Approximation Theory Based Methods for RKHS Bandits

Sho Takemori [1] **Masahiro Sato** [1]

Abstract

The RKHS bandit problem (also called kernelized multi-armed bandit problem) is an online optimization problem of non-linear functions with noisy feedback. Although the problem has been extensively studied, there are unsatisfactory results for some problems compared to the well-studied linear bandit case. Specifically, there is no general algorithm for the adversarial RKHS bandit problem. In addition, high computational complexity of existing algorithms hinders practical application. We address these issues by considering a novel amalgamation of approximation theory and the misspecified linear bandit problem. Using an approximation method, we propose efficient algorithms for the stochastic RKHS bandit problem and the first general algorithm for the adversarial RKHS bandit problem. Furthermore, we empirically show that one of our proposed methods has comparable cumulative regret to IGP-UCB and its running time is much shorter.

1. Introduction

The RKHS bandit problem (also called kernelized multi-armed bandit problem) is an online optimization problem of non-linear functions with noisy feedback. Srinivas et al. (2010) studied a multi-armed bandit problem where the reward function belongs to the reproducing kernel Hilbert space (RKHS) associated with a kernel. In this paper, we call this problem the (stochastic) RKHS bandit problem. Although the problem has been studied extensively, some issues are not completely solved yet. In this paper, we focus on mainly two issues: non-existence of general algorithms for the adversarial RKHS bandit problem and high computational complexity for the stochastic RKHS bandit algorithms.

First, as a non-linear generalization of the classical adversarial linear bandit problem, Chatterji et al. (2019) proposed the adversarial RKHS bandit problem, where a learner interacts with a sequence of any functions from the RKHS with bounded norms. However, they only consider the kernel loss, i.e., a loss function of the form $x \mapsto K(x, x_0)$, where x_0 is a fixed point. Considering functions in the RKHS can be represented as infinite linear combinations of such functions, kernel loss is a very special function in the RKHS. Therefore, there are no algorithms for the adversarial RKHS bandits with general loss (or reward) functions.

Next, we discuss the efficiency of existing methods for the stochastic RKHS bandit problem. We note that most of the existing methods have regret guarantees at the cost of high computational complexity. For example, IGP-UCB (Chowdhury & Gopalan, 2017) requires matrix-vector multiplication of size t for each arm at each round $t = 1, \dots, T$. Therefore, the total computational complexity up to round T is given as $O(|\mathcal{A}|T^3)$, where $\mathcal{A}$ is the set of arms. To address the issue, Calandriello et al. (2020) proposed BBKB and proved its total computational complexity is given as $\widetilde{O}(|\mathcal{A}|T\gamma_T^2 + \gamma_T^4)$, where $\mathcal{A} \subset \Omega$ is the set of arms, Ω is a subset of a Euclidean space $\mathbb{R}^d$, and γ_T is the maximum information gain (Srinivas et al., 2010). If the kernel is a squared exponential kernel, then since $\gamma_T = \widetilde{O}(\log^d(T))$[1] (Srinivas et al., 2010), ignoring the polylogarithmic factor, BBKB's computational complexity is nearly linear in T. However, the coefficient $|\mathcal{A}|$ in the term is large in general.

In this paper, we address these two issues by considering a novel amalgamation of approximation theory (Wendland, 2004) and the misspecified linear bandit problem (Lattimore et al., 2020). That is, we approximately reduce the RKHS bandit problem to the well-studied linear bandit problem. Here, because of an approximation error, the model is a misspecified linear model. Ordinary approximation methods (such as Random Fourier Features or Nyström embedding) basically aim to approximate kernel $K(x, y)$ by an inner product of finite dimensional vectors. However, to reduce the RKHS bandits to the linear bandits, we want to approximate a function f in the RKHS $\mathcal{H}_K(\Omega)$ by a function ϕ in a finite dimensional subspace so that $\|f - \phi\|_{L^\infty(\Omega)}$ is small. Since the usual approximation methods are not appropriate for the purpose, in this paper, we utilize a method developed

[1]FUJIFILM Business Innovation, Kanagawa, Japan. Correspondence to: Sho Takemori <sho.takemori.py@fujifilm.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]In this paper, we use $\widetilde{O}$ notation to ignore $\log^c(T)$ factor, where c is a universal constant.

in the approximation theory literature called the P-greedy algorithm (De Marchi et al., 2005) to minimize the L^∞ error. More precisely, we shall introduce that any function f in the RKHS is approximately equal (in terms of the L^∞ norm) to a linear combination of $D_{q,\alpha}(T)$ functions, where $q, \alpha > 0$ are parameters and $D_{q,\alpha}(T)$ is the number of functions (or equivalently points) returned by the P-greedy algorithm (Algorithm 1) with admissible error $\mathfrak{e} = \frac{\alpha}{T^q}$. If K is sufficiently smooth, $D_{q,\alpha}(T)$ is much smaller than T and $|\mathcal{A}|$. By this approximation, we can tackle the original RKHS bandit problem by applying an algorithm for the misspecified linear bandit problem.

Contributions

To state contributions, we introduce terminology for kernels. In this paper, we consider two types of kernels: kernels with infinite smoothness and those with finite smoothness with smoothness parameter ν (we provide a precise definition in §4). Examples of the former include Rational Quadratic (RQ) and Squared Exponential (SE) kernels and those of the latter include the Matérn kernels with parameter ν. The latter type of kernels also include a general class of kernels that belong to $C^{2\nu}(\Omega \times \Omega)$ with $\nu \in \frac{1}{2}\mathbb{Z}_{>0}$ and satisfy some additional conditions. Let $D_{q,\alpha}(T) \in \mathbb{Z}_{>0}$ be as before. Then, in §4, we shall introduce that $D_{q,\alpha}(T) = O\left((q \log T - \log(\alpha))^d\right)$ if K has infinite smoothness and $D_{q,\alpha}(T) = O\left(\alpha^{-d/\nu} T^{dq/\nu}\right)$ if K has finite smoothness. Our contributions are stated as follows:

1. We apply an approximation method that has not been applied to the RKHS bandit problem and reduce the problem to the well-studied (misspecified) linear bandit problem. This novel reduction method has potential to tackle issues other than the ones we deal with in this paper.

2. We propose APG-EXP3 for the adversarial RKHS bandit problem, where APG stands for an Approximation theory based method using P-Greedy. We prove its expected cumulative regret is upper bounded by $\widetilde{O}\left(\sqrt{T D_{1,\alpha}(T) \log(|\mathcal{A}|)}\right)$, where $\alpha = \log(|\mathcal{A}|)$. To the best of our knowledge, this is the first method for the adversarial RKHS bandit problem with general reward functions.

3. We propose a method for the stochastic RKHS bandit problem called APG-PE and prove its cumulative regret is given as $\widetilde{O}\left(\sqrt{T D_{1/2,\alpha}(T) \log\left(\frac{|\mathcal{A}|}{\delta}\right)}\right)$, with probability at least $1-\delta$ and its total computational complexity is given as $\widetilde{O}\left((|\mathcal{A}| + T) D_{1/2,\alpha}^2(T)\right)$. We note that the total computational complexity is generally much better than that of the state of the art result $\widetilde{O}(|\mathcal{A}| T \gamma_T^2 + \gamma_T^4)$ (Calandriello et al., 2020).

4. We propose APG-UCB as an approximation of IGP-UCB and provide an upper bound of its cumulative regret if $q \geq 1/2$ and prove that its total computational complexity is given as $O(|\mathcal{A}| T D_{q,\alpha}^2(T))$.

   If we take the parameter q so that $q > 3/2$, then we shall show that $R_{\text{APG-UCB}}(T)$ is upper bounded by $4\beta_T^{\text{IGP-UCB}} \sqrt{\gamma_T T} + O(\sqrt{T \gamma_T} T^{(3/2-q)/2} + \gamma_T T^{1-q})$, where we define $\beta_T^{\text{IGP-UCB}}$ in §6. Since the upper bound for the cumulative regret of IGP-UCB is also given as $4\beta_T^{\text{IGP-UCB}} \sqrt{\gamma_T (T+2)}$, APG-UCB has asymptotically the same regret upper bound as that of IGP-UCB in this case. If the kernel has infinite smoothness or finite smoothness with sufficiently large ν (i.e., $\nu > 3d/2$), then this method is more efficient than IGP-UCB, whose computational complexity is $O(|\mathcal{A}| T^3)$.

5. In synthetic environments, we empirically show that APG-UCB has almost the same cumulative regret as that of IGP-UCB and its running time is much shorter.

2. Related Work

First, we review previous works on the adversarial RKHS bandit problem. There are almost no existing results concerning the adversarial RKHS bandit problem except for (Chatterji et al., 2019). They also used an approximation method to solve the problem, but their approximation method can handle only a limited case. Therefore, there are no existing algorithms for the adversarial RKHS bandit problem with general reward functions. Next, we review existing results for the stochastic RKHS bandit problem. Srinivas et al. (2010) studied a multi-armed bandit problem, where the reward function is assumed to be sampled from a Gaussian process or belongs to an RKHS. Chowdhury & Gopalan (2017) improved the result of Srinivas et al. (2010) in the RKHS setting and proposed two methods called IGP-UCB and GP-TS. Valko et al. (2013) considered a stochastic RKHS bandit problem, where the arm set $\mathcal{A}$ is finite and fixed, prpposed a method called SupKernelUCB, and proved a regret upper bound $\widetilde{O}(\sqrt{T \gamma_T \log^3(|\mathcal{A}| T/\delta)})$. To address the computational inefficiency in the stochastic RKHS bandit problem, Mutny & Krause (2018) proposed Thompson Sampling and UCB-type algorithms using an approximation method called Quadrature Fourier Features which is an improved method of Random Fourier Features (Rahimi & Recht, 2008). They proved that the total computational complexity of their methods is given as $\widetilde{O}(|\mathcal{A}| T \gamma_T^2)$. However, their methods can be applied to only a very special class of kernels. For example, among three examples introduced in §3, only SE kernels satisfy their assumption unless $d = 1$. Our methods work for general symmetric positive definite kernels with enough smoothness. Calandriello et al. (2020) proposed a method called BBKB and proved its regret is upper bounded by $55\tilde{C}^3 R_{\text{GP-UCB}}(T)$ with $\tilde{C} > 1$ and its total computational complexity is given as $\widetilde{O}(|\mathcal{A}|T\gamma^2(T) + \gamma^4(T))$. Here we use the maximum information gain instead of the effective dimension since they have the same order up to polylogarithmic factors (Calandriello et al., 2019). If the kernel is an SE kernel, ignoring polylogarithmic factors, their computational complexity is linear in T. However, the term incurs generally large coefficient $|\mathcal{A}|$ in the term unlike APG-PE. Finally, we note that we construct APG-PE from PHASED ELIMINATION (Lattimore et al., 2020), which is an algorithm for the stochastic misspecified linear bandit problem, where PE stands for PHASED ELIMINATION.

3. Problem Formulation

Let Ω be a non-empty subset of a Euclidean space $\mathbb{R}^d$ and $K : \Omega \times \Omega \to \mathbb{R}$ be a symmetric, positive definite kernel on Ω, i.e., $K(x, y) = K(y, x)$ for all $x, y \in \Omega$ and for a pairwise distinct points $\{x_1, \dots, x_n\} \subseteq \Omega$, the kernel matrix $(K(x_i, x_j))_{1 \le i,j \le n}$ is positive definite. Examples of such kernels are Rational Quadratic (RQ), Squared Exponential (SE), and Matérn kernels defined as $K_{\mathrm{RQ}}(x, y) := \left(1 + \frac{s^2}{2\mu l^2}\right)^{-\mu}$, $K_{\mathrm{SE}}(x, y) := \exp\left(-\frac{s^2}{2l^2}\right)$, and $K^{(\nu)}_{\text{Matérn}}(x, y) := \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{s\sqrt{2\nu}}{l}\right)^{\nu} K_\nu \left(\frac{s\sqrt{2\nu}}{l}\right)$. where $s = \|x - y\|_2$ and $l > 0, \mu > d/2$, $\nu > 0$ are parameters, and K_ν is the modified Bessel function of the second kind. As in the previous work (Chowdhury & Gopalan, 2017), we normalize kernel K so that $K(x, x) \le 1$ for all $x \in \Omega$. We note that the above three examples satisfy $K(x, x) = 1$ for any x. We denote by $\mathcal{H}_K(\Omega)$ the RKHS corresponding to the kernel K, which we shall review briefly in §4 and assume that $f \in \mathcal{H}_K(\Omega)$ has bounded norm, i.e., $\|f\|_{\mathcal{H}_K(\Omega)} \le B$. In this paper, we consider the following multi-armed bandit problem with time interval T and arm set $\mathcal{A} \subseteq \Omega$. First, we formulate the stochastic RKHS bandit problem. In each round $t = 1, 2, \dots, T$, a learner selects an arm $x_t \in \mathcal{A}$ and observes noisy reward $y_t = f(x_t) + \varepsilon_t$. Here we assume that noise stochastic process is conditionally R-sub-Gaussian with respect to a filtration $\{\mathcal{F}_t\}_{t=1,2,\dots}$, i.e., $\mathbf{E}\left[\exp(\xi\varepsilon_t) \mid \mathcal{F}_t\right] \le \exp(\xi^2 R^2/2)$ for all $t \ge 1$ and $\xi \in \mathbb{R}$. We also assume that x_t is $\mathcal{F}_t$-measurable and y_t is $\mathcal{F}_{t+1}$-measurable. The objective of the learner is to maximize the cumulative reward $\sum_{t=1}^T f(x_t)$ and regret is defined by $R(T) := \sup_{x\in\mathcal{A}} \sum_{t=1}^T (f(x) - f(x_t))$. In the adversarial (or non-stochastic) bandit RKHS problem, we assume a sequence $f_t \in \mathcal{H}_K(\Omega)$ with $\|f_t\|_{\mathcal{H}_K(\Omega)} \le B$ for $t = 1, \dots, T$ is given. In each round $t = 1, \dots, T$, a learner selects an arm $x_t \in \mathcal{A}$ and observes a reward $f_t(x_t)$. The learner's objective is to minimize the cumulative regret $R(T) := \sup_{x\in\mathcal{A}} \sum_{t=1}^T f_t(x) - \sum_{t=1}^T f_t(x_t)$. In this paper we only consider oblivious adversary, i.e., we assume the adversary chooses a sequence $\{f_t\}_{t=1}^T$ before the game starts.

4. Results from Approximation Theory

In this section, we introduce important results provided by approximation theory. For introduction to this subject, we refer to the monograph (Wendland, 2004). We first briefly review basic properties of the RKHS and introduce classical results regarding the convergence rate of the power function, which are required for the proof of Theorem 6. Then, we introduce the P-greedy algorithm and its convergence rate in Theorem 6, which generalizes the existing result (Santin & Haasdonk, 2017).

4.1. Reproducing Kernel Hilbert Space

Let $F(\Omega) := \{f : \Omega \to \mathbb{R}\}$ be the real vector space of $\mathbb{R}$-valued functions on Ω. Then, there exists a unique real Hilbert space $\left(\mathcal{H}_K(\Omega), \langle\cdot,\cdot\rangle_{\mathcal{H}_K(\Omega)}\right)$ with $\mathcal{H}_K(\Omega) \subseteq F(\Omega)$ satisfying the following two properties: (i) $K(\cdot, x) \in \mathcal{H}_K(\Omega)$ for all $x \in \Omega$. (ii) $\langle f, K(\cdot, x)\rangle_{\mathcal{H}_K(\Omega)} = f(x)$ for all $f \in \mathcal{H}_K(\Omega)$ and $x \in \Omega$. Because of the second property, the kernel K is called reproducing kernel and $\mathcal{H}_K(\Omega)$ is called the reproducing kernel Hilbert space (RKHS).

For a subset $\Omega' \subseteq \Omega$, we denote by $V(\Omega')$ the vector subspace of $\mathcal{H}_K(\Omega)$ spanned by $\{K(\cdot, x) \mid x \in \Omega'\}$. We define an inner product of $V(\Omega')$ as follows. For $f = \sum_{i\in I} a_i K(\cdot, x_i)$ and $g = \sum_{j\in I} b_j K(\cdot, x_j)$ with $|I| < \infty$, we define $\langle f, g\rangle := \sum_{i,j\in I} a_i b_j K(x_i, x_j)$. Since K is symmetric and positive definite, $V(\Omega')$ becomes a pre-Hilbert space with this inner product. Then it is known that RKHS $\mathcal{H}_K(\Omega)$ is isomorphic to the completion of $V(\Omega)$. Therefore, for each $f \in \mathcal{H}_K(\Omega)$, there exists a sequence $\{x_n\}_{i=1}^\infty \subseteq \Omega$ and real numbers $\{a_n\}_{n=1}^\infty$ such that $f = \sum_{n=1}^\infty a_n K(\cdot, x_n)$. Here the convergence is that with respect to the norm of $\mathcal{H}_K(\Omega)$ and because of a special property of RKHS, it is also a pointwise convergence.

4.2. Power Function and its Convergence Rate

Since for any $f \in \mathcal{H}_K(\Omega)$, there exists a sequence of finite sums $\sum_{n=1}^N a_n K(\cdot, x_n)$ that converges to f, it is natural to consider the error between f and the finite sum. A natural notion to capture the error for any $f \in \mathcal{H}_K(\Omega)$ is the power function defined as below. For a finite subset of points $X = \{x_n\}_{n=1}^N \subseteq \Omega$, we denote by $\Pi_{V(X)} : \mathcal{H}_K(\Omega) \to V(X)$ the orthogonal projection to $V(X)$. We note that the function $\Pi_{V(X)} f$ is characterized as the interpolant of f, i.e, $\Pi_{V(X)} f$ is a unique function $g \in V(X)$ satisfying $g(x) = f(x)$ for all $x \in X$. Then the power function

$P_{V(X)} : \Omega \to \mathbb{R}_{\geq 0}$ is defined as:

$$P_{V(X)}(x) = \sup_{f \in \mathcal{H}_K(\Omega) \setminus \{0\}} \frac{|f(x) - (\Pi_{V(X)} f)(x)|}{\|f\|_{\mathcal{H}_K(\Omega)}}.$$

By definition, we have

$$\left| f(x) - \left(\Pi_{V(X)} f\right)(x) \right| \leq \|f\|_{\mathcal{H}_K(\Omega)} P_{V(X)}(x)$$

for any $f \in \mathcal{H}_K(\Omega)$ and $x \in \Omega$.

Since the power function $P_{V(X)}$ represents how well the space $V(X)$ approximates any function in $\mathcal{H}_K(\Omega)$ with a bounded norm, it is intuitively clear that the value of $P_{V(X)}$ is small if X is a "fine" discretization of Ω. The fineness of a finite subset $X = \{x_1, \ldots, x_N\} \subseteq \Omega$ can be evaluated by the fill distance $h_{X,\Omega}$ of X defined as $\sup_{x \in \Omega} \min_{1 \leq n \leq N} \|x - x_n\|_2$. We introduce classical results on the convergence rate of the power function as $h_{X,\Omega} \to 0$. We introduce two kinds of these results: polynomial decay and exponential decay.[2] Before introducing the results, we define smoothness of kernels.

DEFINITION 1. (i) We say (K, Ω) has finite smoothness[3] with a smoothness parameter $\nu \in \frac{1}{2}\mathbb{Z}_{>0}$, if Ω is bounded and satisfies an interior cone condition (see remark below), and satisfies either the following condition (a) or (b): (a) $K \in C^{2\nu}(\Omega^\iota \times \Omega^\iota)$, and all the differentials of K of order 2ν are bounded on $\Omega \times \Omega$. Here Ω^ι denotes the interior. (b) There exists $\Phi : \mathbb{R}^d \to \mathbb{R}$ such that $K(x, y) = \Phi(x - y)$, $\nu + d/2 \in \mathbb{Z}$, Φ has continuous Fourier transformation $\hat{\Phi}$ and $\hat{\Phi}(x) = \Theta((1 + \|x\|_2^2)^{-(\nu + d/2)})$ as $\|x\|_2 \to \infty$.

(ii) We say (K, Ω) has infinite smoothness if Ω is a d-dimensional cube $\{x \in \mathbb{R}^d : |x - a_0|_\infty \leq r_0\}$, $K(x, y) = \phi(\|x - y\|_2)$ with a function $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}$, and there exists a positive integer l_0 and a constant $M > 0$ such that $\varphi(r) := \phi(\sqrt{r})$ satisfies $|\frac{d^l \varphi}{dr^l}(r)| \leq l! M^l$ for any $l \geq l_0$ and $r \in \mathbb{R}_{\geq 0}$.

REMARK 2. (i) Results introduced in this subsection depend on local polynomial reproduction on Ω and such a result is hopeless if Ω is a general bounded set (Wendland, 2004). The interior cone condition is a mild condition that assures such results. For example, if Ω is a cube $\{x : |x - a|_\infty \leq r\}$ or ball $\{x : |x - a|_2 \leq r\}$, then this condition is satisfied. (ii) Since $\hat{\Phi}(x) = c\,(1 + \|x\|_2^2)^{-\nu - d/2}$ with $c > 0$ for $\Phi(x) = \|x\|_2^\nu K_\nu(\|x\|_2)$, Matérn kernels $K^{(\nu)}_{\text{Matérn}}$ have finite smoothness with smoothness parameter ν. In addition, it can be shown that the RQ and SE kernels have infinite smoothness.

[2] We note that more generalized results including in the case of conditionally positive definite kernels and differentials of functions of RKHS are proved (Wendland, 2004, Chapter 11).

[3] By abuse of notation, omitting Ω, we also say "K has finite smoothness".

Algorithm 1 Construction of Newton basis with P-greedy algorithm (c.f. Pazouki & Schaback (2011))

Input: kernel K, admissible error $\mathfrak{e} > 0$, a subset of points $\widehat{\Omega} \subseteq \Omega$,
Output: A subset of points $X_m \subseteq \widehat{\Omega}$ and Newton basis $N_1, \ldots, N_m$ of $V(X_m)$.
$\xi_1 := \operatorname{argmax}_{x \in \widehat{\Omega}} K(x, x)$.
$N_1(x) := \frac{K(x, \xi_1)}{\sqrt{K(\xi_1, \xi_1)}}$.
for $m = 1, 2, 3, \ldots,$ **do**
  $P_m^2(x) := K(x, x) - \sum_{k=1}^m N_k^2(x)$.
  if $\max_{x \in \widehat{\Omega}} P_m^2(x) < \mathfrak{e}^2$ **then**
    return $\{\xi_1, \ldots, \xi_m\}$ and $\{N_1, \ldots, N_m\}$.
  end if
  $\xi_{m+1} := \operatorname{argmax}_{x \in \widehat{\Omega}} P_m^2(x)$.
  $u(x) := K(x, \xi_{m+1}) - \sum_{k=1}^m N_k(\xi_{m+1}) N_k(x)$,
  $N_{m+1}(x) := u(x) / \sqrt{P_m^2(\xi_{m+1})}$.
end for

THEOREM 3 (Wu & Schaback (1993), Wendland (2004) Theorem 11.13). *We assume (K, Ω) has finite smoothness with smoothness parameter ν. Then there exist constants $C > 0$ and $h_0 > 0$ that depend only on ν, d, K and Ω such that $\|P_X\|_{L^\infty(\Omega)} \leq C h_{X,\Omega}^\nu$ for any $X \subseteq \Omega$ with $h_{X,\Omega} \leq h_0$.*

One can apply this result to RQ and SE kernels for any $\nu > 0$, but a stronger result holds for these kernels.

THEOREM 4 (Madych & Nelson (1992), Wendland (2004) Theorem 11.22). *Let $\Omega \subset \mathbb{R}^d$ be a cube and assume K has infinite smoothness. Then, there exist constants $C_1, C_2, h_0 > 0$ depending only on d, Ω, and K such that*

$$\|P_X\|_{L^\infty(\Omega)} \leq C_1 \exp\left(-C_2 / h_{X,\Omega}\right),$$

for any finite subset $X \subseteq \Omega$ with $h_{X,\Omega} \leq h_0$.

REMARK 5. (i) The assumption on Ω can be relaxed, i.e., the set Ω is not necessarily a cube. See (Madych & Nelson, 1992) for details. (ii) In the case of SE kernels, a stronger result holds. More precisely, for sufficiently small $h_{X,\Omega}$, $\|P_X\|_{L^\infty(\Omega)} \leq C_1 \exp\left(C_2 \log(h_{X,\Omega}) / h_{X,\Omega}\right)$ holds.

4.3. P-greedy Algorithm and its Convergence Rate

In a typical application, for a given discretization $\widehat{\Omega} \subseteq \Omega$ and function $f \in \mathcal{H}_K(\Omega)$, we want to find a finite subset $X = \{\xi_1, \ldots, \xi_m\} \subseteq \widehat{\Omega}$ with $|X| \ll |\widehat{\Omega}|$ so that f is close to an element of $V(X)$. Several greedy algorithms are proposed to solve this problem (De Marchi et al., 2005; Schaback & Wendland, 2000; Müller, 2009). Among them, the P-greedy algorithm (De Marchi et al., 2005) is most suitable for our purpose, since the point selection depends only on K and $\widehat{\Omega}$ but not on the function f which is unknown to the learner in the bandit setting.

The P-greedy algorithm first selects a point $\xi_1 \in \widehat{\Omega}$ maximizing $P_{V(\emptyset)}(x) = K(x,x)$ and after selecting points $X_{n-1} = \{\xi_1, \dots, \xi_{n-1}\}$, it selects ξ_n by $\xi_n = \mathrm{argmax}_{x \in \widehat{\Omega}} P_{V(X_{n-1})}(x)$. Following Pazouki & Schaback (2011), we introduce (a variant of) the P-greedy algorithm that simultaneously computes the Newton basis (Müller & Schaback, 2009) in Algorithm 1. If $\widehat{\Omega}$ is finite, this algorithm outputs Newton Basis $N_1, \dots, N_m$ at the cost of $O(|\widehat{\Omega}|m^2)$ time complexity using $O(|\widehat{\Omega}|m)$ space. Newton basis $\{N_1, \dots, N_m\}$ is the Gram-Schmidt orthonormalization of basis $\{K(\cdot, \xi_1), \dots, K(\cdot, \xi_m)\}$. Because of orthonormality, the following equality holds (Santin & Haasdonk, 2017, Lemma 5): $P_{V(X)}^2(x) = K(x,x) - \sum_{i=1}^m N_i^2(x)$, where $X = \{\xi_1, \dots, \xi_m\}$. Seemingly, Algorithm 1 is different from the P-greedy algorithm described above, using this formula, we can see that these two algorithms are identical.

The following theorem is essentially due to Santin & Haasdonk (2017) and we provide a more generalized result.

THEOREM 6 (Santin & Haasdonk (2017)). *Let $K : \Omega \times \Omega \to \mathbb{R}$ be a symmetric positive definite kernel. Suppose that the P-greedy algorithm applied to $\widehat{\Omega} \subseteq \Omega$ with error $\mathfrak{e}$ gives $n_{\mathfrak{e}}$ points $X \subseteq \widehat{\Omega}$ with $|X| = n_{\mathfrak{e}}$. Then the following statements hold:*

(i) Suppose (K, Ω) has finite smoothness with smoothness parameter $\nu > 0$. Then, there exists a constant $\hat{C} > 0$ depending only on d, ν, K, and Ω such that $\|P_{V(X)}\|_{L^\infty(\widehat{\Omega})} < \hat{C} n_{\mathfrak{e}}^{-\nu/d}$.

(ii) Suppose (K, Ω) has infinite smoothness. Then there exist constants $\hat{C}_1, \hat{C}_2 > 0$ depending only on d, K, and Ω such that $\|P_{V(X)}\|_{L^\infty(\widehat{\Omega})} < \hat{C}_1 \exp\left(-\hat{C}_2 n_{\mathfrak{e}}^{1/d}\right)$.

The statements of the theorem are non-trivial in two folds. First, by Theorems 3, 4, if $n \in \mathbb{Z}_{>0}$ is sufficiently large, there exists a subset $X \subseteq \Omega$ with $|X| = n$ that gives the same convergence rate above (e.g. X is a uniform mesh of Ω). This theorem assures the same convergence rate is achieved by the points selected by the P-greedy algorithm. Secondly, it also assures that the same result holds even if the P-greedy algorithm is applied to a subset $\widehat{\Omega} \subseteq \Omega$.

If the kernel has finite smoothness, Santin & Haasdonk (2017) only considered the case when $\mathcal{H}_K(\Omega)$ is norm equivalent to a Sobolev space, which is also norm equivalent to the RKHS associated with a Matérn kernel. One can prove Theorem 6 from Theorems 3, 4 and (DeVore et al., 2013, Corollary 3.3) by the same argument to Santin & Haasdonk (2017).

For later use, we provide a restatement of Theorem 6 as follows.

COROLLARY 7. *Let $\alpha, q > 0$ be parameters and denote by $D = D_{q,\alpha}(T)$ the number of points returned by the P-greedy algorithm with error $\mathfrak{e} = \alpha / T^q$.*

(i) Suppose K has finite smoothness with smoothness parameter $\nu > 0$. Then $D_{q,\alpha}(T) = O\left(\alpha^{-d/\nu} T^{dq/\nu}\right)$.

(ii) Suppose K has infinite smoothness. Then $D_{q,\alpha}(T) = O\left((q \log T - \log(\alpha))^d\right)$.

5. Misspecified Linear Bandit Problem

Since we can approximate $f \in \mathcal{H}_K(\Omega)$ by an element of $V(X)$, where $V(X)$ is a finite dimensional subspace of the RKHS, we study a linear bandit problem where the linear model is misspecified, i.e, the misspecified linear bandit problem (Lattimore et al., 2020; Lattimore & Szepesvári, 2020). In this section, we introduce several algorithms for the stochastic and adversarial misspecified linear bandit problems. It turns out that such algorithms can be constructed by modifying (or even without modification) algorithms for the linear bandit problem. We provide proofs in this section in the supplementary material.

First, we provide a formulation of the stochastic misspecified linear bandit problem suitable for our purpose. Let $\mathcal{A}$ be a set and suppose that there exists a map $x \mapsto \widetilde{x}$ from $\mathcal{A}$ to the unit ball $\{\xi \in \mathbb{R}^D : \|\xi\|_2 \le 1\}$ of a Euclidean space. In each round $t = 1, 2, \dots, T$, a learner selects an action $x_t \in \mathcal{A}$ and the environment reveals a noisy reward $y_t = g(x_t) + \varepsilon_t$, where $g(x) := \langle \theta, \widetilde{x} \rangle + \omega(x)$, $\theta \in \mathbb{R}^D$, and $\omega(x)$ is a biased noise and satisfies $\sup_{x \in \mathcal{A}} \|\omega(x)\| \le \epsilon$ and $\epsilon > 0$ is known to the learner. We also assume that there exists $B > 0$ such that $\sup_{x \in \mathcal{A}} \|g(x)\| \le B$ and $\|\theta\|_2 \le B$. As before, $\{\varepsilon_t\}_{t \ge 1}$ is conditionally R-sub-Gaussian w.r.t a filtration $\{\mathcal{F}_t\}_{t \ge 1}$ and we assume that $\widetilde{x}_t$ is $\mathcal{F}_t$-measurable and y_t is $\mathcal{F}_{t+1}$-measurable. The regret is defined as $R(T) := \sum_{t=1}^T \left(\sup_{x \in \mathcal{A}} g(x) - g(x_t)\right)$. We can formulate the adversarial misspecified linear bandit problem in a similar way. Let $\{g_t\}_{t=1}^T$ be a sequence of functions on $\mathcal{A}$ with $g_t(x) = \langle \theta_t, \widetilde{x} \rangle + \omega_t(x)$, $\theta_t \in \mathbb{R}^D$, and $\sup_{x \in \mathcal{A}} \|\omega_t(x)\| \le \epsilon$, where the map $x \mapsto \widetilde{x}$ is as before. We also assume that there exists $B > 0$ such that $\sup_{x \in \mathcal{A}} |g(x)| \le B$ and $\|\theta_t\|_2 \le B$. In each round, $t = 1, \dots, T$, the learner selects an arm $x_t \in \mathcal{A}$ and observes a reward $g_t(x_t)$. The cumulative regret is defined as $R_T = \sup_{x \in \mathcal{A}} \sum_{t=1}^T g_t(x) - \sum_{t=1}^T g_t(x_t)$.

First, we introduce a modification of LinUCB (Abbasi-Yadkori et al., 2011). To do this, we prepare notation for the stochastic linear bandit problem. Let $\lambda > 0$ and $\delta > 0$ be parameters. We define $A_t := \lambda 1_D + \sum_{s=1}^t \widetilde{x}_s \widetilde{x}_s^{\mathrm{T}}$, $b_t := \sum_{s=1}^t y_s \widetilde{x}_s$, and $\hat{\theta}_t := A_t^{-1} b_t$. Here, 1_D is the identity matrix of size D. For $x \in \mathbb{R}^D$, we define the Mahalanobis norm as $\|x\|_{A_t^{-1}} := \sqrt{x^{\mathrm{T}} A_t^{-1} x}$ and define β_t as

$$\beta_t := \beta(A_t, \delta, \lambda) := R\sqrt{\log \frac{\det \lambda^{-1} A_t}{\delta^2}} + \sqrt{\lambda} B.$$

We note that by the proof of (Abbasi-Yadkori et al., 2011, Lemma 11), computational complexity for updating β_t is $O(D^2)$ at each round.

Lattimore et al. (2020) (see appendix of its arXiv version) considered a modification of LinUCB which selects $x \in \mathcal{A}$ maximizing (modified) UCB $\langle \hat{\theta}_t, \widetilde{x} \rangle + \beta_t \|\widetilde{x}\|_{A_t^{-1}} + \epsilon \sum_{s=1}^{t} |\widetilde{x}^{\mathrm{T}} A_t^{-1} \widetilde{x}_s|$ in $(t+1)$th round and proved the regret of the algorithm is upper bounded by $O(D\sqrt{T}\log(T) + \epsilon T \sqrt{D \log(T)})$. However, computing the above value requires $O(t)$ time for each arm $x \in \mathcal{A}$. Therefore, instead of incurring additional $\sqrt{D}$ factor in the second term in the regret upper bound above, we consider another upper confidence bound which can be easily computed. In $(t+1)$th round, our modification of UCB type algorithm selects arm $x \in \mathcal{A}$ maximizing the modified UCB $\langle \hat{\theta}_t, \widetilde{x} \rangle + \|\widetilde{x}\|_{A_t^{-1}} (\beta_t + \epsilon \psi_t)$, where ψ_t is defined as $\sum_{s=1}^{t} \|\widetilde{x}_s\|_{A_{s-1}^{-1}}$. Then by storing ψ_t in each round, the complexity for computing this value is given as $O(D^2)$ for each $x \in \mathcal{A}$ and as is well-known, one can update A_t^{-1} in $O(D^2)$ time using the Sherman–Morrison formula. By the standard argument, we can prove the following regret bound.

PROPOSITION 8. *Let notation and assumptions be as above. We further assume that* $\lambda \geq 1$. *Then with probability at least* $1 - \delta$, *the regret* $R(T)$ *of the modified UCB algorithm satisfies* $R(T) \leq 2\beta_T \sqrt{T}\sqrt{2 \log \det(\lambda^{-1} A_t)} + 2\epsilon T + 4\epsilon T \log(\det(\lambda^{-1} A_T))$. *In particular, we have*

$$R(T) = \widetilde{O}\left(\sqrt{DT \log(1/\delta)} + D\sqrt{T} + \epsilon DT\right).$$

In the supplementary material, we also introduce a modification of Thompson Sampling.

The regret upper bound provided above does not depend on the arm set $\mathcal{A}$. Moreover, the same results hold even if the arm set changes over time step t (with minor modification of the definition of regret). On the other hand, several authors (Lattimore et al., 2020; Auer, 2002; Valko et al., 2013) studied algorithm whose regret depends on the cardinality of the arm set in the stochastic linear or RKHS setting. In some rounds, such algorithms eliminate arms that are supposed to be non-optimal with a high probability and therefore the arm set should be the same over time. Generally, these algorithms are more complicated than LinUCB or Thompson Sampling. However, recently, Lattimore et al. (2020) proposed a simpler and sophisticated algorithm called PHASED ELIMINATION using Kiefer–Wolfowitz theorem. Furthermore, they showed that it works well for the stochastic misspecified linear bandit problem without modification. More precisely, they proved the following result.

THEOREM 9 (Lattimore et al. (2020); Lattimore & Szepesvári (2020))**.** *Let* $R(T)$ *be the regret PHASED ELIMINATION incurs for the stochastic misspecified linear bandit problem. We further assume that* $\{\varepsilon_t\}$ *is independent* R*-sub-Gaussian. Then, with probability at least* $1 - \delta$, *we have*

$$R(T) = O\left(\sqrt{DT \log\left(\frac{|\mathcal{A}| \log(T)}{\delta}\right)} + \epsilon\sqrt{D}T \log(T)\right).$$

Moreover the total computational complexity up to round T *is given as* $O(D^2|\mathcal{A}| \log\log(D) \log(T) + TD^2)$.

REMARK 10. Although they provided an upper bound for the expected regret, it is not difficult to see that their proof gave a high probability regret upper bound.

Next, we show that EXP3 for adversarial linear bandits (c.f. Lattimore & Szepesvári (2020)) works for the adversarial misspecified linear bandits without modification. We introduce notations for EXP3. Let $\eta > 0$ be a learning rate, γ an exploration parameter, and $\pi_{\exp}$ be an exploration distribution over $\mathcal{A}$. For a distribution π on $\mathcal{A}$, we define a matrix $Q(\pi) := \sum_{x \in \mathcal{A}} \pi(x) \widetilde{x}\widetilde{x}^{\mathrm{T}}$. We also put $\phi_t := g_t(x_t) Q_t^{-1} \widetilde{x}_t$ and $\phi_t(x) := \langle \phi_t, \widetilde{x} \rangle$ for $x \in \mathcal{A}$, where the matrix Q_t is defined later. We define a distribution q_t over $\mathcal{A}$ by $q_t(x) \sim \exp(\eta \sum_{s=1}^{t-1} \phi_s(x))$ and a distribution p_t by $p_t(x) = \gamma \pi_{\exp}(x) + (1-\gamma) q_t(x)$ for $x \in \mathcal{A}$. The matrix Q_t is defined as $Q(p_t)$. We put $\Gamma(\pi_{\exp}) := \sup_{x \in \mathcal{A}} \widetilde{x}^{\mathrm{T}} Q(\pi_{\exp})^{-1} \widetilde{x}$.

PROPOSITION 11. *We assume that* $\{\widetilde{x} \mid x \in \mathcal{A}\}$ *spans* $\mathbb{R}^D$. *We also assume* $\pi_{\exp}$ *satisfies* $\Gamma(\pi_{\exp}) \leq D$ *and we take* $\gamma = B\Gamma(\pi_{\exp})\eta$. *Then applying EXP3 to the adversarial misspecified linear bandit problem, we have the following upper bound for the expected regret:*

$$\mathbf{E}\left[R(T)\right] \leq 2\varepsilon T + eB^2 \eta DT + \frac{2\epsilon T}{B\eta} + \frac{\log|\mathcal{A}|}{\eta}.$$

REMARK 12. By the Kiefer–Wolfowitz theorem, there exists an exploitation distribution $\pi_{\exp}$ such that $\Gamma(\pi_{\exp}) \leq D$.

6. Main Results

Using results from approximation theory explained in §4 and algorithms for the misspecified bandit problem, we provide several algorithms for the stochastic and adversarial RKHS bandit problems. We provide proofs of the results in this section in the supplementary material.

Let $N_1, \ldots, N_D$ be the Newton basis returned by Algorithm 1 with $\mathfrak{e} = \frac{\alpha}{T^q}$ with $q, \alpha > 0$, and $\widehat{\Omega} = \mathcal{A}$. Then, by orthonormality of the Newton basis and the definition of the power function, for any $f \in \mathcal{H}_K(\Omega)$ and $x \in \Omega$, we have

$$|f(x) - \langle \theta_f, \widetilde{x} \rangle| \leq \|f\|_{\mathcal{H}_K(\Omega)} P_{V(X)}(x),$$

where $\theta_f = (\langle f, N_i \rangle)_{1 \leq i \leq D} \in \mathbb{R}^D$ and $\widetilde{x} = (N_i(x))_{1 \leq i \leq D} \in \mathbb{R}^D$. Therefore, if f is an objective function of a RKHS bandit problem, we can regard f as a linearly

Algorithm 2 Approximated RKHS Bandit Algorithm of UCB type (APG-UCB)

Input: Time interval T, admissible error $\mathfrak{e} = \frac{\alpha}{T^q}$, λ, R, B, δ
Using Alg. 1, compute Newton basis $N_1, \dots, N_D$ with admissible error $\mathfrak{e}$ and $\widehat{\Omega} = \mathcal{A}$, and put $\epsilon = B\mathfrak{e}$.
for $x \in \mathcal{A}$ **do**
  $\widetilde{x} := [N_1(x), N_2(x), \dots, N_D(x)]^{\mathrm{T}} \in \mathbb{R}^D$.
end for
for $t = 0, 1, \dots, T-1$ **do**
  $A_t := \lambda 1_D + \sum_{s=1}^t \widetilde{x}_s \widetilde{x}_s^{\mathrm{T}},\ b_t := \sum_{s=1}^t y_s \widetilde{x}_s$.
  $\hat{\theta}_t := A_t^{-1} b_t, \quad \psi_t := \sum_{s=1}^t \|\widetilde{x}_s\|_{A_{s-1}^{-1}}$.
  $x_{t+1} := \operatorname{argmax}_{x \in \mathcal{A}} \left\{ \langle \widetilde{x}, \hat{\theta}_t \rangle + \|\widetilde{x}\|_{A_t^{-1}} (\beta_t + \epsilon \psi_t) \right\}$.
  Select x_{t+1} and observe y_{t+1}.
end for

misspecified model and apply algorithms for misspecified linear bandit problems to solve the original RKHS bandit problems.

In this section, we reduce the RKHS bandit problem to the the misspecified linear bandit problem by the map $x \mapsto \widetilde{x}$ and apply modified LinUCB, PHASED ELIMINATION, and EXP3 to the problem. We call these algorithms APG-UCB, APG-PE and APG-EXP3 respectively and APG-UCB is displayed in Algorithm 2. We denote by $D_{q,\alpha}(T) = D$ the number of points returned by Algorithm 1 with $\mathfrak{e} = \frac{\alpha}{T^q}$. By the results in §4, we have an upper bound of $D_{q,\alpha}(T)$ (Corollary 7).

First, we state the results for APG-UCB.

Theorem 13. *We denote by $R_{\text{APG-UCB}}(T)$ the regret that Algorithm 2 incurs for the stochastic RKHS bandit problem up to time step T and assume that $\lambda \geq 1$ and $q \geq 1/2$. Then with probability at least $1-\delta$, $R_{\text{APG-UCB}}(T)$ is given as*

$$\widetilde{O}\left(\sqrt{T D_{q,\alpha}(T) \log(1/\delta)} + D_{q,\alpha}(T)\sqrt{T}\right)$$

and the total computational complexity of the algorithm is given as $O(|\mathcal{A}| T D_{q,\alpha}^2(T))$.

The admissible error $\mathfrak{e}$ balances the computational complexity and regret minimization. However, this is not clear from Theorem 13. The following theorem provides another upper bound of APG-UCB and it states that if we take smaller error $\mathfrak{e}$, then an upper bound of APG-UCB is almost the same as that of IGP-UCB.

Theorem 14. *We assume $\lambda = 1$ and take parameter q of APG-UCB so that $q > 3/2$. We define $\beta_T^{\text{IGP-UCB}}$ as $B + R\sqrt{2(\gamma_T + 1 + \log(1/\delta))}$. Then with probability at least $1-\delta$, we have $R_{\text{APG-UCB}}(T) \leq b(T)$, where $b(T)$ is given as $4\beta_T^{\text{IGP-UCB}}\sqrt{\gamma_T T} + O(\sqrt{T\gamma_T} T^{(3/2-q)/2} + \gamma_T T^{1-q})$.*

Remark 15. Since the main term of $b(T)$ is $4\beta_T^{\text{IGP-UCB}}\sqrt{\gamma_T T}$ and by the proof in (Chowdhury & Gopalan, 2017), IGP-UCB has the regret upper bound $4\beta_T^{\text{IGP-UCB}}\sqrt{\gamma_T (T+2)}$, APG-UCB has an asymptotically the same regret upper bound as IGP-UCB if we take a small error $\mathfrak{e}$. We note that if ν is sufficiently large compared to d (this is always the case if the kernel has infinite smoothness), then APG-UCB is more efficient than IGP-UCB. We note that for any choice of parameters, the regret upper bound of BBKB is given as $55\widetilde{C}^3 R_{\text{GP-UCB}}(T)$, where $\widetilde{C} \geq 1$.

Next, we state the results for APG-PE.

Theorem 16. *We denote by $R_{\text{APG-PE}}(T)$ the regret that APG-PE with $q = 1/2$ incurs for the stochastic RKHS bandit problem up to time step T. We further assume that $\{\varepsilon_t\}$ is independent R-sub-Gaussian. Then with probability at least $1-\delta$, we have $R_{\text{APG-PE}}(T) = \widetilde{O}\left(\sqrt{T D_{1/2,\alpha}(T) \log\left(\frac{|\mathcal{A}|}{\delta}\right)}\right)$, and its total computational complexity is given as $\widetilde{O}\left((|\mathcal{A}| + T) D_{1/2,\alpha}^2(T)\right)$.*

Finally, we state a result for the adversarial RKHS bandit problem.

Theorem 17. *We denote by $R_{\text{APG-EXP3}}(T)$ the cumulative regret that APG-EXP3 with $\alpha = \log(|\mathcal{A}|)$ and $q = 1$ incurs for the adversarial RKHS bandit problem up to time step T. Then with appropriate choices of the learning rate η and exploration distribution, the expected regret $\mathbf{E}\left[R_{\text{APG-EXP3}}(T)\right]$ is given as $\widetilde{O}\left(\sqrt{T D_{1,\alpha}(T) \log(|\mathcal{A}|)}\right)$.*

7. Discussion

So far, we have emphasized the advantages of our methods. In this section, we discuss limitations of our methods. Here, we focus on Theorem 13 with $q = 1/2$ and Theorem 16. Since we do not see limitations if the kernel has infinite smoothness, in this section we assume the kernel is a Matérn kernel. In our theoretical results, $D_{q,\alpha}(T)$ plays a similar role as the information gain in the theoretical result of BBKB. If the kernel is a Matérn kernel with parameter ν, then, by recent results on the information gain (Vakili et al., 2021), we have $\gamma_T = \widetilde{O}(T^{d/(d+2\nu)})$, which is a nearly optimal result by (Scarlett et al., 2017) and is slightly better than the upper bound of $D_{1/2,\alpha}(T)$. Therefore, in this case the regret upper bound $\widetilde{O}(\sqrt{T} T^{d/(2\nu)})$ of Theorem 13 is slightly worse than the regret upper bound $\widetilde{O}(\sqrt{T} T^{d/(d+2\nu)})$ of BBKB. In addition, similarly SupKernelUCB has nearly optimal regret upper bound if the kernel is Matérn, but regret upper bound of APG-PE is slightly worse in that case.

Inferiority of our method in the Matérn kernel case might be counter-intuitive since it is also proved that the convergence

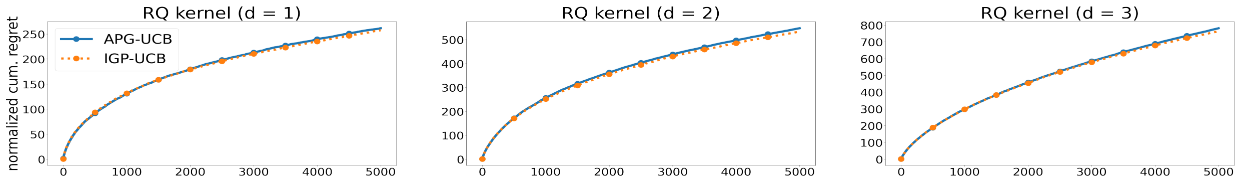

Figure 1. Normalized Cumulative Regret for RQ kernels.

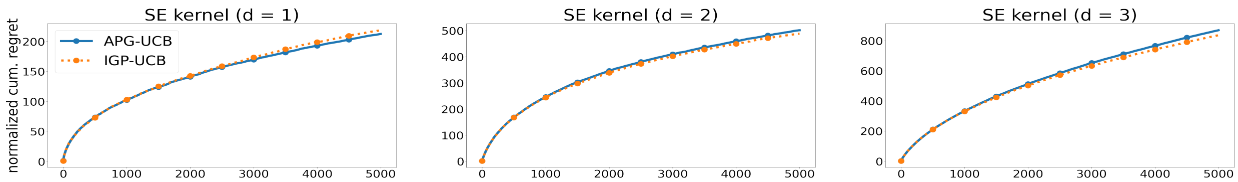

Figure 2. Normalized Cumulative Regret for SE kernels.

rate of the power function for Matérn kernel is optimal (c.f. Schaback (1995)) and Theorem 9 cannot be improved (Lattimore et al., 2020). We explain why a combination of optimal results leads to a non-optimal result. The results on the information gain depend on the eigenvalue decay of the Mercer operator rather than the decay of the power function in the L^∞-norm as in this study. However these two notions are closely related. From the n-width theory (Pinkus, 2012, Chapter IV, Corollary 2.6), eigenvalue decay corresponds to the decay of the power function in the L^2-norm (or more precisely Kolmogorov n-width). The decay in the L^2-norm is derived from that in the L^∞-norm. If the kernel is a Matérn kernel, using a localization trick called Duchon's trick (Wendland, 1997), it can be possible to give a faster decay in the L^p-norm than that in the L^∞-norm if $p < \infty$. Since the norm regarding the misspecified bandit problem is not a L^2 norm but a L^∞ norm, we took the approach proposed in this paper.

8. Experiments

In this section, we empirically verify our theoretical results. We compare APG-UCB to IGP-UCB (Chowdhury & Gopalan, 2017) in terms of cumulative regret and running time for RQ and SE kernels in synthetic environments.

8.1. Environments

We assume the action set is a discretization of a cube $[0,1]^d$ for $d = 1, 2, 3$. We take $\mathcal{A}$ so that $|\mathcal{A}|$ is about 1000. More precisely, we define $\mathcal{A}$ by $\{i/m_d \mid i = 0, 1, \ldots, m_d - 1\}^d$ where $m_1 = 1000, m_2 = 30, m_3 = 10$. We randomly construct reward functions $f \in \mathcal{H}_K(\Omega)$ with $\|f\|_{\mathcal{H}_K(\Omega)} = 1$ as follows. We randomly select points ξ_i (for $1 \le i \le m$) from $\mathcal{A}$ until $m \le 300$ or $\|P_{V(\{\xi_1,\ldots,\xi_m\})}\|_{L^\infty(\mathcal{A})} < 10^{-4}$ and compute orthonormal basis $\{\varphi_1, \ldots, \varphi_m\}$ of $V(\{\xi_1, \ldots, \xi_m\})$. Then, we define $f = \sum_{i=1}^m a_i \varphi_i$, where $[a_1, \ldots, a_m] \in \mathbb{R}^m$ is a random vector with unit norm. We take $l = 0.3\sqrt{d}$ for the RQ kernel and $l = 0.2\sqrt{d}$ for the SE kernel, because the diameter of the d-dimensional cube is $\sqrt{d}$. For each kernel, we generate 10 reward functions as above and evaluate our proposed method and the existing algorithm for time interval $T = 5000$ in terms of mean cumulative regret and total running time. We compute the mean of cumulative regret and running time for these 10 environments. We normalize cumulative regret so that normalized cumulative regret of uniform random policy corresponds to the line through origin with slope 1 in the figure. For simplicity, we assume the kernel, B, and R are known to the algorithms. For the other parameters, we use theoretically suggested ones for both APG-UCB and IGP-UCB. Computation is done by Intel Xeon E5-2630 v4 processor with 128 GB RAM. In supplementary material, we explain the experiment setting in more detail and provide additional experimental results.

8.2. Results

We show the results for normalized cumulative regret in Figures 1 and 2. As suggested by the theoretical results, growth of the cumulative regret of these algorithms is $\widetilde{O}(\sqrt{T})$ ignoring a polylogarithmic factor. Although, convergence rate of the power function of SE kernels is slightly faster than that of RQ kernels (by remark of Theorem 4), empirical results of RQ kernels and SE kernels are similar. In both cases, APG-UCB has almost the same cumulative regret as that of IGP-UCB.

We also show (mean) total running time in Table 1, where we abbreviate APG-UCB as APG and IGP-UCB as IGP. For all dimensions, it took about from five to six thousand seconds for IGP-UCB to complete an experiment for one environment. As shown by the table and figures, running time of our methods is much shorter than that of IGP-UCB while it has almost the same regret as IGP-UCB.

Table 1. Total Running Time (in seconds).

	APG(RQ)	IGP(RQ)	APG(SE)	IGP(SE)
$d = 1$	4.2e-01	5.7e+03	4.0e-01	5.7e+03
$d = 2$	2.7e+00	5.1e+03	2.9e+00	5.1e+03
$d = 3$	3.0e+01	5.7e+03	4.3e+01	5.7e+03

9. Conclusion

By reducing the RKHS bandit problem to the misspecified linear bandit problem, we provide the first general algorithm for the adversarial RKHS bandit problem and several efficient algorithms for the stochastic RKHS bandit problem. We provide cumulative regret upper bounds for them and empirically verify our theoretical results.

10. Acknowledgement

We would like to thank anonymous reviewers for suggestions that improved the paper. We also would like to thank Janmajay Singh and Takafumi J. Suzuki for valuable comments on the preliminarily version of the manuscript.

References

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

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 127–135, 2013.

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

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Calandriello, D., Carratino, L., Lazaric, A., Valko, M., and Rosasco, L. Gaussian process optimization with adaptive sketching: Scalable and no regret. In *Proceedings of the 32nd Conference on Learning Theory*, pp. 533–557. PMLR, 2019.

Calandriello, D., Carratino, L., Valko, M., Lazaric, A., and Rosasco, L. Near-linear time gaussian process optimization with adaptive batching and resparsification. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Chatterji, N., Pacchiano, A., and Bartlett, P. Online learning with kernel losses. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 971–980. PMLR, 2019.

Chowdhury, S. R. and Gopalan, A. On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 844–853, 2017. supplementary material.

De Marchi, S., Schaback, R., and Wendland, H. Near-optimal data-independent point locations for radial basis function interpolation. *Advances in Computational Mathematics*, 23(3):317–330, 2005.

DeVore, R., Petrova, G., and Wojtaszczyk, P. Greedy algorithms for reduced bases in banach spaces. *Constructive Approximation*, 37(3):455–466, 2013.

Hoffman, A., Wielandt, H., et al. The variation of the spectrum of a normal matrix. *Duke Mathematical Journal*, 20(1):37–39, 1953.

Lattimore, T. and Szepesvári, C. *Bandit Algorithm*. Cambridge University Press, 2020.

Lattimore, T., Szepesvari, C., and Weisz, G. Learning with good feature representations in bandits and in rl with a generative model. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Madych, W. and Nelson, S. Bounds on multivariate polynomials and exponential error estimates for multiquadric interpolation. *Journal of Approximation Theory*, 70(1): 94–114, 1992.

Müller, S. *Komplexität und Stabilität von kernbasierten Rekonstruktionsmethoden*. PhD thesis, Fakultät für Mathematik und Informatik, Georg-August-Universität Göttingen, 2009.

Müller, S. and Schaback, R. A newton basis for kernel spaces. *Journal of Approximation Theory*, 161(2):645–655, 2009.

Mutny, M. and Krause, A. Efficient high dimensional bayesian optimization with additivity and quadrature fourier features. In *Advances in Neural Information Processing Systems*, pp. 9005–9016, 2018.

Pazouki, M. and Schaback, R. Bases for kernel-based spaces. *Journal of Computational and Applied Mathematics*, 236 (4):575–588, 2011.

Pinkus, A. *N-widths in Approximation Theory*, volume 7. Springer Science & Business Media, 2012.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pp. 1177–1184, 2008.

Santin, G. and Haasdonk, B. Convergence rate of the data-independent p-greedy algorithm in kernel-based approximation. *Dolomites Research Notes on Approximation*, 10 (Special_Issue), 2017.

Scarlett, J., Bogunovic, I., and Cevher, V. Lower bounds on regret for noisy gaussian process bandit optimization. In *Proceedings of the 30th Conference on Learning Theory*, pp. 1723–1742, 2017.

Schaback, R. Error estimates and condition numbers for radial basis function interpolation. *Advances in Computational Mathematics*, 3(3):251–264, 1995.

Schaback, R. and Wendland, H. Adaptive greedy techniques for approximate solution of large rbf systems. *Numerical Algorithms*, 24(3):239–254, 2000.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 1015–1022, 2010.

Vakili, S., Khezeli, K., and Picheny, V. On information gain and regret bounds in gaussian process bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pp. 82–90. PMLR, 2021.

Valko, M., Korda, N., Munos, R., Flaounas, I., and Cristianini, N. Finite-time analysis of kernelised contextual bandits. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, 2013.

Wendland, H. Sobolev-type error estimates for interpolation by radial basis functions. *Surface fitting and multiresolution methods*, pp. 337–344, 1997.

Wendland, H. *Scattered data approximation*, volume 17. Cambridge University Press, 2004.

Wu, Z.-m. and Schaback, R. Local error estimates for radial basis function interpolation of scattered data. *IMA Journal of Numerical Analysis*, 13(1):13–27, 1993.

Appendix

In this appendix, we provide some results on Thompson Sampling based algorithms in §A, proofs of the results in §B, and detailed experimental setting and additional experimental results in §C.

A. Additional Results for Thompson Sampling

A.1. Misspecified Linear Bandit Problem

We consider a modification of Thompson Sampling (Agrawal & Goyal, 2013). In $(t+1)$th round, we sample μ_t from the multinomial normal distribution $\mathcal{N}\left(\hat{\theta}_t, \left(\beta(A_t, \delta/2, \lambda) + \epsilon\psi_t\right)^2 A_t^{-1}\right)$, and the modified algorithm selects $x \in \mathcal{A}$ that maximizes $\langle \mu_t, \widetilde{x} \rangle$. Then the following result holds.

PROPOSITION 18. *We assume that $\lambda \geq 1$. Then, with probability at least $1-\delta$, the modification of Thompson Sampling algorithm incurs regret upper bound by*

$$\widetilde{O}\left(\sqrt{\log(|\mathcal{A}|)}\left\{D\sqrt{T} + \sqrt{DT\log(1/\delta)} + \log(1/\delta)\sqrt{T} + \left(DT + T\sqrt{D\log(1/\delta)}\right)\epsilon)\right\}\right)$$

We provide proof of the proposition in §B.4.

A.2. Thompson Sampling for the Stochastic RKHS Bandit Problem

We provide a result on a Thompson Sample based algorithm for the stochastic RKHS bandit problem.

THEOREM 19. *We reduce the RKHS bandit problem to the misspecified linear bandit problem, and apply the modified Thompson Sampling introduced above with admissible error $\mathfrak{e} = \frac{\alpha}{T^q}$ with $q \geq 1/2$. We denote by $R_{\textit{APG-TS}}(T)$ its regret and assume that $\lambda \geq 1$. Then with probability at least $1-\delta$, $R_{\textit{APG-TS}}(T)$ is upper bounded by*

$$\widetilde{O}\left(\sqrt{\log(|\mathcal{A}|)}\left(D_{q,\alpha}(T)\sqrt{T} + \sqrt{D_{q,\alpha}(T)T\log(1/\delta)} + \log(1/\delta)\sqrt{T})\right)\right).$$

The total computational complexity of the algorithm is given as $O(|\mathcal{A}|TD^2_{q,\alpha}(T) + TD^3_{q,\alpha}(T))$.

We provide proof of the theorem in §B.5.

B. Proofs

We provide omitted proofs in the main article and §A.

B.1. Proof of Corollary 7

For completeness, we provide a proof of corollary 7.

Proof. For simplicity, we consider only the infinite smoothness case. We use the same notation as in Theorem 6 and Algorithm 1. Denote by D the number of points returned by the algorithm with error $\mathfrak{e} = \alpha/T^q$. Since the statement of the corollary is obvious if $D = 1$, we assume $D > 1$. Because the condition $\max_{x\in\widehat{\Omega}} P_m(x) < \alpha/T^q$ is satisfied only when $m \geq D$, we have $\alpha/T^q \leq \max_{x\in\widehat{\Omega}} P_{D-1}(x)$. If we run the algorithm with error $\mathfrak{e} = \max_{x\in\widehat{\Omega}} P_{D-1}(x) + \epsilon$ with sufficiently small $\epsilon > 0$, then the algorithm returns $D-1$ points. Therefore by the theorem and the inequality above, we have

$$\alpha/T^q \leq \max_{x\in\widehat{\Omega}} P_{D-1}(x) < \hat{C}_1 \exp\left(-\hat{C}_2(D-1)^{1/d}\right).$$

Ignoring constants other than α, T, q, we have the assertion of the corollary. □

B.2. Proof of Proposition 8

For symmetric matrices $P, Q \in \mathbb{R}^{n\times n}$, we write $P \geq Q$ if and only if $P - Q$ is positive semi-definite, i.e., $x^T(P-Q)x \geq 0$ for all $x \in \mathbb{R}^n$. For completeness, we prove the following elementary lemma.

LEMMA 20. *Let $P, Q \in \mathbb{R}^{n \times n}$ be symmetric matrices of size n and assume that $0 < P \le Q$. Then we have $Q^{-1} \le P^{-1}$.*

Proof. It is enough to prove the statement for $U^{\mathrm{T}} P U$ and $U^{\mathrm{T}} Q U$ for some $U \in \mathrm{GL}_n(\mathbb{R})$, where $\mathrm{GL}_n(\mathbb{R})$ is the general linear group of size n. Since P is positive definite, using Cholesky decomposition, one can prove that there exists $U \in \mathrm{GL}_n(\mathbb{R})$ such that $U^{\mathrm{T}} P U = 1_n$ and $U^{\mathrm{T}} Q U = \Lambda$ is a diagonal matrix. Then, the assumption implies that every diagonal entry of Λ is greater than or equal to 1. Now, the statement is obvious. □

Next, we prove that $\langle \widetilde{x}, \theta_t \rangle + \|\widetilde{x}\|_{A_t^{-1}}(\beta_t + \epsilon \psi_t)$ is a UCB up to a constant.

LEMMA 21. *We assume $\lambda \ge 1$. Then, with probability at least $1 - \delta$, we have*

$$|\langle \widetilde{x}, \theta_t \rangle - \langle \widetilde{x}, \theta \rangle| \le \|\widetilde{x}\|_{A_t^{-1}}(\beta_t + \epsilon \psi_t),$$

for any t and $\widetilde{x} \in \mathbb{R}^D$.

Proof. By proof of (Abbasi-Yadkori et al., 2011, Theorem 2), we have

$$\begin{aligned}
|\langle \widetilde{x}, \theta_t \rangle - \langle \widetilde{x}, \theta \rangle| &\le \|\widetilde{x}\|_{A_t^{-1}} \left(\left\| \sum_{s=1}^{t} \widetilde{x}_s (\varepsilon_s + \omega(x_s)) \right\|_{A_t^{-1}} + \lambda^{1/2} \|\theta\| \right) \\
&\le \|\widetilde{x}\|_{A_t^{-1}} \left(\left\| \sum_{s=1}^{t} \widetilde{x}_s \varepsilon_s \right\|_{A_t^{-1}} + \lambda^{1/2} \|\theta\| + \epsilon \sum_{s=1}^{t} \|\widetilde{x}_s\|_{A_t^{-1}} \right).
\end{aligned}$$

By the self-normalized concentration inequality (Abbasi-Yadkori et al., 2011), with probability at least $1 - \delta$, we have

$$|\langle \widetilde{x}, \theta_t \rangle - \langle \widetilde{x}, \theta \rangle| \le \|\widetilde{x}\|_{A_t^{-1}} \left(\beta_t + \epsilon \sum_{s=1}^{t} \|\widetilde{x}_s\|_{A_t^{-1}} \right).$$

Since $A_{s-1} \le A_s$ for any s, by Lemma 20, we have $\sum_{s=1}^{t} \|\widetilde{x}_s\|_{A_t^{-1}} \le \psi_t$. This completes the proof. □

Proof of Proposition 8. We assume $\lambda \ge 1$. Let $x^* := \operatorname{argmax}_{x \in \mathcal{A}} f(x)$ and $(x_t)_t$ be a sequence of arms selected by the algorithm. Denote by E the event on which the inequality in Lemma 21 holds for all t and $\widetilde{x}$. Then on event E, we have

$$\begin{aligned}
f(x^*) - f(x_t) &\le 2\epsilon + \langle \widetilde{x}^*, \theta \rangle - \langle \widetilde{x}_t, \theta \rangle \\
&\le 2\epsilon + \langle \widetilde{x}^*, \theta_t \rangle + \|\widetilde{x}^*\|_{A_t^{-1}}(\beta_t + \epsilon \psi_t) - \langle \widetilde{x}_t, \theta \rangle \\
&\le 2\epsilon + \langle \widetilde{x}_t, \theta_t \rangle + \|\widetilde{x}_t\|_{A_t^{-1}}(\beta_t + \epsilon \psi_t) - \left(\langle \widetilde{x}_t, \theta_t \rangle - \|\widetilde{x}_t\|_{A_t^{-1}}(\beta_t + \epsilon \psi_t) \right) \\
&= 2\epsilon + 2\|\widetilde{x}_t\|_{A_t^{-1}}(\beta_t + \epsilon \psi_t).
\end{aligned}$$

Therefore, on event E,

$$\begin{aligned}
R(T) &\le 2\epsilon T + 2\beta_T \sum_{t=1}^{T} \|\widetilde{x}_t\|_{A_{t-1}^{-1}} + 2\epsilon \sum_{t=1}^{T} \|\widetilde{x}_t\|_{A_{t-1}^{-1}} \psi_t \\
&\le 2\epsilon T + 2\beta_T \sqrt{T} \sqrt{\sum_{t=1}^{T} \|\widetilde{x}_t\|_{A_{t-1}^{-1}}^2} + 2\epsilon \psi_T \sum_{t=1}^{T} \|\widetilde{x}_t\|_{A_{t-1}^{-1}} \\
&= 2\epsilon T + 2\beta_T \sqrt{T} \sqrt{\sum_{t=1}^{T} \|\widetilde{x}_t\|_{A_{t-1}^{-1}}^2} + 2\epsilon \left(\sum_{t=1}^{T} \|\widetilde{x}_t\|_{A_{t-1}^{-1}} \right)^2 \\
&\le 2\epsilon T + 2\beta_T \sqrt{T} \sqrt{\sum_{t=1}^{T} \|\widetilde{x}_t\|_{A_{t-1}^{-1}}^2} + 2\epsilon T \left(\sum_{t=1}^{T} \|\widetilde{x}_t\|_{A_{t-1}^{-1}}^2 \right).
\end{aligned}$$

By assumptions, we have $\|\widetilde{x}\|_{A_{t-1}^{-1}} \le \|\widetilde{x}\|_{A_0^{-1}} = \lambda^{-1/2}\|\widetilde{x}\|_2 \le 1$ for any $x \in \mathcal{A}$. Therefore, by (Abbasi-Yadkori et al., 2011, Lemma 11), the following inequalities hold:

$$\sum_{t=1}^{T} \|\widetilde{x}_t\|^2_{A_{t-1}^{-1}} \le 2\log(\det(\lambda^{-1}A_t)) \le 2D\log\left(1+\frac{T}{\lambda D}\right), \quad \beta_t \le R\sqrt{D\log\left(1+\frac{T}{\lambda D}\right)+2\log(1/\delta)}+\sqrt{\lambda}B. \quad (1)$$

Thus, on event E, we have

$$\begin{aligned} R(T) &\le 2\beta_T\sqrt{T}\sqrt{2\log(\lambda^{-1}A_T)} + 2\epsilon T + 4\epsilon T\log(\det(\lambda^{-1}A_T)) \\ &= \widetilde{O}\left(\epsilon T + (\sqrt{D}+\sqrt{1/\delta})\sqrt{DT} + \epsilon DT\right) \\ &= \widetilde{O}\left(D\sqrt{T} + \sqrt{DT\log(1/\delta)} + \epsilon DT\right). \end{aligned}$$

□

B.3. Proof of Proposition 11

This proposition can be proved by adapting the standard proof of the adversarial linear bandit problem (Lattimore & Szepesvári, 2020; Bubeck & Cesa-Bianchi, 2012). We recall notation for the adversarial misspecified linear bandit problem and EXP3. Let $\mathcal{A} \ni x \mapsto \widetilde{x} \in \{\xi \in \mathbb{R}^D : \|\xi\| \le 1\}$ be a map, $\{g_t\}_{t=1}^T$ be a sequence of reward functions on $\mathcal{A}$ such that $g_t(x) = \langle \theta_t, \widetilde{x}\rangle + \omega_t(x)$ for $x \in \mathcal{A}$, where $\theta_t \in \mathbb{R}^D$, $\sup_{x\in\mathcal{A}}|g_t(x)|, \|\theta_t\| \le B$, and $\sup_{x\in\mathcal{A}}|\omega_t(x)| \le \epsilon$.

Let $\gamma \in (0,1)$ be an exploration parameter, $\eta > 0$ a learning rate, and π_{exp} an exploitation distribution over $\mathcal{A}$. For a distribution π over $\mathcal{A}$, we put $Q(\pi) = \sum_{x\in\mathcal{A}} \pi(x)\widetilde{x}\widetilde{x}^{\mathrm{T}}$. We define $\phi_t = g_x(x_t)Q_t^{-1}\widetilde{x}_t$ and $\phi_t(x) = \langle \phi_t, \widetilde{x}\rangle$ for $x \in \mathcal{A}$, where the matrix Q_t can be computed from the past observations at round t and is defined later. Let q_t be a distribution over $\mathcal{A}$ such that $q_t(x) \sim \exp\left(\eta\sum_{s=1}^{t-1}\phi_s(x)\right)$ and put $p_t(x) = \gamma\pi_{\mathrm{exp}}(x) + (1-\gamma)q_t(x)$ for $x \in \mathcal{A}$. We assume that $Q(\pi_{\mathrm{exp}})$ is non-singular and define $Q_t = Q(p_t)$. For a distribution π over $\mathcal{A}$, we define $\Gamma(\pi) = \sup_{x\in\mathcal{A}} \widetilde{x}^{\mathrm{T}}Q(\pi_{\mathrm{exp}})^{-1}\widetilde{x}$ and in this section we assume $\Gamma(\pi) \le D$.

Let $x_* = \mathrm{argmax}_{x\in\mathcal{A}}\sum_{t=1}^T g_t(x)$ be an optimal arm and regret is defined as $R(T) = \sum_{t=1}^T g_t(x_*) - \sum_{t=1}^T g_t(x_t)$. We have

$$\begin{aligned} \mathbf{E}\left[R(T)\right] &= \mathbf{E}\left[\sum_{t=1}^T (\langle\theta_t,\widetilde{x}_*\rangle - \langle\theta_t,\widetilde{x}_t\rangle)\right] + \mathbf{E}\left[\sum_{t=1}^T(\omega_t(\widetilde{x}_*) - \omega_t(\widetilde{x}_t))\right] \\ &\le 2\epsilon T + \mathbf{E}\left[\sum_{t=1}^T (\langle\theta_t,\widetilde{x}_*\rangle - \langle\theta_t,\widetilde{x}_t\rangle)\right]. \end{aligned} \quad (2)$$

We denote by $\mathcal{H}_{t-1}$ the sigma field generated by $x_1,\dots,x_{t-1}$ and by $\mathbf{E}_{t-1}$ the conditional expectation conditioned on $\mathcal{H}_{t-1}$. We note that $p_t(x), q_t(x)$ for $x \in \mathcal{A}$ and Q_t are $\mathcal{H}_{t-1}$-measurable but ϕ_t is not. Then we have

$$\begin{aligned} \mathbf{E}\left[\sum_{t=1}^T\langle\theta_t,\widetilde{x}_t\rangle\right] &= \mathbf{E}\left[\sum_{t=1}^T \mathbf{E}_{t-1}\left[\langle\theta_t,\widetilde{x}_t\rangle\right]\right] = \mathbf{E}\left[\sum_{t=1}^T\sum_{x\in\mathcal{A}} p_t(x)\langle\theta_t,\widetilde{x}\rangle\right] \\ &= \gamma\mathbf{E}\left[\sum_{t=1}^T\sum_{x\in\mathcal{A}}\pi_{\mathrm{exp}}(x)\langle\theta,\widetilde{x}\rangle\right] + (1-\gamma)\mathbf{E}\left[\sum_{t=1}^T\sum_{x\in\mathcal{A}} q_t(x)\langle\theta_t,\widetilde{x}\rangle\right] \\ &\ge -\gamma BT + (1-\gamma)S. \end{aligned} \quad (3)$$

Here we used $|\langle\theta_t,\widetilde{x}\rangle| \le \|\theta_t\|\|\widetilde{x}\| \le B$ and S is defined as $\mathbf{E}\left[\sum_{t=1}^T\sum_{x\in\mathcal{A}} q_t(x)\langle\theta_t,\widetilde{x}\rangle\right]$. Since $\sum_{t=1}^T\langle\theta_t,\widetilde{x}_*\rangle \le \gamma BT + (1-\gamma)\sum_{t=1}^T\langle\theta_t,\widetilde{x}_*\rangle$, by inequalities (2), (3), we have

$$\mathbf{E}\left[R(T)\right] \le 2\epsilon T + 2\gamma BT + (1-\gamma)\left(\sum_{t=1}^T\langle\theta_t,\widetilde{x}_*\rangle - S\right). \quad (4)$$

We decompose $S = S_1 + S_2$, where

$$S_1 = \mathbf{E}\left[\sum_{t=1}^{T}\sum_{x\in\mathcal{A}} q_t(x)\langle \phi_t, \widetilde{x}\rangle\right], \quad S_2 = \mathbf{E}\left[\sum_{t=1}^{T}\sum_{x\in\mathcal{A}} q_t(x)\langle \theta_t - \phi_t, \widetilde{x}\rangle\right].$$

First, we bound $|S_2|$. To do this, we prove the following lemma.

LEMMA 22. *For any $x \in \mathcal{A}$, the following inequality holds:*

$$|\mathbf{E}_{t-1}\left[\langle \phi_t - \theta_t, \widetilde{x}\rangle\right]| \le \frac{\epsilon\Gamma(\pi_{\mathrm{exp}})}{\gamma}.$$

In particular, we have $|\mathbf{E}\left[\langle \phi_t - \theta_t, \widetilde{x}\rangle\right]| \le \frac{\epsilon\Gamma(\pi_{\mathrm{exp}})}{\gamma}$.

Proof. We note that by conditioning on $\mathcal{H}_{t-1}$, randomness comes only from x_t. By definition of ϕ_t, we have

$$\begin{aligned}
\mathbf{E}_{t-1}\left[\langle \phi_t, \widetilde{x}\rangle\right] &= \mathbf{E}_{t-1}\left[\langle \left(\langle \theta_t, \widetilde{x}_t\rangle + \omega_t(x_t)\right) Q_t^{-1}\widetilde{x}_t, \widetilde{x}\rangle\right] \\
&= \mathbf{E}_{t-1}\left[\widetilde{x}^{\mathrm{T}} Q_t^{-1}\widetilde{x}_t\widetilde{x}_t^{\mathrm{T}}\theta_t\right] + \mathbf{E}_{t-1}\left[\omega_t(x_t)\widetilde{x}^{\mathrm{T}} Q_t^{-1}\widetilde{x}_t\right] \\
&= \langle \theta_t, \widetilde{x}\rangle + \mathbf{E}_{t-1}\left[\omega_t(x_t)\widetilde{x}^{\mathrm{T}} Q_t^{-1}\widetilde{x}_t\right].
\end{aligned}$$

Therefore,

$$|\mathbf{E}_{t-1}\left[\langle \phi_t - \theta_t, \widetilde{x}\rangle\right]| \le \epsilon \mathbf{E}_{t-1}\left[\|\widetilde{x}_t\|_{Q_t^{-1}}\|\widetilde{x}\|_{Q_t^{-1}}\right] \le \frac{\epsilon}{\gamma}\mathbf{E}_{t-1}\left[\|\widetilde{x}_t\|_{Q(\pi_{\mathrm{exp}})^{-1}}\|\widetilde{x}\|_{Q(\pi_{\mathrm{exp}})^{-1}}\right] \le \frac{\epsilon\Gamma(\pi_{\mathrm{exp}})}{\gamma}.$$

Here in the second inequality, we use $\gamma Q(\pi_{\mathrm{exp}}) \le Q_t$ and the last inequality follows from the definition of $\Gamma(\pi_{\mathrm{exp}})$. The second assertion follows from

$$|\mathbf{E}\left[\langle \phi_t - \theta_t, \widetilde{x}\rangle\right]| \le \mathbf{E}\left[|\mathbf{E}_{t-1}\left[\langle \phi_t - \theta_t, \widetilde{x}\rangle\right]|\right] \le \frac{\epsilon\Gamma(\pi_{\mathrm{exp}})}{\gamma}.$$

□

By this lemma, we can bound S_2 as follows.

LEMMA 23. *The following inequality holds:*

$$|S_2| \le \frac{\epsilon T\Gamma(\pi_{\mathrm{exp}})}{\gamma}.$$

Proof. Since $q_t(x)$ is $\mathcal{H}_{t-1}$-measurable for any $x \in \mathcal{A}$, we have

$$S_2 = \mathbf{E}\left[\sum_{t=1}^{T}\sum_{x\in\mathcal{A}} q_t(x)\mathbf{E}_{t-1}\left[\langle \theta_t - \phi_t, \widetilde{x}\rangle\right]\right].$$

Therefore, we have

$$|S_2| \le \mathbf{E}\left[\sum_{t=1}^{T}\sum_{x\in\mathcal{A}} q_t(x)\left|\mathbf{E}_{t-1}\left[\langle \theta_t - \phi_t, \widetilde{x}\rangle\right]\right|\right] \le \frac{\epsilon T\Gamma(\pi_{\mathrm{exp}})}{\gamma}.$$

Here we used Lemma 22 in the last inequality. □

Next, we introduce the following elementary lemma (c.f. Chatterji et al. (2019, Lemma 49)).

LEMMA 24. *Let $\eta > 0$ and X be a random variable. We assume that $\eta X \le 1$ almost surely. Then we have*

$$\mathbf{E}\left[X\right] \ge \frac{1}{\eta}\log\left(\mathbf{E}\left[\exp(\eta X)\right]\right) - (e-2)\eta\mathbf{E}\left[X^2\right].$$

Proof. By $\log(x) \le x - 1$ for $x > 0$ and $\exp(y) \le 1 + y + (e-2)y^2$ for $y \le 1$, we have

$$\log \mathbf{E}\left[\exp(\eta X)\right] \le \mathbf{E}\left[\exp(\eta X)\right] - 1 \le \mathbf{E}\left[\eta X + (e-2)\eta^2 X^2\right].$$

□

To apply the lemma with $X = \langle \phi_t, \widetilde{x} \rangle$ and $\mathbf{E} = \mathbf{E}_{x \sim q_t}$, we prove the following:

LEMMA 25. *Let $x \in \mathcal{A}$ and assume that $\gamma = \eta B \Gamma(\pi_{\exp})$. Then, we have $\eta |\langle \phi_t, \widetilde{x} \rangle| \le 1$.*

Proof. By definition of ϕ_t, we have

$$\eta |g_t(x_t) \widetilde{x}_t^{\mathrm{T}} Q_t^{-1} \widetilde{x}| \le \eta B \|\widetilde{x}_t\|_{Q_t^{-1}} \|\widetilde{x}\|_{Q_t^{-1}} \le \frac{\eta B \Gamma(\pi_{\exp})}{\gamma}.$$

Here, in the last inequality, we use $\gamma Q(\pi_{\exp}) \le Q_t$ and the definition of $\Gamma(\pi_{\exp})$. □

By Lemma 24 and Lemma 25, we obtain the following.

$$S_1 \ge \frac{U_1}{\eta} - (e-2)\eta U_2, \tag{5}$$

where U_1 and U_2 are given as

$$U_1 = \mathbf{E}\left[\sum_{t=1}^{T} \log\left(\sum_{x \in \mathcal{A}} q_t(x) \exp\left(\eta \langle \phi_t, \widetilde{x} \rangle\right)\right)\right], \quad U_2 = \mathbf{E}\left[\sum_{t=1}^{T} \sum_{x \in \mathcal{A}} q_t(x) \langle \phi_t, \widetilde{x} \rangle^2\right].$$

We bound $|U_2|$ as follows.

LEMMA 26. *The following inequality holds:*

$$|U_2| \le \frac{B^2 D T}{1 - \gamma}.$$

Proof. By definition of ϕ_t, we have

$$\begin{aligned}\sum_{x \in \mathcal{A}} q_t(x) \langle \phi_t, \widetilde{x} \rangle^2 &\le B^2 \sum_{x \in \mathcal{A}} q_t(x) \widetilde{x}_t^{\mathrm{T}} Q_t^{-1} \widetilde{x} \widetilde{x}^{\mathrm{T}} Q_t^{-1} \widetilde{x}_t = B^2 \widetilde{x}_t Q_t^{-1} Q(q_t) Q_t^{-1} \widetilde{x}_t \\ &\le \frac{B^2}{1-\gamma} \widetilde{x}_t^{\mathrm{T}} Q_t^{-1} \widetilde{x}_t.\end{aligned}$$

Here the last inequality follows from $(1-\gamma) Q(q_t) \le Q_t$. Therefore,

$$\begin{aligned}\mathbf{E}\left[\sum_{x \in \mathcal{A}} \langle \phi_t, \widetilde{x} \rangle^2\right] &\le \frac{B^2}{1-\gamma} \mathbf{E}\left[\widetilde{x}_t^{\mathrm{T}} Q_t^{-1} \widetilde{x}_t\right] = \frac{B^2}{1-\gamma} \mathbf{E}\left[\mathrm{Tr}\left(\widetilde{x}_t \widetilde{x}_t^{\mathrm{T}} Q_t^{-1}\right)\right] \\ &= \frac{B^2}{1-\gamma} \mathbf{E}\left[\mathrm{Tr}\left(Q_t Q_t^{-1}\right)\right] = \frac{B^2 D}{1-\gamma}.\end{aligned}$$

Here the second equality follows from the fact that Q_t is $\mathcal{H}_{t-1}$-measurable and the linearity of the trace. The assertion of the lemma follows from this. □

Next, we give a lower bound for U_1.

LEMMA 27. *Let $x_0 \in \mathcal{A}$ be any element. Then the following inequality holds:*

$$U_1 \ge \eta \mathbf{E}\left[\sum_{t=1}^{T} \langle \phi_t, \widetilde{x}_0 \rangle\right] - \log(|\mathcal{A}|).$$

Proof. By definition of q_t, we have

$$\begin{aligned}
U_1 &= \mathbf{E}\left[\sum_{t=1}^{T}\left\{\log\left(\sum_{x\in\mathcal{A}}\exp\left(\eta\sum_{s=1}^{t}\langle\phi_s,\widetilde{x}\rangle\right)\right)-\log\left(\sum_{x\in\mathcal{A}}\exp\left(\eta\sum_{s=1}^{t-1}\langle\phi_s,\widetilde{x}\rangle\right)\right)\right\}\right]\\
&= \mathbf{E}\left[\log\left(\sum_{x\in\mathcal{A}}\exp\left(\eta\sum_{s=1}^{T}\langle\phi_s,\widetilde{x}\rangle\right)\right)\right]-\log|\mathcal{A}|\\
&= \eta\mathbf{E}\left[\sum_{t=1}^{T}\langle\phi_t,\widetilde{x}_0\rangle\right]-\log|\mathcal{A}|.
\end{aligned}$$

□

Proof of Proposition 11. We assume $\gamma=\eta B\Gamma(\pi_{\exp})$. By (4), we have

$$\mathbf{E}\left[R(T)\right]\le 2\epsilon T+2\gamma BT+(1-\gamma)\mathbf{E}\left[\sum_{t=1}^{T}\langle\theta_t,\widetilde{x}_*\rangle\right]-(1-\gamma)S_1+(1-\gamma)|S_2|$$

By inequality (5), Lemma 23, and Lemma 27 with $x_0=x_*$, we have

$$\begin{aligned}
\mathbf{E}\left[R(T)\right] &\le 2\epsilon T+2\gamma BT+(1-\gamma)\mathbf{E}\left[\sum_{t=1}^{T}\langle\theta_t-\phi_t,\widetilde{x}_*\rangle\right]+\frac{1-\gamma}{\eta}\log|\mathcal{A}|+(e-2)\eta(1-\gamma)U_2+\frac{\epsilon T\Gamma(\pi_{\exp})}{\gamma}\\
&\le 2\epsilon T+2\gamma BT+\frac{\epsilon T\Gamma(\pi_{\exp})(1-\gamma)}{\gamma}+\frac{\log|\mathcal{A}|}{\eta}+(e-2)\eta B^2DT+\frac{\epsilon T\Gamma(\pi_{\exp})}{\gamma}\\
&\le 2\epsilon T+2\gamma BT+\frac{\log|\mathcal{A}|}{\eta}+(e-2)\eta B^2DT+\frac{2\epsilon T\Gamma(\pi_{\exp})}{\gamma}.
\end{aligned}$$

Here in the second inequality, we used Lemma 22 and Lemma 26. By $\gamma=\eta B\Gamma(\pi_{\exp})$ and $\Gamma(\pi_{\exp})\le D$, we have

$$\begin{aligned}
\mathbf{E}\left[R(T)\right] &\le 2\epsilon T+2B^2\eta T\Gamma(\pi_{\exp})+\frac{2\epsilon T}{B\eta}+\frac{\log|\mathcal{A}|}{\eta}+(e-2)B^2\eta DT\\
&\le 2\epsilon T+eB^2\eta DT+\frac{2\epsilon T}{B\eta}+\frac{\log|\mathcal{A}|}{\eta}.
\end{aligned}$$

□

B.4. Proof of Proposition 18

We assume $\lambda\ge 1$. This can be proved by modifying the proof of (Agrawal & Goyal, 2013). Since most of their arguments can be directly applicable to our case, we omit proofs of some lemmas. Let $(\Psi,\Pr,\mathcal{G})$ be the probability space on which all random variables considered here are defined, where $\mathcal{G}\subset 2^{\Psi}$ is a σ-algebra on Ψ. We put $x^*:=\operatorname{argmax}_{x\in\mathcal{A}}g(x)$ and for $t=1,2,\dots,T$ and $x\in\mathcal{A}$, we put $\Delta(x):=\langle\widetilde{x}^*,\theta\rangle-\langle\widetilde{x},\theta\rangle$. We also put $v_t:=l_t:=\beta(A_{t-1},\delta/(2T),\lambda)+\epsilon\psi_{t-1}$ and $g_t:=\sqrt{4\log(|\mathcal{A}|t)}v_t+l_t$. In each round t, μ_{t-1} is sampled from the multinomial normal distribution $\mathcal{N}\left(\theta_{t-1},\left(\beta(A_{t-1},\delta/2,\lambda)+\epsilon\psi_t\right)^2A_{t-1}^{-1}\right)$. For $t=1,\dots,T$, we define E_t by

$$E_t:=\left\{\psi\in\Psi:|\langle\widetilde{x},\theta_{t-1}\rangle-\langle\widetilde{x},\theta\rangle|\le l_t\|\widetilde{x}\|_{A_{t-1}^{-1}},\quad\forall x\in\mathcal{A}\right\},$$

and define event E'_t by

$$E'_t:=\left\{\psi\in\Psi:|\langle\widetilde{x},\mu_{t-1}\rangle-\langle\widetilde{x},\theta_{t-1}\rangle|\le\sqrt{4\log(|\mathcal{A}|t)}v_t\|\widetilde{x}\|_{A_{t-1}^{-1}},\quad\forall x\in\mathcal{A}\right\}.$$

For an event G, we denote by 1_G the corresponding indicator function. Then by assumptions, we see that $E_t\in\mathcal{F}_{t-1}$, i.e., 1_{E_t} is $\mathcal{F}_{t-1}$-measurable. For a random variable X on Ψ, we say "on event E_t, the conditional expectation (or conditional probability) $\mathbf{E}\left[X\mid\mathcal{F}_{t-1}\right]$ satisfies a property" if and only if $1_{E_t}\mathbf{E}\left[X\mid\mathcal{F}_{t-1}\right]=\mathbf{E}\left[1_{E_t}X\mid\mathcal{F}_{t-1}\right]$ satisfies the property for almost all $\psi\in\Psi$.

Then by Lemma 21 and the proof of (Agrawal & Goyal, 2013, Lemma 1), we have

LEMMA 28. $\Pr(E_t) \geq 1 - \frac{\delta}{2T}$ *and* $\Pr(E_t' \mid \mathcal{F}_{t-1}) \geq 1 - 1/t^2$ *for all* t.

We note that the proof of (Agrawal & Goyal, 2013, Lemma 2) works if $l_t \leq v_t$, i.e., we have the following lemma:

LEMMA 29. *On event* E_t*, we have*

$$\Pr\left(\langle \mu_t, \widetilde{x}^* \rangle > \langle \theta, \widetilde{x}^* \rangle \mid \mathcal{F}_{t-1}\right) \geq p,$$

where $p = \frac{1}{4e\sqrt{\pi}}$.

The main differences of our proof and theirs lie in the definitions of l_t, v_t, x^*, and $\Delta(x)$ (they define $\Delta(x)$ as $\sup_{y \in \mathcal{A}} \langle \theta, \widetilde{y} \rangle - \langle \theta, \widetilde{x} \rangle$ and we consider $x^* = \operatorname{argmax} g(x)$ instead of $\operatorname{argmax}_x \langle \widetilde{x}, \theta \rangle$). However, it can be verified that these differences do not matter in the arguments of Lemma 3, 4 in (Agrawal & Goyal, 2013). In fact, we can prove the following lemma in a similar way to the proof of (Agrawal & Goyal, 2013, Lemma 3).

LEMMA 30. *We define* $C(t)$ *by* $\{x \in \mathcal{A} : \Delta(x) > g_t \|\widetilde{x}\|_{A_{t-1}^{-1}}\}$*. On event* E_t*, we have*

$$\Pr(x \notin C(t) \mid \mathcal{F}_{t-1}) \geq p - \frac{1}{t^2}.$$

Here p *is given in Lemma 29.*

Proof. Because the algorithm selects $x \in \mathcal{A}$ that maximizes $\langle \widetilde{x}, \mu_{t-1} \rangle$, if $\langle \widetilde{x}^*, \mu_{t-1} \rangle > \langle \widetilde{x}, \mu_{t-1} \rangle$ for all $x \in C(t)$, then we have $x_t \notin C(t)$. Therefore, we have

$$\Pr(x_t \notin C(t) \mid \mathcal{F}_{t-1}) \geq \Pr\left(\langle \widetilde{x}^*, \mu_{t-1} \rangle > \langle \widetilde{x}, \mu_{t-1} \rangle, \forall x \in C(t) \mid \mathcal{F}_{t-1}\right). \tag{6}$$

By definitions of $C(t), E_t$, and E_t', on even $E_t \cap E_t'$, we have $\langle \widetilde{x}, \mu_{t-1} \rangle \leq \langle \widetilde{x}, \theta \rangle + g_t \|\widetilde{x}\|_{A_{t-1}^{-1}} < \langle \widetilde{x}^*, \theta \rangle$ for all $x \in C(t)$. Therefore, on $E_t \cap E_t'$, if $\langle \widetilde{x}^*, \mu_{t-1} \rangle > \langle \widetilde{x}^*, \theta \rangle$, we have $\langle \widetilde{x}^*, \mu_{t-1} \rangle > \langle \widetilde{x}, \mu_{t-1} \rangle$ for all $x \in C(t)$. Thus we obtain the following inequalities:

$$\begin{aligned}
&\Pr(\langle \widetilde{x}^*, \mu_{t-1} \rangle > \langle \widetilde{x}, \mu_{t-1} \rangle, \quad \forall x \in C(t) \mid \mathcal{F}_{t-1}) \\
&\geq \Pr(\langle \widetilde{x}^*, \mu_{t-1} \rangle > \langle \widetilde{x}^*, \theta \rangle \mid \mathcal{F}_{t-1}) - \Pr((E_t')^c \mid \mathcal{F}_{t-1}) \\
&\geq p - 1/t^2.
\end{aligned}$$

Here $(E_t')^c$ is the complement of E_t' and we used Lemmas 28, 29 in the last inequality. By inequality (6), we have our assertion. ∎

We can also prove the following lemma in a similar way to the proof of (Agrawal & Goyal, 2013, Lemma 4).

LEMMA 31. *On event* E_t*, we have*

$$\mathbf{E}\left[\Delta(x_t) \mid \mathcal{F}_{t-1}\right] \leq c_1 g_t \mathbf{E}\left[\|x_t\|_{A_{t-1}^{-1}} \mid \mathcal{F}_{t-1}\right] + \frac{c_2 g_t}{t^2},$$

where c_1 *and* c_2 *are universal constants.*

For $t = 1, 2, \ldots, T$, define random variables X_t and Y_t by

$$X_t := \Delta(x_t) 1_{E_t} - c_1 g_t \|x_t\|_{A_{t-1}^{-1}} - \frac{c_2 g_t}{t^2}, \quad Y_t := \sum_{s=1}^{t} X_t.$$

From Lemma 31, we can prove the following lemma.

LEMMA 32. *The process* $\{Y_t\}_{t=0,\ldots,T}$ *is a super-martingale process w.r.t. the filtration* $\{\mathcal{F}_t\}_t$.

Proof of Proposition 18. By Lemma 32 and $\|X_t\| \leq 2(B+\epsilon)+(c_1+c_2)g_t$ (for all t), applying Azuma-Hoeffding inequality, we see that there exists an event G with $\Pr(G) \geq 1 - \delta/2$ such that on G, the following inequality holds:

$$\sum_{t=1}^{T} \Delta(x_t) 1_{E_t} \leq \sum_{t=1}^{T} c_1 g_t \|\widetilde{x}_t\|_{A_{t-1}^{-1}} + \sum_{t=1}^{T} c_2 g_t / t^2 + \sqrt{\left(4T(B+\epsilon)^2 + 2(c_1+c_2)^2 \sum_{t=1}^{T} g_t^2\right) \log(2/\delta)}.$$

Since $g_t \le g_T$ for any t, on the event G, we have

$$\sum_{t=1}^{T} \Delta(x_t) 1_{E_t} \le c_1 g_T \sqrt{T} \sqrt{\sum_{t=1}^{T} \|\widetilde{x}_t\|^2_{A_{t-1}^{-1}}} + c_2 g_T \frac{\pi^2}{6} + \sqrt{T}\sqrt{(4(B+\epsilon)^2 + 2(c_1+c_2)^2 g_T^2)\log(2/\delta)}.$$

By inequalities (1), we have

$$\sqrt{\sum_{t=1}^{T} \|\widetilde{x}_t\|^2_{A_{t-1}^{-1}}} = \widetilde{O}(\sqrt{D}), \quad g_T = \widetilde{O}(\sqrt{\log(|\mathcal{A}|)} v_T) = \widetilde{O}\left(\sqrt{|\log|\mathcal{A}|}(\sqrt{D} + \sqrt{\log(1/\delta)} + \epsilon \psi_T)\right).$$

Since $\psi_T = \sum_{s=1}^{T} \|\widetilde{x}_s\|_{A_{s-1}^{-1}} \le \sqrt{T}\sqrt{\sum_{s=1}^{T} \|\widetilde{x}_s\|^2_{A_{s-1}^{-1}}} = \widetilde{O}(\sqrt{DT})$, we obtain

$$g_T = \widetilde{O}\left(\sqrt{\log A}(\sqrt{D} + \sqrt{\log(1/\delta)} + \epsilon\sqrt{DT})\right).$$

Therefore, on the event G, we have

$$\sum_{t=1}^{T} \Delta(x_t) 1_{E_t} = \widetilde{O}\left(\sqrt{\log(|\mathcal{A}|)}\left\{D\sqrt{T} + \sqrt{DT\log(1/\delta)} + \log(1/\delta)\sqrt{T} + \left(DT + T\sqrt{D\log(1/\delta)}\right)\epsilon\right\}\right).$$

Therefore, on event $\bigcap_{t=1}^{T} E_t \cap G$, we can upper bound the regret as follows:

$$\begin{aligned} R(T) &= \sum_{t=1}^{T}\{g(x^*) - g(x_t)\} \le \epsilon T + \sum_{t=1}^{T} \Delta(x_t) 1_{E_t} \\ &= \widetilde{O}\left(\sqrt{\log(|\mathcal{A}|)}\left\{D\sqrt{T} + \sqrt{DT\log(1/\delta)} + \log(1/\delta)\sqrt{T} + \left(DT + T\sqrt{D\log(1/\delta)}\right)\epsilon\right\}\right). \end{aligned}$$

Since $\Pr(\bigcap_{t=1}^{T} E_t \cap G) \ge 1 - \delta$, we have the assertion of the proposition. □

B.5. Proof of Theorem 13

Since Theorems 16, 19 can be proved in a similar way, we only provide proof of Theorem 13.

Let $\{\xi_1, \dots, \xi_D\}$ and $N_1, \dots, N_D$ be a sequence of points and Newton basis returned by Algorithm 2 with $\mathfrak{e} = \frac{\alpha}{T^q}$, where $D = D_{q,\alpha}(T)$ and $q \ge 1/2$.

We verify the assumptions of the (stochastic) misspecified linear bandit problem hold, i.e., we show there exists $\theta \in \mathbb{R}^D$ such that the following conditions are satisfied for $\widetilde{x} = [N_1(x), \dots, N_D(x)]^{\mathrm{T}}$ and θ:

1. $\|\widetilde{x}\|_2 \le 1$.
2. If x is a $\mathcal{A}$-valued random variable and $\mathcal{F}_t$-measurable, then $\widetilde{x}$ is $\mathcal{F}_t$-measurable .
3. $\|\theta\|_2 \le B$.
4. $\sup_{x\in\mathcal{A}} |f(x) - \langle \theta, \widetilde{x}\rangle| < \epsilon$, where $\epsilon = \alpha B / T^q$.

We put $X_D := \{\xi_1, \dots, \xi_D\}$. Then by definition, Newton basis $N_1, \dots, N_D$ is a basis of $V(X_D)$. We define $\theta_1, \dots, \theta_D \in \mathbb{R}$ by $\Pi_{V(X_D)} f = \sum_{i=1}^{D} \theta_i N_i$ and put $\theta = [\theta_1, \dots, \theta_D]^{\mathrm{T}}$. Since Newton basis is an orthonormal basis of $V(X_D)$, we have

$$\|\theta\|_2 = \left\|\sum_{i=1}^{D} \theta_i N_i\right\|_{\mathcal{H}_K(\Omega)} = \left\|\Pi_{V(X_D)} f\right\|_{\mathcal{H}_K(\Omega)} \le \|f\|_{\mathcal{H}_K(\Omega)} \le B.$$

By the orthonormality, we have $P^2_{V(X)}(x) = K(x,x) - \sum_{i=1}^{m} N_i^2(x)$ (c.f. Santin & Haasdonk (2017, Lemma 5)). Then by assumption, we have $\|\widetilde{x}\|_2^2 = \sum_{i=1}^{m} N_i^2(x) = K(x,x) - P^2_{V(X_D)}(x) \le 1$. Since N_k for $k = 1, \dots, D$ is a linear combination of $K(\cdot, \xi_1), \dots, K(\cdot, \xi_D)$ and K is continuous, $x \mapsto \widetilde{x}$ is continuous. Therefore, $\widetilde{x}$ is $\mathcal{F}_t$-measurable if x is $\mathcal{F}_t$-measurable. By definition of the P-greedy algorithm, we have $\sup_{x \in \mathcal{A}} P_{V(X_D)}(x) < \frac{\alpha}{T^q}$. By this inequality and the definition of the power function, the following inequality holds:

$$\sup_{x \in \mathcal{A}} |f(x) - \langle \theta, \widetilde{x} \rangle| = \sup_{x \in \mathcal{A}} |f(x) - \left(\Pi_{V(X_D)} f\right)(x)| \leq \|f\| \frac{\alpha}{T^q} \leq \frac{\alpha B}{T^q}.$$

Thus, one can apply results of a misspecified linear bandit problem with $\epsilon = \frac{\alpha B}{T^q}$. By applying Proposition 8, with probability at least $1 - \delta$, the regret is upper bounded as follows:

$$R_{\text{APG-UCB}}(T) = \widetilde{O}\left(\sqrt{T D_{q,\alpha}(T) \log(1/\delta)} + D_{q,\alpha}(T)\sqrt{T}\right).$$

Since computing Newton basis requires $O(|\mathcal{A}| D^2)$ time and total complexity of the modified LinUCB is given as $O(|\mathcal{A}| D^2 T)$, we have the assertion of Theorem 13.

B.6. Proof of Theorem 17

For simplicity, by normalization, we assume $B = 1$. We denote by $R_{\text{APG-EXP3}}(T)$ the cumulative regret that APG-EXP3 with $q = 1$ and $\alpha = \log(|\mathcal{A}|)$ incurs up to time step T. We can reduce the adversarial RKHS bandit problem to the adversarial misspecified linear bandit problem as in §B.5. To apply Proposition 11, we need to prove that $\{\widetilde{x} | x \in \mathcal{A}\}$ spans $\mathbb{R}^D$. We denote by $X = \{X_1, \dots, X_D\}$ the points returned by the P-greedy algorithm. Then, since $N_1, \dots, N_D$ is a basis of $V(X)$ and K is positive definite, $\operatorname{rank}\left(N_i(x)\right)_{1 \leq i \leq D, x \in \mathcal{A}} = \operatorname{rank}\left(K(x_i, x)\right)_{1 \leq i \leq D, x \in \mathcal{A}} = D$. Therefore, $\{\widetilde{x} | x \in \mathcal{A}\}$ spans $\mathbb{R}^D$.

By Proposition 11, we have

$$\mathbf{E}\left[R_{\text{APG-EXP3}}(T)\right] \leq 2\epsilon T + e\eta D T + \frac{2\epsilon}{\eta} + \frac{\log(|\mathcal{A}|)}{\eta},$$

where $\epsilon = \frac{\log(|\mathcal{A}|)}{T}$ and $D = D_{1,\log(|\mathcal{A}|)}(T)$. Thus we have $\mathbf{E}\left[R_{\text{APG-EXP3}}(T)\right] \leq 2\log(|\mathcal{A}|) + e\eta D T + \frac{3\log(|\mathcal{A}|)}{\eta}$. By taking $\eta = \sqrt{\frac{\log(|\mathcal{A}|)}{DT}}$, we have the assertion of the theorem.

B.7. Proof of Theorem 14

First, we prove that the P-greedy algorithm (Algorithm 1) also gives a uniform kernel approximation.

LEMMA 33. *Let $N_1, \dots, N_D$ be a Newton basis returned by the P-greedy algorithm 1 with error $\mathfrak{e}$ and $\widehat{\Omega} = \mathcal{A}$. For $x \in \mathcal{A}$, we put $\widetilde{x} := [N_1(x), \dots, N_D(x)]^{\mathrm{T}}$. Then, we have $\sup_{x,y \in \mathcal{A}} |K(x,y) - \langle \widetilde{x}, \widetilde{y} \rangle| \leq \mathfrak{e}$.*

Proof. We denote by X the points returned by the P-greedy algorithm. Then, by definition of the Power function, we have

$$\left|h(x) - \left(\Pi_{V(X)} h\right)(x)\right| \leq \|h\|_{\mathcal{H}_K(\Omega)} \mathfrak{e},$$

for any $h \in \mathcal{H}_K(\Omega)$ and $x \in \mathcal{A}$. We take arbitrary $y \in \mathcal{A}$ and take $h = K(\cdot, y)$. Since $N_1, \dots, N_D$ is an orthonormal basis of $V(X)$, we have

$$\left(\Pi_{V(X)} h\right)(x) = \sum_{i=1}^{D} \langle h, N_i \rangle_{\mathcal{H}_K(\Omega)} N_i(x) = \sum_{i=1}^{D} N_i(y) N_i(x) = \langle \widetilde{x}, \widetilde{y} \rangle.$$

Here, in the second equality, we used the reproducing property. Since $\|h\|_{\mathcal{H}_K(\Omega)} \leq 1$ and x, y are arbitrary, we have our assertion. $\square$

Next, we introduce the following classical result on matrix eigenvalues.

LEMMA 34 (a specail case of the Wielandt-Hoffman theorem Hoffman et al. (1953))**.** *Let $A, B \in \mathbb{R}^{n \times n}$ be symmetric matrices. Denote by $a_1 \leq \cdots \leq a_n$ and $b_1 \leq \cdots \leq b_n$ be the eigenvalues of A and B respectively. Then, we have $\sum_{i=1}^{n} |a_i - b_i|^2 \leq \|A - B\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm.*

By these lemmas, we can prove $\log\det\left(\lambda^{-1} A_T\right)$ is an approximation of the maximum information gain.

LEMMA 35. *We apply APG-UCB with admissible error $\mathfrak{e}$ to the stochastic RKHS bandit, then following inequality holds:*

$$\log\det\left(\lambda^{-1}A_T\right) \le 2\gamma_T + \frac{\mathfrak{e}T^{3/2}}{\lambda}.$$

Proof. We define a $T\times T$ matrix $\widetilde{K}_T$ as $(\langle \widetilde{x}_i, \widetilde{x}_j\rangle)_{1\le i,j\le T}$. Since for any matrix $X\in\mathbb{R}^{n\times m}$, $\det(1_n + XX^{\mathrm{T}}) = \det(1_m + X^{\mathrm{T}}X)$ holds, we have $\det(\lambda^{-1}A_T) = \det\left(1_T + \lambda^{-1}\widetilde{K}_T\right)$. We denote by $\rho_1 \le \cdots \le \rho_T$ the eigenvalues of K_T and $\widetilde{\rho}_1 \le \cdots \le \widetilde{\rho}_T$ those of $\widetilde{K}_T$. Then by the Wielandt-Hoffman theorem (Lemma 34), we have

$$\sqrt{\sum_{i=1}^{T}(\rho_i - \widetilde{\rho}_i)^2} \le \lambda^{-1}\|K_T - \widetilde{K}_T\|_F \le \lambda^{-1}\mathfrak{e}T, \tag{7}$$

where the last inequality follows from Lemma 33. Thus, we have

$$\begin{aligned}
\log\det\left(\lambda^{-1}A_T\right) = \log\det\left(1_T + \lambda^{-1}\widetilde{K}_T\right) &= \sum_{i=1}^{T}\log(\widetilde{\rho}_i) = \sum_{i=1}^{T}\log(\rho_i) + \sum_{i=1}^{T}\log(\widetilde{\rho}_i/\rho_i) \\
&\le \log\det(1_T + \lambda^{-1}K_T) + \sum_{i=1}^{T}\frac{\widetilde{\rho}_i - \rho_i}{\rho_i} \\
&\le \log\det(1_T + \lambda^{-1}K_T) + \sum_{i=1}^{T}|\widetilde{\rho}_i - \rho_i| \\
&\le \log\det(1_T + \lambda^{-1}K_T) + \frac{\mathfrak{e}T^{3/2}}{\lambda}.
\end{aligned}$$

Here in the second inequality, we used $\rho_i \ge 1$ and in the third inequality, we used inequality (7) and the Cauchy-Schwartz inequality. Noting that $\log\det(1_T + \lambda^{-1}K_T) \le 2\gamma_T$ (Chowdhury & Gopalan, 2017), we have our assertion. ☐

We provide a more precise result than Theorem 14. We can prove the following by Proposition 8.

PROPOSITION 36. *We assume that $\lambda^{-1}\log\left(\det(\lambda^{-1}A_T)\right) \le 2\gamma_T + \delta_T$, where $\delta_T = O(T^{a-q})$ with $a\in\mathbb{R}$ and q is the parameter of APG-UCB. We also assume that $\delta_T = O(\gamma_T)$ and $\lambda = 1$. Then with probability at least $1-\delta$, the cumulative regret of APG-UCB is upper bounded by a function $b(T)$, where $b(T)$ is given as*

$$b(T) = 4\beta_T^{\textit{IGP-UCB}}\sqrt{\gamma_T T} + O(\sqrt{T\gamma_T}T^{(a-q)/2} + \gamma_T T^{1-q}), \tag{8}$$

where $\beta_T^{\textit{IGP-UCB}}$ is defined by $B + R\sqrt{2(\gamma_T + 1 + \log(1/\delta))}$.

REMARK 37. We note that the cumulative regret of IGP-UCB is upper bounded by $4\beta_T^{\text{IGP-UCB}}\sqrt{\gamma_T(T+2)}$ by the proof in (Chowdhury & Gopalan, 2017).

If $q > \max(a, 1/2)$, then the first term $4\beta_T^{\text{IGP-UCB}}\sqrt{\gamma_T T}$ in (8) is the main term of $b(T)$. By Lemma 35, we can take $a = 3/2$. Thus, we have the assertion of Theorem 14.

C. Supplement to the Experiments

C.1. Experimental Setting

For each reward function f, we add independent Gaussian noise of mean 0 and standard deviation $0.2\cdot\|f\|_{L^1(\mathcal{A})}$. We use the L^1-norm because even if we normalize f so that $\|f\|_{\mathcal{H}_K(\Omega)} = 1$, the values of the function f can be small. As for the parameters of the kernels, we take $\mu = 2d$ for the RQ kernel because the condition $\mu = \Omega(d)$ is required for positive definiteness. We take $l = 0.3\sqrt{d}$ and $l = 0.2\sqrt{d}$ if the kernel is RQ kernel and SE kernel respectively because the diameter of the d-dimensional cube is $\sqrt{d}$. As for the parameters of the algorithms, we take $B = 1, \delta = 10^{-3}$ and $R = 0.2\cdot\left(\sum_{i=1}^{10}\|f_i\|_{L^1(\mathcal{A})}/10\right)$ for both algorithms, where $f_1,\dots,f_{10}$ are the reward functions used for the experiment. We take $\lambda = 1, \alpha = 5\cdot 10^{-3}, q = 1/2$ for APG-UCB and $\lambda = 1 + 2/T$ for IGP-UCB.

Since exact value of the maximum information gain is not known, when computing UCB for IGP-UCB, we modify IGP-UCB as follows. Using notation of (Chowdhury & Gopalan, 2017), IGP-UCB selects an arm x maximizing $\mu_{t-1}(x) + \beta_t \sigma_{t-1}(x)$, where $\beta_t = B + R\sqrt{2(\gamma_{t-1} + 1 + \log(1/\delta))}$. Since exact value of γ_{t-1} is not known, we use $\frac{1}{2} \ln \det(I + \lambda^{-1} K_{t-1})$ instead of γ_{t-1}. From their proof, it is easy to see that this modification of IGP-UCB have the same guarantee for the regret upper bound as that of IGP-UCB. In addition, by $\ln \det(I + \lambda^{-1} K_t) = \sum_{s=1}^{t} \log(1 + \lambda^{-1} \sigma_{s-1}^2(x_s))$, one can update $\ln \det(I + \lambda^{-1} K_t)$ in $O(t^2)$ time at each round if K_t^{-1} is known. To compute the inverse of the regularized kernel matrix K_t^{-1}, we used a Schur complement of the matrix.

Computation was done by Intel Xeon E5-2630 v4 processor with 128 GB RAM. We computed UCB for each arm in parallel for both algorithms. For matrix-vector multiplication, we used efficient implementation of the dot product provided in `https://github.com/dimforge/nalgebra/blob/dev/src/base/blas.rs`.

C.2. Additional Experimental Results

As shown in the main article and §B.7, the error $\mathfrak{e}$ balances the computational complexity and cumulative regret, i.e., if $\mathfrak{e}$ is smaller, then the cumulative regret is smaller, but the computational complexity becomes larger. In this subsection, we provide additional experimental results by changing α with fixed $q = 1/2$. We also show results for more complicated reward functions, i.e. $l = 0.2\sqrt{d}$ for RQ kernels (μ is the same) and $l = 0.1\sqrt{d}$ for SE kernels.

In Table 2, we show the number of points returned by the P-greedy algorithms for the RQ and SE kernels.

Table 2. The Number of Points Returned by the P-greedy Algorithm with $\epsilon = \frac{5 \cdot 10^{-3}}{\sqrt{T}}$.

	RQ ($l = 0.3\sqrt{d}$)	SE ($l = 0.2\sqrt{d}$)	RQ ($l = 0.2\sqrt{d}$)	SE ($l = 0.1\sqrt{d}$)
$d = 1$	18	15	23	25
$d = 2$	105	108	188	283
$d = 3$	376	457	725	994

In Figures 3, 4 and Tables 3, 4, we show the dependence on the parameter α. In these figures, we denote APG-UCB with parameter α by APG-UCB(α).

In Figures 5, 6 and Tables 5, 6, we also show the dependence on the parameter α for more complicated functions.

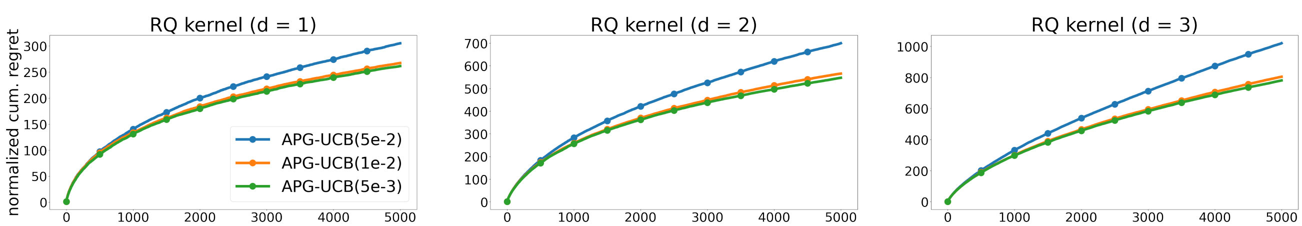

Figure 3. Normalized Cumulative Regret for RQ kernels with $l = 0.3\sqrt{d}$.

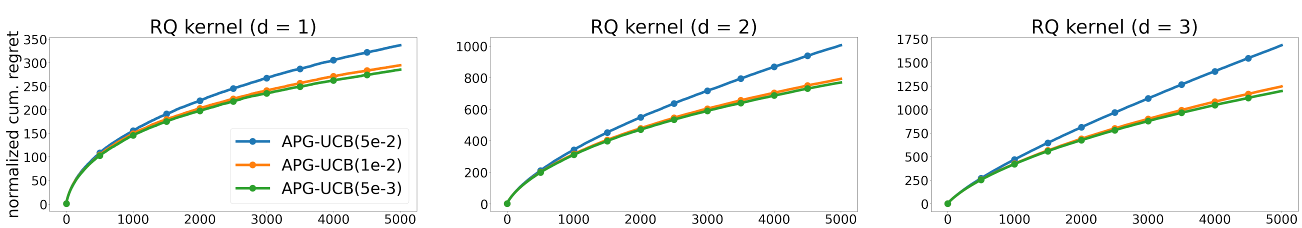

Figure 4. Normalized Cumulative Regret for RQ kernels with $l = 0.2\sqrt{d}$.

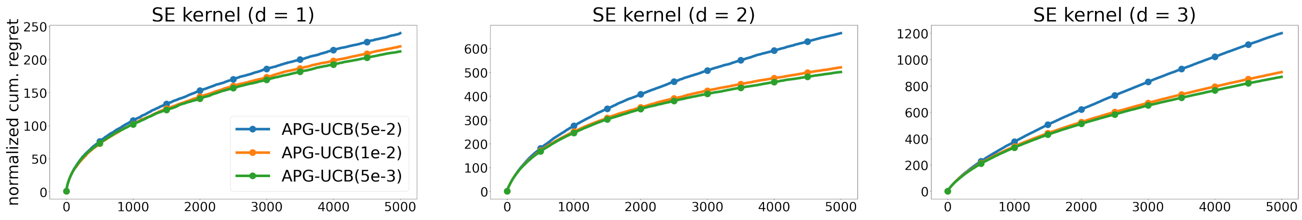

Figure 5. Normalized Cumulative Regret for SE kernels with $l = 0.2\sqrt{d}$.

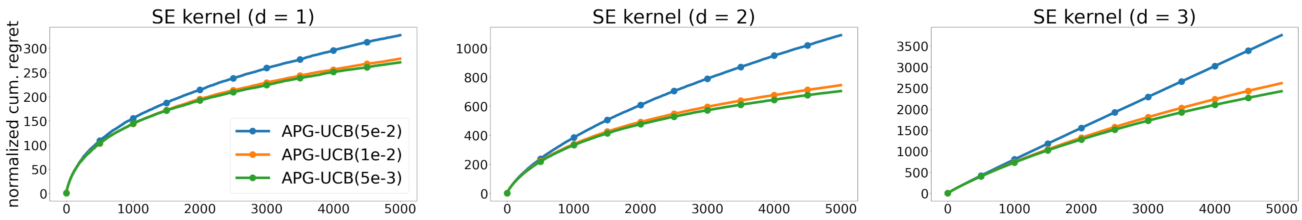

Figure 6. Normalized Cumulative Regret for SE kernels with $l = 0.1\sqrt{d}$.

Table 3. Total Running Time for RQ Kernels with $l = 0.3\sqrt{d}$.

	APG-UCB(5e-2)	APG-UCB(1e-2)	APG-UCB(5e-3)
d = 1 (RQ)	3.91e-01	4.06e-01	4.23e-01
d = 2 (RQ)	1.36e+00	2.39e+00	2.76e+00
d = 3 (RQ)	1.19e+01	2.40e+01	2.98e+01

Table 4. Total Running Time for SE Kernels with $l = 0.2\sqrt{d}$.

	APG-UCB(5e-2)	APG-UCB(1e-2)	APG-UCB(5e-3)
d = 1 (SE)	3.84e-01	4.04e-01	4.02e-01
d = 2 (SE)	1.69e+00	2.59e+00	2.89e+00
d = 3 (SE)	2.13e+01	3.51e+01	4.30e+01

Table 5. Total Running Time for RQ Kernels with $l = 0.2\sqrt{d}$.

	APG-UCB(5e-2)	APG-UCB(1e-2)	APG-UCB(5e-3)
d = 1 (RQ)	4.49e-01	4.84e-01	4.96e-01
d = 2 (RQ)	3.84e+00	6.01e+00	7.39e+00
d = 3 (RQ)	4.87e+01	8.76e+01	1.07e+02

Table 6. Total Running Time for SE Kernels with $l = 0.1\sqrt{d}$.

	APG-UCB(5e-2)	APG-UCB(1e-2)	APG-UCB(5e-3)
d = 1 (SE)	4.72e-01	4.88e-01	5.08e-01
d = 2 (SE)	9.59e+00	1.40e+01	1.61e+01
d = 3 (SE)	1.77e+02	2.02e+02	2.02e+02