(\lambda) \\ &= \frac{N_{\Sigma,2}(\lambda)}{2} - \frac{N_{\Sigma,2}(\lambda)}{2t}.\end{aligned} \tag{34}$$

$$
\begin{aligned}
\textcircled{3} &= \frac{1}{t^2}\sum_{\ell=2}^{t}\sum_{s=1}^{\ell-1}\mathbb{E}\left[\hat{\tau}_s\hat{\tau}_\ell \operatorname{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)\right] \\
&= \frac{1}{t^2}\sum_{\ell=2}^{t}\sum_{s=1}^{\ell-1}\mathbb{E}\left[\mathbb{E}[\hat{\tau}_s\hat{\tau}_\ell \operatorname{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}\right.\right. \\
&\qquad \left.\left.\times k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)|\mathcal{F}_{\ell-1},X_\ell]\right] \\
&= \frac{1}{t^2}\sum_{\ell=2}^{t}\sum_{s=1}^{\ell-1}\mathbb{E}\left[\hat{\tau}_s\mathbb{E}(\hat{\tau}_\ell|\mathcal{F}_{\ell-1},X_\ell) \operatorname{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}\right.\right. \\
&\qquad \left.\left.\times k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)\right] \\
&= \frac{1}{t^2}\sum_{\ell=2}^{t}\sum_{s=1}^{\ell-1}\mathbb{E}\left[\hat{\tau}_s \operatorname{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)\right].
\end{aligned} \tag{35}
$$

Now, using law of iterated expectations, we have

$$
\begin{aligned}
(35) &= \frac{1}{t^2}\sum_{\ell=2}^{t}\sum_{s=1}^{\ell-1}\mathbb{E}\left[\mathbb{E}[\hat{\tau}_s \operatorname{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\right)\big|\mathcal{F}_s]\right] \\
&= \frac{1}{t^2}\sum_{\ell=2}^{t}\sum_{s=1}^{\ell-1}\mathbb{E}\left[\hat{\tau}_s \operatorname{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}\mathbb{E}[k(\cdot,X_\ell)\otimes k(\cdot,X_\ell)\big|\mathcal{F}_s]\right)\right] \\
&= \frac{1}{t^2}\sum_{\ell=2}^{t}\sum_{s=1}^{\ell-1}\mathbb{E}\left[\mathbb{E}\left(\hat{\tau}_s \operatorname{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}\Sigma\right)|\mathcal{F}_{s-1},X_s\right)\right] \\
&= \frac{1}{t^2}\sum_{\ell=2}^{t}\sum_{s=1}^{\ell-1}\mathbb{E}\left[\operatorname{Tr}\left(\mathbb{E}(\hat{\tau}_s|\mathcal{F}_{s-1},X_s)(\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}\Sigma\right)\right] \\
&= \frac{1}{t^2}\sum_{\ell=2}^{t}\sum_{s=1}^{\ell-1}\mathbb{E}\left[\operatorname{Tr}\left(\mathbb{E}(\hat{\tau}_s|\mathcal{F}_{s-1},X_s)(\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}\Sigma\right)\right] \\
&= \frac{1}{t^2}\sum_{\ell=2}^{t}\sum_{s=1}^{\ell-1}\mathbb{E}\left[\operatorname{Tr}\left((\Sigma+\lambda I)^{-1}k(\cdot,X_s)\otimes k(\cdot,X_s)(\Sigma+\lambda I)^{-1}\Sigma\right)\right] \\
&= \frac{1}{t^2}\sum_{\ell=2}^{t}(\ell-1)N_{\Sigma,2}(\lambda) = \frac{1}{t^2}\left[\frac{t(t-1)}{2}\right]N_{\Sigma,2}(\lambda) \\
&= \frac{N_{\Sigma,2}(\lambda)}{2} - \frac{N_{\Sigma,2}(\lambda)}{2t}.
\end{aligned} \tag{36}
$$

Putting together (30), (34) and (36) in (27), and the result in (26), we get,

$$\begin{aligned}\mathbb{E}\|B_t\|_{HS}^2 &\leq \frac{1}{t^2}\left(\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\right)N_{\Sigma,1}(\lambda)\frac{\kappa}{\lambda}+2\left(\frac{N_{\Sigma,2}(\lambda)}{2}-\frac{1}{2t}N_{\Sigma,2}(\lambda)\right)-N_{\Sigma,2}(\lambda)\\ &\leq \frac{1}{t^2}\left(\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\right)N_{\Sigma,1}(\lambda)\frac{\kappa}{\lambda}. \end{aligned} \tag{37}$$

Using Assumption $(\mathcal{A}_3)$, we obtain,

$$\mathbb{E}\|B_t\|_{HS}^2 \leq A_1(\bar{C},\alpha)\frac{\kappa}{t^2}\left(\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\right)\lambda^{-(1+1/\alpha)},\ \alpha>1,$$

where $A_1(\bar{C},\alpha)$ is a constant depending only on $\bar{C}$ and α. Therefore,

$$P\left(\|B_t\|_{HS}\geq\frac{1}{2}\right)\leq 4\mathbb{E}\|B_t\|_{HS}^2\leq\frac{4\kappa}{t^2}A_1(\bar{C},\alpha)\lambda^{-(1+1/\alpha)}\left(\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\right).$$

Thus for $\delta>0$, choosing

$$\lambda\geq\left[\frac{4\kappa A_1(\bar{C},\alpha)}{t^2\delta}\left(\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\right)\right]^{\alpha/(1+\alpha)},\ \alpha>1,$$

yields

$$P\left(\|B_t\|_{HS}\geq\frac{1}{2}\right)\leq\delta,$$

implying that with probability at least $1-\delta$,

$$\mathcal{S}_1\leq\sqrt{2}. \tag{38}$$

We now bound $\mathcal{S}_2$ as

$$\begin{aligned}\mathcal{S}_2 &= \left\|(\Sigma+\lambda I)^{-1/2}\left[-\lambda f_1+\frac{1}{t}\sum_{s=1}^{t}\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s}k(\cdot,X_s)e_s\right]\right\|_{\mathcal{H}}\\ &\leq \left\|(\Sigma+\lambda I)^{-1/2}\frac{1}{t}\sum_{s=1}^{t}\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s}k(\cdot,X_s)e_s\right\|_{\mathcal{H}}+\lambda\|(\Sigma+\lambda I)^{-1/2}f_1\|_{\mathcal{H}}. \end{aligned} \tag{39}$$

For the second term in (39), using Assumption $(\mathcal{A}_4)$, we have

$$\begin{aligned}\left\|(\Sigma+\lambda I)^{-1/2}f_1\right\|_{\mathcal{H}} &\leq \left\|(\Sigma+\lambda I)^{-1/2}\Sigma^{\gamma_1}h\right\|_{\mathcal{H}}\leq\left\|(\Sigma+\lambda I)^{-1/2}\Sigma^{\gamma_1}\right\|_{\infty}\|\Sigma^{-\gamma_1}f_1\|_{\mathcal{H}}\\ &\leq\sup_i\frac{\eta_i^{\gamma_1}}{(\eta_i+\lambda)^{1/2}}\|\Sigma^{-\gamma_1}f_1\|_{\mathcal{H}}\leq\sup_{x\geq0}\left[\frac{x^{2\gamma_1}}{x+\lambda}\right]^{1/2}\|\Sigma^{-\gamma_1}f_1\|_{\mathcal{H}}\\ &\leq\lambda^{\gamma_1-\frac{1}{2}}\|\Sigma^{-\gamma_1}f_1\|_{\mathcal{H}}, \end{aligned} \tag{40}$$

where the last inequality follows by noting that for $0 < \gamma_1 \leq 1/2$,

$$\left(\sup_{x\geq 0} \frac{x^{2\gamma_1}}{x+\lambda}\right)^{1/2} = \left(\sup_{x\geq 0} \left(\frac{x}{x+\lambda}\right)^{2\gamma_1} \frac{1}{(x+\lambda)^{1-2\gamma_1}}\right)^{1/2} \leq \lambda^{\gamma_1 - \frac{1}{2}}.$$

For any $\xi > 0$, applying Chebyshev's inequality to the first term in (39) yields

$$\begin{aligned}
&P\left(\left\|(\Sigma+\lambda I)^{-1/2}\frac{1}{t}\sum_{s=1}^{t}\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s}k(\cdot,X_s)e_s\right\|_{\mathcal{H}} \geq \xi\right) \\
&\leq \frac{1}{\xi^2}\mathbb{E}\left\|(\Sigma+\lambda I)^{-1/2}\frac{1}{t}\sum_{s=1}^{t}\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s}k(\cdot,X_s)e_s\right\|_{\mathcal{H}}^2 \\
&= \frac{1}{\xi^2}\mathbb{E}\left\langle (\Sigma+\lambda I)^{-1/2}\frac{1}{t}\sum_{s=1}^{t}\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s}k(\cdot,X_s)e_s, (\Sigma+\lambda I)^{-1/2}\frac{1}{t}\sum_{\ell=1}^{t}\frac{I\{\hat{a}_\ell=1\}}{\hat{\pi}_\ell}k(\cdot,X_\ell)e_\ell\right\rangle_{\mathcal{H}} \\
&= \frac{1}{\xi^2}\mathbb{E}\left\langle (\Sigma+\lambda I)^{-1}, \frac{1}{t^2}\sum_{s=1}^{t}\sum_{\ell=1}^{t}\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s}\frac{I\{\hat{a}_\ell=1\}}{\hat{\pi}_\ell}e_s e_\ell k(\cdot,X_s)\otimes k(\cdot,X_\ell)\right\rangle_{HS} \\
&= \frac{1}{\xi^2}\left\langle (\Sigma+\lambda I)^{-1}, \mathbb{E}\left(\frac{1}{t^2}\sum_{s=1}^{t}\sum_{\ell=1}^{t}\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s}\frac{I\{\hat{a}_\ell=1\}}{\hat{\pi}_\ell}e_s e_\ell k(\cdot,X_s)\otimes k(\cdot,X_\ell)\right)\right\rangle_{HS}.
\end{aligned} \tag{41}$$

In the following, we simplify the expectation term in (41) by considering three cases for the double sum: (1) $\ell = s$, (2) $\ell > s$ and (3) $\ell < s$. Recall $\hat{\tau}_s = I\{\hat{a}_s = 1\}/\hat{\pi}_s$, where $\hat{\pi}_s = P(\hat{a}_s = 1|\mathcal{F}_{s-1}, X_s)$. Consider

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{t^2}\sum_{s=1}^{t}\sum_{\ell=1}^{t}\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s}\frac{I\{\hat{a}_\ell=1\}}{\hat{\pi}_\ell}e_s e_\ell k(\cdot,X_s)\otimes k(\cdot,X_\ell)\right) \\
&= \mathbb{E}\left(\frac{1}{t^2}\sum_{s=1}^{t}\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s^2}e_s^2 k(\cdot,X_s)\otimes k(\cdot,X_s)\right) + \mathbb{E}\left(\frac{1}{t^2}\sum_{\ell=1}^{t-1}\sum_{s=\ell+1}^{t}\hat{\tau}_s\hat{\tau}_\ell e_s e_\ell k(\cdot,X_s)\otimes k(\cdot,X_\ell)\right) \\
&\qquad + \mathbb{E}\left(\frac{1}{t^2}\sum_{\ell=2}^{t}\sum_{s=1}^{\ell-1}\hat{\tau}_s\hat{\tau}_\ell e_s e_\ell k(\cdot,X_s)\otimes k(\cdot,X_\ell)\right) \\
&= \textcircled{4} + \textcircled{5} + \textcircled{6},
\end{aligned}$$

where

$$\begin{aligned}
\textcircled{4} &= \mathbb{E}\left(\frac{1}{t^2}\sum_{s=1}^{t}\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s^2}e_s^2 k(\cdot,X_s)\otimes k(\cdot,X_s)\right) \\
&= \frac{1}{t^2}\sum_{s=1}^{t}\mathbb{E}\left[\mathbb{E}\left(\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s^2}e_s^2 k(\cdot,X_s)\otimes k(\cdot,X_s)\middle|\mathcal{F}_{s-1},X_s,\hat{a}_s\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{t^2}\sum_{s=1}^{t}\mathbb{E}\left[\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s^2}k(\cdot,X_s)\otimes k(\cdot,X_s)\mathbb{E}\left(e_s^2\middle|\mathcal{F}_{s-1},X_s,\hat{a}_s\right)\right]\\
&\preceq \frac{\sigma^2}{t^2}\sum_{s=1}^{t}\mathbb{E}\left[\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s^2}k(\cdot,X_s)\otimes k(\cdot,X_s)\right]\\
&= \frac{\sigma^2}{t^2}\sum_{s=1}^{t}\mathbb{E}\left[\mathbb{E}\left[\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s^2}k(\cdot,X_s)\otimes k(\cdot,X_s)\middle|\mathcal{F}_{s-1},X_s\right]\right]\\
&= \frac{\sigma^2}{t^2}\sum_{s=1}^{t}\mathbb{E}\left[\frac{1}{\hat{\pi}_s}k(\cdot,X_s)\otimes k(\cdot,X_s)\right] \preceq \frac{\sigma^2}{t^2}\left(\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\right)\Sigma, \qquad (42)
\end{aligned}
$$

where the above inequality follows from $(\mathcal{A}_1)$, and the fact that $\hat{\pi}_s \geq \epsilon_s/(L-1)$. To bound (5), define $D_{s\ell} := \{\hat{a}_s = \hat{a}_\ell = 1\}$. Then the complement of this event consists of situations when the two arms are not the same or either (or both) are not arm 1, i.e., one or both of $\hat{\tau}_s$ and $\hat{\tau}_\ell$ will be zero. Therefore,

$$
\begin{aligned}
(5) &= \mathbb{E}\left(\frac{1}{t^2}\sum_{\ell=1}^{t-1}\sum_{s=\ell+1}^{t}\hat{\tau}_s\hat{\tau}_\ell e_s e_\ell k(\cdot,X_s)\otimes k(\cdot,X_\ell)\right)\\
&= \frac{1}{t^2}\sum_{\ell=1}^{t-1}\sum_{s=\ell+1}^{t}\mathbb{E}\left[E\left(\hat{\tau}_s\hat{\tau}_\ell e_s e_\ell k(\cdot,X_s)\otimes k(\cdot,X_\ell)|\mathcal{F}_{s-1},X_s,D_{s\ell}\right)\right]\\
&= \frac{1}{t^2}\sum_{\ell=1}^{t-1}\sum_{s=\ell+1}^{t}\mathbb{E}\left[\frac{1}{\hat{\pi}_s\hat{\pi}_\ell}\mathbb{E}(e_s e_\ell|\mathcal{F}_{s-1},X_s,D_{s\ell})k(\cdot,X_s)\otimes k(\cdot,X_\ell)\right]\\
&= \frac{1}{t^2}\sum_{\ell=1}^{t-1}\sum_{s=\ell+1}^{t}\mathbb{E}\left[\frac{1}{\hat{\pi}_s\hat{\pi}_\ell}\mathbb{E}(e_s|\mathcal{F}_{s-1},X_s,\hat{a}_s=1)\mathbb{E}(e_\ell|\mathcal{F}_{s-1},X_s,\hat{a}_\ell=1)k(\cdot,X_s)\otimes k(\cdot,X_\ell)\right] \qquad (43)\\
&= 0, \qquad (44)
\end{aligned}
$$

where (43) follows from $(\mathcal{A}_2)$, i.e., errors and covariates at time t are independent for a given arm. Similar to (5), we obtain

$$
\begin{aligned}
(6) &= \mathbb{E}\left(\frac{1}{t^2}\sum_{\ell=2}^{t}\sum_{s=1}^{\ell-1}\hat{\tau}_s\hat{\tau}_\ell e_s e_\ell k(\cdot,X_s)\otimes k(\cdot,X_\ell)\right)\\
&= \frac{1}{t^2}\sum_{\ell=2}^{t}\sum_{s=1}^{\ell-1}\mathbb{E}\left[E\left(\hat{\tau}_s\hat{\tau}_\ell e_s e_\ell k(\cdot,X_s)\otimes k(\cdot,X_\ell)|\mathcal{F}_{\ell-1},X_\ell,D_{s\ell}\right)\right]\\
&= \frac{1}{t^2}\sum_{\ell=2}^{t}\sum_{s=1}^{\ell-1}\mathbb{E}\left[\frac{1}{\hat{\pi}_s\hat{\pi}_\ell}\mathbb{E}(e_s e_\ell|\mathcal{F}_{\ell-1},X_\ell,D_{s\ell})k(\cdot,X_s)\otimes k(\cdot,X_\ell)\right]\\
&= \frac{1}{t^2}\sum_{\ell=2}^{t}\sum_{s=1}^{\ell-1}\mathbb{E}\left[\frac{1}{\hat{\pi}_s\hat{\pi}_\ell}\mathbb{E}(e_s|\mathcal{F}_{\ell-1},X_\ell,\hat{a}_s=1)\mathbb{E}(e_\ell|\mathcal{F}_{\ell-1},X_\ell,\hat{a}_\ell=1)k(\cdot,X_s)\otimes k(\cdot,X_\ell)\right]\\
&= 0. \qquad (45)
\end{aligned}
$$

Combining (42), (44) and (45) in (41), we obtain

$$P\left(\left\|(\Sigma+\lambda I)^{-1/2}\frac{1}{t}\sum_{s=1}^{t}\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s}k(\cdot,X_s)e_s\right\|_{\mathcal{H}}\geq\xi\right)\leq\frac{1}{\xi^2}\left\langle(\Sigma+\lambda I)^{-1},\frac{\sigma^2}{t^2}\left(\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\right)\Sigma\right\rangle_{HS}$$

$$=\frac{\sigma^2}{t^2\xi^2}\left(\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\right)\mathrm{Tr}\left[(\Sigma+\lambda I)^{-1}\Sigma\right]=\frac{\sigma^2}{t^2\xi^2}\left(\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\right)N_{\Sigma,1}(\lambda) \tag{46}$$

$$\leq\frac{\sigma^2A_1(\bar{C},\alpha)}{t^2\xi^2}\left(\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\right)\lambda^{-1/\alpha}, \tag{47}$$

where we used $(\mathcal{A}_3)$ in the last inequality. Combining (40) and (47) in (39), and choosing

$$\xi=\left[\frac{\sigma^2A_1(\bar{C},\alpha)}{\delta t^2}\left(\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\right)\lambda^{-1/\alpha}\right]^{1/2},$$

yields that with probability at least $1-\delta$,

$$\mathcal{S}_2\leq\left[\frac{\sigma^2A_1(\bar{C},\alpha)}{\delta t^2}\left(\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\right)\lambda^{-1/\alpha}\right]^{1/2}+\lambda^{\gamma_1+\frac{1}{2}}\|\Sigma^{-\gamma_1}f_1\|_{\mathcal{H}}. \tag{48}$$

Using (38) and (48) in (25) yields that with probability at least $1-2\delta$, we have

$$\|\hat{f}_{1,t}-f_1\|_{\mathcal{H}}\leq\left[\frac{2\sigma^2A_1(\bar{C},\alpha)}{\delta t^2}\left(\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\right)\lambda^{-(1+1/\alpha)}\right]^{1/2}+\sqrt{2}\lambda^{\gamma_1}\|\Sigma^{-\gamma_1}f_1\|_{\mathcal{H}}$$

$$\leq\sqrt{2}\max\{C_0,C_1\}\left[\left(\frac{1}{\delta t^2}\left(\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)\lambda^{-(1+1/\alpha)}\right)^{1/2}+\lambda^{\gamma_1}\right], \tag{49}$$

where $C_0=\sqrt{\sigma^2(L-1)A_1(\bar{C},\alpha)}$ and $C_1=\|\Sigma^{-\gamma_1}f_1\|_{\mathcal{H}}$. The result follows by choosing

$$\lambda=\lambda_{1,t}:=\left[\frac{1}{\delta t^2}\left(\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)\right]^{\alpha/(2\gamma_1\alpha+\alpha+1)}$$

in (49). Also, the same proof works for all arms $i=1,\ldots,L$, by defining $\hat{\pi}_s=P(\hat{a}_s=i|\mathcal{F}_{s-1},X_s)$.

Next, we derive the bound for the estimation error in expectation. From the above, note that, for $i=1,\ldots,L$ and any $\delta>0$,

$$P(\|\hat{f}_{i,t}-f_i\|_{\mathcal{H}}\leq\theta_t)\geq1-2\delta,$$

where,

$$\theta_t=2\sqrt{2}\max\{C_0,C_i\}\left[\frac{1}{\delta t^2}\left(\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)\right]^{\gamma_i\alpha/(2\gamma_i\alpha+\alpha+1)}.$$

For any $\theta > 0$, the above inequality can be alternately written as

$$P(\|\hat{f}_{i,t} - f_i\|_{\mathcal{H}} > \theta) \leq \min\left\{1, \left(\frac{2\sqrt{2}\max\{C_0, C_i\}}{\theta}\right)^{\frac{2\gamma_i\alpha+\alpha+1}{\gamma_i\alpha}} \left[\frac{1}{t^2}\left(\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)\right]\right\}.$$

The expectation result therefore follows by using $\mathbb{E}(\|\hat{f}_{i,t} - f_i\|_{\mathcal{H}}^{1+\zeta}) = \int_0^\infty P(\|\hat{f}_{i,t} - f_i\|_{\mathcal{H}} > \theta)\theta^\zeta d\theta$.

8.2 Proof of Theorem 2

Since most steps in the proof of Theorem 2 follow that of the proof of Theorem 1, we only highlight the differences when $\mathcal{H}$ is finite-dimensional. Again, without loss of generality, we assume that $i = 1$ and the proof for other arms follows similarly. Recall, $\hat{\pi}_s := P(\hat{a}_s = 1|\mathcal{F}_{s-1}, X_s)$. Note that $(\mathcal{A}_5)$ implies $\mathcal{H}$ is finite-dimensional. Define $d := \dim(\mathcal{H})$. Then

$$N_{\Sigma,1}(\lambda) = \mathrm{Tr}((\Sigma + \lambda I)^{-1}\Sigma) = \sup_i \frac{\eta_i(\Sigma)d}{(\eta_i(\Sigma) + \lambda)} \leq d, \text{ and } \|(\Sigma + \lambda I)^{-1}\|_\infty \leq \frac{1}{\eta}. \tag{50}$$

Therefore (25) modifies to

$$\begin{aligned}
\|\hat{f}_{1,t} - f_1\|_{\mathcal{H}} &\leq \|(\hat{\Sigma}_{1,t} + \lambda I)^{-1/2}\|_\infty \|(\hat{\Sigma}_{1,t} + \lambda I)^{-1/2}(\Sigma + \lambda I)^{1/2}\|_\infty \\
&\qquad \times \left\|(\Sigma + \lambda I)^{-1/2}\left[-\lambda f_1 + \frac{1}{t}\sum_{s=1}^{t}\frac{I\{\hat{a}_s = 1\}}{\hat{\pi}_s}k(\cdot, X_s)\epsilon_s\right]\right\|_{\mathcal{H}} \\
&\leq \|(\Sigma + \lambda I)^{-1/2}\|_\infty \|(\hat{\Sigma}_{1,t} + \lambda I)^{-1/2}(\Sigma + \lambda I)^{1/2}\|_\infty^2 \\
&\qquad \times \left\|(\Sigma + \lambda I)^{-1/2}\left[-\lambda f_1 + \frac{1}{t}\sum_{s=1}^{t}\frac{I\{\hat{a}_s = 1\}}{\hat{\pi}_s}k(\cdot, X_s)\epsilon_s\right]\right\|_{\mathcal{H}} \\
&\leq \frac{\mathcal{S}_1^2\mathcal{S}_2}{\sqrt{\eta}}.
\end{aligned} \tag{51}$$

We now bound $\mathcal{S}_1$ and $\mathcal{S}_2$. Bounding $\mathcal{S}_1$ proceeds exactly as in the proof of Theorem 1 until (29) yielding

$$\begin{aligned}
\textcircled{1} &\leq \frac{1}{t^2}\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\mathrm{Tr}\left((\Sigma + \lambda I)^{-1}\Sigma\right)\|(\Sigma + \lambda I)^{-1}\|_\infty \sup_{X_s}\|k(\cdot, X_s) \otimes k(\cdot, X_s)\|_\infty \\
&\leq \frac{1}{t^2}\left(\sum_{s=1}^{t}\frac{L-1}{\epsilon_s}\right)\frac{d\kappa}{\eta},
\end{aligned} \tag{52}$$

where we use (50) in the last line of the above inequality. Putting together (52), (34) and (36) in (27), and the result in (26), we get the following analog of (37):

$$\mathbb{E}\|B_t\|_{HS}^2 \leq \frac{(L-1)d\kappa}{\eta}\left(\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right).$$

Therefore, using Chebyshev's inequality,

$$P\left(\|B_t\|_{HS} \geq \frac{1}{2}\right) \leq 4\mathbb{E}\|B_t\|_{HS}^2 \leq \frac{4(L-1)d\kappa}{\eta}\left(\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right) \leq \delta.$$

Thus for $\delta > 0$, with probability at least $1-\delta$, we obtain

$$\mathcal{S}_1^2 \leq 2.$$

To bound $\mathcal{S}_2$, we bound the second term in (39) as

$$\lambda\|(\Sigma+\lambda I)^{-1/2}f_1\|_{\mathcal{H}} \leq \lambda\|(\Sigma+\lambda I)^{-1/2}\|_{\infty}\|f_1\|_{\mathcal{H}} \leq \frac{\lambda}{\sqrt{\eta}}\|f_1\|_{\mathcal{H}}. \tag{53}$$

For bounding the first term of (39), we follow the same steps as in the proof of Theorem 1 until (46). By using (50) in (46), we obtain

$$P\left(\left\|(\Sigma+\lambda I)^{-1/2}\frac{1}{t}\sum_{s=1}^{t}\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s}k(\cdot,X_s)e_s\right\|_{\mathcal{H}} \geq \xi\right) \leq \frac{(L-1)\sigma^2 d}{\xi^2}\left(\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right),$$

i.e., with probability at least $1-\delta$,

$$\left\|(\Sigma+\lambda I)^{-1/2}\frac{1}{t}\sum_{s=1}^{t}\frac{I\{\hat{a}_s=1\}}{\hat{\pi}_s}k(\cdot,X_s)e_s\right\|_{\mathcal{H}} \leq \sqrt{\frac{(L-1)\sigma^2 d}{\delta}}\left(\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)^{\frac{1}{2}}. \tag{54}$$

Combining (53) and (54) in (39), we obtain that with probability at least $1-\delta$,

$$\mathcal{S}_2 \leq \sqrt{\frac{(L-1)\sigma^2 d}{\delta}}\left(\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)^{\frac{1}{2}} + \frac{\lambda}{\sqrt{\eta}}\|f_1\|_{\mathcal{H}}.$$

Using these bounds on $\mathcal{S}_1$ and $\mathcal{S}_2$ in (51), we obtain that with probability at least $1-2\delta$,

$$\begin{aligned}\|\hat{f}_{1,t}-f_1\|_{\mathcal{H}} &\leq \left[\frac{4(L-1)\sigma^2 d}{\delta\eta}\left(\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)\right]^{1/2} + \frac{2\lambda}{\eta}\|f_1\|_{\mathcal{H}} \\ &\leq 2\max\{\tilde{C}_0,\tilde{C}_1\}\left[\left(\frac{1}{\delta t^2}\left(\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)\right)^{1/2} + \lambda\right],\end{aligned}$$

where $\tilde{C}_0 := \sqrt{(L-1)\sigma^2 d/\eta}$ and $\tilde{C}_1 := \|f_1\|_{\mathcal{H}}/\eta$. The result, therefore, follows by choosing

$$\lambda = \lambda_t := \left[\frac{1}{\delta t^2}\left(\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)\right]^{1/2}.$$

Note, that same proof works for all arms $i = 1,\ldots,L$, by defining $\hat{\pi}_s = P(\hat{a}_s = i|\mathcal{F}_{s-1}, X_s)$. The expectation bound follows the same idea as in the proof of Theorem 1.

8.3 Proof of Theorem 3

The randomization error in the regret decomposition in (16) can be bounded as

$$\begin{aligned}
&\sum_{t=1}^{T} |f_{A_t}(X_t) - f_{\hat{a}_t}(X_t)| = \sum_{t=1}^{T} |\langle f_{A_t} - f_{\hat{a}_t}, k(\cdot, X_t)\rangle_{\mathcal{H}}| \\
&= \sum_{t=1}^{T} I\{\hat{a}_t \neq A_t\} |\langle f_{A_t} - f_{\hat{a}_t}, k(\cdot, X_t)\rangle_{\mathcal{H}}| \leq \kappa \sum_{t=1}^{T} I\{\hat{a}_t \neq A_t\} \|f_{A_t} - f_{\hat{a}_t}\|_{\mathcal{H}} \\
&\leq \kappa \left(\sum_{t=1}^{T} I\{\hat{a}_t \neq A_t\}\right)^{1/p} \left(\sum_{t=1}^{T} \|f_{A_t} - f_{\hat{a}_t}\|_{\mathcal{H}}^q\right)^{1/q}, \qquad (55) \\
&\leq \kappa \left(\sum_{t=1}^{T} I\{\hat{a}_t \neq A_t\}\right)^{1/p} \left(T \sup_{\substack{a,a' \in \mathcal{A} \\ a \neq a'}} \|f_a - f_{a'}\|_{\mathcal{H}}^q\right)^{1/q} \\
&= \kappa T^{1/q} \left(\sum_{t=1}^{T} I\{\hat{a}_t \neq A_t\}\right)^{1/p} \sup_{\substack{a,a' \in \mathcal{A} \\ a \neq a'}} \|f_a - f_{a'}\|_{\mathcal{H}}, \qquad (56)
\end{aligned}$$

where we obtain (55) by Hölder's inequality with p and q such that $\frac{1}{p} + \frac{1}{q} = 1$, for $p, q \in [1, \infty]$. Next, we bound the first term in (56) below. By the law of iterated expectations, we obtain

$$\mathbb{E}[I\{\hat{a}_t \neq A_t\}] = \mathbb{E}\left[\mathbb{E}(I\{\hat{a}_t \neq A_t\}|\mathcal{F}_{t-1}, X_t)\right] = \mathbb{E}[P(\hat{a}_t \neq A_t|\mathcal{F}_{t-1}, X_t)] = \frac{\epsilon_t}{L-1}. \qquad (57)$$

Therefore, for any $\xi > 0$, Chebyshev's inequality yields

$$\begin{aligned}
&P\left(\left|\sum_{t=1}^{T} I\{\hat{a}_t \neq A_t\} - \sum_{t=1}^{T} \frac{\epsilon_t}{L-1}\right| \geq \xi\right) \leq \frac{1}{\xi^2}\mathbb{E}\left[\sum_{t=1}^{T}\left(I\{\hat{a}_t \neq A_t\} - \frac{\epsilon_t}{L-1}\right)\right]^2 \\
&= \frac{1}{\xi^2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=1}^{T}\left(I\{\hat{a}_t \neq A_t\}I\{\hat{a}_s \neq A_s\} - \frac{\epsilon_t}{L-1}I\{\hat{a}_s \neq A_s\} - \frac{\epsilon_s}{L-1}I\{\hat{a}_t \neq A_t\} + \frac{\epsilon_t\epsilon_s}{(L-1)^2}\right)\right].
\end{aligned}$$

Now, we simplify the expectation in the r.h.s. of the above inequality by considering cases: (i) $s = t$, (ii) $s < t$ and (iii) $s > t$ as

$$P\left(\left|\sum_{t=1}^{T} I\{\hat{a}_t \neq A_t\} - \sum_{t=1}^{T} \frac{\epsilon_t}{L-1}\right| \geq \xi\right) \leq \textcircled{7} + \textcircled{8} + \textcircled{9},$$

where

$$\begin{aligned}
⑦ &:= \frac{1}{\xi^2}\mathbb{E}\left[\sum_{t=1}^{T}\left(1-\frac{2\epsilon_t}{L-1}\right)I\{\hat{a}_t \neq A_t\} + \frac{\epsilon_t^2}{(L-1)^2}\right] \\
&\overset{(57)}{=} \frac{1}{\xi^2}\sum_{t=1}^{T}\left[\left(1-\frac{2\epsilon_t}{L-1}\right)\left(\frac{\epsilon_t}{L-1}\right) + \frac{\epsilon_t^2}{(L-1)^2}\right] \\
&= \frac{1}{\xi^2}\sum_{t=1}^{T}\frac{\epsilon_t}{L-1} - \frac{\epsilon_t^2}{(L-1)^2} = \frac{1}{\xi^2}\sum_{t=1}^{T}\frac{\epsilon_t}{L-1}\left[1-\frac{\epsilon_t}{L-1}\right],
\end{aligned}$$

and

$$\begin{aligned}
⑧ &:= \frac{1}{\xi^2}\sum_{t=2}^{T}\sum_{s=1}^{t-1}\Big[\mathbb{E}\big[I\{\hat{a}_t \neq A_t\}I\{\hat{a}_s \neq A_s\}\big] - \frac{\epsilon_t}{L-1}\mathbb{E}\big[I\{\hat{a}_s \neq A_s\}\big] - \frac{\epsilon_s}{L-1}\mathbb{E}\big[I\{\hat{a}_t \neq A_t\}\big] \\
&\qquad + \frac{\epsilon_t\epsilon_s}{(L-1)^2}\Big] \\
&= 0,
\end{aligned}$$

which follows from (57) and for $s < t$,

$$\begin{aligned}
\mathbb{E}\big[I\{\hat{a}_t \neq A_t\}I\{\hat{a}_s \neq A_s\}\big] &= \mathbb{E}\big[\mathbb{E}(I\{\hat{a}_t \neq A_t\}I\{\hat{a}_s \neq A_s\} \mid \mathcal{F}_{t-1}, X_t)\big] \\
&= \mathbb{E}\big[I\{\hat{a}_s \neq A_s\}P(\hat{a}_t \neq A_t|\mathcal{F}_{t-1}, X_t)\big] \\
&= \mathbb{E}\Big[I\{\hat{a}_s \neq A_s\} \times \frac{\epsilon_t}{L-1}\Big] \\
&\overset{(57)}{=} \left(\frac{\epsilon_s}{L-1}\right)\left(\frac{\epsilon_t}{L-1}\right) = \mathbb{E}[I\{\hat{a}_t \neq A_t\}]\mathbb{E}[I\{\hat{a}_s \neq A_s\}].
\end{aligned}$$

Through similar calculation, it follows that

$$\begin{aligned}
⑨ &:= \frac{1}{\xi^2}\left[\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\mathbb{E}[I\{\hat{a}_t \neq A_t\}I\{\hat{a}_s \neq A_s\}] - \frac{\epsilon_t}{L-1}\mathbb{E}[I\{\hat{a}_s \neq A_s\}] - \frac{\epsilon_s}{L-1}\mathbb{E}[I\{\hat{a}_t \neq A_t\}]\right. \\
&\qquad \left. + \frac{\epsilon_t\epsilon_s}{(L-1)^2}\right] \\
&= 0.
\end{aligned}$$

Therefore, we obtain

$$P\left(\left|\sum_{t=1}^{T}I\{\hat{a}_t \neq A_t\} - \sum_{t=1}^{T}\frac{\epsilon_t}{L-1}\right| \geq \xi\right) \leq \frac{1}{\xi^2}\sum_{t=1}^{T}\frac{\epsilon_t}{L-1}\left[1-\frac{\epsilon_t}{L-1}\right] \leq \frac{1}{\xi^2}\sum_{t=1}^{T}\frac{\epsilon_t}{L-1},$$

which implies for any $\delta > 0$, with probability at least $1-\delta$,

$$\sum_{t=1}^{T}\frac{\epsilon_t}{L-1} - \left[\frac{1}{\delta}\sum_{t=1}^{T}\frac{\epsilon_t}{L-1}\right]^{1/2} \leq \sum_{t=1}^{T}I\{\hat{a}_t \neq A_t\} \leq \sum_{t=1}^{T}\frac{\epsilon_t}{L-1} + \left[\frac{1}{\delta}\sum_{t=1}^{T}\frac{\epsilon_t}{L-1}\right]^{1/2}. \quad (58)$$

Using (58) in (56) yields that with probability at least $1-\delta$,

$$\sum_{t=1}^{T}|f_{A_t}(X_t)-f_{\hat{a}_t}(X_t)|\leq \kappa T^{1-\frac{1}{p}}\left[\sum_{t=1}^{T}\frac{\epsilon_t}{L-1}+\left\{\frac{1}{\delta}\sum_{t=1}^{T}\frac{\epsilon_t}{L-1}\right\}^{1/2}\right]^{1/p}\sup_{\substack{a,a'\in\mathcal{A}\\ a\neq a'}}\|f_a-f_{a'}\|_{\mathcal{H}}. \tag{59}$$

Next, we construct an upper bound for the cumulative estimation error in the regret decomposition in (16). By recalling $\mathcal{A}:=\{1,\ldots,L\}$, consider

$$\sup_{i\in\mathcal{A}}|(f_i(X_t)-\hat{f}_i(X_t))|=\sup_{i\in\mathcal{A}}|\langle f_i-\hat{f}_{i,t},k(\cdot,X_t)\rangle_{\mathcal{H}}|\leq\kappa\sup_{i\in\mathcal{A}}\|f_i-\hat{f}_{i,t}\|_{\mathcal{H}}, \tag{60}$$

where we will use Theorem 1 to bound (60). To this end, by union bounding, we have

$$\begin{aligned}
&P\left(\sup_{i\in\mathcal{A}}\|f_i-\hat{f}_{i,t}\|_{\mathcal{H}}\geq b_t\right)\leq\sum_{i=1}^{L}P\left(\|f_i-\hat{f}_{i,t}\|_{\mathcal{H}}\geq b_t\right)\\
&\leq\left(\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)\sum_{i=1}^{L}\left(\frac{2\sqrt{2}\max\{C_0,C_i\}}{b_t}\right)^{1/w_i}\\
&\leq L\left(\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)\max_{i\in\mathcal{A}}\left(\frac{2\sqrt{2}\max\{C_0,C_i\}}{b_t}\right)^{1/w_i}\\
&\leq L\left(\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)\max_{i\in\mathcal{A}}\left(\frac{\Theta}{b_t}\right)^{1/w_i}\\
&\leq L\left(\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)\times\begin{cases}\left(\frac{\Theta}{b_t}\right)^{\max_{i\in\mathcal{A}}1/w_i}, & b_t<\Theta\\ \left(\frac{\Theta}{b_t}\right)^{\min_{i\in\mathcal{A}}1/w_i}, & b_t\geq\Theta\end{cases},
\end{aligned}$$

where $\Theta:=\max_{i\in\mathcal{A}}2\sqrt{2}\max\{C_0,C_i\}$ and $w_i=\gamma_i\alpha/(2\gamma_i\alpha+\alpha+1)$. This means with probability at least $1-\delta$,

$$\sup_{i\in\mathcal{A}}\|f_i-\hat{f}_{i,t}\|_{\mathcal{H}}\leq\begin{cases}\Theta\Delta_t^{\min_{i\in\mathcal{A}}w_i}, & \Delta_t<1\\ \Theta\Delta_t^{\max_{i\in\mathcal{A}}w_i}, & \Delta_t\geq 1\end{cases}=\begin{cases}\Theta\Delta_t^{\frac{(\min_{i\in\mathcal{A}}\gamma_i)\alpha}{2(\min_{i\in\mathcal{A}}\gamma_i)\alpha+\alpha+1}}, & \Delta_t<1\\ \Theta\Delta_t^{\frac{(\max_{i\in\mathcal{A}}\gamma_i)\alpha}{2(\max_{i\in\mathcal{A}}\gamma_i)\alpha+\alpha+1}}, & \Delta_t\geq 1\end{cases}, \tag{61}$$

where

$$\Delta_t:=\frac{L}{\delta t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}$$

and used the fact that $h(x)=\frac{x\alpha}{2x\alpha+\alpha+1}$ is a strictly increasing function of x for all $\alpha>0$. The result follows by using (61) in (60) and combining it with (59) in (16), while noting that $\lambda_{i,t}$ is given by the choice in (8) in Theorem 1 but with δ replaced by δ/L.

8.4 Proof of Theorem 4

We bound the randomization error exactly as in the proof of Theorem 3 in Section 8.3. For bounding the cumulative estimation error, instead of Theorem 1, we use the bound from Theorem 2 in (60). Then using the same union bounding idea as in the proof of Theorem 3, we obtain

$$P\left(\sup_{i\in\mathcal{A}}\|f_i-\hat{f}_{i,t}\|_{\mathcal{H}}\geq b_t\right)\leq\sum_{i=1}^{L}P\left(\|f_i-\hat{f}_{i,t}\|_{\mathcal{H}}\geq b_t\right)$$
$$\leq\left(\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)\sum_{i=1}^{L}\left(\frac{4\max\{\tilde{C}_0,\tilde{C}_i\}}{b_t}\right)^2\leq L\left(\frac{1}{t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)\left(\frac{4\max\{\tilde{C}_0,\tilde{C}_*\}}{b_t}\right)^2,$$

which implies that with probability at least $1-\delta$,

$$\sup_{i\in\mathcal{A}}\|f_i-\hat{f}_{i,t}\|_{\mathcal{H}}\leq 4\max\{\tilde{C}_0,\tilde{C}_*\}\left(\frac{L}{\delta t^2}\sum_{s=1}^{t}\frac{1}{\epsilon_s}\right)^{1/2}.$$

The result follows by using the above bound in (60) and combining it with (59) in (16).

8.5 Proof of Theorem 5

Since $L=2$, $a\in\{0,1\}$. Recall, $A_s=\arg\max_{a\in\{0,1\}}\hat{f}_{a,s-1}(X_s)$. Note that the regret in Definition 1 can be written as

$$R_T=\sum_{s=1}^{T}|f_1(X_s)-f_0(X_s)|I\{\hat{a}_s\neq a_s^*\},$$

where

$$I\{\hat{a}_s\neq a_s^*\}=I\{\hat{a}_s\neq a_s^*,a_s^*=A_s\}+I\{\hat{a}_s\neq a_s^*,a_s^*\neq A_s\}\leq I\{\hat{a}_s\neq A_s\}+I\{A_s\neq a_s^*\}.$$

Therefore,

$$\mathbb{E}R_T\leq\mathbb{E}\sum_{s=1}^{T}|f_1(X_s)-f_0(X_s)|I\{\hat{a}_s\neq A_s\}+\mathbb{E}\sum_{s=1}^{T}|f_1(X_s)-f_0(X_s)|I\{A_s\neq a_s^*\}$$
$$=R_T^{(1)}+R_T^{(2)},\tag{62}$$

where

$$R_T^{(1)}:=\mathbb{E}\sum_{s=1}^{T}|f_1(X_s)-f_0(X_s)|I\{\hat{a}_s\neq A_s\}$$

is the error due to exploration (or randomization) and

$$R_T^{(2)}:=\mathbb{E}\sum_{s=1}^{T}|f_1(X_s)-f_0(X_s)|I\{A_s\neq a_s^*\}$$

is the cumulative estimation error. Next, we bound these two terms as follows.

$$\begin{aligned} R_T^{(2)} &= \mathbb{E}\sum_{s=1}^{T} |f_1(X_s) - f_0(X_s)| I\{A_s \neq a_s^*\} \\ &= \sum_{s=1}^{T} \mathbb{E}\left[|f_1(X_s) - f_0(X_s)| I\{A_s \neq a_s^*\}\right] \\ &= \sum_{s=1}^{T} \mathbb{E}\left[\mathbb{E}\left[|f_1(X_s) - f_0(X_s)| I\{A_s \neq a_s^*\} \Big| \mathcal{F}_{s-1}\right]\right]. \end{aligned} \tag{63}$$

We only consider the inner expectation from here onwards and find an upper bound for that. Let $\hat{f}_{s-1} = \hat{f}_{1,s-1} - \hat{f}_{0,s-1}$ and $f_- = f_1 - f_0$. We have that

$$\begin{aligned} &\mathbb{E}\left[|f_1(X_s) - f_0(X_s)| I\{A_s \neq a_s^*\} \Big| \mathcal{F}_{s-1}\right] \\ &= -\mathbb{E}\left[\left(I\{\hat{f}_{s-1}(X_s) \geq 0\} - I\{f_-(X_s) \geq 0\}\right) I\{f_-(X_s) \neq 0\} f_-(X_s) \Big| \mathcal{F}_{s-1}\right] \geq 0. \end{aligned} \tag{64}$$

Similarly, we have that

$$\mathbb{E}\left[\left(I\{\hat{f}_{s-1}(X_s) \geq 0\} - I\{f_-(X_s) \geq 0\}\right) I\{f_-(X_s) \neq 0\} \hat{f}_{s-1}(X_s) \Big| \mathcal{F}_{s-1}\right] \geq 0. \tag{65}$$

Therefore using the fact that both (64) and (65) are non-negative, we get

$$\begin{aligned} &\sum_{s=1}^{T} \mathbb{E}\left[|f_1(X_s) - f_0(X_s)| I\{A_s \neq a_s^*\} \Big| \mathcal{F}_{s-1}\right] \\ &\leq \sum_{s=1}^{T} \mathbb{E}\left[\left(I\{\hat{f}_{s-1}(X_s) \geq 0\} - I\{f_-(X_s) \geq 0\}\right) I\{f_-(X_s) \neq 0\} (\hat{f}_{s-1}(X_s) - f_-(X_s)) \Big| \mathcal{F}_{s-1}\right] \\ &= S_1 + S_2, \end{aligned} \tag{66}$$

where for $\theta > 0$,

$$\begin{aligned} S_1 := \sum_{s=1}^{T} \mathbb{E}\Big[I\{0 < |f_-(X_s)| \leq T^{-\theta}\} \left(I\{\hat{f}_{s-1}(X_s) \geq 0\} - I\{f_-(X_s) \geq 0\}\right) \\ \times I\{f_-(X_s) \neq 0\} (\hat{f}_{s-1} - f_-)(X_s) \Big| \mathcal{F}_{s-1}\Big], \end{aligned}$$

and

$$\begin{aligned} S_2 := \sum_{s=1}^{T} \mathbb{E}\Big[I\{|f_-(X_s)| > T^{-\theta}\} \left(I\{\hat{f}_{s-1}(X_s) \geq 0\} - I\{f_-(X_s) \geq 0\}\right) \\ \times I\{f_-(X_s) \neq 0\} (\hat{f}_{s-1} - f_-)(X_s) \Big| \mathcal{F}_{s-1}\Big]. \end{aligned}$$

Note that

$$S_1 \leq \sum_{s=1}^{T} \mathbb{E}\left[I\{0 < |f_-(X_s)| \leq T^{-\theta}\}(\hat{f}_{s-1} - f_-)(X_s) \Big| \mathcal{F}_{s-1} \right]$$

$$\leq \sum_{s=1}^{T} C\kappa T^{-\theta} \|\hat{f}_{s-1} - f_-\|_{\mathcal{H}} \tag{67}$$

$$\leq C\kappa T^{-\theta} \sum_{s=1}^{T} \|\hat{f}_{s-1} - f_-\|_{\mathcal{H}} \leq C\kappa T^{-\theta} \sum_{s=1}^{T} \left[\|\hat{f}_{1,s-1} - f_1\|_{\mathcal{H}} + \|\hat{f}_{0,s-1} - f_0\|_{\mathcal{H}} \right], \tag{68}$$

where (67) follows from $(\mathcal{A}_6)$ and the fact that $\sup_{x\in\mathcal{X}} k(x,x) \leq \kappa$. The last inequality follows from the definition of $\hat{f}_{s-1}$ and f_-. Now, taking the expectation of S_1 in (68) we get that,

$$\mathbb{E}S_1 \leq C\kappa T^{-\theta} \sum_{s=1}^{T} \left[\mathbb{E}\|\hat{f}_{1,s-1} - f_1\|_{\mathcal{H}} + \mathbb{E}\|\hat{f}_{0,s-1} - f_0\|_{\mathcal{H}} \right]$$

$$\leq 2C\kappa T^{-\theta} \max_{i\in\{0,1\}} B(C_0, \|\Sigma^{-\gamma} f_i\|_{\mathcal{H}}, \gamma, 0, \alpha) \sum_{t=1}^{T} \left[\frac{1}{t^2} \sum_{s=1}^{t} \frac{1}{\epsilon_s} \right]^{w}, \tag{69}$$

where the last inequality follows from Theorem 1 with $C_0 = \sqrt{\sigma^2 A_1(\bar{C}, \alpha)}$ and $w = \gamma\alpha/(2\gamma\alpha + \alpha + 1)$.

Next, to construct an upper bound for S_2, we use the fact that

$$I\{|(\hat{f}_{s-1} - f_-)(x)| > |f_-(x)|\} \geq I\{\hat{f}_{s-1}(x) \geq 0\} - I\{f_-(x) \geq 0\}.$$

For $\zeta \geq 0$, we obtain

$$S_2 \leq \sum_{s=1}^{T} \mathbb{E}\left[I\{|f_-(X_s)| > T^{-\theta}\} I\{|(\hat{f}_{s-1} - f_-)(X_s)| > |f_-(X_s)|\} |(\hat{f}_{s-1} - f_-)(X_s)| \Big| \mathcal{F}_{s-1} \right]$$

$$\leq \sum_{s=1}^{T} \mathbb{E}\left[I\{|f_-(X_s)| > T^{-\theta}\} \frac{|(\hat{f}_{s-1} - f_-)(X_s)|^{1+\zeta}}{|f_-(X_s)|^{\zeta}} \Big| \mathcal{F}_{s-1} \right]$$

$$\leq T^{\theta\zeta} \sum_{s=1}^{T} \mathbb{E}\left[|(\hat{f}_{s-1} - f_-)(X_s)|^{1+\zeta} \Big| \mathcal{F}_{s-1} \right] \leq \kappa^{1+\zeta} T^{\theta\zeta} \sum_{s=1}^{T} \|\hat{f}_{s-1} - f_-\|_{\mathcal{H}}^{1+\zeta}. \tag{70}$$

Taking expectation of S_2, choice of $\lambda_{i,t}$ as in (8) and using (11) we obtain

$$\mathbb{E}S_2 \leq 2 \max_{i\in\{0,1\}} B(C_0, \|\Sigma^{-\gamma} f_i\|_{\mathcal{H}}, \gamma, \zeta, \alpha) \kappa^{1+\zeta} T^{\theta\zeta} \sum_{s=1}^{T} \left[\frac{1}{t^2} \sum_{s=1}^{t} \frac{1}{\epsilon_s} \right]^{w(1+\zeta)}, \tag{71}$$

for $0 \leq \zeta < \frac{\gamma\alpha+\alpha+1}{\gamma\alpha}$. Combining (69), (71), and (66) in (63), we obtain

$$R_T^{(2)} \leq \mathbb{E}S_1 + \mathbb{E}S_2$$

$$\lesssim T^{-\theta} \sum_{t=1}^{T} \left[\frac{1}{t^2} \sum_{s=1}^{t} \frac{1}{\epsilon_s} \right]^{w} + T^{\theta\zeta} \sum_{t=1}^{T} \left[\frac{1}{t^2} \sum_{s=1}^{t} \frac{1}{\epsilon_s} \right]^{w(1+\zeta)}. \tag{72}$$

Now, we bound $R_T^{(1)}$ as

$$\begin{aligned} R_T^{(1)} &= \mathbb{E}\sum_{s=1}^{T} |f_1(X_s) - f_0(X_s)| I\{\hat{a}_s \neq A_s\} \\ &\leq \kappa \|f_1 - f_0\|_{\mathcal{H}} \sum_{s=1}^{T} \mathbb{E} I\{\hat{a}_s \neq A_s\} \leq \kappa \|f_1 - f_0\|_{\mathcal{H}} \sum_{s=1}^{T} P(\hat{a}_s \neq A_s) \\ &= \kappa \|f_1 - f_0\|_{\mathcal{H}} \sum_{s=1}^{T} \frac{\epsilon_s}{2}. \end{aligned} \tag{73}$$

Combining (72) and (73) in (62) yields the result.

8.6 Proof of Theorem 6

Recall, $A_s = \arg\max_{a\in\{0,1\}} \hat{f}_{a,s-1}(X_s)$. We repeat the same steps of the proof of Theorem 5 but specialized to the setting of a finite-dimensional RKHS. Again, we re-write the cumulative regret for the proposed strategy as,

$$R_T = \sum_{s=1}^{T} |f_1(X_s) - f_0(X_s)| I\{\hat{a}_s \neq a_s^*\}.$$

We then break the expected regret into $R_T^{(1)}$, the error due to exploration (or randomization) and $R_T^{(2)}$, the cumulative estimation error. Let us first consider $R_T^{(2)}$. Following exactly the same steps as in the proof of Theorem 5, we then split $R_T^{(2)}$ into two parts, S_1 and S_2 respectively, as in (66). Using Assumption $(\mathcal{A}_6)$ and the fact that $\sup_{x\in\mathcal{X}} k(x,x) \leq \kappa$, we get

$$\begin{aligned} S_1 &\leq \sum_{s=1}^{T} \mathbb{E}\left[I\{0 < |f_-(X_s)| \leq T^{-\theta}\} (\hat{f}_{s-1} - f_-)(X_s) \big| \mathcal{F}_{s-1} \right] \\ &\leq \sum_{s=1}^{T} C\kappa T^{-\theta} \|\hat{f}_{s-1} - f_-\|_{\mathcal{H}} \leq C\kappa T^{-\theta} \sum_{s=1}^{T} \|\hat{f}_{s-1} - f_-\|_{\mathcal{H}} \\ &\leq C\kappa T^{-\theta} \sum_{s=1}^{T} \left[\|\hat{f}_1 - f_1\|_{\mathcal{H}} + \|\hat{f}_0 - f_0\|_{\mathcal{H}} \right], \end{aligned}$$

yielding

$$\mathbb{E}S_1 \leq 2C\kappa T^{-\theta} \max_{i\in\{0,1\}} B(\tilde{C}_0, \tilde{C}_i, 0, \eta) \sum_{t=1}^{T} \left[\frac{1}{t^2} \sum_{s=1}^{t} \frac{1}{\epsilon_s} \right]^{1/2},$$

where the last inequality from (13) in Theorem 2. For S_2, similar to (70), for the choice of $\lambda_{i,t}$ as in (22) we get that for $0 \leq \zeta < 1$,

$$\mathbb{E}S_2 \leq \kappa^{1+\zeta} T^{\theta\zeta} \sum_{s=1}^{T} \mathbb{E}\|\hat{f}_{s-1} - f_-\|_{\mathcal{H}}^{1+\zeta} \leq 2 \max_{i\in\{0,1\}} B(\tilde{C}_0, \tilde{C}_i, \zeta, \eta) \kappa^{1+\zeta} T^{\theta\zeta} \sum_{t=1}^{T} \left[\frac{1}{t} \sum_{s=1}^{t} \frac{1}{\epsilon_s} \right]^{(1+\zeta)/2},$$

where the last inequality follows from (13). The result follows by using the bound in (73) for $R_T^{(1)}$ and using these bounds in $\mathbb{E}R_T \leq R_T^{(1)} + R_T^{(2)}$.

Acknowledgments

BKS is partially supported by National Science Foundation (NSF) CAREER Award DMS-1945396.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Agarwal, A., Dudik, M., Kale, S., Langford, J., and Schapire, R. (2012). Contextual bandit learning with predictable rewards. In Lawrence, N. D. and Girolami, M., editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 19–26, La Palma, Canary Islands. PMLR.

Arya, S. and Yang, Y. (2020). Randomized allocation with nonparametric estimation for contextual multi-armed bandits with delayed rewards. *Statistics & Probability Letters*, 164:108818.

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

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

Bather, J. A. (1985). On the allocation of treatments in sequential medical trials. *International Statistical Review / Revue Internationale de Statistique*, 53(1):1–13.

Bietti, A., Agarwal, A., and Langford, J. (2021). A contextual bandit bake-off. *Journal of Machine Learning Research*, 22(133):1–49.

Bogunovic, I. and Krause, A. (2021). Misspecified Gaussian process bandit optimization. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3004–3015. Curran Associates, Inc.

Cai, X. and Scarlett, J. (2021). On lower bounds for standard and robust Gaussian process bandit optimization. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1216–1226. PMLR.

Camilleri, R., Jamieson, K., and Katz-Samuels, J. (2021). High-dimensional experimental design and kernel bandits. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 1227–1237. PMLR.

Chen, H., Lu, W., and Song, R. (2021). Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association*, 116(533):240–255. PMID: 33737759.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In Gordon, G., Dunson, D., and Dudik, M., editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 208–214, Fort Lauderdale, FL, USA. PMLR.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. *21st Annual Conference on Learning Theory*, pages 355–366.

Dann, C., Mansour, Y., Mohri, M., Sekhari, A., and Sridharan, K. (2022). Guarantees for epsilon-greedy reinforcement learning with function approximation. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 4666–4689. PMLR.

Dimakopoulou, M., Zhou, Z., Athey, S., and Imbens, G. (2019). Balanced linear contextual bandits. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Goldenshluger, A. and Zeevi, A. (2013). A linear response bandit problem. *Stochastic Systems*, 3(1):230 – 261.

Gopalan, A., Mannor, S., and Mansour, Y. (2014). Thompson sampling for complex online problems. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 100–108, Bejing, China. PMLR.

Hu, Y., Kallus, N., and Mao, X. (2022). Smooth contextual bandits: Bridging the parametric and nondifferentiable regret regimes. *Operations Research*, 70(6):3261–3281.

Janz, D., Burt, D., and Gonzalez, J. (2020). Bandit optimization of functions in the Matérn kernel RKHS. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2486–2495. PMLR.

Kleinberg, R., Slivkins, A., and Upfal, E. (2008). Multi-armed bandits in metric spaces. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, pages 681–690, New York, NY, USA. Association for Computing Machinery.

Krause, A. and Ong, C. (2011). Contextual Gaussian process bandit optimization. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Langford, J. and Zhang, T. (2007). The epoch-greedy algorithm for multi-armed bandits with side information. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

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

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 661–670, New York, NY, USA. Association for Computing Machinery.

Magureanu, S., Combes, R., and Proutiere, A. (2014). Lipschitz bandits: Regret lower bound and optimal algorithms. In Balcan, M. F., Feldman, V., and Szepesvári, C., editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 975–999, Barcelona, Spain. PMLR.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., and Murphy, S. A. (2017). Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462.

Qian, W., Ing, C.-K., and Liu, J. (2023). Adaptive algorithm for multi-armed bandit problem with high-dimensional covariates. *Journal of the American Statistical Association*, pages 1–13.

Qian, W. and Yang, Y. (2016). Kernel estimation and model combination in a bandit problem with covariates. *Journal of Machine Learning Research*, 17(1):5181–5217.

Rigollet, P. and Zeevi, A. (2010). Nonparametric bandits with covariates. *Conference on Learning Theory (COLT)*, page 54.

Rudi, A., Canas, G. D., and Rosasco, L. (2013). On the sample complexity of subspace learning. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411.

Scarlett, J., Bogunovic, I., and Cevher, V. (2017). Lower bounds on regret for noisy Gaussian process bandit optimization. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1723–1742. PMLR.

Slivkins, A. (2014). Contextual bandits with similarity information. *Journal of Machine Learning Research*, 15(1):2533–2568.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 1015–1022, Madison, WI, USA. Omnipress.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT press.

Tewari, A. and Murphy, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer.

Vakili, S., Ahmed, D., Bernacchia, A., and Pike-Burke, C. (2023). Delayed feedback in kernel bandits. *arXiv preprint arXiv:2302.00392*.

Vakili, S., Khezeli, K., and Picheny, V. (2021a). On information gain and regret bounds in Gaussian process bandits. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 82–90. PMLR.

Vakili, S., Scarlett, J., and Javidi, T. (2021b). Open problem: Tight online confidence intervals for RKHS elements. In Belkin, M. and Kpotufe, S., editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4647–4652. PMLR.

Valko, M., Korda, N., Munos, R., Flaounas, I., and Cristianini, N. (2013). Finite-time analysis of kernelised contextual bandits. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, pages 654–663, Arlington, Virginia, USA. AUAI Press.

Villar, S. S., Bowden, J., and Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199.

Yang, Y. and Zhu, D. (2002). Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics*, 30(1):100–121.

Zenati, H., Bietti, A., Diemert, E., Mairal, J., Martin, M., and Gaillard, P. (2022). Efficient kernelized ucb for contextual bandits. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 5689–5720. PMLR.

Zhou, D., Li, L., and Gu, Q. (2020). Neural contextual bandits with UCB-based exploration. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11492–11502. PMLR.

Zhu, Y., Zhou, D., Jiang, R., Gu, Q., Willett, R., and Nowak, R. (2021). Pure exploration in kernel and neural bandits. *Advances in Neural Information Processing Systems*, 34:11618–11630.