Sharper Guarantees for Misspecified Kernelized Bandit Optimization

Davide Maran
Politecnico di Milano
davide.maran@polimi.it

Csaba Szepesvári
University of Alberta
szepesva@ualberta.ca

Abstract

Existing guarantees for misspecified kernelized bandit optimization pay for misspecification through kernel complexity: in generic offline bounds, the misspecification level ε is multiplied by $\sqrt{d_{\text{eff}}}$, where d_{eff} is the kernel effective dimension, while in online regret bounds, the corresponding penalty is $\sqrt{\gamma_n}\, n\varepsilon$, where γ_n is the maximum information gain after n rounds of interaction. In this work, we show that, for a large class of kernels, the misspecification amplification can be reduced to logarithmic or polylogarithmic growth. In the offline setting, we first prove high-probability simple-regret bounds whose misspecification term is governed by a spectral Lebesgue constant. This yields logarithmic amplification for one-dimensional monotone spectra and polylogarithmic amplification for multivariate Fourier-diagonal product kernels. In the online setting, we modify a domain-splitting algorithm and prove a cumulative regret bound of $\widetilde{\mathcal{O}}(\sqrt{\gamma_n n} + n\varepsilon)$ under mild localized eigendecay assumptions, removing the extra $\sqrt{\gamma_n}$ factor from the misspecification term. The common principle is localization: spectral localization controls the Lebesgue constant of the offline approximation operator, while domain splitting implements the spatial analogue of this mechanism in the online setting, preventing local misspecification errors from being amplified globally.

1 Introduction

In stochastic bandits and reinforcement learning, the learner must control errors pointwise to identify high-value actions; in contrast, supervised learning often succeeds with average-error guarantees. This difference makes misspecification substantially more harmful in these settings [Lattimore and Szepesvári, 2020]. The issue is not merely statistical noise but geometric amplification: even in linear approximation with adaptive data collection, pointwise error can scale as $\Theta(\sqrt{d}\,\varepsilon)$ [Du et al., 2020, Lattimore et al., 2020]. This may lead to a paradoxical situation where increasing model complexity actually hurts performance (even for infinite data, regardless of overfitting).

This phenomenon is not merely conceptual: it is formalized by lower- and upper-bound analyses in misspecified linear settings [Du et al., 2020, Lattimore et al., 2020]. More recently, [Maran and Szepesvári, 2025] identified the Lebesgue constant of the random-design projection operator as the exact geometric quantity governing the amplification of misspecification in least-squares regression, and showed that this dependence is intrinsic to squared-loss minimization. The natural next question is whether a comparable operator-theoretic picture can be developed for kernels, whose model class is infinite-dimensional and whose spectral structure matters. In the worst case, kernelized guarantees are still pessimistic [Bogunovic and Krause, 2021]. Our main message is that this worst-case behavior is often avoidable under additional spectral structure: in the offline setting, square-root amplification can be improved to logarithmic or polylogarithmic growth, depending on the kernel class.

Specifically, we investigate both offline optimization, where the agent operates on a fixed dataset of input queries, and the online setting, where the agent must adaptively interact with the environment

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: High-level comparison of leading misspecification-dependent terms under the assumptions listed in each row.

Setting	Assumptions	Baseline / prior guarantee	This work
Offline	i.i.d. spectral-basis setting; misspecification in $\|f - f_\star\|_\infty$	Generic bound $\Lambda(\mathsf{P}_\tau) \le \sqrt{d_{\text{eff}}(k\|\tau)}$ (Thm. 3.5, proved here), yielding $\sqrt{d_{\text{eff}}}\,\varepsilon$	1D monotone spectrum: $\log(e + \kappa/\tau)\,\varepsilon$; Fourier-diagonal product kernels on $[-1,1]^m$: $\log^{2m-1}(e + \kappa^m/\tau)\,\varepsilon$
Online	Adaptive kernelized bandits with subdomain eigendecay and bounded eigenfunctions	[Bogunovic and Krause, 2021]: $\widetilde{\mathcal{O}}(\gamma_n\sqrt{n} + \sqrt{\gamma_n}\,n\varepsilon)$	$\widetilde{\mathcal{O}}(\sqrt{\gamma_n n} + n\varepsilon)$; removes the extra $\sqrt{\gamma_n}$ factor from the misspecification term

to minimize regret. In the offline case, we prove logarithmic misspecification amplification in the one-dimensional monotone-spectrum setting and extend the analysis to multivariate product kernels, where the amplification remains polylogarithmic, for example of order $\log^{2m-1}(e + \kappa^m/\tau)$. In the online case, we demonstrate that under a localized eigendecay assumption—restricted to specific subdomains—the misspecification term in the regret can be reduced to $n\varepsilon$ up to logarithmic factors, removing the extra $\sqrt{\gamma_n}$ factor present in prior work. Both results instantiate the same high-level principle: misspecification becomes less harmful when the approximation problem can be localized. The two settings realize this principle differently:

- In the offline analysis, localization takes a spectral form. This is the kernel counterpart of the viewpoint of [Maran and Szepesvári, 2025]: the misspecification term is governed by the Lebesgue constant of the approximation operator. The new question in the kernel setting is when spectral structure forces this constant to be much smaller than the generic $\sqrt{d_{\text{eff}}}$ bound.
- In the bandit setting, localization takes a spatial form. We show that the domain-splitting algorithm π-GP-UCB proposed by [Janz et al., 2020] is sufficient for this purpose, albeit requiring a different bonus construction, refined hyperparameter selection, and a partially novel analysis. Conceptually, this is the online analogue of spatial-localization mechanisms that reduce global misspecification amplification: the algorithm fits local KRR estimators and controls information gain region by region.

For readability, we summarize the main quantitative comparison with the most relevant baseline guarantees in Table 1. The table is meant as a comparison of leading misspecification terms; stochastic and complexity terms are kept implicit. The detailed derivations are developed in Sections 3 and 4.

2 Problem Definition

Let $\mathbb{R}$ denote the set of reals, $\mathcal{X} \subseteq [-1,1]^m$ a measurable input space, and $f : \mathcal{X} \to \mathbb{R}$ an unknown measurable *target* (or regression) function that we wish to optimize by means of a kernel function $k : \mathcal{X}^2 \to \mathbb{R}$. In this paper, k will always satisfy the following.

Assumption 2.1. The kernel $k : \mathcal{X}^2 \to \mathbb{R}$ is positive semidefinite, stationary (i.e. there exists $\mathscr{K} : \mathcal{X} - \mathcal{X} \to \mathbb{R}$ such that $k(x,y) = \mathscr{K}(x-y)$ for all $x, y \in \mathcal{X}$), and satisfies $\sup_{u \in \mathcal{X}-\mathcal{X}} |\mathscr{K}(u)| \le \kappa$ for some $\kappa \ge 1$.

The agent interacts with f by querying points $\boldsymbol{x} = (x_t)_{t=1}^n \subset \mathcal{X}$ and receiving labels $\boldsymbol{y} = (y_t)_{t=1}^n \subset \mathbb{R}$ such that

$$f(x_t) \coloneqq \mathbb{E}[y_t | x_t]$$

almost surely for each $t \in [n] := \{1, \dots, n\}$. Regarding the choice of the query points, we study two different scenarios.

1. (Offline optimization) data $(\boldsymbol{x}, \boldsymbol{y}) = ((x_1, y_1), \dots, (x_n, y_n))$ are collected as independent pairs, with the inputs sampled from a common probability distribution $\mathcal{D}$ supported on $\mathcal{X}$ and the labels being R-subgaussian conditionally on the input.
2. (Online optimization) the agent sequentially queries points x_t, receiving a label y_t that is conditionally independent of the past and R-subgaussian given the input.

Most results in this paper build on approximating f via kernel ridge regression (KRR). For $\tau > 0$, the KRR estimator on $(\boldsymbol{x}, \boldsymbol{y})$ is given by

$$\mathsf{P}_{\boldsymbol{x},\tau}\boldsymbol{y}(x) := \underbrace{[k(x_1, x), \ldots, k(x_n, x)]}_{\Psi_{\boldsymbol{x}}(x)^\top}(K + n\tau I)^{-1}[y_1, \ldots, y_n]^\top, \tag{1}$$

where K is the kernel matrix such that $K_{ij} = k(x_i, x_j)$. In the offline part of the paper, we adopt the normalized regularization convention $K + n\tau I$ rather than $K + \lambda I$. This estimator is closely linked to the posterior variance of the process, which is given by

$$\sigma_{\boldsymbol{x},\tau}(x) := \sqrt{k(x, x) - \Psi_{\boldsymbol{x}}(x)^\top (K + n\tau I)^{-1}\Psi_{\boldsymbol{x}}(x)}.$$

The two parameterizations are equivalent: the regularization λ in the $K + \lambda I$ convention corresponds to $\tau = \lambda/n$ here, so in particular $\lambda = \Theta(1)$ there becomes $\tau = \Theta(1/n)$ here. By the Moore-Aronszajn Theorem [Aronszajn, 1950], there is a unique Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$ with the reproducing property: $\langle g, k(x, \cdot)\rangle_{\mathcal{H}} = g(x)$ for all $g \in \mathcal{H}$ and $x \in \mathcal{X}$. In this paper, we do not assume $f \in \mathcal{H}$. Instead, throughout the paper, we let $f_\star$ denote any function in $\mathcal{H}$ (e.g., a near-best sup-norm approximation of f), and aim to find an optimization procedure for f whose performance depends on $\mathcal{H}$, $\|f_\star\|_{\mathcal{H}}$ and, crucially,

$$\boxed{\varepsilon := \sup_{x\in\mathcal{X}} |f(x) - f_\star(x)|}.$$

In the rest of this work, regret is measured relative to a finite competitor set $\mathcal{X}^\circ \subset \mathcal{X}$, and we let $x_* \in \operatorname{argmax}_{x\in\mathcal{X}^\circ} f(x)$ (note that the algorithms may still query arbitrary points in $\mathcal{X}$). We use a finite benchmark to avoid imposing compactness or continuity assumptions only to guarantee an optimizer, and to keep the focus on misspecification amplification rather than routine discretization arguments. For compact continuous domains, one can take $\mathcal{X}^\circ$ to be a sufficiently fine net and add the usual approximation term between the true maximizer and the best point in the net.

Definition 2.2 (Simple regret)**.** For an algorithm $\mathscr{A}$, the simple regret under a budget of n tries is defined as

$$r_n^{\mathscr{A}} = f(x_*) - f(x_n).$$

This notion of regret applies to the first scenario (offline optimization). For the second interaction model, we use the following notion against the same competitor set.

Definition 2.3 (Online regret)**.** For an algorithm $\mathscr{A}$, the online regret over n interaction steps is defined as

$$R_n^{\mathscr{A}} = \sum_{t=1}^{n} f(x_*) - f(x_t).$$

When $\mathcal{X}^\circ$ is used as a fine discretization, it may be much larger than n; accordingly, our offline guarantees only scale logarithmically with $|\mathcal{X}^\circ|$ when a union bound over competitors is needed. The online confidence bounds used later are uniform over the queried regions, so the online regret bound does not pay such a factor explicitly. In the next section, we focus on the scenario of offline optimization.

3 Offline Optimization

In this section, we study offline optimization under the assumption that the query points come i.i.d. from the distribution $\mathcal{D}$. Our goal is to isolate, for KRR, the same kind of operator-theoretic mechanism that [Maran and Szepesvári, 2025] identified for random-design least squares: misspecification amplification is controlled by the $L^\infty \to L^\infty$ norm of the relevant approximation operator. Throughout this section, we assume that k admits a pointwise Mercer representation with measurable eigenfunctions $\phi_i : \mathcal{X} \to \mathbb{R}$ that are orthonormal in $L^2(\mathcal{X}; \mathcal{D})$, namely

$$k(x, y) = \sum_{i=1}^{\infty} \mu_i \phi_i(x)\phi_i(y). \tag{2}$$

The expansion holds for every $x, y \in \mathcal{X}$. The sequence (μ_i) gives the nonnegative eigenvalues of the integral operator induced by k; we call them the kernel eigenvalues and retain their given ordering,

including any zero entries. They allow us to define the effective dimension, which, for any $\tau > 0$, takes the following form

$$d_{\text{eff}}(k|\tau) := \sum_{i=1}^{\infty} \frac{\mu_i}{\tau + \mu_i}. \tag{3}$$

In the kernel setting, the role of the approximation operator is played by the following regularized *population operator*.

$$\mathsf{P}_\tau f(x) := \sum_{i=1}^{\infty} \frac{\mu_i f_i \phi_i(x)}{\tau + \mu_i} \qquad f_i := \int_{\mathcal{X}} \phi_i(z) f(z) \mathrm{d}\mathcal{D}(z). \tag{4}$$

The link between this population operator and the uniform error is captured by the following quantity, which is known in approximation theory as the *Lebesgue constant*.

$$\Lambda(\mathsf{P}_\tau) := \|\mathsf{P}_\tau\|_{\infty\to\infty} = \sup_{\|g\|_\infty > 0} \frac{\|\mathsf{P}_\tau g\|_\infty}{\|g\|_\infty}. \tag{5}$$

This is the kernel analogue of the projection-operator norm that appears in [Maran and Szepesvári, 2025], and it allows us to characterize the pointwise error of KRR. To quantify the difference between the empirical and population operators, define the population leverage function

$$\ell_\tau(x) := \sum_{i=1}^{\infty} \frac{\mu_i}{\mu_i + \tau} \phi_i(x)^2, \qquad L \geq \max\left\{1, \sup_{z\in\mathcal{X}} \ell_\tau(z)\right\},$$

where L is deterministic. Such an L is finite because $\ell_\tau(z) \leq k(z,z)/\tau \leq \kappa/\tau$. If $\sup_{i,z} |\phi_i(z)| \leq C_\phi$, then $\ell_\tau(z) \leq C_\phi^2 d_{\text{eff}}(k|\tau)$, so one may take $L = \max\{1, C_\phi^2 d_{\text{eff}}(k|\tau)\}$.

Theorem 3.1. *Under Assumption 2.1 and the pointwise Mercer representation (2), let $(x_t, y_t)_{t=1}^n$ be independent offline observations with $x_t \sim \mathcal{D}$ and conditionally mean-zero, R-subgaussian noise $y_t - f(x_t)$. Fix $\tau > 0$, $\delta \in (0,1)$, and $f_\star \in \mathcal{H}$ independently of the observations, and set*

$$\varepsilon := \|f - f_\star\|_\infty < \infty, \qquad u := \log \frac{64L}{\delta}.$$

Assume that $n \geq 64Lu$. Then, for every fixed $x \in \mathcal{X}$, with probability at least $1-\delta$,

$$|\mathsf{P}_{\boldsymbol{x},\tau} \boldsymbol{y}(x) - f(x)| \leq R\sigma_{\boldsymbol{x},\tau}(x)\sqrt{\frac{2u}{n\tau}} + \sigma_{\boldsymbol{x},\tau}(x)\|f_\star\|_{\mathcal{H}} + (1 + \Lambda(\mathsf{P}_\tau) + o_{n,\tau}(x,\delta))\,\varepsilon,$$

where

$$o_{n,\tau}(x,\delta) := 3\,(1 + \Lambda(\mathsf{P}_\tau))\sqrt{\frac{\ell_\tau(x)u}{n}} + 24\frac{L\sqrt{\ell_\tau(x)}\,u}{n}.$$

On the same event, $\sigma_{\boldsymbol{x},\tau}(x)^2 \leq 2\tau\ell_\tau(x)$.

The three terms in the previous theorem have standard interpretations in nonparametric statistics. First, $R\sigma_{\boldsymbol{x},\tau}(x)\sqrt{2u/(n\tau)}$ captures stochastic uncertainty due to the label noise and scales with the subgaussian parameter and the failure probability. Second, $\sigma_{\boldsymbol{x},\tau}(x)\|f_\star\|_{\mathcal{H}}$ represents target complexity through the RKHS norm of $f_\star$. Lastly, the misspecification term consists of a population contribution controlled by $1 + \Lambda(\mathsf{P}_\tau)$ and the explicit empirical remainder $o_{n,\tau}(x,\delta)$. In this sense, Theorem 3.1 is the KRR counterpart of the random-design least-squares picture of [Maran and Szepesvári, 2025]; the rest of the section asks when kernel structure makes this amplification substantially smaller than the generic $\sqrt{d_{\text{eff}}}$ scale. From this pointwise error bound, a result for the simple regret immediately follows.

Corollary 3.2. *Under the assumptions of Theorem 3.1, let $\mathcal{X}^\circ \subseteq \mathcal{X}$ be finite, nonempty, and fixed independently of the observations. Consider the plug-in maximization algorithm $\mathscr{A}$ that returns*

$$x_n \in \operatorname*{argmax}_{x\in\mathcal{X}^\circ} \mathsf{P}_{\boldsymbol{x},\tau} \boldsymbol{y}(x).$$

Set $u^\circ := \log(64L|\mathcal{X}^\circ|/\delta)$ and assume

$$n \geq \max\left\{64Lu^\circ, \frac{24L^{3/2}u^\circ}{1 + \Lambda(\mathsf{P}_\tau)}\right\}.$$

Then, with probability at least $1-\delta$,

$$r_n^{\mathscr{A}} \lesssim \left(R\sqrt{\frac{u^\circ}{n\tau}} + \|f_\star\|_{\mathcal{H}}\right) \max_{x\in\mathcal{X}^\circ} \sigma_{\boldsymbol{x},\tau}(x) + (1 + \Lambda(\mathsf{P}_\tau))\,\varepsilon.$$

Proof. Apply Theorem 3.1 at each point of $\mathcal{X}^\circ$ with failure probability $\delta/|\mathcal{X}^\circ|$, and take a union bound. The sample-size conditions and $\ell_\tau(x) \le L$ give, simultaneously on this set,

$$1 + \Lambda(\mathsf{P}_\tau) + o_{n,\tau}\left(x, \frac{\delta}{|\mathcal{X}^\circ|}\right) \le (1 + \Lambda(\mathsf{P}_\tau)) \left(1 + 3\sqrt{\frac{Lu^\circ}{n}}\right) + \frac{24L^{3/2}u^\circ}{n}$$
$$\le 3\,(1 + \Lambda(\mathsf{P}_\tau))\,.$$

Since x_n maximizes the estimate over $\mathcal{X}^\circ$,

$$r_n^{\mathscr{A}} \le 2 \max_{x \in \mathcal{X}^\circ} |\mathsf{P}_{\boldsymbol{x},\tau}\boldsymbol{y}(x) - f(x)|.$$

Substituting the uniform error estimate proves the claim. □

In the previous result, as in the rest of the paper, we use the notation $a_n \lesssim b_n$ to indicate that there is a universal constant c such that $a_n \le cb_n$ for all n. We write $a_n \lesssim_m b_n$ when the implicit constant is allowed to depend on the input dimension m. Under the displayed sample-size conditions, apart from the stochastic and epistemic uncertainty terms, the error is governed by $(1 + \Lambda(\mathsf{P}_\tau))\,\varepsilon$. Along choices $\tau = \tau_n$ for which the statistical terms vanish, this determines the remaining regret guarantee. The remainder of this section studies the Lebesgue constant and relates it to standard complexity measures.

We use the following standard spectral size measure.

Definition 3.3 (D1: Polynomial effective dimension)**.** A kernel is said to have polynomial effective dimension of degree $s > 0$ if, for some constant $\mathrm{C}_2 > 0$ and every $\tau > 0$,

$$d_{\text{eff}}(k|\tau) \le \mathrm{C}_2 \tau^{-1/s}.$$

Another important assumption that will often be invoked in our analysis is the following.

Definition 3.4 (D2: Polynomial eigendecay)**.** A kernel satisfies polynomial eigendecay of degree $s > 0$ if, for some constant $\mathrm{C}_1 > 0$,

$$\mu_i \le \mathrm{C}_1 i^{-s}.$$

For the same s, one has D2 $\implies$ D1. A key feature of D2 that distinguishes it from D1 is that it is *not* invariant under reordering of the eigenfunctions. While the effective dimension treats all eigenvalues in the same way, polynomial eigendecay imposes a bound that explicitly depends on the current index i. When the eigenfunctions are interchangeable, one can sort them by eigenvalue, but when they possess an intrinsic ordering, as for Fourier features, this distinction becomes crucial.

3.1 Bounding the Lebesgue constant

Previous analyses of misspecified linear/kernelized optimization rely on reducing the error amplification to the maximal leverage/posterior variance (see Proposition 4.5 in [Lattimore et al., 2020] and Lemma 2 in [Bogunovic and Krause, 2021]). Their approach gives an upper bound on the Lebesgue constant, which in the most favorable case reduces to the following.

Theorem 3.5. *If the eigenfunctions are bounded by one, then*

$$\Lambda(\mathsf{P}_\tau) \le \sqrt{d_{\mathit{eff}}(k|\tau)}.$$

The proof is given in Section B. The posterior-variance estimate in Theorem 3.1 also yields a bound directly in terms of the effective dimension. Suppose that $\sup_{i,z} |\phi_i(z)| \le 1$, and set$L := 1 \vee d_{\text{eff}}(k|\tau), u := \log \frac{64L}{\delta}$. If $n \ge \max\left\{64Lu, \frac{24L^{3/2}u}{1+\Lambda(\mathsf{P}_\tau)}\right\}$, then, for every fixed $x \in \mathcal{X}$, with probability at least $1 - \delta$,

$$|\mathsf{P}_{\boldsymbol{x},\tau}\boldsymbol{y}(x) - f(x)| \le 2R\sqrt{\frac{d_{\text{eff}}(k|\tau)u}{n}} + \|f_\star\|_{\mathcal{H}}\sqrt{2\tau d_{\text{eff}}(k|\tau)} + 3\,(1 + \Lambda(\mathsf{P}_\tau))\,\varepsilon. \tag{6}$$

The simpler sufficient condition $n \ge 64L^{3/2}u$ recovers the sample-size scaling $n = \widetilde{\Omega}((1 \vee d_{\text{eff}}(k|\tau))^{3/2})$. Equation (6) holds for every $\tau > 0$ satisfying the displayed sample-size conditions;

its proof is given in Section B. Repeating the strategy of Corollary 3.2 and using Theorem 3.5, when $\tau \asymp 1/n$ and the preceding sample-size conditions hold with u replaced by $u^\circ = \log(64L|\mathcal{X}^\circ|/\delta)$, we obtain

$$r_n^{\mathscr{A}} \lesssim (R + \|f_\star\|_{\mathcal{H}}) \sqrt{\frac{d_{\text{eff}}(k|\tau)u^\circ}{n}} + \left(1 + \sqrt{d_{\text{eff}}(k|\tau)}\right)\varepsilon.$$

While this bound has the advantage of being very general, it is not satisfactory for many applications. In the case **D1** of a kernel with polynomial effective dimension, the previous result only shows that $\Lambda(\mathsf{P}_\tau) \lesssim \tau^{-1/2s}$. Thus, if τ is driven to zero polynomially fast with n, the misspecification amplification still grows polynomially. In the rest of this section, we bound the Lebesgue constant directly rather than through the posterior variance, which gives tighter rates under mild assumptions.

Lebesgue constant under generalized Bochner-type assumptions The results in this subsection are inspired by Bochner's theorem, which ensures that the eigenfunctions of periodic stationary kernels correspond to the Fourier basis. The argument relies on spectral control of the Mercer basis rather than on periodicity itself. We isolate this condition first.

Assumption 3.6 (Spectral Lebesgue growth)**.** In the one-dimensional case, define

$$B_i := \sup_{x \in \mathcal{X}} \left\| \sum_{j=1}^{i} \phi_j(x)\phi_j(\cdot) \right\|_{L^1(\mathcal{D})}.$$

We say that the kernel has logarithmic spectral Lebesgue growth if $B_i \lesssim \log(e+i)$ for all $i \geq 1$.

This assumption is the compact-domain analogue of the Dirichlet-kernel estimate used below. It is weaker than requiring periodicity or an exact Fourier basis, but it is also not automatic for an arbitrary restriction of a stationary Euclidean kernel to a bounded domain. One canonical setting where Assumption 3.6 holds is the periodic case.

Theorem 3.7 (Fourier diagonalization on the torus [Seeger, 2004])**.** *Let $k(x,y) = \mathscr{K}(x-y)$ be a kernel satisfying Assumption 2.1. If $\mathcal{X} = [-1,1]^m$, $\mathcal{D}$ is the uniform measure over the domain and the profile $\mathscr{K}$ is periodic, then the eigenfunctions ϕ_i in equation (2) correspond to the Fourier basis.*

In this periodic setting, the partial sums in Assumption 3.6 are Dirichlet kernels, so $B_i \lesssim \log(e+i)$. Standard stationary kernels on $\mathbb{R}^m$ also admit Fourier representations by the Euclidean form of Bochner's theorem, but after restriction to a compact domain their Mercer eigenfunctions need not be Fourier modes. Thus ordinary squared-exponential, Matérn, Wendland, or spectral-mixture kernels do not satisfy Assumption 3.6 merely by stationarity. Periodized versions of such kernels provide a direct ML-facing surrogate. Given a stationary kernel $k(x,y) = \mathscr{K}(x-y)$ on $\mathbb{R}^m$ and a period $2L$, one can define $\mathscr{K}_L^{\text{per}}(t) := \sum_{\ell \in (2L\mathbb{Z})^m} \mathscr{K}(t+\ell)$, whenever the sum is absolutely convergent. The resulting kernel is positive semidefinite on the torus, has Fourier eigenfunctions, and therefore satisfies the spectral-basis condition above. On a fixed domain such as $[-1,1]^m$, it differs from the original kernel only through the wrap-around tails

$$\sup_{x,y \in [-1,1]^m} |k_L^{\text{per}}(x,y) - k(x,y)| \leq \sum_{\ell \in (2L\mathbb{Z})^m \setminus \{0\}} \sup_{t \in [-2,2]^m} |\mathscr{K}(t+\ell)|.$$

For kernels whose correlations decay with distance, this term becomes small when L is large relative to the domain and the kernel length scale. In this sense, periodization buys exact Fourier diagonalization without substantially changing the kernel on the region where data are observed.

To improve readability, we first state the result of this section for $m = 1$, postponing the product-kernel multivariate case to the end of the section. Let $\{a_i\}_{i=1}^\infty$ be a bounded sequence of real numbers. We define the first finite difference operator as $\Delta^1 a_i := a_{i+1} - a_i$, which serves as a discrete analogue of the first derivative. The Lebesgue constant of the KRR operator is closely linked to the ℓ^1 norm of the discrete derivative of its sequence of eigenvalues, as the following theorem shows.

Theorem 3.8. *Define, for every integer $i \geq 1$, $h_i = \frac{\mu_i}{\tau + \mu_i}$. Then, $\Lambda(\mathsf{P}_\tau) \leq h_1 B_1 + \sum_{i=1}^\infty \left|\Delta^1 h_i\right| B_i$. In particular, under logarithmic spectral Lebesgue growth,*

$$\Lambda(\mathsf{P}_\tau) \lesssim \sum_{i=1}^{\infty} \left|\Delta^1 h_i\right| \log(e+i).$$

The proof is given in Section B. Theorem 3.8 shows that if the sequence h_i is sufficiently regular, then the Lebesgue constant is small. This only depends on the sequence of eigenvalues, since the regularization parameter τ does not depend on i. In particular, kernels with monotone spectra, meaning that higher-order Fourier features have smaller eigenvalues, satisfy a simple logarithmic bound.

Corollary 3.9. *Suppose that the kernel has logarithmic spectral Lebesgue growth and non-increasing eigenvalues. Then,*

$$\Lambda(\mathsf{P}_\tau) \lesssim \log(e + \kappa/\tau).$$

The proof is given in Section B. For periodic Matérn kernels, whose Fourier eigenvalues are proportional to $(1 + i^2)^{-\nu-1/2}$ (equation 4.14 in [Seeger, 2004]), the previous corollary is enough. Nonetheless, for most classes, including Wendland and mixture kernels, the eigenvalue decay is not monotonic (see Section A).

While Corollary 3.9 does not apply to kernels with non-monotone eigenvalue sequences, it is natural to ask whether one can reduce to Theorem 3.8 by slightly changing the spectrum of k without significantly altering the corresponding RKHS $\mathcal{H}$. This motivates the following.

> **Question.** Equation (6) and Theorem 3.8 imply that the KRR error depends on
>
> $$\|f\|_{\mathcal{H}}, \qquad d_{\text{eff}}(k|\tau), \qquad \sum_{i=1}^{\infty} |\Delta^1 h_i| \, B_i.$$
>
> Is it possible, under **D2** or **D1**, to modify the spectrum of k (obtaining another kernel $\overline{k}$) so as to reduce the third quantity without increasing the other two?

First, we show a positive answer under **D2**.

Theorem 3.10. *Assume that k satisfies polynomial eigendecay **D2** for some $s > 0$. Then, there is a kernel $\overline{k}$ with the same Mercer eigenfunctions and non-increasing eigenvalues that also satisfies **D2** with the same exponent s and such that, for any $f \in \mathcal{H}$,*

$$\|f\|_{\overline{\mathcal{H}}} \leq \|f\|_{\mathcal{H}}.$$

The proof, given in Section B, relies on a simple envelope construction: set $\overline{\mu}_i := \max_{j \geq i} \mu_j$ while keeping the same eigenfunctions, which makes the sequence non-increasing. Assuming **D2** is crucial in this context, as it extends easily to the new sequence of eigenvalues $\overline{\mu}_i$. As noted above, since **D2** $\implies$ **D1**, the previous result also entails that the new kernel $\overline{k}$ has polynomial effective dimension with the same exponent s as the original kernel. If the original Mercer basis has logarithmic spectral Lebesgue growth, then the same is true for $\overline{k}$, because the basis has not changed. Therefore, Theorem 3.10 answers one part of the previous question.

Is it possible to control the effective dimension of $\overline{k}$ by assuming only a polynomial bound on the effective dimension of k? The next result gives a partial negative answer: **D1** alone is not enough to guarantee a strictly better asymptotic rate for the smoothness term. This obstruction applies not only to our simple smoothing technique, but to any candidate kernel $\overline{k}$ that preserves the Mercer basis.

Theorem 3.11. *For every $s > 1$, there is a Fourier-diagonal kernel k satisfying **D1** of degree s such that*

$$\limsup_{\tau \to 0} \frac{\Lambda(\mathsf{P}_\tau)}{\sqrt{d_{\text{eff}}(k|\tau)}} \gtrsim 1$$

and there is no kernel $\overline{k}$ with the same Mercer eigenfunctions which satisfies all the following three conditions (i) *Norm equivalence: for some constant $C > 0$, for every $f \in \mathcal{H}$, $\|f\|_{\overline{\mathcal{H}}} \leq C\|f\|_{\mathcal{H}}$* (ii) $\sum_{i=1}^{\infty} \left|\Delta^1 \overline{h}_i\right| \lesssim (1/\tau)^{1/s'}$ *for some $s' > s$* (iii) $d_{\text{eff}}(\overline{k}|\tau)$ *is polynomial in $1/\tau$ (i.e. **D1**).*

The proof is given in Section B. The previous result shows that there exist kernels that are intrinsically ill-conditioned with respect to the Lebesgue constant. The kernel k constructed above satisfies $\Lambda(\mathsf{P}_\tau) \asymp \sqrt{d_{\text{eff}}(k|\tau)}$, i.e., the worst-case behavior permitted by Theorem 3.5, and one cannot in general find a same-basis kernel $\overline{k}$ that both preserves a comparable RKHS norm and polynomial effective dimension and also yields a strictly better asymptotic rate for $\sum_{i=1}^{\infty} \left|\Delta^1 \overline{h}_i\right|$. In particular, **D1** alone does not force logarithmic or polylogarithmic control of this quantity.

Multivariate product kernels. The preceding one-dimensional result has a multivariate analogue for Fourier-diagonal product kernels. As discussed above, Fourier diagonalization can be enforced by periodizing stationary kernels, with only a wrap-around perturbation on the observed domain when correlations decay. To state the result, let k_0 be a one-dimensional kernel on $[-1,1]$, and define the product kernel k_Π on $[-1,1]^m$ by

$$k_\Pi(x,y) := \prod_{j=1}^{m} k_0(x_j, y_j). \tag{7}$$

Theorem 3.12. *Let $k_0(x,y) = \mathscr{K}_0(x-y)$ be a one-dimensional kernel on $[-1,1]$ satisfying Assumption 2.1. Assume that the Mercer eigenfunctions of k_0 are the Fourier basis and that its eigenvalues are non-increasing. Then the population KRR operator associated with the product kernel k_Π in (7) satisfies*

$$\Lambda(\mathsf{P}_\tau) \lesssim_m \log^{2m-1}(e + \kappa^m/\tau).$$

The proof, given in Section B.2, follows from a multivariate version of Theorem 3.8: the one-dimensional finite difference is replaced by a mixed finite difference over the coordinates of the Mercer multi-index. This is still only polylogarithmic in the regularization scale. The dependence of the logarithmic exponent on m is the expected fixed-dimensional cost of working with product spectra. It comes from bounding the total absolute mixed variation of the product multiplier near the hyperbolic surface $\prod_j \mu_{i_j} \simeq \tau$.

4 Online Optimization

In this section, we turn to the online problem. Here, the sample sequence $(x_t)_{t=1}^n$ is not fixed in advance, but is chosen sequentially in order to optimize f. This setting focuses on the online regret objective (see Definition 2.3). The analysis of the previous section does not apply directly, as the exploration-exploitation dilemma requires visiting a highly skewed distribution of points, whereas the offline Lebesgue-constant analysis is tied to a fixed input distribution and its associated Mercer basis. To remain consistent with the kernel-bandit literature cited below, in this section we switch back to the standard unnormalized regularization convention, so the regional KRRs use $K_t^A + \lambda I$ (equivalently $V_t^A = \lambda I + \sum_{s \le t} \phi(x_s)\phi(x_s)^\top$).

The problem of misspecification in this setting was extensively studied by [Bogunovic and Krause, 2021], who prove an upper bound for the regret of their algorithm EC-GP-UCB of the form

$$R_n^{\text{EC-GP-UCB}} \le \widetilde{\mathcal{O}}(\gamma_n\sqrt{n} + \varepsilon\sqrt{\gamma_n}n).$$

Amplifying the misspecification with $\sqrt{\gamma_n}$ is particularly harmful for many kernels, such as the Matérn family with parameters $m \gg \nu$, where γ_n can be nearly linear in n. In such a case, only $\varepsilon \le n^{-1/2}$ would guarantee a non-vacuous bound. On the other hand, the result cannot be improved in full generality. In fact, kernelized bandits generalize linear bandits of dimension $d \approx \gamma_n$, where the regret may grow as $\varepsilon\sqrt{d}n$, as [Lattimore et al., 2020] shows. Section C.3 discusses a related obstruction for nested finite-dimensional linear approximations, where vanishing approximation error does not by itself rule out floors for feature-based bandit algorithms. We therefore impose additional structural conditions on the kernel, natural ones that have appeared before in this literature. Under sup-norm misspecification, an $n\varepsilon$ contribution is in general inevitable; the main goal is to avoid additional multiplicative factors in front of this term.

Definition 4.1 (**D2+**: Polynomial eigendecay on subdomains)**.** (Definition 1 in [Vakili and Olkhovskaya, 2023]) A kernel satisfies polynomial eigendecay on subdomains of degree $(\alpha, s) > 0$ for some constant $\mathrm{C}_3 > 0$ if, when restricted to a hypercube of side $\rho \le 1$, its eigenvalues satisfy $\mu_i \le \mathrm{C}_3\rho^\alpha i^{-s}$.

This definition strengthens the polynomial eigendecay condition **D2** by ensuring that the eigenvalues decrease whenever one restricts the domain of the kernel.

Definition 4.2 (**D3**: Bounded eigenfunctions on subdomains)**.** (cf. Lemma 2 of [Vakili and Olkhovskaya, 2023]) A kernel satisfies bounded eigenfunctions on subdomains for some constant $\mathrm{C}_4 > 0$ if, when restricted to a hypercube of side $\rho \le 1$, its eigenfunctions satisfy $\sup_i \|\phi_i\|_\infty \le \mathrm{C}_4$.

Algorithm 1 π-MISSPEC-GPUCB

procedure π-MISSPEC-GPUCB$(n, b, \lambda, B, \varepsilon, \delta)$
  Start with $\mathcal{A}_1 \leftarrow \{[-1,1]^m\}$
  for $t = 1, \dots, n$ **do**
    $A_t, x_t \leftarrow \arg\max_{A \in \mathcal{A}_t,\, x \in A} \mathrm{UCB}_t(A, x)$ in (8)
    Pull x_t and receive y_t
    if $|N_t^{A_t}| \geq \rho_{A_t}^{-b}$ **then**
      Split A_t into 2^m sub-regions $\mathcal{A}_t^\circ$
      $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t \cup \mathcal{A}_t^\circ \setminus \{A_t\}$
      Fit KRR estimator $\widehat{\mu}_t^A$ on all $A \in \mathcal{A}_{t+1} \setminus \mathcal{A}_t$
    else
      $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t$
      Update KRR estimator $\widehat{\mu}_t^{A_t}$
    end if
  end for
end procedure

Assuming bounded eigenfunctions is standard in the literature; see, for instance, [Salgia et al., 2023, Chatterji et al., 2019, Vakili et al., 2021, Riutort-Mayol et al., 2023, Liu and Li, 2023]. This assumption is satisfied for all kernels that admit the Fourier basis as eigenvectors, including most of those examined in the previous section. It is well known that Matérn kernels enjoy **D2+** with $s = \frac{2\nu+m}{m}$ and $\alpha = 2\nu$; this fact is central to the analysis in [Yang et al., 2020, Vakili and Olkhovskaya, 2023]. [Vakili et al., 2021] claims also that **D3** holds for Matérn kernels; while several papers use this result, the recent work [Janz et al., 2026] questions its validity (see their note 1) highlighting a potential gap in their argument (while at the same time time providing a proof for the polynomial eigendecay not passing trough bounded eigenfunctions).

4.1 Domain splitting approach

In the random-design linear setting, [Maran and Szepesvári, 2025] makes the localization principle explicit through the Lebesgue constant of the projection operator. The same principle must be implemented differently in the online setting, because the offline Lebesgue-constant analysis does not extend directly to adaptive sampling. Instead, we exploit domain splitting to keep the number of samples per region under control. This makes the misspecification contribution in the regional confidence bounds depend on the local sample size, while subdomain eigendecay keeps the corresponding regional information gain small. A similar idea was already developed, for purposes unrelated to misspecification, by [Janz et al., 2020]. Their algorithm π-GP-UCB relies on a dyadic splitting of each region whenever the number of data points it contains exceeds a given threshold. This creates 2^m new regions, on each of which KRR is fit anew. While they work only with Matérn kernels, this idea remains valid under our assumptions.

At each time step t, Algorithm 1 maintains a collection of regions $\mathcal{A}_t$ partitioning the space $\mathcal{X} = [-1,1]^m$. It then fits a separate KRR on each region $A \in \mathcal{A}_t$, with mean estimate $\widehat{\mu}_{t-1}^A$ and posterior variance $\sigma_{t-1}^A(x)$. The algorithm then maximizes the following upper-confidence bound jointly over the region and the point in the region:

$$\mathrm{UCB}_t(A, x) = \widehat{\mu}_{t-1}^A(x) + \beta_t^A \sigma_{t-1}^A(x). \tag{8}$$

The previous equation contains an exploration bonus that, for an upper bound B on the function norm $\|f_\star\|_{\mathcal{H}}$ and the maximal information gain $\gamma_{N_{t-1}^A}$ associated with the points available in the region before round t, takes the form

$$\beta_t^A := \frac{R}{\lambda^{1/2}} \sqrt{bm \log(4t/\delta) + 1 + \gamma_{N_{t-1}^A}} + B + \varepsilon \sqrt{\frac{N_{t-1}^A}{\lambda}}. \tag{9}$$

Here, b is a parameter that determines how many points are needed to split a region; together with λ it will be specified in the regret bound. The main feature of this upper bound relative to other works on kernelized bandits is the presence of the term $\varepsilon\sqrt{N_{t-1}^A/\lambda}$, which inflates the confidence width proportionally to the accumulated misspecification in region A. With this exploration bonus, we obtain the following regret bound.

Theorem 4.3. *Consider the online model of Definition 2.3 on* $[-1,1]^m$. *Under Assumption 2.1, let* $f_\star \in \mathcal{H}$ *satisfy* $\|f_\star\|_{\mathcal{H}} \leq B$ *and* $\|f - f_\star\|_\infty \leq \varepsilon$, *and suppose the kernel satisfies* **D2+** *and* **D3**. *Then running Algorithm 1 with parameters* $\lambda = 1$ *and* $b = \alpha$ *yields, with probability at least* $1-\delta$, *a regret bound of order*

$$R_n^{\mathscr{A}} \leq \widetilde{\mathcal{O}}\left(n^{\frac{2m+\alpha}{2m+2\alpha}} + n\varepsilon\right).$$

The proof is given in Section C. Under the assumptions of the theorem, we have in general $\gamma_n = \widetilde{\Theta}(n^{\frac{m}{m+\alpha}})$. Therefore, the regret bound can equivalently be written as $\widetilde{\mathcal{O}}(\sqrt{\gamma_n n}+n\varepsilon)$, which improves over the bound $\gamma_n\sqrt{n} + \sqrt{\gamma_n}n\varepsilon$ from [Bogunovic and Krause, 2021]. This is particularly relevant for the Matérn kernel, where tight bounds for maximal information gain give

$$\gamma_n = \widetilde{\Theta}(n^{\frac{m}{m+2\nu}}) = \widetilde{\Theta}(n^{\frac{m}{m+\alpha}}),$$

so that our theorem recovers the optimal well-specified regret rate $\widetilde{\Theta}(n^{\frac{m+\nu}{m+2\nu}})$, plus the misspecification term. For this kernel, the choice $b = \alpha = 2\nu$ does not correspond to the one by [Janz et al., 2020], which takes $b = \frac{m+2\nu}{m+1}$. This reflects a different way to bound the information gain.

5 Conclusions

We studied misspecified kernelized optimization through a common localization perspective. In the offline setting, localization acts through the approximation operator and yields logarithmic misspecification amplification under monotone spectra and polylogarithmic amplification for multivariate product kernels, improving on the generic square-root control. The key technical ingredient is an explicit control of $\Lambda(\mathsf{P}_\tau)$ via discrete spectral smoothness. This yields sharp bounds under monotone spectra, and a constructive monotone-envelope reduction covers the polynomial-eigendecay case.

In the online setting, localization acts through the query space. We analyzed a domain-splitting UCB strategy and derived regret of order $\widetilde{\mathcal{O}}(\sqrt{\gamma_n n} + n\varepsilon)$ under subdomain eigendecay assumptions. Relative to prior misspecified kernel-bandit bounds, this removes an additional $\sqrt{\gamma_n}$ penalty from the misspecification-dependent term.

Finally, the lower-bound result shows that polynomial effective dimension alone does not imply logarithmic or polylogarithmic control. Identifying the minimal structural assumptions sufficient for sharper online guarantees remains open.

References

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Ilija Bogunovic and Andreas Krause. Misspecified Gaussian process bandit optimization. *Advances in neural information processing systems*, 34:3004–3015, 2021.

Romain Camilleri, Kevin Jamieson, and Julian Katz-Samuels. High-dimensional experimental design and kernel bandits. In *International Conference on Machine Learning*, pages 1227–1237. PMLR, 2021.

Niladri Chatterji, Aldo Pacchiano, and Peter Bartlett. Online learning with kernel losses. In *International Conference on Machine Learning*, pages 971–980. PMLR, 2019.

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

Jialin Dong and Lin Yang. Does sparsity help in learning misspecified linear bandits? In *International Conference on Machine Learning*, pages 8317–8333. PMLR, 2023.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2020.

Abdelhamid Ezzerg, Ilija Bogunovic, and Jeremias Knoblauch. Robust bayesian optimisation with unbounded corruptions. *arXiv preprint arXiv:2511.15315*, 2025.

Gerald B Folland. *Fourier analysis and its applications*, volume 4. American Mathematical Soc., 2009.

P. G. Guest. The spacing of observations in polynomial regression. *The Annals of Mathematical Statistics*, 29(1):294–299, 1958.

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9.1–9.24. JMLR Workshop and Conference Proceedings, 2012.

David Janz, David Burt, and Javier González. Bandit optimisation of functions in the Matérn kernel RKHS. In *International Conference on Artificial Intelligence and Statistics*, pages 2486–2495. PMLR, 2020.

David Janz, Arya Akhavan, and Alexandre B Tsybakov. Maximum effective dimension and information gain. *arXiv preprint arXiv:2608.24450*, 2026.

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Tor Lattimore, Csaba Szepesvári, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International conference on machine learning*, pages 5662–5670. PMLR, 2020.

Zihan Li and Jonathan Scarlett. Gaussian process bandit optimization with few batches. In *International Conference on Artificial Intelligence and Statistics*, pages 92–107. PMLR, 2022.

Zejian Liu and Meng Li. On the estimation of derivatives using plug-in kernel ridge regression estimators. *Journal of Machine Learning Research*, 24(266):1–37, 2023.

Davide Maran and Csaba Szepesvári. Beyond least squares: Uniform approximation and the hidden cost of misspecification. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Davide Maran, Alberto Maria Metelli, Matteo Papini, and Marcello Restelli. Local linearity: the key for no-regret reinforcement learning in continuous MDPs. *arXiv preprint arXiv:2410.24071*, 2024.

Gabriel Parra and Felipe Tobar. Spectral mixture kernels for multi-output gaussian processes. *Advances in Neural Information Processing Systems*, 30, 2017.

Gabriel Riutort-Mayol, Paul-Christian Bürkner, Michael R Andersen, Arno Solin, and Aki Vehtari. Practical Hilbert space approximate Bayesian Gaussian processes for probabilistic programming. *Statistics and Computing*, 33(1):17, 2023.

Sudeep Salgia, Sattar Vakili, and Qing Zhao. A domain-shrinking based bayesian optimization algorithm with order-optimal regret performance. *Advances in Neural Information Processing Systems*, 34:28836–28847, 2021.

Sudeep Salgia, Sattar Vakili, and Qing Zhao. Random exploration in bayesian optimization: Order-optimal regret and computational efficiency. *arXiv preprint arXiv:2310.15351*, 2023.

Matthias Seeger. Gaussian processes for machine learning. *International journal of neural systems*, 14(02):69–106, 2004.

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

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

Sattar Vakili and Julia Olkhovskaya. Kernelized reinforcement learning with order optimal regret bounds. *Advances in Neural Information Processing Systems*, 36:4225–4247, 2023.

Sattar Vakili, Kia Khezeli, and Victor Picheny. On information gain and regret bounds in Gaussian process bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 82–90. PMLR, 2021.

Michal Valko, Nathaniel Korda, Rémi Munos, Ilias Flaounas, and Nelo Cristianini. Finite-time analysis of kernelised contextual bandits. *arXiv preprint arXiv:1309.6869*, 2013.

Holger Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in computational Mathematics*, 4(1):389–396, 1995.

Justin Whitehouse, Aaditya Ramdas, and Steven Z Wu. On the sublinear regret of GP-UCB. *Advances in Neural Information Processing Systems*, 36:35266–35276, 2023.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael Jordan. Provably efficient reinforcement learning with kernel and neural function approximations. *Advances in Neural Information Processing Systems*, 33:13903–13916, 2020.

A Related Work

Misspecification amplification As noted in the introduction, controlling misspecification amplification is a well-studied problem in sequential decision-making under uncertainty, including bandits and reinforcement learning, for algorithms based on linear and richer function approximation models. For linear function approximation, [Du et al., 2020] established the first $\sqrt{d}$ amplification lower bound, which was improved by [Lattimore et al., 2020], who also established an upper bound of $\sqrt{d}$ based on optimal design. These lower bounds hold for a worst-case feature map, while *allowing the learner to choose the data distribution*. Several approaches to circumvent this bound have been explored in different contexts. [Maran et al., 2024] shows how to remove it in the case of a locally linear feature map; [Dong and Yang, 2023] improves the $\sqrt{d}$ amplification under sparsity. Very recently, [Maran and Szepesvári, 2025] showed that, in misspecified linear regression under random design, the Lebesgue constant of the projection operator is the exact quantity governing amplification of the uniform approximation error, and that this dependence is intrinsic for empirical squared-loss minimizers. Our Theorem 3.1 is the kernel KRR analogue of that perspective, while the rest of Section 3 studies when kernel spectral structure forces the corresponding Lebesgue constant to be only logarithmic or polylogarithmic.

In the broader context of kernel methods, the work most closely related to ours is [Bogunovic and Krause, 2021], which proves a regret upper bound for misspecified kernelized bandits. Their result provides the worst-case bound that serves as the starting point of our analysis. A related paper that studies kernelized bandits beyond the realizability assumption is [Ezzerg et al., 2025], which focuses on corrupted observations — a distinct robustness setting in which only weaker guarantees can be achieved.

Kernel methods in sequential decision-making Kernelized bandits build on the GP-UCB and kernelized UCB lines of work [Srinivas et al., 2010, Chowdhury and Gopalan, 2017, Valko et al., 2013], where regret is typically controlled through posterior variance or information gain. From the viewpoint of [Maran and Szepesvári, 2025], the main challenge here is to recover a comparable localization mechanism under adaptive sampling. While several algorithms achieving order-optimal regret bounds are now known [Salgia et al., 2021, Camilleri et al., 2021, Li and Scarlett, 2022], only the domain-splitting approach of [Janz et al., 2020] appears well-suited to addressing misspecification amplification as well. The reason is structural: our analysis needs a mechanism that localizes the data and keeps the number of samples per region under control, so that both the misspecification contribution in the confidence widths and the regional information gain can be controlled locally. The strategy of [Janz et al., 2020] provides exactly this ingredient. By contrast, other order-optimal approaches treat the dependence induced by GP-UCB-style sampling more globally [Whitehouse et al., 2023], and do not offer an equally direct route to the same localized misspecification control.

Stationary kernels with non-monotone spectrum While standard analyses of kernelized methods typically assume eigenvalue decay, Section 3.1 shows that the Lebesgue constant also depends on eigenvalue *smoothness*. While the common class of Matérn kernels has a monotonically decreasing spectrum, this property does not hold in general. Wendland kernels [Wendland, 1995] provide compactly supported, positive definite radial functions with strong locality and sparsity properties, and are widely used in approximation theory and large-scale settings. Their compact support induces an oscillatory spectral density, leading to non-monotone eigenvalue behavior[1]. Another class with this feature is given by spectral mixture kernels [Parra and Tobar, 2017], which explicitly model the spectrum as a mixture, enabling flexible, multimodal frequency representations.

B Proofs of Section 3

Theorem 3.1. *Under Assumption 2.1 and the pointwise Mercer representation (2), let $(x_t, y_t)_{t=1}^n$ be independent offline observations with $x_t \sim \mathcal{D}$ and conditionally mean-zero, R-subgaussian noise $y_t - f(x_t)$. Fix $\tau > 0$, $\delta \in (0,1)$, and $f_\star \in \mathcal{H}$ independently of the observations, and set*

$$\varepsilon := \|f - f_\star\|_\infty < \infty, \qquad u := \log \frac{64L}{\delta}.$$

Assume that $n \geq 64Lu$. Then, for every fixed $x \in \mathcal{X}$, with probability at least $1-\delta$,

$$|\mathsf{P}_{\boldsymbol{x},\tau}\boldsymbol{y}(x) - f(x)| \leq R\sigma_{\boldsymbol{x},\tau}(x)\sqrt{\frac{2u}{n\tau}} + \sigma_{\boldsymbol{x},\tau}(x)\|f_\star\|_{\mathcal{H}} + (1 + \Lambda(\mathsf{P}_\tau) + o_{n,\tau}(x,\delta))\,\varepsilon,$$

where

$$o_{n,\tau}(x,\delta) := 3\,(1+\Lambda(\mathsf{P}_\tau))\sqrt{\frac{\ell_\tau(x)u}{n}} + 24\frac{L\sqrt{\ell_\tau(x)}\,u}{n}.$$

On the same event, $\sigma_{\boldsymbol{x},\tau}(x)^2 \leq 2\tau\ell_\tau(x)$.

Proof. Fix $x \in \mathcal{X}$ independently of the observations, and abbreviate

$$B := \|f_\star\|_{\mathcal{H}}, \qquad \Lambda := \Lambda(\mathsf{P}_\tau), \qquad \alpha := \sqrt{\frac{Lu}{n}} \leq \frac{1}{8}.$$

All unlabelled inner products and norms below are those of the feature Hilbert space, and $X \sim \mathcal{D}$ denotes a generic design point.

Feature representation and posterior variance. Define

$$\psi(z) := (\sqrt{\mu_j}\phi_j(z))_{j\geq 1}, \qquad \Sigma := \mathbb{E}[\psi(X)\otimes\psi(X)] = \mathrm{diag}(\mu_j),$$
$$A := \Sigma + \tau I, \qquad \widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^n \psi(x_i)\otimes\psi(x_i), \qquad A_n := \widehat{\Sigma} + \tau I.$$

The pointwise Mercer representation makes ψ a feature map for k. Its measurable coordinates and $\|\psi(z)\|^2 = k(z,z) \leq \kappa$ ensure that the Hilbert-space expectations below are well-defined. For a bounded deterministic function g, write

$$\widehat{\mathsf{P}}_\tau g(z) := \left\langle \psi(z), A_n^{-1}\frac{1}{n}\sum_{i=1}^n \psi(x_i)g(x_i)\right\rangle.$$

This is the empirical KRR prediction on the labels $(g(x_i))_{i=1}^n$. The population counterpart has coefficient $A^{-1}\mathbb{E}[\psi(X)g(X)]$. The Woodbury identity gives

$$\sigma_{\boldsymbol{x},\tau}(x)^2 = \tau\langle\psi(x), A_n^{-1}\psi(x)\rangle.$$

Let

$$a(z) := A^{-1/2}\psi(z), \qquad S := \mathbb{E}[a(X)\otimes a(X)] = A^{-1/2}\Sigma A^{-1/2}.$$

[1] Page 5 of [Wendland, 1995] provides an explicit expression for the Fourier transform of the kernel in terms of an integral involving Bessel functions, whose oscillatory nature implies that the spectrum is not monotone.

Then

$$\|a(z)\|^2 = \ell_\tau(z), \qquad 0 \preceq S \preceq I, \qquad \operatorname{tr}(S) = d_{\text{eff}}(k|\tau) = \mathbb{E}[\ell_\tau(X)] \le L.$$

For the equivalent kernel $G_\tau(z,t) := \langle a(z), a(t)\rangle$, we have

$$\mathsf{P}_\tau g(z) = \mathbb{E}[G_\tau(z,X)g(X)],$$

and

$$\mathbb{E}[G_\tau(z,X)^2] = \langle a(z), Sa(z)\rangle \le \ell_\tau(z), \qquad |G_\tau(z,t)| \le \sqrt{L\ell_\tau(z)}.$$

Consequently, since $\mathcal{D}$ is a probability measure,

$$\Lambda \le \sup_{z\in\mathcal{X}} \mathbb{E}|G_\tau(z,X)| \le \sqrt{L}.$$

Concentration of the empirical inverse. Set

$$\Delta := A^{-1/2}(\widehat{\Sigma} - \Sigma)A^{-1/2}, \qquad M := (I+\Delta)^{-1} = A^{1/2}A_n^{-1}A^{1/2}.$$

For $Z_i := a(x_i)\otimes a(x_i) - S$, we have $\Delta = n^{-1}\sum_{i=1}^n Z_i$ and

$$\|Z_i\|_{\text{op}} \le L, \qquad \mathbb{E}[Z_i^2] = \mathbb{E}[\|a(X)\|^2 a(X)\otimes a(X)] - S^2 \preceq LS, \qquad \operatorname{tr}\mathbb{E}[Z_i^2] \le L^2.$$

The matrix Bernstein inequality of [Hsu et al., 2012, Lemma 32], applied to both signs with $r = v = L$, intrinsic-dimension parameter L, and tail parameter $2u$, yields an event $\mathcal{E}_{\text{cov}}$ with failure probability at most $2Le^{-u}$ on which

$$\|\Delta\|_{\text{op}} \le 2\sqrt{\frac{Lu}{n}} + \frac{2Lu}{3n} = 2\alpha + \frac{2}{3}\alpha^2 =: \eta \le \frac{25}{12}\alpha < \frac{1}{2}.$$

The same bound holds in the countable feature space by applying the inequality to finite-dimensional compressions and passing to the limit; the displayed estimates are independent of the compression dimension. On $\mathcal{E}_{\text{cov}}$,

$$\|M\|_{\text{op}} \le 2, \qquad \|M - I\|_{\text{op}} \le \frac{\eta}{1-\eta} \le \frac{25}{6}\alpha.$$

In particular,

$$\sigma_{\boldsymbol{x},\tau}(x)^2 = \tau\langle a(x), Ma(x)\rangle \le 2\tau\ell_\tau(x).$$

Population residual and exact decomposition. Set

$$h := f - f_\star, \qquad p := \mathsf{P}_\tau h, \qquad r := h - p, \qquad \beta_\tau := A^{-1}\mathbb{E}[\psi(X)h(X)].$$

Then $p(z) = \langle \psi(z), \beta_\tau\rangle$. The function p minimizes $\mathbb{E}[(h(X) - g(X))^2] + \tau\|g\|_{\mathcal{H}}^2$ over $g \in \mathcal{H}$. Comparing its objective with that of $g = 0$ gives

$$\|r\|_{L^2(\mathcal{D})}^2 + \tau\|p\|_{\mathcal{H}}^2 \le \|h\|_{L^2(\mathcal{D})}^2 \le \varepsilon^2, \qquad \|r\|_\infty \le (1+\Lambda)\varepsilon.$$

The population normal equation also gives

$$\mathbb{E}[\psi(X)r(X)] = \tau\beta_\tau.$$

Define

$$w := \frac{1}{n}\sum_{i=1}^n a(x_i)r(x_i) - \mathbb{E}[a(X)r(X)].$$

Subtracting the population coefficient from the empirical ridge coefficient and using the preceding normal equation yields

$$\begin{aligned} A_n^{-1}\frac{1}{n}\sum_{i=1}^n \psi(x_i)h(x_i) - \beta_\tau &= A_n^{-1}\left(\frac{1}{n}\sum_{i=1}^n \psi(x_i)r(x_i) - \tau\beta_\tau\right) \\ &= A_n^{-1}A^{1/2}w. \end{aligned}$$

Therefore the following decomposition is exact:

$$\widehat{\mathsf{P}}_\tau h(x) - \mathsf{P}_\tau h(x) = \langle a(x), Mw\rangle = \langle a(x), w\rangle + \langle a(x), (M-I)w\rangle.$$

Scalar fluctuation and empirical-inverse remainder. For $U_i := G_\tau(x, x_i)r(x_i)$,

$$\langle a(x), w\rangle = \frac{1}{n}\sum_{i=1}^{n}(U_i - \mathbb{E}U_i),$$

where

$$\mathrm{Var}(U_i) \le (1+\Lambda)^2\varepsilon^2\ell_\tau(x), \qquad |U_i - \mathbb{E}U_i| \le 2(1+\Lambda)\varepsilon\sqrt{L\ell_\tau(x)}.$$

Scalar Bernstein's inequality gives, outside an event of probability at most $2e^{-u}$,

$$\begin{aligned}|\langle a(x), w\rangle| &\le (1+\Lambda)\varepsilon\left(\sqrt{\frac{2\ell_\tau(x)u}{n}} + \frac{4\sqrt{L\ell_\tau(x)}\,u}{3n}\right)\\ &\le 3(1+\Lambda)\varepsilon\sqrt{\frac{\ell_\tau(x)u}{n}},\end{aligned}$$

where the last inequality uses $\alpha \le 1/8$.

For the centered vectors $W_i := a(x_i)r(x_i) - \mathbb{E}[a(X)r(X)]$, the residual bounds imply

$$\mathbb{E}\|W_i\|^2 \le \mathbb{E}[\ell_\tau(X)r(X)^2] \le L\|r\|^2_{L^2(\mathcal{D})} \le L\varepsilon^2, \qquad \|W_i\| \le 2\sqrt{L}(1+\Lambda)\varepsilon.$$

The vector Bernstein inequality of [Hsu et al., 2012, Lemma 31] therefore gives, with failure probability at most e^{-u},

$$\begin{aligned}\|w\| &\le \varepsilon\sqrt{\frac{L}{n}}(1+\sqrt{8u}) + \frac{8\sqrt{L}(1+\Lambda)\varepsilon u}{3n}\\ &= \varepsilon\alpha\left(\frac{1}{\sqrt{u}} + \sqrt{8} + \frac{8}{3}\frac{1+\Lambda}{\sqrt{L}}\alpha\right) \le 5\varepsilon\alpha.\end{aligned}$$

Here we used $u \ge 1$, $L \ge 1$, $\alpha \le 1/8$, and $1+\Lambda \le 2\sqrt{L}$. Intersecting with $\mathcal{E}_{\mathrm{cov}}$ gives

$$\begin{aligned}|\langle a(x), (M-I)w\rangle| &\le \sqrt{\ell_\tau(x)}\|M-I\|_{\mathrm{op}}\|w\|\\ &\le \frac{125}{6}\varepsilon\sqrt{\ell_\tau(x)}\alpha^2 \le 24\varepsilon\frac{L\sqrt{\ell_\tau(x)}\,u}{n}.\end{aligned}$$

This estimate does not require independence between M and w. The use of $\|r\|_{L^2(\mathcal{D})} \le \varepsilon$ in the vector variance bound avoids an additional factor $1+\Lambda$ in this remainder. Combining the two fluctuation estimates with $|\mathsf{P}_\tau h(x) - h(x)| \le (1+\Lambda)\varepsilon$ yields

$$|\widehat{\mathsf{P}}_\tau h(x) - h(x)| \le \big(1 + \Lambda + o_{n,\tau}(x,\delta)\big)\varepsilon.$$

RKHS bias, observation noise, and conclusion. Let $\beta_\star$ be the feature coefficient of $f_\star$, so that $f_\star(z) = \langle\psi(z), \beta_\star\rangle$ and $\|\beta_\star\| = B$. The empirical ridge bias is

$$\widehat{\mathsf{P}}_\tau f_\star(x) - f_\star(x) = -\tau\langle\psi(x), A_n^{-1}\beta_\star\rangle.$$

Since $A_n \succeq \tau I$, spectral calculus gives $A_n^{-2} \preceq \tau^{-1}A_n^{-1}$, and hence

$$|\widehat{\mathsf{P}}_\tau f_\star(x) - f_\star(x)| \le \tau B\|A_n^{-1}\psi(x)\| \le B\sigma_{\boldsymbol{x},\tau}(x).$$

Write $\xi_i := y_i - f(x_i)$ and

$$N(x) := \left\langle \psi(x), A_n^{-1}\frac{1}{n}\sum_{i=1}^{n}\psi(x_i)\xi_i\right\rangle.$$

Conditional on the design, the variables ξ_i are independent and R-subgaussian. Thus $N(x)$ is mean-zero subgaussian with variance proxy at most

$$\begin{aligned}\frac{R^2}{n}\langle\psi(x), A_n^{-1}\widehat{\Sigma}A_n^{-1}\psi(x)\rangle &\le \frac{R^2}{n}\langle\psi(x), A_n^{-1}\psi(x)\rangle\\ &= \frac{R^2\sigma_{\boldsymbol{x},\tau}(x)^2}{n\tau}.\end{aligned}$$

The inequality uses the quadratic-form comparison

$$A_n^{-1}\widehat{\Sigma}A_n^{-1} = A_n^{-1} - \tau A_n^{-2} \preceq A_n^{-1}.$$

Consequently, with conditional failure probability at most $2e^{-u}$,

$$|N(x)| \le R\sigma_{\boldsymbol{x},\tau}(x)\sqrt{\frac{2u}{n\tau}}.$$

Finally, by linearity,

$$\mathsf{P}_{\boldsymbol{x},\tau}\boldsymbol{y}(x) - f(x) = \widehat{\mathsf{P}}_\tau f_\star(x) - f_\star(x) + N(x) + \widehat{\mathsf{P}}_\tau h(x) - h(x).$$

Combining the preceding bounds proves the claimed error estimate. All required events, including the posterior-variance estimate, hold jointly outside probability at most

$$2Le^{-u} + 2e^{-u} + e^{-u} + 2e^{-u} = (2L+5)e^{-u} = \frac{2L+5}{64L}\delta < \delta.$$

□

B.1 Proofs of Section 3.1

Proof of equation (6). Under $\sup_{i,z}|\phi_i(z)| \le 1$, we have $\ell_\tau(x) \le d_{\text{eff}}(k|\tau) \le L$. The two sample-size conditions imply

$$\begin{aligned}1 + \Lambda(\mathsf{P}_\tau) + o_{n,\tau}(x,\delta) &\le \big(1+\Lambda(\mathsf{P}_\tau)\big)\left(1 + 3\sqrt{\frac{Lu}{n}}\right) + \frac{24L^{3/2}u}{n}\\ &\le 3\big(1+\Lambda(\mathsf{P}_\tau)\big).\end{aligned}$$

On the event of Theorem 3.1, $\sigma_{\boldsymbol{x},\tau}(x)^2 \le 2\tau\ell_\tau(x)$, so

$$R\sigma_{\boldsymbol{x},\tau}(x)\sqrt{\frac{2u}{n\tau}} \le 2R\sqrt{\frac{d_{\text{eff}}(k|\tau)u}{n}}, \qquad \sigma_{\boldsymbol{x},\tau}(x)\|f_\star\|_{\mathcal{H}} \le \|f_\star\|_{\mathcal{H}}\sqrt{2\tau d_{\text{eff}}(k|\tau)}.$$

Substitution into Theorem 3.1 proves the result. □

Theorem 3.5. *If the eigenfunctions are bounded by one, then*

$$\Lambda(\mathsf{P}_\tau) \le \sqrt{d_{\mathit{eff}}(k|\tau)}.$$

Proof. For a function f, let

$$f_i \coloneqq \int_{\mathcal{X}} f(x)\phi_i(x)\,\mathrm{d}\mathcal{D}(x).$$

The Lebesgue constant corresponds to

$$\begin{aligned}\sup_{\|f\|_\infty=1, x\in\mathcal{X}} |\mathsf{P}_\tau f(x)| &= \sup_{\|f\|_\infty=1, x\in\mathcal{X}} \left|\sum_{i=1}^\infty \frac{\mu_i f_i \phi_i(x)}{\tau+\mu_i}\right|\\ &\le \sup_{\|f\|_{L^\infty}=1, x\in\mathcal{X}} \sqrt{\sum_{i=1}^\infty \frac{\mu_i^2\phi_i(x)^2}{(\tau+\mu_i)^2}}\sqrt{\sum_{i=1}^\infty f_i^2}\\ &\le \sqrt{\sum_{i=1}^\infty \frac{\mu_i^2}{(\tau+\mu_i)^2}}\\ &\le \sqrt{\sum_{i=1}^\infty \frac{\mu_i}{(\tau+\mu_i)}} = \sqrt{d_{\text{eff}}(k|\tau)},\end{aligned}$$

where the first inequality follows from Cauchy-Schwarz together with $|\phi_i(x)| \le 1$, and the second uses $\sum_{i=1}^\infty f_i^2 = \|f\|_{L^2}^2 \le \|f\|_{L^\infty}^2 = 1$. □

Lemma B.1. *Let k be a kernel with Mercer decomposition (2). Then,*

$$\Lambda(\mathsf{P}_\tau) = \sup_{x\in\mathcal{X}} \left\| \sum_{i=1}^{\infty} \frac{\mu_i \phi_i(x)}{\tau+\mu_i} \phi_i(\cdot) \right\|_{L^1}.$$

Proof. By definition,

$$\begin{aligned}
\Lambda(\mathsf{P}_\tau) &= \sup_{x\in\mathcal{X}} \sup_{\|f\|_\infty=1} \left| \sum_{i=1}^{\infty} \frac{\mu_i \phi_i(x)}{\tau+\mu_i} \int_{\mathcal{X}} \phi_i(y) f(y)\, \mathrm{d}\mathcal{D}(y) \right| \\
&= \sup_{x\in\mathcal{X}} \sup_{\|f\|_\infty=1} \left| \int_{\mathcal{X}} \sum_{i=1}^{\infty} \frac{\mu_i \phi_i(x)}{\tau+\mu_i} \phi_i(y) f(y)\, \mathrm{d}\mathcal{D}(y) \right| \\
&= \sup_{x\in\mathcal{X}} \left\| \sum_{i=1}^{\infty} \frac{\mu_i \phi_i(x)}{\tau+\mu_i} \phi_i(\cdot) \right\|_{1},
\end{aligned}$$

where the last equality uses the duality between L^1 and L^∞. □

Theorem 3.8. *Define, for every integer $i \geq 1$, $h_i = \frac{\mu_i}{\tau+\mu_i}$. Then, $\Lambda(\mathsf{P}_\tau) \leq h_1 B_1 + \sum_{i=1}^{\infty} \left|\Delta^1 h_i\right| B_i$. In particular, under logarithmic spectral Lebesgue growth,*

$$\Lambda(\mathsf{P}_\tau) \lesssim \sum_{i=1}^{\infty} \left|\Delta^1 h_i\right| \log(e+i).$$

Proof. By Lemma B.1, the Lebesgue constant of the population KRR can be written as follows

$$\Lambda(\mathsf{P}_\tau) = \sup_{x\in\mathcal{X}} \left\| \sum_{i=1}^{\infty} h_i \phi_i(x) \phi_i(\cdot) \right\|_{L^1}.$$

For each fixed x, let

$$S_i^x(\cdot) := \sum_{j=1}^{i} \phi_j(x) \phi_j(\cdot).$$

Abel summation gives

$$\sum_{i=1}^{\infty} h_i \phi_i(x) \phi_i(\cdot) = -\sum_{i=1}^{\infty} \Delta^1 h_i S_i^x(\cdot),$$

where the boundary term vanishes because $h_i \to 0$. Therefore, taking L^1 norms gives

$$\Lambda(\mathsf{P}_\tau) \leq \sum_{i=1}^{\infty} \left|\Delta^1 h_i\right| B_i \leq h_1 B_1 + \sum_{i=1}^{\infty} \left|\Delta^1 h_i\right| B_i,$$

which completes the proof. □

Lemma B.2. *Let k be a kernel satisfying Assumption 2.1 such that its eigenvalues are non-increasing. Then it satisfies polynomial eigendecay* **D2** *with $C_1 = \kappa$ and $s = 1$.*

Proof. By monotonicity, for any fixed $i \in \mathbb{N}$,

$$\mu_i \leq i^{-1} \sum_{j=1}^{i} \mu_j \leq i^{-1} \sum_{j=1}^{\infty} \mu_j = \int_{\mathcal{X}} k(x,x) d\mathcal{D}(x) \leq \frac{\kappa}{i},$$

where the trace equality uses Assumption 2.1. □

Corollary 3.9. *Suppose that the kernel has logarithmic spectral Lebesgue growth and non-increasing eigenvalues. Then,*

$$\Lambda(\mathsf{P}_\tau) \lesssim \log(e + \kappa/\tau).$$

Proof. Since the kernel has logarithmic spectral Lebesgue growth, there is a constant C_B such that $B_i \le C_B \log(e+i)$ for all i. By Lemma B.2, the non-increasing spectrum satisfies $\mu_i \le \kappa/i$. More generally, if $\mu_i \le C_1 i^{-s}$, Theorem 3.8 and monotonicity give

$$\begin{aligned}\Lambda(\mathsf{P}_\tau) &\le C_B \sum_{i=1}^{\infty} \left|\Delta^1 h_i\right| \log(e+i) + C_B \\ &\lesssim C_B \left(1 + \sum_{i=1}^{\infty} \frac{h_i}{i}\right).\end{aligned}$$

Now, take $i^* := \lfloor (C_1/\tau)^{1/s} \rfloor + 2$. The previous sum splits into two terms. First, as $h_i \le 1$,

$$\sum_{i=1}^{i^*} \frac{h_i}{i} \le \log(i^*) \lesssim_s \log(e + C_1/\tau).$$

Let us focus on the second one. For $i \ge i^*$,

$$\sum_{i=i^*}^{\infty} \frac{h_i}{i} \le \frac{1}{\tau} \sum_{i=i^*}^{\infty} \frac{\mu_i}{i} \le \frac{C_1}{\tau} \sum_{i=i^*}^{\infty} i^{-s-1} \le \frac{C_1}{s\tau (i^*-1)^s} \le \frac{1}{s}.$$

The claim follows with $C_1 = \kappa$ and $s = 1$. This completes the proof. □

Theorem 3.10. *Assume that k satisfies polynomial eigendecay D2 for some $s > 0$. Then, there is a kernel $\overline{k}$ with the same Mercer eigenfunctions and non-increasing eigenvalues that also satisfies D2 with the same exponent s and such that, for any $f \in \mathcal{H}$,*

$$\|f\|_{\overline{\mathcal{H}}} \le \|f\|_{\mathcal{H}}.$$

Proof. Let $\overline{k}$ be the kernel with the same Mercer eigenfunctions as k and spectrum given as follows, $\forall i \in \mathbb{N}$, $\overline{\mu}_i := \max_{j \ge i} \mu_j$. Then,

- For any $f \in \mathcal{H}$,
$$\|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{\infty} \frac{f_i^2}{\mu_i} \ge \sum_{i=1}^{\infty} \frac{f_i^2}{\overline{\mu}_i} = \|f\|_{\overline{\mathcal{H}}}^2.$$
- By D2, for any i,
$$\overline{\mu}_i = \max_{j \ge i} \mu_j \le \max_{j \ge i} \mathrm{C}_1 j^{-s} \le \mathrm{C}_1 i^{-s}.$$

□

Theorem 3.11. *For every $s > 1$, there is a Fourier-diagonal kernel k satisfying D1 of degree s such that*

$$\limsup_{\tau \to 0} \frac{\Lambda(\mathsf{P}_\tau)}{\sqrt{d_{\mathit{eff}}(k|\tau)}} \gtrsim 1$$

and there is no kernel $\overline{k}$ with the same Mercer eigenfunctions which satisfies all the following three conditions (i) *Norm equivalence: for some constant $C > 0$, for every $f \in \mathcal{H}$, $\|f\|_{\overline{\mathcal{H}}} \le C\|f\|_{\mathcal{H}}$* (ii) $\sum_{i=1}^{\infty} \left|\Delta^1 \overline{h}_i\right| \lesssim (1/\tau)^{1/s'}$ *for some $s' > s$* (iii) $d_{\mathit{eff}}(\overline{k}|\tau)$ *is polynomial in $1/\tau$ (i.e. D1).*

Proof. The proof is divided into two well-separated parts. First, we build the kernel k; then we prove the desired trade-off.

Part 1: building the kernel

Work on $\mathcal{X} = [-1, 1]$ with the uniform measure, and fix any ordering of the real Fourier basis. The basis is orthonormal in $L^2(\mathcal{D})$ and uniformly bounded in L^∞. Let $\{U_i\}_{i=1}^{\infty}$ be an infinite sequence of i.i.d. random variables with distribution $\mathrm{Bern}(1/2)$. Let k be a random kernel with eigenvalues

$$\mu_i = \begin{cases} \log_2(i)^{-s} U_i & \log_2(i) \in \mathbb{N} \\ \log_2(i+1)^{-s} U_i & \log_2(i+1) \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases}. \tag{10}$$

Since $s > 1$, the eigenvalue sequence is summable almost surely; hence the preceding display defines a bounded positive semidefinite Fourier-diagonal kernel. By Lemma B.1, if

$$G_\tau(x,y) \coloneqq \sum_{i=1}^{\infty} \frac{\mu_i}{\tau+\mu_i}\phi_i(x)\phi_i(y),$$

then $\Lambda(\mathsf{P}_\tau) \geq \mathbb{E}_X \mathbb{E}_Y |G_\tau(X,Y)|$, where X, Y are independent and uniformly distributed on $\mathcal{X}$. Now, fix $x, y \in \mathcal{X}$. The value of $G_\tau(x,y)$ is a sum of independent bounded random variables, so

$$\begin{aligned} V_\tau(x,y) &\coloneqq \operatorname{Var}\left(G_\tau(x,y)\right) \\ &= \sum_{i=1}^{\infty} \operatorname{Var}\left(\frac{\mu_i}{\tau+\mu_i}\right)\phi_i^2(x)\phi_i^2(y). \end{aligned}$$

By the fact that $G_\tau(x,y)$ is a sum of independent bounded variables, there is a constant such that $\mathbb{E}[|G_\tau(x,y) - \mathbb{E}\, G_\tau(x,y)|] \geq c\sqrt{V_\tau(x,y)}$. Therefore, $\mathbb{E}[|G_\tau(x,y)|] \gtrsim \sqrt{V_\tau(x,y)}$. The uniform boundedness and orthonormality of the Fourier basis imply

$$\mathop{\mathbb{E}}_{X,Y} V_\tau(X,Y) \asymp \sum_{\ell=1}^{\infty} \frac{\ell^{-2s}}{(\ell^{-s}+\tau)^2}, \qquad \sup_{x,y} V_\tau(x,y) \lesssim \sum_{\ell=1}^{\infty} \frac{\ell^{-2s}}{(\ell^{-s}+\tau)^2}.$$

Hence $\mathbb{E}_{X,Y}\sqrt{V_\tau(X,Y)} \geq \mathbb{E}_{X,Y} V_\tau(X,Y)/\sqrt{\sup_{x,y} V_\tau(x,y)} \gtrsim \sqrt{\sum_{\ell=1}^{\infty} \frac{\ell^{-2s}}{(\ell^{-s}+\tau)^2}}$. This formula allows us to estimate the expected Lebesgue constant of this random operator:

$$\begin{aligned} \mathbb{E}[\Lambda(\mathsf{P}_\tau)] &\geq \mathbb{E}\left[\mathop{\mathbb{E}}_{X,Y} |G_\tau(X,Y)|\right] \\ &\gtrsim \sqrt{\sum_{\ell=1}^{\infty} \frac{\ell^{-2s}}{(\ell^{-s}+\tau)^2}} \\ &\geq \sqrt{\int_1^{\infty} \frac{1}{(1+\tau x^s)^2} dx} \\ &= \tau^{-\frac{1}{2s}} \sqrt{\int_{\tau^{1/s}}^{\infty} \frac{1}{(1+t^s)^2} \mathrm{d}t} \gtrsim \tau^{-\frac{1}{2s}}. \end{aligned}$$

At the same time, Theorem 3.5 ensures that

$$\Lambda(\mathsf{P}_\tau) \leq \sqrt{d_{\text{eff}}(k|\tau)}.$$

The effective dimension of this kernel can be upper bounded as follows.

$$\begin{aligned} d_{\text{eff}}(k|\tau) &= \sum_{i=1}^{\infty} \frac{\mu_i}{\tau+\mu_i} \\ &\leq \sum_{i=1}^{\infty} 1(\log_2(i) \in \mathbb{N}) \frac{\log_2(i)^{-s}}{\log_2(i)^{-s}+\tau} \\ &\quad + \sum_{i=1}^{\infty} 1(\log_2(i+1) \in \mathbb{N}) \frac{\log_2(i+1)^{-s}}{\log_2(i+1)^{-s}+\tau} \\ &\leq 2\sum_{\ell=1}^{\infty} \frac{\ell^{-s}}{\ell^{-s}+\tau} \lesssim \tau^{-\frac{1}{s}}. \end{aligned}$$

As there are positive constants, say c_1 and c_2, such that

$$\mathbb{E}[\Lambda(\mathsf{P}_\tau)] \geq c_1 \tau^{-\frac{1}{2s}} \qquad \Lambda(\mathsf{P}_\tau) \leq c_2 \tau^{-\frac{1}{2s}} \ a.s.,$$

this means that, for $\tau_n = 1/n$,

$$\mathbb{P}(E_n) \coloneqq \mathbb{P}\left(\Lambda(\mathsf{P}_{\tau_n}) \geq \frac{c_1}{2}\tau_n^{-\frac{1}{2s}}\right) \geq \frac{c_1}{2c_2 - c_1} =: c_3.$$

By continuity from above,

$$\mathbb{P}\left(\bigcap_{n\in\mathbb{N}}\bigcup_{\ell=n}^{\infty} E_\ell\right) = \lim_n \mathbb{P}\left(\bigcup_{\ell=n}^{\infty} E_\ell\right) \geq \lim_n c_3 = c_3.$$

This shows that $E_{\text{lim}} \coloneqq \bigcap_{n\in\mathbb{N}}\bigcup_{\ell=n}^{\infty} E_\ell$ has positive probability over the realizations of the sequence (10). By the strong law for stationary 1-dependent sequences, we also know that the set of realizations E_{LLN} such that

$$E_{\text{LLN}} \coloneqq \left\{\lim_{m\to\infty} \frac{\sum_{j=1}^m U_{2^j}\cdot U_{2^{j+1}}}{m} = 1/4\right\}, \tag{11}$$

satisfies $\mathbb{P}(E_{\text{LLN}}) = 1$. This identity will be used in the second part of the proof. The set $E^* \coloneqq E_{\text{LLN}} \cap E_{\text{lim}}$ therefore has positive probability. We choose the kernel k as any element of E^*. By definition of E^*, $\Lambda(\mathsf{P}_{\tau_n}) \gtrsim \tau_n^{-1/(2s)}$ for some subsequence of $\{\tau_n\}_{n\in\mathbb{N}}$. This proves that, for the kernel k constructed above,

$$\limsup_{\tau\to 0} \frac{\Lambda(\mathsf{P}_\tau)}{\sqrt{d_{\text{eff}}(k|\tau)}} \gtrsim 1.$$

Part 2: proving the trade-off

Now, assume we build another kernel $\overline{k}$ with associated space $\overline{\mathcal{H}}$,

$$\overline{k}(x,y) = \sum_{i=1}^{\infty} \overline{\mu}_i \phi_i(x)\phi_i(y) \qquad \overline{h}_i = \frac{\overline{\mu}_i}{\tau+\overline{\mu}_i},$$

such that, for some constant C

$$\sup_{f\in\mathcal{H}} \frac{\|f\|_{\overline{\mathcal{H}}}}{\|f\|_{\mathcal{H}}} \leq C < +\infty.$$

By definition of norm in $\mathcal{H}$, this means that for any i,

$$\overline{\mu}_i = \|\phi_i\|_{\overline{\mathcal{H}}}^{-2} \geq C^{-2}\|\phi_i\|_{\mathcal{H}}^{-2} = C^{-2}\mu_i.$$

Now, consider all values of $j \in \mathbb{N}$ such that both $U_{2^j} = 1$ and $U_{2^{j+1}} = 1$ in the definition of k. By Eq. (11), the number of such pairs for $j \leq J$ is asymptotic to $J/4$. Taking any j satisfying this property, two things may happen for $\overline{k}$:

- There is $2^j < i^* < 2^{j+1}$ such that $\overline{\mu}_{i^*} \leq C^{-2}\frac{j^{-s}}{2}$. We refer to this as **scenario A**. In such a case, it follows that

$$\begin{aligned}\sum_{i=2^j}^{2^{j+1}} \left|\Delta^1 \overline{h}_i\right| &\geq \overline{h}_{2^j} - \overline{h}_{i^*} \\ &\geq \frac{C^{-2}j^{-s}}{C^{-2}j^{-s}+\tau} - \frac{C^{-2}j^{-s}/2}{C^{-2}j^{-s}/2+\tau} \\ &= \frac{\tau C^2 j^s}{(\tau C^2 j^s+1)(2\tau C^2 j^s+1)}.\end{aligned}$$

- There is no $2^j < i < 2^{j+1}$ such that $\overline{\mu}_i \leq C^{-2}\frac{j^{-s}}{2}$. We call this **scenario B**. In this case, one has

$$\sum_{i=2^j}^{2^{j+1}} \overline{h}_i \geq 2^{j-1}\frac{C^{-2}j^{-s}}{C^{-2}j^{-s}+\tau}.$$

Now, choose $J = \tau^{-1/s}$. By Eq. (11), for τ sufficiently small, the number of $j \leq J$ such that both $U_{2^j} = 1$ and $U_{2^{j+1}} = 1$ in the definition of k is $\asymp J/4$. This means that at least one of **scenario A** or **B** happens at least $J/8$ times.

(A) If this scenario happens at least $J/8$ times, we have

$$\sum_{i=1}^{\infty} |\Delta^1 \overline{h}_i| \geq \frac{J}{8} \frac{\tau C^2 J^s}{(\tau C^2 J^s + 1)(2\tau C^2 J^s + 1)}$$
$$\geq \frac{\tau^{-1/s}}{8} \frac{C^2}{(C^2+1)(2C^2+1)} \gtrsim \tau^{-1/s}.$$

This contradicts item 2) in the statement of Theorem 3.11 for every $s' > s$.

(B) Let j^* be the maximum, which cannot be lower than $J/8 = \tau^{-1/s}/8$. In this case, the effective dimension of $\overline{k}$ is

$$d_{\text{eff}}(\overline{k}|\tau) = \sum_{i=1}^{\infty} \overline{h}_i \geq \sum_{i=2^{j^*}}^{2^{j^*+1}} \overline{h}_i$$
$$\geq 2^{j^*-1} \frac{C^{-2} j^{*-s}}{C^{-2} j^{*-s} + \tau} \asymp 2^{\frac{(1/\tau)^{1/s}}{8}}.$$

Since the effective dimension grows exponentially in $\tau^{-1/s}$, conclusion 3) is false.

□

B.2 Proofs for multivariate product kernels

In the multivariate Fourier-diagonal setting, Mercer's decomposition takes the form

$$k(x,y) = \sum_{\boldsymbol{i} \in \mathbb{N}^m} \mu_{\boldsymbol{i}} \prod_{j=1}^{m} \phi_{i_j}(x_j)\phi_{i_j}(y_j), \qquad \boldsymbol{i} = (i_1, \ldots, i_m), \tag{12}$$

where $\phi_{i_j}(x_j)$ is the i_j-th single-variate Fourier feature applied to the j-th coordinate. Throughout this subsection, $\mathbb{N} = \{1, 2, \ldots\}$, and sums over $\mathbb{N}^m$ are understood as limits of rectangular partial sums when the summand is signed, and as the usual extended sums when the summand is nonnegative. Let $\boldsymbol{e}_j$ be the j-th coordinate vector, define $\Delta_j h_{\boldsymbol{i}} = h_{\boldsymbol{i}+\boldsymbol{e}_j} - h_{\boldsymbol{i}}$, and set $\Delta^{\mathbf{1}} = \Delta_1 \cdots \Delta_m$. One should think of Δ_j as a discrete partial derivative in the j-th coordinate, and $\Delta^{\mathbf{1}}$ as the mixed first-order difference, analogous to $\partial_1 \cdots \partial_m$. The following result gives a bound on the Lebesgue constant of the population KRR operator in this setting.

Theorem B.3 (Multivariate Lebesgue bound)**.** *Let k be a kernel on $[-1,1]^m$ whose Mercer decomposition has the Fourier-diagonal form (12). Then the Lebesgue constant of the corresponding population KRR operator is bounded as follows:*

$$\Lambda(\mathsf{P}_\tau) \lesssim_m \sum_{\boldsymbol{i} \in \mathbb{N}^m} \left| \Delta^{\mathbf{1}} \left(\frac{\mu_{\boldsymbol{i}}}{\tau + \mu_{\boldsymbol{i}}} \right) \right| \prod_{j=1}^{m} \log(e + i_j).$$

For the proof, we need one summation-by-parts lemma. For $\boldsymbol{r}, \boldsymbol{i} \in \mathbb{N}^m$, we write $\boldsymbol{r} \geq \boldsymbol{i}$ when $r_j \geq i_j$ for every j. For an array $c_{\boldsymbol{i}}$, define the discrete antiderivative in the k-th coordinate by

$$(I_k c)_{\boldsymbol{i}} \coloneqq \sum_{\ell=1}^{i_k} c_{i_1, \ldots, i_{k-1}, \ell, i_{k+1}, \ldots, i_m}.$$

Lemma B.4 (Multidimensional Abel summation)**.** *Let $(a_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{N}^m}$ be a real array and $(b_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{N}^m}$ an array of L^1 functions. Assume that $\Delta^{\mathbf{1}} a \in \ell^1(\mathbb{N}^m)$ and that $a_{\boldsymbol{i}} \to 0$ whenever $\|\boldsymbol{i}\|_\infty \to \infty$. If*

$$\sum_{\boldsymbol{r} \in \mathbb{N}^m} |\Delta^{\mathbf{1}} a_{\boldsymbol{r}}| \, \|(I_m \cdots I_1 b)_{\boldsymbol{r}}\|_{L^1} < \infty,$$

then

$$\sum_{\boldsymbol{i} \in \mathbb{N}^m} a_{\boldsymbol{i}} b_{\boldsymbol{i}} = (-1)^m \sum_{\boldsymbol{r} \in \mathbb{N}^m} \Delta^{\mathbf{1}} a_{\boldsymbol{r}} (I_m \cdots I_1 b)_{\boldsymbol{r}}$$

with convergence in L^1. In particular, if $b_{\boldsymbol{i}}(x) = \prod_{j=1}^m \psi_{j,i_j}(x_j)$ on a product space and $\Psi_{j,i} = \sum_{\ell=1}^i \psi_{j,\ell}$, then

$$(I_m \cdots I_1 b)_{\boldsymbol{i}}(x) = \prod_{j=1}^m \Psi_{j,i_j}(x_j).$$

Proof. First, telescoping in each coordinate gives the tail representation

$$a_{\boldsymbol{i}} = (-1)^m \sum_{\boldsymbol{r} \geq \boldsymbol{i}} \Delta^{\mathbf{1}} a_{\boldsymbol{r}}.$$

Indeed, the finite rectangular tail sum equals the alternating sum of a over the outer corners of the rectangle, and the boundary terms vanish as the upper limits tend to infinity by the assumed decay of a. Substituting this representation gives

$$\begin{aligned}
\sum_{\boldsymbol{i} \in \mathbb{N}^m} a_{\boldsymbol{i}} b_{\boldsymbol{i}} &= (-1)^m \sum_{\boldsymbol{i} \in \mathbb{N}^m} \sum_{\boldsymbol{r} \geq \boldsymbol{i}} \Delta^{\mathbf{1}} a_{\boldsymbol{r}} b_{\boldsymbol{i}} \\
&= (-1)^m \sum_{\boldsymbol{r} \in \mathbb{N}^m} \Delta^{\mathbf{1}} a_{\boldsymbol{r}} \sum_{\boldsymbol{i} \leq \boldsymbol{r}} b_{\boldsymbol{i}} \\
&= (-1)^m \sum_{\boldsymbol{r} \in \mathbb{N}^m} \Delta^{\mathbf{1}} a_{\boldsymbol{r}} (I_m \cdots I_1 b)_{\boldsymbol{r}}.
\end{aligned}$$

The identity $\sum_{\boldsymbol{i} \leq \boldsymbol{r}} b_{\boldsymbol{i}} = (I_m \cdots I_1 b)_{\boldsymbol{r}}$ holds because applying $I_1, \ldots, I_m$ successively sums b over the rectangular box $1 \leq i_j \leq r_j$ in each coordinate. The assumed L^1 summability of the last series justifies the exchange of sums and the convergence in L^1. Finally, when $b_{\boldsymbol{i}}$ has product form, each I_k acts only on the k-th factor, so applying $I_1, \ldots, I_m$ gives the product of the one-dimensional partial sums. □

Proof of Theorem B.3. By Lemma B.1, the Lebesgue constant of the population KRR can be written as follows

$$\Lambda(\mathsf{P}_\tau) = \sup_{x \in \mathcal{X}} \left\| \sum_{\boldsymbol{i} \in \mathbb{N}^m} \frac{\mu_{\boldsymbol{i}}}{\tau + \mu_{\boldsymbol{i}}} \prod_{j=1}^m \phi_{i_j}(x_j) \phi_{i_j}(\cdot) \right\|_{L^1} \leq \left\| \sum_{\boldsymbol{i} \in \mathbb{N}^m} \frac{\mu_{\boldsymbol{i}}}{\tau + \mu_{\boldsymbol{i}}} \prod_{j=1}^m \phi_{i_j}(\cdot) \right\|_{L^1}.$$

We take the inner function and set $h_{\boldsymbol{i}} = \mu_{\boldsymbol{i}}/(\mu_{\boldsymbol{i}} + \tau)$. If the right-hand side in the statement is infinite, there is nothing to prove. Otherwise, the summability condition in Lemma B.4 is satisfied. Moreover $h_{\boldsymbol{i}} \to 0$ whenever any coordinate $i_j \to \infty$, since the Fourier coefficients of the kernel vanish at infinity. Letting, for any $i \in \mathbb{N}$,

$$\Phi_i(x) = \sum_{\ell=1}^i \phi_\ell(x).$$

Lemma B.4 gives

$$\sum_{\boldsymbol{i} \in \mathbb{N}^m} h_{\boldsymbol{i}} \prod_{j=1}^m \phi_{i_j}(\cdot) = (-1)^m \sum_{\boldsymbol{i} \in \mathbb{N}^m} \Delta^{\mathbf{1}} h_{\boldsymbol{i}} \prod_{j=1}^m \Phi_{i_j}(\cdot).$$

As in the single-variate case, the term $\Phi_i(\cdot)$ corresponds to the Dirichlet kernel, whose L^1 norm is $\mathcal{O}(\log(e+i))$ [Folland, 2009]. Therefore, taking L^1 norms gives

$$\Lambda(\mathsf{P}_\tau) \lesssim_m \sum_{\boldsymbol{i} \in \mathbb{N}^m} \left| \Delta^{\mathbf{1}} \left(\frac{\mu_{\boldsymbol{i}}}{\tau + \mu_{\boldsymbol{i}}} \right) \right| \prod_{j=1}^m \log(e + i_j),$$

which completes the proof. □

Lemma B.5. *Let $m \geq 1$, $\tau > 0$, and $D = z\frac{d}{dz}$. Then, for all $z > 0$,*

$$\left| D^m \left(\frac{z}{\tau + z} \right) \right| \lesssim_m \min\{z/\tau, \tau/z\}.$$

Proof. Put $u = z/\tau$. Since $D = u\frac{d}{du}$, it is enough to prove the corresponding bound for $(u\frac{d}{du})^m \frac{u}{1+u}$. We claim by induction on m that

$$g_m(u) \coloneqq \left(u\frac{d}{du}\right)^m \frac{u}{1+u} = \frac{uP_m(u)}{(1+u)^{m+1}},$$

where P_m is a polynomial of degree at most $m-1$ whose coefficients depend only on m. The case $m = 1$ has $P_1 = 1$. If the claim holds for m, applying $u\frac{d}{du}$ gives the same form with

$$P_{m+1}(u) = (P_m(u) + uP_m'(u))\,(1+u) - (m+1)uP_m(u),$$

which has degree at most m. Thus the claim follows. For $u \le 1$, g_m is upper bounded by a constant times u; for $u \ge 1$, it is bounded by a constant times u^{-1}. Taking the larger of the constants and using that for $0 \le u \le 1$, $1/u \ge 1 \ge u$, while for $u \ge 1$, $u \ge 1 \ge 1/u$ gives the desired result. □

Lemma B.6. *Let $(\lambda_i)_{i\ge1}$ be non-increasing and satisfy $\lambda_i \le C_1 i^{-s}$ for some $C_1, s > 0$. For*

$$H_\tau(y_1, \dots, y_m) \coloneqq \frac{\prod_{j=1}^m y_j}{\tau + \prod_{j=1}^m y_j},$$

one has

$$\sum_{\boldsymbol{i}\in\mathbb{N}^m} \left|\Delta^{\mathbf{1}} H_\tau(\lambda_{i_1}, \dots, \lambda_{i_m})\right| \prod_{j=1}^m \log(e + i_j) \lesssim_m \frac{\log^{2m-1}(e + C_1^m/\tau)}{s^m}.$$

Proof. Let $I_i = [\lambda_{i+1}, \lambda_i]$. By applying the fundamental theorem of calculus in each coordinate,

$$\left|\Delta^{\mathbf{1}} H_\tau(\lambda_{i_1}, \dots, \lambda_{i_m})\right| \le \int_{\prod_j I_{i_j}} \left|\frac{\partial^m H_\tau}{\partial y_1 \cdots \partial y_m}(y)\right| dy.$$

If $y_j \in I_{i_j}$, then $\log(e + i_j) \lesssim s^{-1} \log(e + C_1/y_j)$. Therefore

$$\begin{aligned}\sum_{\boldsymbol{i}\in\mathbb{N}^m} &\left|\Delta^{\mathbf{1}} H_\tau(\lambda_{i_1}, \dots, \lambda_{i_m})\right| \prod_{j=1}^m \log(e + i_j) \\ &\lesssim_m \frac{1}{s^m} \int_{(0,C_1]^m} \left|\frac{\partial^m H_\tau}{\partial y_1 \cdots \partial y_m}(y)\right| \prod_{j=1}^m \log(e + C_1/y_j)\, dy.\end{aligned}$$

Put $z = \prod_{j=1}^m y_j$ and $D = z\frac{d}{dz}$. Since $y_j\partial_{y_j} = D$ on functions of z, we have

$$\frac{\partial^m H_\tau}{\partial y_1 \cdots \partial y_m}(y) = \frac{1}{z} D^m \left(\frac{z}{\tau + z}\right).$$

By Lemma B.5, the rational function on the right satisfies

$$\left|D^m \left(\frac{z}{\tau + z}\right)\right| \lesssim_m \min\{z/\tau, \tau/z\}.$$

We now set $y_j = C_1 e^{-r_j}$, $r_j \ge 0$, and $A = C_1^m/\tau$. Then $z/\tau = Ae^{-\sum_j r_j}$ and the preceding integral is bounded, up to a constant depending only on m, by

$$\int_{\mathbb{R}_+^m} \min\left\{Ae^{-\sum_j r_j}, A^{-1}e^{\sum_j r_j}\right\} \prod_{j=1}^m (1 + r_j)\, dr.$$

Using the simplex identity

$$\int_{\{r\in\mathbb{R}_+^m : \sum_j r_j = t\}} \prod_{j=1}^m (1 + r_j)\, d\sigma(r) \lesssim_m (1 + t)^{2m-1},$$

this is at most

$$\int_0^\infty \min\{Ae^{-t}, A^{-1}e^t\}(1 + t)^{2m-1}\, dt \lesssim_m \log^{2m-1}(e + A).$$

This proves the claim. □

Theorem 3.12. *Let $k_0(x,y) = \mathscr{K}_0(x-y)$ be a one-dimensional kernel on $[-1,1]$ satisfying Assumption 2.1. Assume that the Mercer eigenfunctions of k_0 are the Fourier basis and that its eigenvalues are non-increasing. Then the population KRR operator associated with the product kernel k_Π in (7) satisfies*

$$\Lambda(\mathsf{P}_\tau) \lesssim_m \log^{2m-1}(e + \kappa^m/\tau).$$

Proof. As in Corollary 3.9, Lemma B.2 gives polynomial eigendecay $\mu_i^\circ \le \kappa/i$ for the single-variate spectrum. By Theorem B.3, we have

$$\Lambda(\mathsf{P}_\tau) \lesssim_m \sum_{\boldsymbol{i}\in\mathbb{N}^m} \left|\Delta^{\mathbf{1}}\left(\frac{\mu_{\boldsymbol{i}}}{\tau+\mu_{\boldsymbol{i}}}\right)\right| \prod_{j=1}^m \log(e+i_j).$$

As the kernel factorizes k_Π, letting μ_i° be the eigenvalues of k_0, one has $\mu_{\boldsymbol{i}} = \prod_{j=1}^m \mu_{i_j}^\circ$. The desired bound then follows by applying Lemma B.6 with $\lambda_i = \mu_i^\circ$, $C_1 = \kappa$, and $s = 1$. This completes the proof. □

C Proofs of Section 4

We collect the additional notation used in this section in Section C.

Symbol	Meaning
A	Region of the input space $\mathcal{X} = [-1,1]^m$.
$\mathcal{A}_t$	Active regions at step t
$\widetilde{\mathcal{A}}_n$	$= \bigcup_{s=1}^n \mathcal{A}_s$
$\mathbb{T}(A)$	Maximum number of queries to points in region A during the process.
N_t^A	Number of queries to points in region A after t rounds.
$\boldsymbol{x}_t^A$	Queries in region A after t rounds
$\boldsymbol{y}_t^A$	Responses in region A after t rounds
$\widehat{\mu}_{t-1}^A$	KRR estimator trained on $(\boldsymbol{x}_{t-1}^A, \boldsymbol{y}_{t-1}^A)$
β_t^A	Exploration bonus on region A at step t, see (9).
σ_{t-1}^A	Posterior variance on region A at step t.

C.1 Preliminary results

Lemma C.1. *Given $\delta \in (0,1)$, for all $t \le n$, for all $A \in \mathcal{A}_t$, for all $x \in A$, we have*

$$|\widehat{\mu}_{t-1}^A(x) - f_\star(x)| \le \beta_t^A \sigma_{t-1}^A(x),$$

with probability $1-\delta$.

Proof. Let μ_{t-1}^A be the posterior mean obtained from the modified dataset $(\boldsymbol{x}_{t-1}^A, f_\star(\boldsymbol{x}_{t-1}^A) + \boldsymbol{y}_{t-1}^A - f(\boldsymbol{x}_{t-1}^A))$. Then, from Lemma 2 of [Bogunovic and Krause, 2021],

$$|\widehat{\mu}_{t-1}^A(x) - \mu_{t-1}^A(x)| \le \varepsilon \sqrt{\frac{N_{t-1}^A}{\lambda}} \sigma_{t-1}^A(x).$$

Since the modified labels have conditional mean $f_\star$ and preserve the same subgaussian noise level, Lemma 5 of [Janz et al., 2020] ensures

$$|\mu_{t-1}^A(x) - f_\star(x)| \le \left(\frac{R}{\lambda^{1/2}}\sqrt{(bm)\log(4t/\delta) + 1 + \gamma_{N_{t-1}^A}} + B\right)\sigma_{t-1}^A(x).$$

The proof is completed by triangle inequality and the definition of β_t^A. □

Lemma C.2. *The total number of regions does not exceed $|\widetilde{\mathcal{A}}_n| \le 9^m n^{\frac{m}{m+b}}$.*

Proof. Let us call $\mathcal{B}_n$ the regions that have split through the process. By definition,

$$|\widetilde{\mathcal{A}}_n| \leq (2^m + 1)|\mathcal{B}_n|,$$

where $+1$ takes into account the initial region. We can therefore focus on bounding $\mathcal{B}_n$. Let us denote K_ℓ the number of regions in $\mathcal{B}_n$ of side $2^{-\ell}$. By the geometry of the space,

$$K_\ell \leq 2^{m\ell}.$$

Moreover, by the splitting rule, $2^{b\ell}$ points are necessary to split a region of side $2^{-\ell}$. Therefore, the cardinality of $\mathcal{B}_n$ can be written as

$$|\mathcal{B}_n| = \sum_{\ell=1}^{\infty} K_\ell, \text{ such that } \quad K_\ell \leq 2^{m\ell}, \quad \sum_{\ell=1}^{\infty} 2^{b\ell} K_\ell \leq n.$$

The former is an optimization problem that can be solved explicitly. In fact, under the constraint $\sum_{\ell=1}^{\infty} 2^{b\ell} K_\ell \leq n$, where $2^{b\ell}$ is strictly increasing in ℓ, the allocation that maximizes $\sum_{\ell=1}^{\infty} K_\ell$ sets $K_\ell = 2^{m\ell}$ until the budget n is reached, for some L. Let $L = \lceil \log_2(n)/(b+m) \rceil$. Then,

$$\sum_{\ell=1}^{L} 2^{(b+m)\ell} \geq 2^{(b+m)L} \geq n.$$

This proves that in the optimal allocation $K_\ell > 0$ only for, at most, $\ell \leq L$. Thus, the number of split regions does not exceed

$$\begin{aligned} |\mathcal{B}_n| &= \sum_{\ell=1}^{L} 2^{m\ell} \leq 2^{m(L+1)} \\ &= 2^{m(\lceil \log_2(n)/(b+m) \rceil + 1)} \leq 2^{2m + m\lfloor \log_2(n)/(b+m) \rfloor} \\ &= 4^m n^{\frac{m}{m+b}}. \end{aligned}$$

Therefore,

$$|\widetilde{\mathcal{A}}_n| \leq (2^m + 1)|\mathcal{B}_n| \leq 9^m n^{\frac{m}{m+b}}.$$

□

C.2 Regret bound

Theorem 4.3. *Consider the online model of Definition 2.3 on $[-1,1]^m$. Under Assumption 2.1, let $f_\star \in \mathcal{H}$ satisfy $\|f_\star\|_{\mathcal{H}} \leq B$ and $\|f - f_\star\|_\infty \leq \varepsilon$, and suppose the kernel satisfies* **D2+** *and* **D3**. *Then running Algorithm 1 with parameters $\lambda = 1$ and $b = \alpha$ yields, with probability at least $1 - \delta$, a regret bound of order*

$$R_n^{\mathscr{A}} \leq \widetilde{\mathcal{O}}\left(n^{\frac{2m+\alpha}{2m+2\alpha}} + n\varepsilon \right).$$

Proof. Let $A_t^\star$ be the unique active region in $\mathcal{A}_t$ containing $x^\star$. By Lemma C.1 and the definition of (A_t, x_t),

$$\begin{aligned} f(x^\star) - f(x_t) &\leq f_\star(x^\star) - f_\star(x_t) + 2\varepsilon \\ &\leq \widehat{\mu}_{t-1}^{A_t^\star}(x^\star) + \beta_t^{A_t^\star} \sigma_{t-1}^{A_t^\star}(x^\star) - \widehat{\mu}_{t-1}^{A_t}(x_t) + \beta_t^{A_t} \sigma_{t-1}^{A_t}(x_t) + 2\varepsilon \\ &\leq 2\beta_t^{A_t} \sigma_{t-1}^{A_t}(x_t) + 2\varepsilon. \end{aligned}$$

Therefore,

$$\begin{aligned} R_n^{\mathscr{A}} &= \sum_{t=1}^{n} f(x^\star) - f(x_t) \\ &\leq \sum_{t=1}^{n} \left(2\beta_t^{A_t} \sigma_{t-1}^{A_t}(x_t) + 2\varepsilon \right) \end{aligned}$$

$$\leq \sqrt{n \sum_{t=1}^{n} (2\beta_t^{A_t} \sigma_{t-1}^{A_t}(x_t))^2} + 2n\varepsilon,$$

where the last step uses Cauchy-Schwarz. We can rewrite this sum as a sum over the cover elements. For each region $A \in \widetilde{\mathcal{A}}_n$, define the monotone upper envelope

$$\overline{\beta}_A \coloneqq \frac{R}{\lambda^{1/2}} \sqrt{bm \log(4n/\delta) + 1 + \gamma^A_{\mathbb{T}(A)}} + B + \varepsilon \sqrt{\frac{\mathbb{T}(A)}{\lambda}}.$$

For every round in which region A is used, $\beta_t^A \leq \overline{\beta}_A$.

$$\begin{aligned}
\sum_{t=1}^{n} (2\beta_t^{A_t} \sigma_{t-1}^{A_t}(x_t))^2 &\leq \sum_{t=1}^{n} \sum_{A \in \widetilde{\mathcal{A}}_n} \mathbb{I}\{x_t \in A\} \left(2\overline{\beta}_A \sigma_{t-1}^A(x_t)\right)^2 \\
&= \sum_{A \in \widetilde{\mathcal{A}}_n} \sum_{t=1}^{n} \mathbb{I}\{x_t \in A\} \left(2\overline{\beta}_A \sigma_{t-1}^A(x_t)\right)^2 \\
&\leq 16\lambda \sum_{A \in \widetilde{\mathcal{A}}_n} \overline{\beta}_A^2 \gamma^A_{\mathbb{T}(A)},
\end{aligned}$$

where the last inequality comes from Lemma 7 in [Janz et al., 2020] and the definition of $\mathbb{T}(A)$, which ensures that

$$\sum_{t=1}^{n} \mathbb{I}\{x_t \in A\} \sigma_{t-1}^A(x_t)^2 \leq 4\lambda \gamma^A_{\mathbb{T}(A)}.$$

As the kernel satisfies both **D2+** and **D3**, Lemma 2 from [Vakili and Olkhovskaya, 2023] applies for $\eta = 0$. This ensures that the maximum information gain on the hypercube A is bounded as

$$\gamma^A_{\mathbb{T}(A)} \lesssim \log(\mathbb{T}(A))^{\frac{s-1}{s}} \mathbb{T}(A)^{1/s} \rho_A^{\alpha/s}.$$

By line 6 of the algorithm, $\mathbb{T}(A) \leq \rho_A^{-b}$. Therefore, as we are running the algorithm for $b = \alpha$,

$$\gamma^A_{\mathbb{T}(A)} \lesssim \log(\mathbb{T}(A))^{\frac{s-1}{s}} \rho_A^{-\alpha/s} \rho_A^{\alpha/s} \lesssim \log(n)^{\frac{s-1}{s}}.$$

The other term in the product can also be reduced to the information gain on A. Indeed,

$$\begin{aligned}
\overline{\beta}_A^2 &\lesssim \left(\frac{R}{\lambda^{1/2}} \sqrt{bm \log(n/\delta) + 1 + \gamma^A_{\mathbb{T}(A)}} + B \right)^2 + \varepsilon^2 \frac{\mathbb{T}(A)}{\lambda} \\
&\leq \frac{R^2}{\lambda} \log(n/\delta) + \gamma^A_{\mathbb{T}(A)} + B^2 + \varepsilon^2 \frac{\mathbb{T}(A)}{\lambda} \\
&\lesssim \frac{R^2}{\lambda} \log(n/\delta) + \log(n)^{\frac{s-1}{s}} + B^2 + \varepsilon^2 \frac{\mathbb{T}(A)}{\lambda}.
\end{aligned}$$

Combining these estimates gives

$$\begin{aligned}
\sum_{t=1}^{n} (2\beta_t^{A_t} \sigma_{t-1}^{A_t}(x_t))^2 &\leq 16\lambda \sum_{A \in \widetilde{\mathcal{A}}_n} \overline{\beta}_A^2 \gamma^A_{\mathbb{T}(A)} \\
&\lesssim \lambda \log(n)^2 \sum_{A \in \widetilde{\mathcal{A}}_n} \left(\frac{R^2}{\lambda} \log(1/\delta) + B^2 + \varepsilon^2 \frac{\mathbb{T}(A)}{\lambda} \right) \\
&\lesssim \lambda \log(n)^2 \left[\sum_{A \in \widetilde{\mathcal{A}}_n} \left(\frac{R^2}{\lambda} \log(1/\delta) + B^2 \right) + \sum_{A \in \widetilde{\mathcal{A}}_n} \varepsilon^2 \frac{\mathbb{T}(A)}{\lambda} \right] \\
&= \lambda \log(n)^2 |\widetilde{\mathcal{A}}_n| \left(\frac{R^2}{\lambda} \log(1/\delta) + B^2 \right) + \varepsilon^2 \log(n)^2 \sum_{A \in \widetilde{\mathcal{A}}_n} \mathbb{T}(A).
\end{aligned}$$

Let $L = \lceil \log_2(n)/(b+m) \rceil$. By the proof of Lemma C.2, no region deeper than L can be created. Since the regions form a dyadic partition, every query x_t belongs to at most one region of $\widetilde{\mathcal{A}}_n$ at each depth. Therefore,

$$\begin{aligned}\sum_{A \in \widetilde{\mathcal{A}}_n} \mathbb{T}(A) &= \sum_{A \in \widetilde{\mathcal{A}}_n} \sum_{t=1}^{n} \mathbb{I}\{x_t \in A\} \\ &= \sum_{t=1}^{n} \sum_{A \in \widetilde{\mathcal{A}}_n} \mathbb{I}\{x_t \in A\} \\ &\leq n(L+1) \lesssim n \log(n).\end{aligned}$$

Hence,

$$\sum_{t=1}^{n} (2\beta_t^{A_t} \sigma_{t-1}^{A_t}(x_t))^2 \lesssim \log(n)^2 |\widetilde{\mathcal{A}}_n| \left(R^2 \log(1/\delta) + \lambda B^2\right) + n\varepsilon^2 \log(n)^3.$$

We now apply Lemma C.2, which ensures, for $b = \alpha$, $|\widetilde{\mathcal{A}}_n| \leq 9^m n^{\frac{m}{m+\alpha}}$. This results in the following regret bound:

$$\begin{aligned}R_n &\lesssim \sqrt{n \left(\log(n)^2 |\widetilde{\mathcal{A}}_n| \left(R^2 \log(1/\delta) + \lambda B^2\right) + n\varepsilon^2 \log(n)^3 \right)} + 2n\varepsilon \\ &\leq 3^m \log(n)\, n^{\frac{2m+\alpha}{2m+2\alpha}} \sqrt{R^2 \log(1/\delta) + \lambda B^2} + n\varepsilon \log(n)^{3/2} + 2n\varepsilon.\end{aligned}$$

□

C.3 Nested Classes and a Persistent Regret Floor

In statistical learning, when a simple model fails to adequately capture the true environment, the natural remedy is to increase the model's complexity. Consider a learner trying to optimize a true, unknown continuous reward function $f : \mathcal{X} \to \mathbb{R}$. By employing a sequence of increasingly expressive, nested linear function classes $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$ (such as polynomial bases of increasing degree or Fourier features), universal approximation theorems guarantee that the uniform approximation error $\varepsilon_m(f) = \inf_{g \in \mathcal{F}_m} \|f - g\|_\infty$ will vanish as the algebraic dimension $m \to \infty$. Intuitively, increasing the complexity of our linear feature space should strictly improve our ability to learn the environment. Can we leverage this vanishing error in the sequential decision-making setting of stochastic bandits? In continuous or large action spaces, algorithms must rely on function approximation to generalize across arms. In the infinite-sample limit, it is natural to consider the behavior of the per-round expected regret. Since the per-round regret may not converge, we consider the upper and lower asymptotic per-round regrets, defined respectively by the limit superior and limit inferior:

$$\overline{R}_{\mathscr{A}}(f, \mathcal{F}_m) \coloneqq \limsup_{n \to \infty} \frac{R_n^{\mathscr{A}}(f, \mathcal{F}_m)}{n}, \quad \underline{R}_{\mathscr{A}}(f, \mathcal{F}_m) \coloneqq \liminf_{n \to \infty} \frac{R_n^{\mathscr{A}}(f, \mathcal{F}_m)}{n}.$$

As the time horizon n grows indefinitely, the statistical noise averages out, and the algorithm's performance is ultimately expected to be bottlenecked by the systematic bias $\varepsilon_m(f)$. Because this approximation error naturally disappears as we move up the nested hierarchy of linear spaces, one might reasonably hope that the asymptotic regret floor will vanish along with it.

To investigate whether this hope holds true, we turn to the workhorse of misspecified linear bandits: Misspecified Phased Elimination (e.g., Lattimore et al., 2020). Let $\mathscr{A}$ denote this standard algorithm. It operates in discrete epochs, maintaining an active set of plausible optimal actions. In each epoch, it queries a core set of exploratory actions—constructed via a G-optimal design over the active feature vectors—to bound the variance of its least-squares estimator. Crucially, to remain robust against the approximation error $\varepsilon_m(f)$, the algorithm artificially increases the size of its confidence intervals by an additive factor of $\sqrt{m}\,\varepsilon_m(f)$ (here, quite generously, we allow the algorithm to know the misspecification level). By Remark D.1 of Lattimore et al. [2020], we know that this specific algorithm achieves an upper asymptotic per-round regret bounded by the systematic misspecification:

$$\overline{R}_{\mathscr{A}}(f, \mathcal{F}_m) \leq C\sqrt{m}\,\varepsilon_m(f)$$

where $C > 0$ is a universal constant.[2]

Here, we immediately encounter a tension: the vanishing approximation error $\varepsilon_m(f)$ is amplified by $\sqrt{m}$, representing the statistical penalty for exploring a larger feature space. The origin of the $\sqrt{m}$ factor is that the algorithm's confidence intervals must be wide enough to contain the true reward function. Since the algorithm can measure f at only finitely many points, it extrapolates to the rest of the action space through a least-squares estimator built from its linear model. In the worst case, the approximation error of the resulting estimate can be as large as $\sqrt{m}\,\varepsilon_m(f)$, explaining the additive inflation of the confidence width and, in turn, the extra $\sqrt{m}$ factor in the regret bound.

The above upper bound suggests that the asymptotic regret floor may not vanish, even as the approximation error $\varepsilon_m(f)$ goes to zero. Intuitively, when

$$c_f := \liminf_{m\to\infty} \sqrt{m}\,\varepsilon_m(f) > 0\,, \tag{13}$$

the algorithm's confidence intervals will never shrink below a positive threshold. For polynomial features on an interval, the G-optimal designs used by phased elimination are Chebyshev-type designs [Kiefer and Wolfowitz, 1960, Guest, 1958]: they place support throughout the active interval rather than concentrating all mass near the empirically best arm. Consequently, whenever the confidence floor keeps an interval of suboptimal arms active whose average gap is bounded away from zero, the algorithm keeps sampling such arms and incurs a non-vanishing regret floor.

One might first wonder whether the first ingredient of this obstruction, namely Eq. (13), can occur at all. If it can, one may further ask whether it occurs for ordinary nested approximation classes, or only as a pathological possibility. As it happens, Eq. (13) holds already in a simple approximation-theoretic example: take $\mathcal{X} = [0, 1]$, $f(x) = 1 - \sqrt{x}$, so that the optimum action is at $x = 0$, and let $\mathcal{F}_m$ be the space of polynomials of degree at most m.

By Jackson's Theorem for polynomial approximation, the uniform approximation error of an α-Hölder continuous function f on $[0, 1]$ by a degree-m polynomial scales as $\mathcal{O}(1/m^\alpha)$. Since $f(x) = 1 - \sqrt{x}$ is exactly $1/2$-Hölder continuous, the theorem yields $\varepsilon_m(f) \le C/m^{1/2}$. Furthermore, classical results by de La Vallée Poussin confirm that the best uniform polynomial approximation of $\sqrt{x}$ is bounded below by $c/\sqrt{m}$. Therefore:

$$\varepsilon_m(f) = \Theta\left(\frac{1}{\sqrt{m}}\right)$$

and thus Eq. (13) is satisfied with $c_f = \liminf_{m\to\infty} \sqrt{m}\,\varepsilon_m(f) > 0$.

Thus, we have the following surprising conclusion: there exist natural reward functions f and nested linear function classes $\mathcal{F}_m$ such that, even though the approximation error $\varepsilon_m(f)$ vanishes as $m \to \infty$, for a reasonable bandit algorithm, the lower asymptotic per-round regret $\underline{R}_{\mathscr{A}}(f, \mathcal{F}_m)$ fails to vanish.

Note also that $f(x) = 1 - \sqrt{x}$, as a $1/2$-Hölder continuous function, is not particularly pathological. Simple discretization-based bandit algorithms can achieve sublinear (vanishing per-round) regret on this problem. Thus, the persistent regret floor is not an inherent property of the problem, but rather a consequence of the algorithm's reliance on linear function approximation and its associated confidence interval inflation. This raises the question of whether algorithms that rely on linear function approximation are fundamentally limited in their ability to leverage vanishing approximation error:

Open Problem: The Universal Nested-Class Paradox Does there exist a sequence of nested linear feature maps $\phi_m : \mathcal{X} \to \mathbb{R}^m$ such that for m large enough, ϕ_m is injective, and such that, for any bandit algorithm $\mathscr{A}$ that is given access to the feature map ϕ_m and the misspecification error $\varepsilon_m(f)$ as input, and which satisfies $R_n^{\mathscr{A}}(f, \mathcal{F}_m) = \widetilde{\mathcal{O}}(m\sqrt{n} + \sqrt{m}n\varepsilon_m(f))$ as $n \to \infty$ for all $f \in \mathcal{F}_m$, there exists a bounded reward function $f : \mathcal{X} \to [0, 1]$ satisfying:

1. The uniform approximation error vanishes: $\lim_{m\to\infty} \varepsilon_m(f) = 0$.
2. The lower asymptotic per-round regret fails to vanish: $\liminf_{m\to\infty} \underline{R}_{\mathscr{A}}(f, \mathcal{F}_m) > 0$.

[2] A small gap here is that the bound as stated in Lattimore et al. [2020] assumes the algorithm also knows the horizon n, but this can be easily circumvented by a standard doubling trick.

Above, $\mathcal{F}_m = \{x \mapsto \langle w, \phi_m(x) \rangle : w \in \mathbb{R}^m\}$ and $\widetilde{\mathcal{O}}$ hides logarithmic factors in m and n.

Note that in nonparametric regression, series estimators — a form of linear function approximation — achieve minimax-optimal rates in both the L^2 and L^∞ norms, up to logarithmic factors [Tsybakov, 2009], suggesting that any such limitation is specific to the bandit setting.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Proofs are provided for every claim made.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We clearly state the assumptions behind any result. Moreover, comparison with the state of the art is performed, enhancing in which scenarios our theorem do not apply and previous work does.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All proofs are reported in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: No experiments included.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: No experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [N/A]

Justification: No experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [N/A]

Justification: No experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: No experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Standard theory paper, nothing to declare.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper focuses on a well-known theoretical problem in optimizing with kernelized methods. There are no particular direct applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Same as before.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: No existing asset is used, related papers are quoted.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: As before.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourching.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.