Random Exploration in Bayesian Optimization: Order-Optimal Regret and Computational Efficiency

Sudeep Salgia [1] **Sattar Vakili** [2] **Qing Zhao** [3]

Abstract

We consider Bayesian optimization using Gaussian Process models, also referred to as kernel-based bandit optimization. We study the methodology of exploring the domain using random samples drawn from a distribution. We show that this random exploration approach achieves the optimal error rates. Our analysis is based on novel concentration bounds in an infinite dimensional Hilbert space established in this work, which may be of independent interest. We further develop an algorithm based on random exploration with domain shrinking and establish its order-optimal regret guarantees under both noise-free and noisy settings. In the noise-free setting, our analysis closes the existing gap in regret performance under a mild assumption on the underlying function and thereby *partially resolves a COLT open problem*. The proposed algorithm also enjoys a computational advantage over prevailing methods due to the random exploration that obviates the expensive optimization of a non-convex acquisition function for choosing the query points at each iteration.

1. Introduction

1.1. GP-based Bayesian Optimization

We consider the problem of sequential optimization of an unknown, possibly non-convex, function $f : \mathcal{X} \to \mathbb{R}$. The learner sequentially chooses a query point $x_t \in \mathcal{X}$ at each time t and observes the function value (potentially subject to noise) at x_t. The learning objective is to approach a global maximizer x^* of the function through a sequence of query points $\{x_t\}_{t=1}^T$ chosen sequentially in time. In addition to the convergence of $\{x_t\}_{t=1}^T$ to x^*, an online measure of the learning efficiency is the *cumulative regret*

$$R(T) = \sum_{t=1}^{T} \left[f(x^*) - f(x_t) \right]. \tag{1}$$

The above problem finds a wide range of applications including hyperparameter optimization (Li et al., 2016), experimental design (Greenhill et al., 2020), recommendation systems (Vanchinathan et al., 2014) and robotics (Lizotte et al., 2007). An approach that has proven to be particularly effective is Bayesian Optimization (BO) using Gaussian Process (GP) models (a.k.a. kernel-based bandit optimization). The unknown objective function f is assumed to live in a Reproducing Kernel Hilbert Space (RKHS) associated with a known kernel. Within the GP-based BO framework, f is viewed as a realization of a Gaussian process over $\mathcal{X}$. With each new query x_t, the learner sharpens the posterior distribution and uses it as a proxy for f for subsequent optimization. We point out that such a Bayesian approach is equally applicable to a frequentist formulation where f is *deterministic* as considered in this work. In this case, the GP model of f is fictitious and internal to the algorithm.

Under the assumption of noise-free query feedback, BO techniques were used for optimization as early as 1964 (Kushner, 1964). GP-based BO was popularized through the work of Močkus et al. (1978). Since then, a number of approaches have been developed and analyzed over the years, often under certain conditions on the kernels and functional characteristics around x^* (see Sec. 1.3 for a detailed discussion). Surprisingly, despite the long history, an algorithm with guaranteed order-optimal regret performance remains open as discussed in Vakili (2022).

GP-based BO under noisy query was studied much more recently, following the pioneering work by Srinivas et al. (2010) where they proposed the celebrated GP-UCB algorithm. Extensive studies since then have fully characterized the achievable learning performance, both in terms of information-theoretic lower bounds (Scarlett et al., 2017) and the design of algorithms such as SupKernel-UCB (Valko et al., 2013), GP-ThreDS (Salgia et al., 2021), BPE (Li

[1]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA [2]MediaTek Research, Cambridge, UK [3]Department of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA. Correspondence to: Sudeep Salgia <ssalgia@andrew.cmu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

& Scarlett, 2022), and RIPS (Camilleri et al., 2021) that achieve the optimal performance.

Under both the noise-free and noisy settings, a key practical concern for GP-based algorithms is their computational cost. The major computational bottleneck of prevailing GP-based algorithms is the maximization of an *acquisition* function for choosing the query point at each time instant. The acquisition functions are often non-convex and computationally expensive to maximize. To achieve low regret order, such an optimization often needs to be carried out with increasing accuracy as time goes, resulting in a high overall computational requirement.

1.2. Main Results

We explore a new design methodology for GP-based BO: an open-loop exploration of the domain using query points sampled at random from an arbitrary probability distribution supported over the domain. We show that this random exploration approach, while simplistic in nature, leads to order-optimal regret guarantees under both noise-free and noisy feedback models, thus closing the long standing regret gap in the noise-free setting. Moreover, the non-adaptive nature of random sampling bypasses the expensive step of optimizing a non-convex acquisition function, offering a computationally efficient solution without sacrificing learning efficiency.

Random exploration, while not new to many problems (see Sec. 1.3), has not been considered or analyzed for GP-based BO. It stands in sharp contrast to the prevailing exploratory query strategy in GP-based BO: the maximum posterior variance (MPV) sampling. Under MPV, the learning algorithm at each time queries the point with the highest posterior variance conditioned on past observations, i.e., a greedy approach to maximal uncertainty reduction. Surprisingly, we show that the simple, non-adaptive scheme of random exploration achieves the same order of predictive performance as MPV sampling, which is known to be order-optimal. In particular, we show that the worst-case posterior variance corresponding to n randomly drawn points is bounded with high probability by $\tilde{\mathcal{O}}(\gamma_n/n)$ and $\tilde{\mathcal{O}}(n^{1-\beta})$ under noisy and noise-free feedback models, where γ_n is the maximal information gain from n query points and $\beta > 1$ is the order of the polynomial eigendecay of the kernel (see Sec. 2 for their definitions).

A simpler solution is often more demanding when it comes to establishing optimality in performance. The drastically different nature of random exploration from MPV demands different analytical techniques in characterizing its predictive performance. The tightest bound on the worst-case predictive error of MPV sampling, derived in Wenzel et al. (2021), was obtained using the results on scattered data interpolation (i.e., approximating an unknown function using a given set of points) of functions in Sobolev spaces that provide bounds on the worst-case estimation error of the best interpolant based on the fill distance of the given set of points (Wendland, 2004; Narcowich et al., 2006; Brenner et al., 2008; Arcangéli et al., 2012; Wenzel et al., 2021). Since RKHSs of Matérn kernels are norm-equivalent to Sobolev spaces, these results also immediately translate to estimation errors for function interpolation in RKHSs. The analytical techniques used in these studies require various technical assumptions on the regularity of the function domain and its boundary. These technical assumptions on the function domain present major challenges in incorporating MPV sampling with effective optimization techniques such as domain shrinking/elimination, hindering its potential applicability in designing algorithms with optimal regret. In contrast, in analyzing random exploration, we establish the concentration of the spectrum of the sample covariance operator to that of the true covariance operator that holds *universally* for all compact domains. The crux of our analysis builds upon a careful treatment of the infinite-dimensional operators to separately ensure the concentration of the initial spectrum (consisting of the larger eigenvalues) and the tail spectrum, which allows us to obtain optimal convergence rate. The simplicity of random exploration in its implementation and the generality in its guaranteed predictive performance as established in this work make this exploration strategy an attractive alternative to MPV. We believe that the tools and techniques established here are of independent interest for extending the methodology of random exploration to other problem fields.

Built upon the above key results on random exploration, we develop and analyze a new algorithm for GP-based BO. Referred to as Random Exploration with Domain Shrinking (REDS), this algorithm integrates the exploration strategy of random sampling with the optimization technique of domain shrinking (Li & Scarlett, 2022; Salgia et al., 2021). Under the noise-free feedback model, we show that REDS incurs a cumulative regret of $\tilde{\mathcal{O}}(\max\{T^{(3-\beta)/2}, 1\})$, which closes the gap to the known lower bound established in Tuo & Wang (2020) and hence resolves the longstanding open problem. The generality of random exploration, both in terms of the design methodology and performance guarantee is the reason behind the optimal regret performance of REDS. In particular, the order-optimal predictive performance of random exploration that holds universally over all compact domain enables a seamless integration of this exploration strategy with domain shrinking. Similarly, in the noisy setting, we show that REDS offers a cumulative regret of $\tilde{\mathcal{O}}(\sqrt{T\gamma_T})$, which is order-optimal up to logarithmic factors.

The computational advantage of REDS is evident due to the simplicity of random exploration. We further demonstrate this with empirical studies where we compare REDS with

BPE (Li & Scarlett, 2022) and GP-ThreDS (Salgia et al., 2021), all offering optimal regret performance. GP-ThreDS was shown to be computationally more efficient than prevailing algorithms such as GP-UCB. We show that REDS offers a significant speed-up in running time over both algorithms without compromising the regret performance. As shown in Table 1, REDS offers a $\sim 15\times$ and $\sim 100\times$ speed-up in runtime over GP-ThreDS and BPE, respectively.

1.3. Related Work

For GP-based BO with noise-free feedback, a number of algorithms such as GP-EI (Močkus, 1975), EGO (Jones et al., 1998), knowledge-gradient policy (Frazier et al., 2008), and GP-PI (Kushner, 1964; Törn & Žilinskas, 1989; Jones, 2001) have been proposed, which have since become classical. We refer the reader to the excellent tutorial by Brochu et al. (2010) for a more detailed description of the classical approaches. Despite their good empirical performance and popularity, theoretical guarantee on the convergence of these algorithms has only been established relatively recently. Vazquez & Bect (2010) showed that EI converges almost surely for any function drawn from a GP prior of finite smoothness. Grünewälder et al. (2010) established the convergence rate of a computationally infeasible version of EI. Later, Bull (2011) established convergence rates for the computationally feasible version, showing that GP-EI achieves the optimal *simple* regret for Matérn kernels with smoothness $\nu < 1$, which does not translate to optimal cumulative regret performance. More recently, De Freitas et al. (2012) proposed the Branch and Bound algorithm that achieves a constant cumulative regret in Bayesian setting under additional assumptions on the differentiability of the kernel and the behaviour around the unique global maximum, which in practice are difficult to verify. In contrast, REDS requires no such additional assumptions and is analyzed in the frequentist setting. Lyu et al. (2020) showed that for kernels with a polynomial eigendecay with parameter β (See Definition 2.2), the GP-UCB algorithm achieves a regret of $\mathcal{O}(T^{\frac{1+\beta}{2\beta}})$, which is sub-optimal, as shown in Vakili (2022).

The idea of using random sampling has been explored in related fields. The reconstruction of square integrable functions using random samples is a well-studied problem (Bohn & Griebel, 2017; Bastian Bohn, 2017; Bohn, 2018; Smale & Zhou, 2004; Cohen et al., 2013; Chkifa et al., 2015; Cohen & Migliorati, 2017). In particular, a series of studies considers efficient reconstruction of functions in RKHS using random samples drawn from the domain (Kämmerer et al., 2021; Krieg & Ullrich, 2021a;b; Moeller & Ullrich, 2021). Despite certain similarities in the problem setup, an important point of distinction is that these studies focus on bounding the L_2 error of the reconstruction. In this work, we focus on bounding the sup-norm (or equivalently, L_∞ norm) of the estimation error, which is larger than the L_2 norm and more challenging than bounding the L_2 norm. Since the analysis of algorithms requires a bound on the sup-norm of the estimation error, existing results are not applicable here.

2. Problem Statement

2.1. RKHS and Mercer's Theorem

Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$ and ϱ a finite Borel measure supported on $\mathcal{X}$. A measure ϱ is said to be *supported* on $\mathcal{X}$ if $\varrho(\mathcal{Y}) > 0$ for all open sets $\mathcal{Y} \subset \mathcal{X}$. For $\mathcal{X} \subset \mathbb{R}^d$, this is equivalent to ϱ being *absolutely continuous* w.r.t. the Lebesgue measure. Let $L_2(\varrho, \mathcal{X})$ denote the Hilbert space of (real) functions defined over $\mathcal{X}$ that are square-integrable w.r.t. ϱ[1].

Consider a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. A Hilbert space $\mathcal{H}_k$ of functions on $\mathcal{X}$ equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ is called a Reproducing Kernel Hilbert Space (RKHS) with reproducing kernel k if the following conditions are satisfied: (i) $\forall\, x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{H}_k$; (ii) $\forall\, x \in \mathcal{X}, \forall\, f \in \mathcal{H}_k$, $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k}$. For simplicity, we use ψ_x to denote $k(\cdot, x)$. The inner product induces the RKHS norm, $\|f\|^2_{\mathcal{H}_k} = \langle f, f \rangle_{\mathcal{H}_k}$. WLOG, we assume that $k(x, x) = \|\psi_x\|^2_{\mathcal{H}_k} \leq 1$. For brevity, we drop the subscript of $\mathcal{H}_k$ from the inner product for the rest of the paper.

Mercer's Theorem provides an alternative representation for RKHSs through the eigenvalues and eigenfunctions of a kernel integral operator defined over $L_2(\varrho, \mathcal{X})$ using the kernel k.

Theorem 2.1. *(Steinwart & Christmann, 2008, Theorem 4.49) Let $\mathcal{X}$ be a compact metric space, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous kernel and ϱ be a finite Borel measure supported on $\mathcal{X}$. Then, there exists an orthonormal system of functions $\{\varphi_j\}_{j\in\mathbb{N}}$ in $L_2(\varrho, \mathcal{X})$ and a sequence of non-negative values $\{\lambda_j\}_{j\in\mathbb{N}}$ satisfying $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, such that $k(x, x') = \sum_{j\in\mathbb{N}} \lambda_j \varphi_j(x)\varphi_j(x')$ holds for all $x, x' \in \mathcal{X}$ and the convergence is absolute and uniform over $x, x' \in \mathcal{X}$. Moreover, $\{(\lambda_j, \varphi_j)\}_{j\in\mathbb{N}}$ corresponds to the eigensystem of the kernel integral operator $T_k : L_2(\varrho) \to L_2(\varrho)$ given by $T_k f = \int_{\mathcal{X}} k(\cdot, x) f(x) d\varrho(x)$ for all $f \in L_2(\varrho)$.*

Consequently, the Mercer representation (Steinwart & Christmann, 2008, Thm. 4.51) of the RKHS of k is given as

$$\mathcal{H}_k = \left\{ f := \sum_{j\in\mathbb{N}} \alpha_j \lambda_j^{\frac{1}{2}} \varphi_j : \|f\|^2_{\mathcal{H}_k} = \sum_{j\in\mathbb{N}} \alpha_j^2 < \infty \right\}.$$

This also implies that $\{\upsilon_j\}_{j\in\mathbb{N}}$ with $\upsilon_j = \sqrt{\lambda_j}\varphi_j$ is an orthonormal basis for $\mathcal{H}_k$. The following definition charac-

[1]To be rigorous, each $f \in L_2(\varrho, \mathcal{X})$ represents the class of functions that are equivalent ϱ-everywhere.

terizes a class of kernels based on their eigendecay profile corresponding to their Mercer representation.

Definition 2.2. Let $\{\lambda_j\}_{j\in\mathbb{N}}$ denote the eigenvalues of a kernel k arranged in the descending order. The kernel k is said to satisfy the polynomial eigendecay condition with a parameter $\beta > 1$ if, for some universal constant $C > 0$, we have $\lambda_j \leq Cj^{-\beta}$ for all $j \in \mathbb{N}$.

The above class of kernels encompasses a large number of kernels including the widely used Matérn family. We make the following assumption on the kernel k which is commonly adopted in the literature (Vakili et al., 2021b; Chatterji et al., 2019; Riutort-Mayol et al., 2023).

Assumption 2.3. The eigenfunctions $\{\varphi_j\}_{j\in\mathbb{N}}$ corresponding to k are continuous and hence bounded on $\mathcal{X}$, i.e., there exists $F > 0$ such that $\sup_{x\in\mathcal{X}} |\varphi_j(x)| \leq F$ for all $j \in \mathbb{N}$.

2.2. Problem Formulation

We consider the problem of optimizing a fixed and unknown function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subset \mathbb{R}^d$ is a compact domain and $f \in \mathcal{H}_k$ with $\|f\|_{\mathcal{H}_k} \leq B$. A sequential optimization algorithm chooses a point $x_t \in \mathcal{X}$ at each time t and observes $y_t = f(x_t) + \varepsilon_t$. In the noise-free setting, $\varepsilon_t \equiv 0$ for all t. For the noisy setting, we assume that $\{\varepsilon_t\}_{t=1}^T$ are independent, zero-mean, R-sub Gaussian random variables for some fixed constant $R \geq 0$, i.e., $\mathbb{E}[\exp(\zeta\varepsilon_t)] \leq \exp(\zeta^2R^2/2)$, for all $\zeta \in \mathbb{R}$ and $t \leq T$. The performance of the sequential algorithm is measured using the notion of cumulative regret, as defined in Eqn. (1).

2.3. Preliminaries on Gaussian Processes

Under the GP model, the unknown function f is treated hypothetically as a realization of GP$(0, k)$, a Gaussian Process over $\mathcal{X}$ with zero mean and $k(\cdot,\cdot)$ as the covariance kernel. The noise terms ε are also viewed as zero mean Gaussian variables with variance τ. The conjugate property of GPs with Gaussian noise allows for a closed form expression of the posterior distribution. Specifically, let $\mathcal{Z}_t = \{(x_i, y_i)\}_{i=1}^t$ denote a collection of points and their corresponding observations obtained according to the model described in Sec. 2.2. Then, conditioned on $\mathcal{Z}_t$, the posterior distribution of f is also a GP with the following mean and covariance functions:

$$\mu_{t,\tau}(x) = k_{X_t,x}^\top (K_{X_t,X_t} + \tau I_t)^{-1} Y_t, \tag{2}$$

$$k_{t,\tau}(x,\bar{x}) = k(x,\bar{x}) - k_{X_t,x}^\top (K_{X_t,X_t} + \tau I_t)^{-1} k_{X_t,\bar{x}}, \tag{3}$$

where $k_{X_t,x} = [k(x_1,x),\dots k(x_t,x)]^\top$, $Y_t = [y_1,\dots,y_t]^\top$, $K_{X_t,X_t} = [k(x_i,x_j)]_{i,j=1}^t$ and I_t is the $t\times t$ identity matrix. The posterior variance at a point x is given as $\sigma_{t,\tau}^2(x) = k_{t,\tau}(x,x)$. The expression for posterior mean and variance in the noise-free setting is simply obtained by setting $\tau = 0$ in the above relations.

The posterior mean and variance computed using the GP model above are powerful tools to predict the values of the unknown function f and to quantify the uncertainty in the prediction. In particular, the prediction error at a point $x \in \mathcal{X}$, $|f(x) - \mu_{t,\tau}(x)|$, can be upper bounded by $\alpha\sigma_{t,\tau}(x)$, for a certain scaling factor $\alpha > 0$ that depends on the feedback model (Vakili et al., 2021a). Lastly, we define the information gain of a set of points $X_n = \{x_1, x_2, \dots, x_n\}$ as

$$\tilde{\gamma}_{X_n,\tau} := \frac{1}{2}\log\left(\det\left(I_t + \tau^{-1}K_{X_n,X_n}\right)\right). \tag{4}$$

Similarly, we define the maximal information gain as $\gamma_{n,\tau} := \sup_{X_n\subset\mathcal{X}^n} \tilde{\gamma}_{X_n,\tau}$. Maximal information gain is an important term that corresponds to the effective dimension of the kernel and helps characterize the regret of the algorithms. It depends only on the kernel and τ.

3. The Predictive Performance of Random Exploration

The following theorem characterizes the predictive variance, and consequently the predictive error, of a set of randomly sampled points from the domain.

Theorem 3.1. *Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$, ϱ be a finite Borel measure supported on $\mathcal{X}$, and $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a continuous kernel satisfying the polynomial eigendecay condition with parameter $\beta > 1$ (Defn. 2.2). Let $X_n = \{x_1, x_2, \dots, x_n\}$ denote a collection of n i.i.d. points drawn from $\mathcal{X}$ according to ϱ. Let $\sigma_{n,0}^2$ and $\sigma_{n,\tau}^2$ denote, respectively, the posterior variance conditioned on X_n in the noise-free setting and the noisy setting with a noise variance of $\tau > 0$. Then, for a given $\delta \in (0,1)$, there exists a constant $\overline{N}(\delta, k, \varrho, \tau) > 0$, such that, with probability at least $1-\delta$, for all $n > \overline{N}(\delta, k, \varrho, \tau)$,*

$$\sup_{x\in\mathcal{X}} \sigma_{n,\tau}^2(x) = \mathcal{O}\left(\frac{\tau\gamma_{n,\tau}}{n}\right) = \tilde{\mathcal{O}}((n/\tau)^{\frac{1}{\beta}-1}),$$

$$\sup_{x\in\mathcal{X}} \sigma_{n,0}^2(x) = \tilde{\mathcal{O}}(n^{1-\beta}).$$

The above obtained bounds on the worst-case posterior variance under the random exploration scheme are order-optimal (up to polylogarithmic factors), matching the existing lower bounds (Scarlett et al., 2017; Tuo & Wang, 2020). The above theorem also improves upon the best known results for noisy scattered data approximation. In particular, for the class of Matérn kernels with smoothness ν (i.e., $\beta = (2\nu + d)/d$), Theorem 3.1 implies a worst-case predictive error of $\tilde{\mathcal{O}}(n^{-\frac{\nu}{2\nu+d}})$, improving upon the bound of $\tilde{\mathcal{O}}(n^{-\frac{\nu}{2\nu+2d}})$ established by Wynne et al. (2021, Corollary 3).

The constant $\overline{N}(\delta, k, \varrho, \tau)$ is related to the kernel k and measure ϱ through two fundamental functions, $N(R)$ and $T(R)$, which are given as follows for any $R \in \mathbb{N}$:

$$N(R) := \sup_{x \in \mathcal{X}} \sum_{j=1}^{R} \varphi_j^2(x),$$

$$T(R) := \sup_{x \in \mathcal{X}} \sum_{j=R+1}^{\infty} \lambda_j \varphi_j^2(x) = \sup_{x \in \mathcal{X}} \sum_{j=R+1}^{\infty} \upsilon_j^2(x).$$

They are referred to as the spectral functions of the kernel (see Gröchenig (2020) and references therein) because of their dependence on the eigensystem corresponding to the kernel k induced by the measure ϱ. Both $N(R)$ and $T(R)$ are fundamental quantities that appear in the analysis of reconstruction and estimation of functions in general L_2 spaces. The function $N(R)$ corresponds to the inverse of the infimum of the Christoffel function (Dunkl & Xu, 2014) in the special case of reconstruction using orthogonal polynomials. Under Assumption 2.3 and the condition of polynomial eigendecay (Def. 2.2), $\overline{N}(\delta, k, \varrho, \tau)$ can be shown to be bounded as $\mathcal{O}(\max\{F^4, (F^2/\tau)^{\frac{1}{\beta-1}}\} \log(F/\delta))$. The dependence of $\overline{N}(\delta, k, \varrho, \tau)$ on δ is mild, as evident from the previous expression. Lastly, $\overline{N}(\delta, k, \varrho, \tau)$ is inversely proportional to τ. Note that Theorem 3.1 ensures that a smaller value of τ results in a tighter bound on the posterior variance, which in turn requires a larger number of samples. We refer the interested reader to the Appendix A for a more detailed discussion of $\overline{N}(\delta, k, \varrho, \tau)$ and its dependence on $N(R)$ and $T(R)$. For brevity, we drop the arguments and use the notation $\overline{N}$ in the rest of the paper.

We provide a sketch of the proof of Theorem 3.1 below and refer the reader to Appendix A for a detailed proof.

Proof. The main idea of the proof is to relate the worst-case posterior variance conditioned on X_n to $\tilde{\gamma}_{X_n,\tau}$. This relation is established in two parts. In the first part, we establish that as the number of samples grow, the spectrum of random operator $\hat{\mathbf{Z}}$ concentrates to that of $\mathbf{Z}$, where $\hat{\mathbf{Z}}, \mathbf{Z} : \mathcal{H}_k \to \mathcal{H}_k$ are defined as follows:

$$\hat{\mathbf{Z}} g := \left[\sum_{i=1}^{n} \langle g, \psi_{x_i} \rangle \psi_{x_i} \right] + \tau g; \quad \mathbf{Z} := \mathbb{E}_{X_n}[\hat{\mathbf{Z}}],$$

where $\{x_1, x_2, \ldots, x_n\}$ denotes the random ensemble of points drawn according to the measure ϱ. The concentration in spectral norm allows us to approximate the expression of $\sigma_{n,\tau}^2(x) = \tau \langle \psi_x, \hat{\mathbf{Z}}^{-1} \psi_x \rangle$ as $\sigma_{n,\tau}^2(x) \approx \tau \langle \psi_x, \mathbf{Z}^{-1} \psi_x \rangle$, i.e., by replacing the sample covariance operator, $\hat{\mathbf{Z}}$, with the true covariance operator, $\mathbf{Z}$. Here, A^{-1} denotes the inverse of an operator A, i.e., $A \circ A^{-1} = A^{-1} \circ A = \mathbf{Id}$ and $\mathbf{Id}$ denotes the identity operator. Thus, this step allows us to obtain a deterministic bound on posterior variance, which is easier to understand and analyze. We establish the required relation using the following two lemmas:

Lemma 3.2. *For all $n \geq \overline{N}$, the following relation holds with probability $1 - \delta/2$:*

$$\|\mathbf{Z}^{-\frac{1}{2}} \hat{\mathbf{Z}} \mathbf{Z}^{-\frac{1}{2}} - \mathbf{Id}\|_2 \leq 1/9.$$

Lemma 3.3. *If the relation $\|\mathbf{Z}^{-\frac{1}{2}} \hat{\mathbf{Z}} \mathbf{Z}^{-\frac{1}{2}} - \mathbf{Id}\|_2 \leq b$ is true for some $b \in (0, 1/3)$, then following is true $\forall\, x \in \mathcal{X}$:*

$$\langle \psi_x, \hat{\mathbf{Z}}^{-1} \psi_x \rangle \leq \frac{\sqrt{1-b}}{\sqrt{1-b} - \sqrt{2b}} \cdot \langle \psi_x, \mathbf{Z}^{-1} \psi_x \rangle.$$

Lemma 3.2 forms the cornerstone of the proof of the theorem. The result is established by bounding the expression $|\langle g, (\mathbf{Z}^{-1/2} \hat{\mathbf{Z}} \mathbf{Z}^{-1/2} - \mathbf{Id}) g \rangle|$ for an arbitrary g with $\|g\|_{\mathcal{H}_k} = 1$. We bound the above expression by decomposing it into a sum of three terms. Each of the three terms is then carefully bounded using a combination of Matrix-Chernoff inequality (Tropp, 2012, Theorem 1.1), a result for spectral norm concentration based on non-commutative Khinchtine inequality (Buchholz, 2001; 2005; Moeller & Ullrich, 2021) and Bernstein inequality. Lemma 3.3 is established using a combination the structure of covariance matrices, the Cauchy-Schwarz inequality and the relation between the operator norm and 2-norm. We would like to emphasize that both the above lemmas are true in general for *all* eigendecay profiles and even without Assumption 2.3 being true.

In the second part, we show that, with high probability, the information gain of the (random) set X_n is lower bounded by $n \cdot \sup_{x \in \mathcal{X}} \langle \psi_x, \mathbf{Z}^{-1} \psi_x \rangle$, upto a multiplicative constant. The above idea is formalized in the following lemma.

Lemma 3.4. *For all $n \geq \overline{N}$, the following relation holds with probability $1 - \delta/2$:*

$$\tilde{\gamma}_{X_n,\tau} \geq \frac{13}{54F^2} \cdot n \cdot \sup_{x \in \mathcal{X}} \langle \psi_x, \mathbf{Z}^{-1} \psi_x \rangle.$$

Thus $\langle \psi_x, \mathbf{Z}^{-1} \psi_x \rangle$ serves as the bridge for connecting the posterior variance to maximal information gain.

The result for the noisy case follows immediately from the above lemmas by noting that $\gamma_{X_n,\tau} \leq \gamma_{n,\tau}$. For the noise-free setting, the results do not carry forward immediately as the above analysis does not hold for $\tau = 0$. To circumvent this issue, we use the fact that $\sigma_{n,\tau}^2(x)$ is an increasing function of τ. Thus, we obtain a bound on $\sigma_{n,0}^2(x)$ by using the bound on $\sigma_{n,\tau^*}^2(x)$, where τ^* is a carefully chosen value that not only allows us to use the analysis from the noisy case but also ensures that σ_{n,τ^*}^2 is a close representation of $\sigma_{n,0}^2$ to guarantee tightest possible bounds. □

Remark 3.5. We would like to emphasize that the above result holds for samples generated under *every* finite Borel measure ϱ supported on $\mathcal{X}$. However, the quality of the estimate changes with the choice of the measure through the leading constant in the bound in Theorem 3.1.

4. The REDS algorithm

In this section, we present the proposed algorithm and analyze its regret performance.

4.1. REDS with Noise-Free Feedback

REDS integrates random exploration with domain shrinking. It proceeds in epochs, maintaining an active region $\mathcal{X}_r$ of the domain during each epoch $r \geq 1$. The sequence of active regions $\{\mathcal{X}_r\}_r$ shrinks across epochs, i.e., $\mathcal{X}_r \subseteq \mathcal{X}_{r-1} \subseteq \ldots \mathcal{X}_1 = \mathcal{X}$, while ensuring $x^* \in \mathcal{X}_r$ for all r with high probability. During the rth epoch, REDS samples N_r points, uniformly at random from the set $\mathcal{X}_r$, where $N_r = N_1 \cdot 2^{r-1}$ and the initial batch size N_1 is an input to the algorithm. If $\mathcal{X}_r$ consists of multiple disjoint regions, then we carry out this step for each region separately.

Using the observations from these points, REDS computes the posterior mean and variance function over $\mathcal{X}_r$, denoted by μ_r and σ_r^2 respectively, using the Equations (2) and (3) with $\tau = 0$. The posterior mean and variance are then used to obtain $\mathcal{X}_{r+1}$, an improved localization of x^*, as follows:

$$\mathcal{X}_{r+1} = \left\{ x \in \mathcal{X}_r \;\middle|\; \mathrm{UCB}_r(x) \geq \sup_{x' \in \mathcal{X}_r} \mathrm{LCB}_r(x') \right\}.$$

Here, $\mathrm{UCB}(x) = \mu_r(x) + B\sigma_r(x)$ and $\mathrm{LCB}(x) = \mu_r(x) - B\sigma_r(x)$ correspond to upper and lower bounds on the estimate of f. A pseudocode for the algorithm is provided in Algorithm 1.

4.2. REDS under noisy feedback

The REDS algorithm can be extended to operate under noisy feedback with the following two minor modifications to Algorithm 1. First, the posterior mean and variance $(\mu_{r,\tau},\ \sigma_{r,\tau}^2)$ in each epoch should be computed using a noise variance $\tau > 0$ (Line 9 of Algorithm 1). Second, the upper and lower confidence bounds, i.e., UCB and LCB (Line 10 of Algorithm 1), should be updated to the following:

$$\mathrm{UCB}_{r,\tau,\delta}(x) := \mu_{r,\tau}(x) + \alpha_{\tau,\delta}\sigma_{r,\tau}(x) + c_{T,\tau,\delta} \quad (5)$$
$$\mathrm{LCB}_{r,\tau,\delta}(x) := \mu_{r,\tau}(x) - \alpha_{\tau,\delta}\sigma_{r,\tau}(x) - c_{T,\tau,\delta}, \quad (6)$$

where $\alpha_{\tau,\delta} = B + R\sqrt{(2/\tau)\log(|\mathcal{D}_T|/\delta)}$, $c_{T,\tau,\delta} = \frac{2B}{T} + R\sqrt{\frac{2}{T\tau}\log\left(\frac{4T}{\delta}\right)}$ and $\mathcal{D}_T$ is defined in Assumption 4.1.

Algorithm 1 Random Exploration with Domain Shrinking

1: **Input**: N_1, the initial batch size.
2: Set $\mathcal{X}_1 \leftarrow \mathcal{X}$, $t_{\text{curr}} \leftarrow 0$, $r \leftarrow 1$
3: **for** $t = t_{\text{curr}} + 1, t_{\text{curr}} + 2, \ldots, t_{\text{curr}} + N_r$ **do**
4: Sample a point x_t uniformly at random from $\mathcal{X}_r$ and observe y_t
5: **if** $t > T$ **then**
6: Terminate
7: **end if**
8: **end for**
9: Construct μ_r and σ_r based on observations $\{(x_t, y_t : t \in \{t_{\text{curr}}+1, t_{\text{curr}}+2, \ldots, N_r\}\}$ using Eqn (2) and (3) with $\tau = 0$.
10: Set $\mathcal{X}_{r+1} = \{x \in \mathcal{X}_r \mid \mathrm{UCB}_r(x) \geq \sup_{x' \in \mathcal{X}_r} \mathrm{LCB}_r(x')\}$
11: $t_{\text{curr}} \leftarrow t_{\text{curr}} + N_r$, $N_{r+1} \leftarrow 2N_r$
12: $r \leftarrow r + 1$

4.3. Performance Analysis

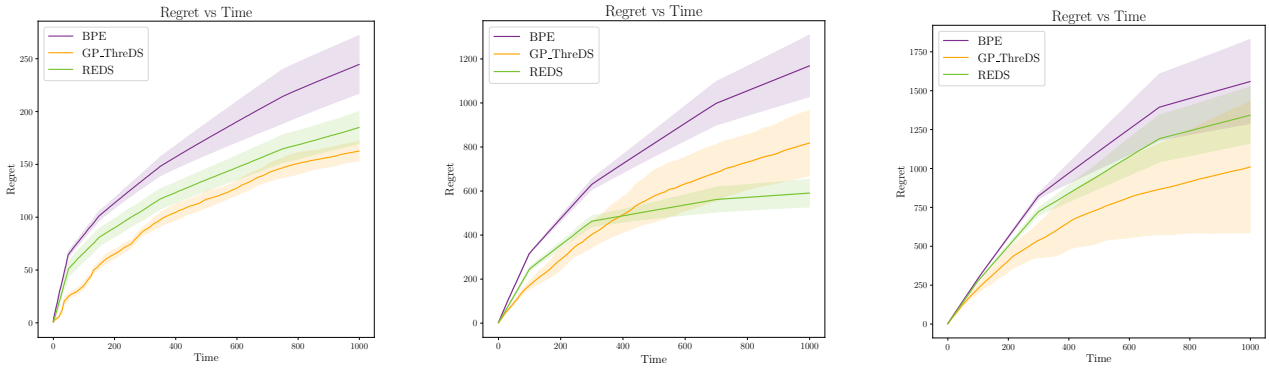
For the analysis of the REDS algorithm, we need to make the following two additional assumptions.

Assumption 4.1. For all $n \in \mathbb{N}$, there exists a discretization $\mathcal{D}_n$ of $\mathcal{X}$ such that for all $f \in \mathcal{H}_k$, $|f(x) - f([x]_{\mathcal{D}_n})| \leq \|f\|_{\mathcal{H}_k}/n$ and $|\mathcal{D}_n| = \mathrm{poly}(n)$[2], where $[x]_{\mathcal{D}_n} = \arg\min_{y \in \mathcal{D}_n} \|x - y\|_2$, is the point in $\mathcal{D}_n$ that is closest to x.

Assumption 4.2. Let $\mathcal{L}_\eta^f = \{x \in \mathcal{X} | f(x) \geq \eta\}$ denote the level set of f for $\eta \in [-B, B]$. Let $\mathcal{X}'$ be a subset of $\mathcal{L}_\eta^f$ with a finite number of connected components. Let $\mathrm{UCB}_t(x; \mathcal{X}')$ denote the upper confidence bound on the function f at any point $x \in \mathcal{X}'$, constructed using t points sampled uniformly at random from each component of $\mathcal{X}'$. For any $\eta' \geq \eta$, we define $\mathcal{L}_{\eta'}^{\mathrm{UCB}_t} = \{x \in \mathcal{X}' | \mathrm{UCB}_t(x; \mathcal{X}') \geq \eta'\}$ to be the level set of the upper confidence bound at level η'. Let $\eta_0 := \sup f - \varepsilon_0$ for some fixed, known $\varepsilon_0 > 0$. We assume the following for all $\mathcal{X}'$ and $t \geq \overline{N}$.

1. For any $\eta \geq \eta_0$, the number of connected components in $\mathcal{L}_\eta^f$ are at most M_f.
2. For any $\eta' \geq \eta \geq \eta_0$, the number of connected components of $\mathcal{L}_{\eta'}^{\mathrm{UCB}_t}$ are at most M more than those of $\{x \in \mathcal{X}' : f(x) \geq \eta'\}$ with probability $1 - \delta/(2\log_2 T)$, where the probability is taken over the randomness in query points and noise (if any).
3. For each such component of $\mathcal{L}_{\eta'}^{\mathrm{UCB}_t}$, there exists a bi-Lipschitzian map between each such component and $\mathcal{X}$ with normalized Lipschitz constant pair $L, L' < \infty$.

[2]The notation $f(x) = \mathrm{poly}(x)$ is equivalent to $f(x) = \mathcal{O}(x^k)$ for some $k \in \mathbb{N}$.



(a) Branin (b) Hartmann-4D (c) Hartmann-6D

Figure 1: Cumulative regret averaged over 10 Monte Carlo runs for all algorithms across different benchmark functions. The shaded region represents the error bars upto one standard deviation. As evident from the plots, the regret of REDS is comparable to that of BPE and GP-ThreDS.

Assumption 4.1 is only required for the noisy case and is a standard assumption adopted in the literature. The existence of such a discretization has been justified and adopted in previous studies (Srinivas et al., 2010; Chowdhury & Gopalan, 2017; Vakili et al., 2021a; Salgia et al., 2022) and is a mild assumption on the kernel. Specifically, the popular class of kernels like Squared Exponential and Matérn kernels are known to be Lipschitz continuous, in which case a ε-cover of the domain with $\varepsilon = \mathcal{O}(1/n)$ is sufficient to show the existence of such a discretization. Assumption 4.2 is an assumption on the regularity of the level sets of the function f and the UCB. The existence of a bi-Lipschitzian map between two sets implies topological similarity between the two sets. Intuitively, this assumption ensures that the shape of the level-sets is not "too arbitrary". Note that such an assumption on the level sets of UCB is relatively mild as the RKHS endows smoothness properties to the UCB which translate to a degree of topological regularity of level sets (Alberti et al., 2011; Lee, 2010). We require Assumption 4.2 in conjunction with separate sampling of disjoint regions for simplicity of analysis. We believe that with refined analysis techniques the need for this assumption, along with the need to separate sampling of disjoint regions, can be eliminated. We leave developing such refined analysis techniques to future work.

The following theorem characterizes the regret performance of REDS under noise-free feedback.

Theorem 4.3. *Assume that the kernel k satisfies the polynomial eigendecay condition with parameter $\beta > 1$ and function f satisfies Assumption 4.2. For a given $\delta \in (0,1)$, if REDS algorithm is run with $N_1 \geq \max\{C_{L_f,L'_f}\overline{N}(\delta/4\log_2(T)), \tilde{N}_{\varepsilon_0}\}$ and noise-free feedback, then the regret incurred by REDS satisfies,*

$$R(T) = \tilde{\mathcal{O}}(\max\{T^{\frac{3-\beta}{2}}, 1\}).$$

with probability at least $1-\delta$. Here, C_{L_f,L'_f} is a constant that depends only on L_f and L'_f and $\tilde{N}_{\varepsilon_0}$ is a constant that depends only on ε_0[3] and is independent of T.

The following is an immediate corollary of the above theorem for the case of Matérn kernels.

Corollary 4.4. *Let k be the Matérn kernel with smoothness $\nu > 0$. For a given $\delta \in (0,1)$, if REDS algorithm is run with $N_1 \geq \max\{C_{L_f,L'_f}\overline{N}(\delta/4\log_2(T)), \tilde{N}_{\varepsilon_0}\}$ under noise-free feedback on a function $f \in \mathcal{H}_k$ satisfying Assumption 4.2, then the regret incurred by REDS satisfies,*

$$R(T) = \begin{cases} \tilde{\mathcal{O}}(T^{1-\nu/d}) & \text{if } \nu < d, \\ \mathcal{O}((\log T)^{5/2}) & \text{if } \nu = d, \\ \mathcal{O}((\log T)^{3/2}) & \text{if } \nu > d. \end{cases}.$$

with probability at least $1-\delta$. Here, C_{L_f,L'_f} is a constant that depends only on L_f and L'_f and $\tilde{N}_{\varepsilon_0}$ is a constant that depends only on ε_0 and is independent of T.

This matches the result conjectured in Vakili (2022) upto logarithmic factors, *resolving the open problem*.

The following theorem characterizes the regret performance of REDS in the noisy feedback setting.

Theorem 4.5. *Consider the noisy observation model described in Sec. 2.2 and assume that Assumptions 4.1 and 4.2 hold. For a given $\delta \in (0,1)$, if REDS algorithm is run with*

[3]Please refer to Appendix B for additional details.

$N_1 \geq \max\{C_{L_f,L'_f}\overline{N}(\delta/(6\log_2 T)), \tilde{N}'_{\varepsilon_0}\}$ and UCB and LCB functions as defined in Eqns. (5) and (6) with parameter $\delta' = \delta/(6\log_2 T)$, then the regret incurred by REDS satisfies,

$$R(T) = \tilde{\mathcal{O}}(\sqrt{T\gamma_T}\log(T/\delta)).$$

with probability at least $1-\delta$. Here $\tilde{N}'_{\varepsilon_0}$ is a constant that depends only on ε_0 and is independent of T.

As shown by the above theorem, REDS achieves order-optimal regret (upto logarithmic factors) even under the noisy feedback model.

The proofs of both Theorems 4.3 and 4.5 follow a similar blueprint. A key aspect of both the proofs is to ensure that as Theorem 3.1 is invoked across the sets $\{\mathcal{X}_r\}_{r\in\mathbb{N}}$, the leading constant in Theorem 3.1, which has an implicit dependence on the domain through the constant F, remains bounded and is independent of T. The following lemma shows that for all functions f satisfying Assumption 4.2, the leading constant only depends on the function and the initial domain.

Lemma 4.6. *Let $f \in \mathcal{H}_k$ be such that Assumption 4.2 holds. Let $\mathcal{X}'$ denote a path connected component of any level set of f and $X' \subset \mathcal{X}'$ be a set of n points drawn uniformly at random from $\mathcal{X}'$. Then for $n \geq C_{L,L'_f}\overline{N}(\delta)$, the following relations holds with probability $1-\delta$:*

$$\sup_{x\in\mathcal{X}'} \sigma^2_{X',\tau}(x) \leq C'_{L,L'_f} \cdot F^2\tau \cdot \frac{\gamma_{n,\tau}}{n}$$
$$\sup_{x\in\mathcal{X}'} \sigma^2_{X',0}(x) \leq C'_{L,L'_f} \cdot F^2 \cdot n^{1-\beta}$$

where F and $\overline{N}(\delta)$ represent, respectively, the constants in Assumption 2.3 and Theorem 3.1 corresponding to the uniform measure on $\mathcal{X}$, and $C_{L,L'_f}, C'_{L_f,L'_f}$ are constants that depend only on L_f, L'_f.

At a high level, the above lemma ensures that under the regularity condition on the topology of level sets (Assumption 4.2), Theorem 3.1 can be applied across level sets of f by just paying the penalty of a constant that depends only on f. The proof is based on the inclusion of RKHSs over subsets along with a change of measure argument. We refer the reader to Appendix B for a detailed proof of Lemma 4.6 and Theorems 4.3 and 4.5.

5. Empirical Studies

We compare the computational efficiency of REDS against algorithms with order-optimal regret performance, namely BPE (Li & Scarlett, 2022) and GP-ThreDS (Salgia et al., 2021) through an empirical study. We compare the regret performance and the running time of the three algorithms for three commonly used benchmark functions in Bayesian Optimization, namely, Branin (Azimi et al., 2012; Picheny et al., 2013), Hartmann-4D (Picheny et al., 2013) and Hartmann-6D (Picheny et al., 2013). All the functions are defined over $\mathcal{X} = [0,1]^d$, with $d = 2, 4, 6$ for Branin, Hartmann-4D and Hartmann-6D respectively.

	BPE	GP-ThreDS	REDS
Branin	29.84	4.37	**0.32**
Hartmann-4D	38.45	7.59	**0.47**
Hartmann-6D	119.71	19.33	**1.19**

Table 1: Time taken (in seconds) by different algorithms across the different benchmark functions.

For all the experiments, we use the Square exponential kernel. The length scale was set to 0.2 for Branin and 1 for Hartmann-4D and Hartmann-6D functions. We corrupted the observations with a zero mean Gaussian noise to the with a standard deviation of 0.2. The value of τ was also set to 0.2. All the algorithms were run for $T = 1000$ time steps. We recorded the cumulative regret and time taken by different algorithms for 10 Monte Carlo runs for each benchmark function. We defer the reader to Appendix C for additional details of the experimental setup and the choice of hyperparameters for the algorithms.

The cumulative regret for all the algorithms over different functions is plotted in Figure 1. The shaded region corresponds to the error bars for one standard deviation on both sides of the mean. We tabulate the (mean) running time for all the algorithms over different functions in Table 1. The values in Table 1 refer to the wall clock time taken by the algorithms. As evident from the plots in Figure 1, the regret incurred by REDS is comparable to that of other algorithms for all benchmark functions. Thus, our algorithm based on non-adaptive randomized samples offers the same regret performance as the best performing algorithms based on adaptive sampling. Moreover, REDS offers the comparable regret performance at a much lower computational cost and runtime. Specifically, REDS offers about a $15\times$ and $100\times$ speedup in terms of runtime over the GP-ThreDS and BPE (See Table 1). The significant improvement in runtime without loss of performance in regret demonstrates the practical benefits of our proposed methodology of random sampling.

6. Conclusion

In this work, we studied the methodology of exploring the domain using random samples drawn from a distribution supported on a compact domain. We showed that this non-adaptive approach offers the optimal-order of worst case predictive error for RKHS function in both noisy and noise-free feedback settings. The proposed approach offers a simple alternative for designing Bayesian Optimization al-

gorithms which typically involve choosing points through a computationally expensive step of optimizing a non-convex acquisition function. Based on this methodology, we developed a algorithm that achieves order-optimal regret in both noisy and noise-free settings, *partially resolving a COLT open problem*. We demonstrated the computational advantage of the proposed approach through an empirical study, where the proposed algorithm achieved upto a $100\times$ runtime speed up over state-of-the-art algorithms.

Acknowledgements

The authors would like to thank the Reviewers and the Area Chair for their constructive feedback and discussion, which has greatly helped improve the quality of the paper.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alberti, G., Bianchini, S., and Crippa, G. Structure of level sets and Sard-type properties of lipschitz maps. *Annali della Scuola Normale Superiore di Pisa. Classe di Scienze. Serie V*, 4, 08 2011. doi: 10.2422/2036-2145.201107_006.

Arcangéli, R., López de Silanes, M. C., and Torrens, J. J. Extension of sampling inequalities to Sobolev semi-norms of fractional order and derivative data. *Numerische Mathematik*, 121(3):587–608, 2012. ISSN 0029599X. doi: 10.1007/s00211-011-0439-3.

Azimi, J., Jalali, A., and Fern, X. Z. Hybrid batch bayesian optimization. In *Proceedings of the 29th International Conference on Machine Learning, ICML*, volume 2, pp. 1215–1222, 2012. ISBN 9781450312851.

Bastian Bohn. *Error analysis of regularized and unregularized least-squares regression on discretized function spaces*. PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, 2017. URL https://hdl.handle.net/20.500.11811/7094.

Bohn, B. On the convergence rate of sparse grid least squares regression. In *Sparse Grids and Applications*, pp. 19–41. Springer International Publishing, 2018. ISBN 978-3-319-75426-0.

Bohn, B. and Griebel, M. Error estimates for multivariate regression on discretized function spaces. *SIAM Journal on Numerical Analysis*, 55(4):1843–1866, 2017.

Brenner, S. C., Scott, L. R., and Scott, L. R. *The mathematical theory of finite element methods*, volume 3. Springer, 2008.

Brochu, E., Cora, V. M., and De Freitas, N. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning, 2010.

Buchholz, A. Operator khintchine inequality in non-commutative probability. *Mathematische Annalen*, 319 (1):1–16, 2001.

Buchholz, A. Optimal constants in khintchine type inequalities for fermions, rademachers and q-gaussian operators. *Bulletin of The Polish Academy of Sciences Mathematics*, 53:315–321, 2005. URL https://api.semanticscholar.org/CorpusID:55683104.

Bull, A. D. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12:2879–2904, 2011. ISSN 15324435.

Camilleri, R., Katz-Samuels, J., and Jamieson, K. High-Dimensional Experimental Design and Kernel Bandits. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021. URL https://arxiv.org/abs/2105.05806v1http://arxiv.org/abs/2105.05806.

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

Chkifa, A., Cohen, A., Migliorati, G., Nobile, F., and Tempone, R. Discrete least squares polynomial approximation with random evaluations- application to parametric and stochastic elliptic pdes. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 49(3):815–837, 2015.

Chowdhury, S. R. and Gopalan, A. On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning, ICML*, volume 2, pp. 1397–1422, 2017. ISBN 9781510855144.

Cohen, A. and Migliorati, G. Optimal weighted least-squares methods. *The SIAM journal of computational mathematics*, 3:181–203, 2017.

Cohen, A., Davenport, M., and Leviatan, D. On the stability and accuracy of least squares approximations. *Foundations of Computational Mathematics*, 13:819–834, 2013.

De Freitas, N., Smola, A. J., and Zoghi, M. Exponential regret bounds for Gaussian process bandits with deterministic observations. In *Proceedings of the 29th International*

Conference on Machine Learning, ICML, volume 2, pp. 1743–1750, 2012. ISBN 9781450312851.

Dunkl, C. F. and Xu, Y. *Orthogonal Polynomials of Several Variables*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2 edition, 2014. doi: 10.1017/CBO9781107786134.

Frazier, P. I., Powell, W. B., and Dayanik, S. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.

Greenhill, S., Rana, S., Gupta, S., Vellanki, P., and Venkatesh, S. Bayesian optimization for adaptive experimental design: A review. *IEEE access*, 8:13937–13948, 2020.

Gröchenig, K. Sampling, marcinkiewicz–zygmund inequalities, approximation, and quadrature rules. *Journal of Approximation Theory*, 257:105455, 2020.

Grünewälder, S., Audibert, J.-Y., Opper, M., and Shawe-Taylor, J. Regret bounds for gaussian process bandit problems. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 273–280, 2010.

Jones, D. R. A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, 21:345–383, 2001.

Jones, D. R., Schonlau, M., and Welch, W. J. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 1998.

Kämmerer, L., Ullrich, T., and Volkmer, T. Worst-case recovery guarantees for least squares approximation using random samples. *Constructive Approximation*, 54(2):295–352, 2021.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences, 2018.

Krieg, D. and Ullrich, M. Function values are enough for l2-approximation. *Foundations of Computational Mathematics*, 21:1141–1151, 2021a. doi: https://doi.org/10.1007/s10208-020-09481-w.

Krieg, D. and Ullrich, M. Function values are enough for l2-approximation: Part ii. *Journal of Complexity*, 66, 2021b. ISSN 0885-064X. doi: https://doi.org/10.1016/j.jco.2021.101569.

Kushner, H. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering*, 86:97–106, 1964.

Lee, J. *Introduction to Topological Manifolds*. Springer, 2010.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization, 2016.

Li, Z. and Scarlett, J. Gaussian process bandit optimization with few batches. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, AISTATS*, 2022.

Lizotte, D. J., Wang, T., Bowling, M. H., and Schuurmans, D. Automatic gait optimization with gaussian process regression. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI)*, volume 7, pp. 944–949, 2007.

Lyu, Y., Yuan, Y., and Tsang, I. W. Efficient batch black-box optimization with deterministic regret bounds, 2020.

Močkus, J. On bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference*, pp. 400–404, Berlin, Heidelberg, 1975. Springer Berlin Heidelberg. ISBN 978-3-540-37497-8.

Moeller, M. and Ullrich, T. L 2-norm sampling discretization and recovery of functions from rkhs with finite trace. *Sampling Theory, Signal Processing, and Data Analysis*, 19(2):13, 2021.

Močkus, J., Tiesis, V., and Žilinskas, A. *Towards Global Optimization*, volume 2, chapter The application of Bayesian methods for seeking the extremum, pp. 117–129. Elsevier, 09 1978. ISBN 0-444-85171-2.

Narcowich, F. J., Ward, J. D., and Wendland, H. Sobolev error estimates and a bernstein inequality for scattered data interpolation via radial basis functions. *Constructive Approximation*, 24:175–186, 2006.

Ostrowski, A. M. A quantitative formulation of slyvester's law of inertia. *Proceedings of the National Academy of Sciences*, 45(5):740–744, 1959. doi: 10.1073/pnas.45.5.740. URL https://www.pnas.org/doi/abs/10.1073/pnas.45.5.740.

Picheny, V., Wagner, T., and Ginsbourger, D. A benchmark of kriging-based infill criteria for noisy optimization. *Structural and Multidisciplinary Optimization*, 48(3):607–626, 2013. ISSN 1615147X. doi: 10.1007/s00158-013-0919-4. URL https://link.springer.com/article/10.1007/s00158-013-0919-4.

Riutort-Mayol, G., Bürkner, P.-C., Andersen, M. R., Solin, A., and Vehtari, A. Practical hilbert space approximate bayesian gaussian processes for probabilistic programming. *Statistics and Computing*, 33(1):17, 2023.

Rudin, W. *Real and complex analysis, 3rd ed.* McGraw-Hill, Inc., USA, 1987. ISBN 0070542341.

Salgia, S., Vakili, S., and Zhao, Q. A domain-shrinking based Bayesian optimization algorithm with order-optimal regret performance. In *Proceedings of the 35th Annual Conference on Neural Information Processing Systems*, volume 34, 2021.

Salgia, S., Vakili, S., and Zhao, Q. Collaborative Learning in Kernel-based Bandits for Distributed Users, 2022.

Scarlett, J., Bogunovic, I., and Cehver, V. Lower Bounds on Regret for Noisy Gaussian Process Bandit Optimization. In *Conference on Learning Theory*, volume 65, pp. 1–20, 2017.

Smale, S. and Zhou, D.-X. Shannon sampling and function reconstruction from point values. *Bulletin of The American Mathematical Society*, 41:279–306, 2004. doi: 10.1090/S0273-0979-04-01025-0.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning, ICML*, pp. 1015–1022, 2010. ISBN 9781605589077. doi: 10.1109/TIT.2011.2182033.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer, 2008. doi: https://doi.org/10.1007/978-0-387-77242-4.

Törn, A. and Žilinskas, A. *Global Optimization*. Springer Berlin, Heidelberg, 1989.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012.

Tuo, R. and Wang, W. Kriging prediction with isotropic matérn correlations: Robustness and experimental designs. *The Journal of Machine Learning Research*, 21(1): 7604–7641, 2020.

Vakili, S. Open problem: Regret bounds for noise-free kernel-based bandits. In *Proceedings of 35th Conference on Learning Theory (COLT)*, volume 178, pp. 5624–5629, 2022.

Vakili, S., Bouziani, N., Jalali, S., Bernacchia, A., and Shiu, D.-s. Optimal order simple regret for Gaussian process bandits. In *Proceedings of the 35th Annual Conference on Neural Information Processing Systems*, 2021a.

Vakili, S., Khezeli, K., and Picheny, V. On information gain and regret bounds in Gaussian process bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, AISTATS*, 2021b.

Valko, M., Korda, N., Munos, R., Flaounas, I., and Cristianini, N. Finite-time analysis of kernelised contextual bandits. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence, UAI*, pp. 654–663, 2013.

Vanchinathan, H. P., Nikolic, I., De Bona, F., and Krause, A. Explore-exploit in top-n recommender systems via gaussian processes. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pp. 225–232, 2014.

Vazquez, E. and Bect, J. Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *Journal of Statistical Planning and Inference*, 140(11):3088–3095, 2010. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2010.04.018.

Wasserman, L. Lecture notes on statistical methods for machine learning, 2008. URL https://www.stat.cmu.edu/~larry/=sml/Concentration.pdf.

Wendland, H. *Scattered Data Approximation*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511617539.

Wenzel, T., Santin, G., and Haasdonk, B. A novel class of stabilized greedy kernel approximation algorithms: Convergence, stability and uniform point distribution. *Journal of Approximation Theory*, 262, 2021. ISSN 10960430. doi: 10.1016/j.jat.2020.105508.

Wynne, G., Briol, F.-X., and Girolami, M. Convergence guarantees for gaussian process means with misspecified likelihoods and smoothness. *The Journal of Machine Learning Research*, 22(1):5468–5507, 2021.

A. Proof of Theorem 3.1

We begin with setting up some notation that will be used throughout the proof. Throughout the appendix, we will represent the elements of $\mathcal{H}_k$ as infinite dimensional vectors and operators over these function spaces as infinite dimensional matrices. We adopt such a convention for ease for presentation while keeping in mind that despite the matrix representation, the actual operation is over elements of $\mathcal{H}_k$. Recall that we defined the sample covariance operator $\hat{\mathbf{Z}}$ for a randomly chosen sample $X_n = \{x_1, x_2, \dots, x_n\}$ and its expected value $\mathbf{Z} = \mathbb{E}[\hat{\mathbf{Z}}]$ as follows for any $g \in \mathcal{H}_k$:

$$\hat{\mathbf{Z}}g := \left[\sum_{i=1}^{n} \langle g, \psi_{x_i}\rangle \psi_{x_i}\right] + \tau g$$
$$\mathbf{Z} := \mathbb{E}[\hat{\mathbf{Z}}].$$

In the matrix-vector notation, the operators (equivalently, matrices) are given as:

$$\hat{\mathbf{Z}} := \left(\sum_{i=1}^{n} \psi_{x_i}\psi_{x_i}^\top\right) + \tau\mathbf{Id}$$
$$\mathbf{Z} = \mathbb{E}[\hat{\mathbf{Z}}] = \mathbb{E}\left[\sum_{i=1}^{n} \psi_{x_i}\psi_{x_i}^\top\right] + \tau\mathbf{Id}$$
$$= n\mathbb{E}[\psi_{x_1}\psi_{x_1}^\top] + \tau\mathbf{Id} = n\mathbf{\Lambda} + \tau\mathbf{Id},$$

where $\mathbf{Id}$ is the identity matrix (operator) and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots)$ is the diagonal matrices consisting of the eigenvalues of the kernel k corresponding to the measure ϱ. If we define $\mathbf{\Psi}_n := [\psi_{x_1}, \psi_{x_2}, \dots, \psi_{x_n}]$, then we can also write $\hat{\mathbf{Z}} = \mathbf{\Psi}_n\mathbf{\Psi}_n^\top + \tau\mathbf{Id}$. Consequently, the posterior variance at any point $x \in \mathcal{X}$ is given as:

$$\sigma_{n,\tau}^2(x) = \tau\psi_x^\top\hat{\mathbf{Z}}^{-1}\psi_x.$$

For any $R \in \mathbb{N}$, we define the following two quantities that will be relevant during our analysis:

$$N(R) := \sup_{x\in\mathcal{X}} \sum_{j=1}^{R} \varphi_j^2(x), \tag{7}$$

$$T(R) := \sup_{x\in\mathcal{X}} \sum_{j=R+1}^{\infty} \lambda_j\varphi_j^2(x) = \sup_{x\in\mathcal{X}} \sum_{j=R+1}^{\infty} \upsilon_j^2(x). \tag{8}$$

Recall that $\{\varphi_j\}_{j\in\mathbb{N}}$ are eigenfunctions of the kernel operator and form an orthonormal system in $L_2(\varrho, \mathcal{X})$ and $\{\upsilon\}_j$ are an orthonormal basis for $\mathcal{H}_k$. The term $N(R)$ is often referred to as the spectral function (see (Gröchenig, 2020) and references therein) and in case of orthogonal polynomials, it is the inverse of the infimum of the Christoffel function (Dunkl & Xu, 2014). Both $N(R)$ and $T(R)$ are fundamental quantities that appear in the analysis of reconstruction and estimation of functions.

Lastly, based on $N(R)$ and $T(R)$, for a given kernel k, measure ϱ and $\delta \in (0, 1)$, we define the following terms for any $n \in \mathbb{N}$ and $\tau > 0$:

$$\mathcal{R}_{k,\varrho}^{(1)}(n, \tau, \delta) := \left\{R \in \mathbb{N} : N(R) \le \frac{n}{1944\log(6n/\delta)}\right\}$$
$$\mathcal{R}_{k,\varrho}^{(2)}(n, \tau, \delta) := \left\{R \in \mathbb{N} : \max\{42T(R), n\lambda_{R+1}\}\log\left(\frac{12}{\delta}\right) \le \frac{\tau}{27}\right\}$$
$$\mathcal{R}_{k,\varrho}(n, \tau, \delta) := \mathcal{R}_{k,\varrho}^{(1)}(n, \tau, \delta) \cap \mathcal{R}_{k,\varrho}^{(2)}(n, \tau, \delta)$$
$$\overline{N}(k, \varrho, \delta, \tau) := \max\left\{\min\left\{n : \mathcal{R}_{k,\varrho}(n, \tau, \delta) \neq \emptyset\right\}, \lceil 729 \cdot F^4 \cdot \log(12/\delta)\rceil\right\}$$

The dependence on k and ϱ is implicit through $\{\varphi_j\}_{j\in\mathbb{N}}$ and $\{\lambda_j\}_{j\in\mathbb{N}}$ used to define $N(R)$ and $T(R)$. For brevity of notation, going forward, we drop the explicit description of dependence on k and ϱ.

We are now ready to prove the theorem. We first prove the statement of the theorem, assuming that the lemmas hold, followed by the proofs of the lemmas.

We begin with result for the noisy case, where $\tau > 0$ is fixed (independent of n). From Lemma 3.2, we know that for $n \geq \overline{N}$, $\|\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id}\|_2 \leq 1/9$ holds with probability $1-\delta$. Using this result along Lemma 3.3, we can conclude that $\psi_x^\top \hat{\mathbf{Z}}^{-1}\psi_x \leq 2\psi_x^\top \mathbf{Z}^{-1}\psi_x$ holds for all x. Thus, we have,

$$\begin{aligned}
\sigma_{n,\tau}^2(x) &= \tau\psi_x^\top \hat{\mathbf{Z}}^{-1}\psi_x \\
&\leq 2\tau\psi_x^\top \mathbf{Z}^{-1}\psi_x \\
&\leq \frac{108F^2}{13}\cdot\tau\cdot\frac{\tilde{\gamma}_{X_n,\tau}}{n} \\
&\leq \frac{108F^2}{13}\cdot\tau\cdot\frac{\gamma_{n,\tau}}{n},
\end{aligned} \tag{9}$$

as required. The third line in the above expression follows from Lemma 3.4. We would like to emphasize that the polynomial eigendecay condition is not necessary to obtain the above relation. It is only necessary to bound the information gain in terms on n. Under the polynomial eigendecay condition with parameter $\beta > 1$, the above equation can also be written as

$$\sigma_{n,\tau}^2(x) \leq C_0 \cdot \left(\frac{n}{\tau}\right)^{\frac{1}{\beta}-1}\log(n),$$

where we used the bound on information gain from Vakili et al. (2021b, Corollary 1) and C_0 is an appropriately chosen constant independent of n and τ.

We now consider the noise-free case. Since information gain is only defined for $\tau > 0$, we cannot directly extend the analysis as used in the noisy case by substituting $\tau = 0$. To circumvent this issue, we carefully choose $\tau^* > 0$, such that σ_{n,τ^*}^2 is a close representation of $\sigma_{n,0}^2$. We choose τ^* to be dependent on n such that τ^* goes to 0 as n becomes larger. This allows σ_{n,τ^*}^2 to faithfully represent the value of $\sigma_{n,0}^2$ over the range of n. Specifically, we choose $\tau^* = c' n^{1-\beta}(\log(n/\delta))^\beta$ for $c' \geq C(1944F^2)^\beta$, where C is the constant in Assumption 2.3. The condition on constant c' ensures that $\overline{N}(k, \varrho, \delta, \tau^*)$ exists. Since all conditions of the analysis for $\tau > 0$ (noisy case) are satisfied, we can directly invoke the result for $\tau > 0$. Using the bound on $\sigma_{n,\tau}^2$ and the monotonicity of $\sigma_{n,\tau}^2$ as a function of τ, we obtain,

$$\sigma_{n,0}^2(x) \leq \sigma_{n,\tau^*}^2(x) \leq C_1 \cdot n^{1-\beta}(\log(n/\delta))^\beta, \tag{10}$$

where C_1 is a constant independent of n.

In the following subsections, we prove Lemmas 3.2, 3.3 and 3.4.

A.1. Proof of Lemma 3.2

Since we are interested in bounding the 2-norm of the operator $\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id}$, we will focus on finding an upper bound on $g^\top(\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id})g$ that holds uniformly for all functions g in the unit ball in RKHS, i.e., $\{g : \|g\|_{\mathcal{H}_k} \leq 1\}$. The high level idea is to separately consider the contribution of component of g that belongs to the subspace spanned by eigenfunctions corresponding to the "large" eigenvalues, i.e., head of the spectrum and those corresponding to the "small" eigenvalues, i.e., tail of the spectrum.

Throughout the proof, we fix a $R \in \mathcal{R}_{n,\tau}$. The existence of such an R is guaranteed by the assumption $n > \overline{N}$. For the analysis, we define two projection operators, $\mathbf{P}$ and $\mathbf{Q}$. We define $\mathbf{P}$ as the projection operator onto the subspace spanned by $\{\upsilon_j\}_{j=1}^R$, i.e., for any $g = \sum_{j\in\mathbb{N}} g_j\upsilon_j \in \mathcal{H}_k$, $\mathbf{P}g = \sum_{j=1}^R g_j\upsilon_j$. Note that $\mathbf{P}$ is an orthogonal projection operator. Similarly, we define $\mathbf{Q} = \mathrm{Id} - \mathbf{P}$.

We also introduce some additional notation for the ease of presentation. We define $\mathbf{L}$ to be the diagonal matrix (operator) whose j^{th} entry is $\dfrac{\lambda_j}{n\lambda_j + \tau}$. Similarly, let $\omega_i = \mathbf{\Lambda}^{-1/2}\psi_{x_i}$ for $i = 1, 2, \ldots, n$. Using this notation, we can rewrite the

matrix $\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id}$ as

$$\begin{aligned}\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id} &= \mathbf{Z}^{-1/2}\left(\sum_{i=1}^{n}\psi_{x_i}\psi_{x_i}^{\top} + \tau\mathbf{Id}\right)\mathbf{Z}^{-1/2} - \mathbf{Id} \\ &= \sum_{i=1}^{n}(\mathbf{Z}^{-1/2}\psi_{x_i})(\mathbf{Z}^{-1/2}\psi_{x_i})^{\top} + \tau\mathbf{Z}^{-1} - \mathbf{Id} \\ &= \sum_{i=1}^{n}(\mathbf{L}^{1/2}\omega_i)(\mathbf{L}^{1/2}\omega_i)^{\top} - n\mathbf{L}.\end{aligned}$$

For any $g \in \mathcal{H}_k$, we have the following decomposition:

$$\begin{aligned}|g^{\top}(\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id})g| &= |(\mathbf{P}g + \mathbf{Q}g)^{\top}(\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id})(\mathbf{P}g + \mathbf{Q}g)| \\ &\leq |(\mathbf{P}g)^{\top}(\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id})(\mathbf{P}g)| + |(\mathbf{Q}g)^{\top}(\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id})(\mathbf{Q}g)| \\ &\quad + |(\mathbf{P}g)^{\top}(\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id}) + (\mathbf{Q}g)^{\top}(\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id})(\mathbf{P}g)| \\ &\leq \underbrace{|g^{\top}\mathbf{P}(\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id})\mathbf{P}g|}_{:=E_1} + \underbrace{|g^{\top}\mathbf{Q}(\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id})\mathbf{Q}g|}_{:=E_2} \\ &\quad + 2\underbrace{|g^{\top}\mathbf{P}(\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id})\mathbf{Q}g|}_{:=E_3}. \end{aligned} \tag{11}$$

We separately bound the terms E_1, E_2 and E_3, beginning we E_1. We have,

$$\begin{aligned}E_1 &= |g^{\top}\mathbf{P}(\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id})\mathbf{P}g| \\ &= \left|(\mathbf{P}g)^{\top}\mathbf{P}\left(\sum_{i=1}^{n}(\mathbf{L}^{1/2}\omega_i)(\mathbf{L}^{1/2}\omega_i)^{\top} - nL\mathbf{L}\right)\mathbf{P}(\mathbf{P}g)\right| \\ &= \left|(\mathbf{P}g)^{\top}\left(\sum_{i=1}^{n}(\mathbf{P}\mathbf{L}^{1/2}\mathbf{P}\omega_i)(\mathbf{P}\mathbf{L}^{1/2}\mathbf{P}\omega_i)^{\top} - n\mathbf{P}\mathbf{L}\mathbf{P}\right)(\mathbf{P}g)\right| \\ &= n\left|(\mathbf{P}g)^{\top}\mathbf{P}\mathbf{L}^{1/2}\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{P}\omega_i)(\mathbf{P}\omega_i)^{\top} - \mathbf{P}\right)\mathbf{P}\mathbf{L}^{1/2}\mathbf{P}(\mathbf{P}g)\right| \\ &\leq n\left\|\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{P}\omega_i)(\mathbf{P}\omega_i)^{\top} - \mathbf{P}\right)\right\|_2 \cdot \|\mathbf{P}\mathbf{L}^{1/2}\mathbf{P}(\mathbf{P}g)\|_{\mathcal{H}_k}^2 \\ &\leq n\left\|\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{P}\omega_i)(\mathbf{P}\omega_i)^{\top} - \mathbf{P}\right)\right\|_2 \cdot (g^{\top}\mathbf{P}\mathbf{L}\mathbf{P}g) \\ &\leq \left\|\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{P}\omega_i)(\mathbf{P}\omega_i)^{\top} - \mathbf{P}\right)\right\|_2 \cdot (n\|\mathbf{L}\|_2) \cdot \|\mathbf{P}g\|_{\mathcal{H}_k}^2. \end{aligned} \tag{12}$$

In the above equations, we used the fact that for any diagonal matrix D, $\mathbf{P}D = D\mathbf{P} = \mathbf{P}D\mathbf{P}$ and that $\mathbf{P}^2 = \mathbf{P}$. Firstly, note that $\|\mathbf{L}\|_2 = \max_{j\in\mathbb{N}} \lambda_j/(n\lambda_j + \tau) \leq 1/n$. Consequently, $n\|\mathbf{L}\|_2 \leq 1$. Secondly, to bound the first term on the RHS, we denote $\mathbf{P}\omega_i := A_i$ for all $i = 1, 2, \ldots, n$. We have, $\mathbb{E}[A_iA_i^{\top}] = \mathbf{P}\mathbb{E}[\omega_i\omega_i^{\top}]\mathbf{P} = \mathbf{P}\mathbf{\Lambda}^{-1/2}\mathbb{E}[\psi_{x_i}\psi_{x_i}^T]\mathbf{\Lambda}^{-1/2}\mathbf{P} = \mathbf{P}\mathbf{\Lambda}^{-1/2}\mathbf{\Lambda}\mathbf{\Lambda}^{-1/2}\mathbf{P} = \mathbf{P}$. Moreover, for all A_i's, only the top $R \times R$ sub-matrix has non-zero entries, implying it is sufficient to bound the 2-norm of that finite sub-matrix to bound the first term on the RHS. We use Matrix-Chernoff inequality (Tropp, 2012, Theorem 1.1) to bound the 2-norm of this finite dimensional submatrix.

For all $i = 1, 2, \ldots, n$, let $[A_i]_R \in \mathbb{R}^R$ denote the R-dimensional vector corresponding to the first R coordinates of A_i. Thus, we are interested in applying the Matrix-Chernoff inequality to bound the following expression:

$$E_{11} := \left\|\left(\frac{1}{n}\sum_{i=1}^{n}[A_i]_R[A_i]_R^{\top} - I_R\right)\right\|_2,$$

where I_R denotes the R dimensional identity matrix. Here, we used the fact that the relevant $R \times R$ sub-matrix of $\mathbf{P}$, or equivalently $\mathbb{E}[[A_1]_R[A_1]_R^\top]$, corresponds to I_R. To invoke the Matrix-Chernoff inequality, we need bounds on the maximum and minimum eigenvalue of $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}[A_i]_R[A_i]_R^\top\right]$ and a bound on $\|[A_i]_R[A_i]_R^\top/n\|_2$ that holds almost surely for all $i = 1, 2, \ldots, n$. Since $\mathbb{E}[[A_1]_R[A_1]_R^\top] = I_R$, $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}[A_i]_R[A_i]_R^\top\right] = I_R$ implying that both the maximum and minimum eigenvalues are 1. For any $i = 1, 2, \ldots, n$, we have,

$$\frac{\|[A_i]_R[A_i]_R^\top\|_2}{n} \leq \frac{1}{n}\text{trace}([A_i]_R[A_i]_R^\top) \leq \frac{1}{n}\text{trace}([A_i]_R^\top[A_i]_R) \leq \frac{1}{n}\|\mathbf{P}\omega_i\|_{\mathcal{H}_k}^2 \leq \frac{1}{n}\sum_{j=1}^{R}\varphi_j^2(x_i) \leq \frac{N(R)}{n}.$$

On invoking the Matrix-Chernoff inequality with these results, we obtain that the following relation is true with probability $1 - \delta/6$:

$$E_{11} \leq \sqrt{\frac{3N(R)\log(3R/\delta)}{n}}. \tag{13}$$

On combining the above bound with Eqn. (12) along with noting that $n\|L\|_2 \leq 1$, we can conclude that:

$$E_1 \leq \sqrt{\frac{3N(R)\log(3R/\delta)}{n}} \cdot \|\mathbf{P}g\|_{\mathcal{H}_k}^2. \tag{14}$$

We would like to mention that the above bound is only valid when the RHS in Eqn. (13) is less than 1. However, this condition is satisfied by the choice of $n > \overline{N}$.

We now consider the second term, E_2. We have,

$$\begin{aligned} E_2 &= |g^\top \mathbf{Q}(\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id})\mathbf{Q}g| \\ &= \left|(\mathbf{Q}g)^\top \left(\sum_{i=1}^{n}(\mathbf{Q}\mathbf{L}^{1/2}\omega_i)(\mathbf{Q}\mathbf{L}^{1/2}\omega_i)^\top - n\mathbf{Q}\mathbf{L}\mathbf{Q}\right)(\mathbf{Q}g)\right| \qquad (15) \\ &\leq n \underbrace{\left\|\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{Q}\mathbf{L}^{1/2}\omega_i)(\mathbf{Q}\mathbf{L}^{1/2}\omega_i)^\top - \mathbf{Q}\mathbf{L}\mathbf{Q}\right)\right\|_2}_{:=E_{21}} \cdot \|\mathbf{Q}g\|_{\mathcal{H}_k}^2. \qquad (16) \end{aligned}$$

Note that the term E_{21} has a similar structure as E_{11} except for the fact that E_{21} involves infinite-dimensional vectors as opposed to finite-dimensional vectors. Thus, to bound E_{21} we use a result from Moeller & Ullrich (2021, Proposition 3.8) which is spectral concentration inequality for infinite-dimensional vectors derived using non-commutative Khinchtine inequality (Buchholz, 2001; 2005; Moeller & Ullrich, 2021). From Proposition 3.8 in Moeller & Ullrich (2021), we can conclude that the following relation holds with probability at least $1 - \delta/6$:

$$\left\|\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{Q}\mathbf{L}^{1/2}\omega_i)(\mathbf{Q}\mathbf{L}^{1/2}\omega_i)^\top - \mathbf{Q}\mathbf{L}\mathbf{Q}\right)\right\|_2 \leq \max\left\{\frac{42}{n}\log\left(\frac{12}{\delta}\right)B_1, B_2\right\}, \tag{17}$$

where $B_1 = \max_{i=1,2,\ldots,n}\|\mathbf{Q}\mathbf{L}^{1/2}\omega_i\|_{\mathcal{H}_k}^2$ and $B_2 = \|\mathbf{Q}\mathbf{L}\mathbf{Q}\|_2$. We can further bound the terms B_1 and B_2 as follows.

$$\begin{aligned} B_1 &= \max_{i=1,2,\ldots,n}\|\mathbf{Q}\mathbf{L}^{1/2}\omega_i\|_{\mathcal{H}_k}^2 = \max_{i=1,2,\ldots,n}\sum_{j=R+1}^{\infty}\frac{\lambda_j}{n\lambda_j + \tau}\varphi_j^2(x_i) \leq \sup_{x\in\mathcal{X}}\frac{1}{\tau}\sum_{j=R+1}^{\infty}\lambda_j\varphi_j^2(x) = \frac{T(R)}{\tau} \\ B_2 &= \|\mathbf{Q}\mathbf{L}\mathbf{Q}\|_2 = \max_{j\in\mathbb{N}, j>R}\frac{\lambda_j}{n\lambda_j + \tau} \leq \frac{\lambda_{R+1}}{\tau}. \end{aligned}$$

On plugging this into Eqn. (17), we obtain the following bound on E_{21}.

$$E_{21} \leq \frac{1}{\tau}\left\{\frac{42}{n}\log\left(\frac{12}{\delta}\right)T(R), \lambda_{R+1}\right\}. \tag{18}$$

Combining Eqn. (16) and (18) yields us,

$$E_2 \leq \frac{1}{\tau}\left\{42\log\left(\frac{12}{\delta}\right)T(R), n\lambda_{R+1}\right\}\|\mathbf{Q}g\|^2_{\mathcal{H}_k}. \tag{19}$$

We now move onto the third term, E_3, which contains the cross terms. For brevity of notation, we define $\zeta_i := \mathbf{P}\mathbf{L}^{1/2}\omega_i$ and $\xi_i := \mathbf{Q}\mathbf{L}^{1/2}\omega_i$ for all $i = 1, 2, \ldots, n$. Note that $\zeta_i^\top \xi_j = 0$ for all $i, j = 1, 2, \ldots, n$. Since $\mathbf{P}$ and $\mathbf{Q}$ commute with $\mathbf{L}$, a diagonal matrix, it is straightforward to note that $\mathbf{PLQ} = 0$. Using this relation along with the definition of $\{\zeta_i\}_{i=1}^n$ and $\{\xi_i\}_{i=1}^n$, we can rewrite E_3 as follows:

$$\begin{aligned}
E_3 &= |g^\top \mathbf{P}(\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id})\mathbf{Q}g| \\
&= \left|g^\top \mathbf{P}\left(\sum_{i=1}^n (\mathbf{L}^{1/2}\omega_i)(\mathbf{L}^{1/2}\omega_i)^\top - n\mathbf{L}\right)\mathbf{Q}g\right| \\
&= \left|\sum_{i=1}^n (g^\top \mathbf{P}\mathbf{L}^{1/2}\omega_i)(g^\top \mathbf{Q}\mathbf{L}^{1/2}\omega_i)^\top\right| \\
&= \left|\sum_{i=1}^n \underbrace{(g^\top\zeta_i)(g^\top\xi_i)}_{:=W_i}\right|.
\end{aligned} \tag{20}$$

We use Bernstein inequality to bound the sum of the random variables W_i, for which we need the values of $\mathbb{E}[W_i]$, $\mathbb{E}[W_i^2]$ and an upper bound on $|W_i|$ that holds almost surely. We begin with $\mathbb{E}[W_i]$. We have,

$$\mathbb{E}[W_i] = \mathbb{E}[(g^\top\zeta_i)(g^\top\xi_i)] = g^\top\mathbb{E}[\zeta_i\xi_i^\top]g = 0. \tag{21}$$

For an upper bound on $|W_i|$, note that for any g with $\|g\|_{\mathcal{H}_k} = 1$, $|W_i|$ is maximized for the choice of $g = \psi_{x_i}$. Thus,

$$\begin{aligned}
|W_i| &= \|g\|^2_{\mathcal{H}_k}\left(\frac{g^\top\zeta_i}{\|g\|_{\mathcal{H}_k}}\right)\left(\frac{g^\top\xi_i}{\|g\|_{\mathcal{H}_k}}\right) \\
&\leq \|g\|^2_{\mathcal{H}_k}(\psi_{x_i}^\top\zeta_i)(\psi_{x_i}^\top\xi_i) \\
&\leq \|g\|^2_{\mathcal{H}_k}\|\zeta_i\|^2_{\mathcal{H}_k}\|\xi_i\|^2_{\mathcal{H}_k} \\
&\leq \|g\|^2_{\mathcal{H}_k}\cdot\left(\sum_{j=1}^R \frac{\lambda_j}{n\lambda_j+\tau}\varphi_j^2(x_i)\right)\cdot\left(\sum_{j=R+1}^\infty \frac{\lambda_j}{n\lambda_j+\tau}\varphi_j^2(x_i)\right) \\
&\leq \|g\|^2_{\mathcal{H}_k}\cdot\left(\frac{1}{n}\sum_{j=1}^R \varphi_j^2(x_i)\right)\cdot\left(\frac{1}{\tau}\sum_{j=R+1}^\infty \lambda_j\varphi_j^2(x_i)\right) \\
&\leq \|g\|^2_{\mathcal{H}_k}\cdot\frac{N(R)}{n}\cdot\frac{T(R)}{\tau}.
\end{aligned} \tag{22}$$

From the above expressions, we can also conclude that $|g^\top\zeta_i| \leq \|g\|_{\mathcal{H}_k}\cdot\frac{N(R)}{n}$ and $|g^\top\xi_i| \leq \|g\|_{\mathcal{H}_k}\cdot\frac{T(R)}{\tau}$. We use these relations to obtain a bound on $\mathbb{E}[W_i^2]$. We have,

$$\begin{aligned}
\mathbb{E}[W_i^2] &= \mathbb{E}[(g^\top\zeta_i)^2(g^\top\xi_i)^2] \\
&\leq \|g\|^2_{\mathcal{H}_k}\cdot\min\left\{\mathbb{E}\left[(g^\top\zeta_i)^2\right]\left(\frac{T(R)}{\tau}\right)^2, \mathbb{E}\left[(g^\top\xi_i)^2\right]\left(\frac{N(R)}{n}\right)^2\right\} \\
&\leq \|g\|^2_{\mathcal{H}_k}\cdot\min\left\{(g^\top\mathbf{P}L\mathbf{P}g)\cdot\left(\frac{T(R)}{\tau}\right)^2, (g^\top\mathbf{Q}L\mathbf{Q}g)\cdot\left(\frac{N(R)}{n}\right)^2\right\} \\
&\leq \|g\|^2_{\mathcal{H}_k}\cdot\min\left\{\frac{\|\mathbf{P}g\|^2_{\mathcal{H}_k}}{n}\cdot\left(\frac{T(R)}{\tau}\right)^2, \frac{\lambda_{R+1}\|\mathbf{Q}g\|^2_{\mathcal{H}_k}}{\tau}\cdot\left(\frac{N(R)}{n}\right)^2\right\}.
\end{aligned} \tag{23}$$

In the last step, we used the bounds on $\|\mathbf{L}\|_2$ and $\|\mathbf{QLQ}\|_2$ derived in the earlier part of the proof. Lastly, since $\mathbb{E}[W_i] = 0$, $\text{Var}(W_i) = \mathbb{E}[W_i^2]$. On applying Bernstein inequality (Wasserman, 2008, Lemma 7.37) using the relations from Eqns. (21), (22) and (23), we can conclude that the following relation holds with probability $1 - \delta/6$:

$$
\begin{aligned}
E_3 &= \left| \sum_{i=1}^{n} (g^\top \zeta_i)(g^\top \xi_i) \right| \\
&\leq \|g\|_{\mathcal{H}_k} \cdot \sqrt{2n \log\left(\frac{6}{\delta}\right) \min\left\{ \frac{\|\mathbf{P}g\|_{\mathcal{H}_k}^2}{n} \cdot \left(\frac{T(R)}{\tau}\right)^2, \frac{\lambda_{R+1}\|\mathbf{Q}g\|_{\mathcal{H}_k}^2}{\tau} \cdot \left(\frac{N(R)}{n}\right)^2 \right\}} \\
&\qquad + \|g\|_{\mathcal{H}_k}^2 \cdot \frac{2N(R)}{3n} \cdot \frac{T(R)}{\tau} \cdot \log\left(\frac{6}{\delta}\right). \qquad (24)
\end{aligned}
$$

On plugging the results from Eqns. (14), (19) and (24) into Eqn. (11), we obtain

$$
\begin{aligned}
\|\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id}\|_2 &= \sup_{g:\|g\|_{\mathcal{H}_k} \leq 1} |g^\top (\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id})g| \\
&\leq \sup_{g:\|g\|_{\mathcal{H}_k} \leq 1} \Bigg[\sqrt{\frac{3N(R)\log(6R/\delta)}{n}} \|\mathbf{P}g\|_{\mathcal{H}_k}^2 + \frac{1}{\tau} \max\left\{ 42 \log\left(\frac{12}{\delta}\right) T(R), n\lambda_{R+1} \right\} \|\mathbf{Q}g\|_{\mathcal{H}_k}^2 \\
&\qquad + 2\|g\|_{\mathcal{H}_k} \sqrt{2n \log\left(\frac{6}{\delta}\right) \min\left\{ \frac{\|\mathbf{P}f\|_{\mathcal{H}_k}^2}{n} \cdot \left(\frac{T(R)}{\tau}\right)^2, \frac{\lambda_{R+1}\|\mathbf{Q}f\|_{\mathcal{H}_k}^2}{\tau} \cdot \left(\frac{N(R)}{n}\right)^2 \right\}} \\
&\qquad + \|g\|_{\mathcal{H}_k}^2 \cdot \frac{4N(R)}{3n} \cdot \frac{T(R)}{\tau} \cdot \log\left(\frac{6}{\delta}\right) \Bigg] \\
&\leq \Bigg[\sqrt{\frac{3N(R)\log(6R/\delta)}{n}} + \frac{1}{\tau} \max\left\{ 42 \log\left(\frac{12}{\delta}\right) T(R), n\lambda_{R+1} \right\} \\
&\qquad + 2\sqrt{2n \log\left(\frac{6}{\delta}\right) \min\left\{ \frac{1}{n} \cdot \left(\frac{T(R)}{\tau}\right)^2, \frac{\lambda_{R+1}}{\tau} \cdot \left(\frac{N(R)}{n}\right)^2 \right\}} + \frac{4N(R)T(R)}{3n\tau} \log\left(\frac{6}{\delta}\right) \Bigg]
\end{aligned}
$$

On plugging in any value of $R \in \mathcal{R}(n, \tau, \delta)$ and using the definition of $\mathcal{R}_{n,\tau,\delta}$ along with the relation $n \geq \overline{N}$, we can conclude that $\|\mathbf{Z}^{-1/2}\hat{\mathbf{Z}}\mathbf{Z}^{-1/2} - \mathbf{Id}\|_2 \leq 1/9$ with probability at least $1 - \delta/2$. The overall probability on the bound is obtained using a union bound for the relations on E_1, E_2 and E_3.

A.2. Proof of Lemma 3.3

We begin the proof by showing that we can relate the $\psi_x^\top \hat{\mathbf{Z}}^{-1}\psi_x$ to $\psi_x^\top \mathbf{Z}^{-1}\psi_x$ through the operator norm of $\mathbf{M} := \hat{\mathbf{Z}}^{-1/2}(\mathbf{Z} - \hat{\mathbf{Z}})\mathbf{Z}^{-1/2}$. Specifically, we show if that operator norm of $\mathbf{M}$ is small, then $\psi_x^\top \hat{\mathbf{Z}}^{-1}\psi_x$ and $\psi_x^\top \mathbf{Z}^{-1}\psi_x$ are within a constant factor of each other. Lastly, we use the condition on $\|\mathbf{Z}^{-\frac{1}{2}}\hat{\mathbf{Z}}\mathbf{Z}^{-\frac{1}{2}} - \mathbf{Id}\|_2$ to bound the $\|\mathbf{M}\|_{\text{op}}$, the operator norm of $\mathbf{M}$, to obtain the required result.

We begin with considering the following expression.

$$
\begin{aligned}
\left| \psi_x^\top (\hat{\mathbf{Z}}^{-1} - \mathbf{Z}^{-1})\psi_x \right| &= \left| \psi_x^\top \hat{\mathbf{Z}}^{-1}(\mathbf{Z} - \hat{\mathbf{Z}})\mathbf{Z}^{-1}\psi_x \right| \\
&= \left| \psi_x^\top \hat{\mathbf{Z}}^{-1/2} \cdot \hat{\mathbf{Z}}^{-1/2}(\mathbf{Z} - \hat{\mathbf{Z}})\mathbf{Z}^{-1/2} \cdot \mathbf{Z}^{-1/2}\psi_x \right| \\
&\leq \|\hat{\mathbf{Z}}^{-1/2}\psi_x\|_{\mathcal{H}_k} \|\mathbf{Z}^{-1/2}\psi_x\|_{\mathcal{H}_k} \|\hat{\mathbf{Z}}^{-1/2}(\mathbf{Z} - \hat{\mathbf{Z}})\mathbf{Z}^{-1/2}\|_{\text{op}} \\
&\leq \sqrt{(\psi_x^\top \hat{\mathbf{Z}}^{-1}\psi_x)} \cdot \sqrt{(\psi_x^\top \mathbf{Z}^{-1}\psi_x)} \cdot \|\mathbf{M}\|_{\text{op}}. \qquad (25)
\end{aligned}
$$

Consider the scenario where the relation $\|\mathbf{M}\|_{\text{op}} \leq c$ is satisfied for some $c \in (0, 1)$. We claim that under this scenario, we have, $\psi_x^\top \hat{\mathbf{Z}}^{-1}\psi_x \leq (1 - c)^{-1} \cdot \psi_x^\top \mathbf{Z}^{-1}\psi_x$. To show this claim, we consider Eqn. (25). If $\psi_x^\top \mathbf{Z}^{-1}\psi_x \geq \psi_x^\top \hat{\mathbf{Z}}^{-1}\psi_x$, the

claim follows immediately. For the other case, we have,

$$\begin{aligned}
\psi_x \hat{\mathbf{Z}}^{-1}\psi_x - \psi_x \mathbf{Z}^{-1}\psi_x &\leq \sqrt{(\psi_x^\top \hat{\mathbf{Z}}^{-1}\psi_x)} \cdot \sqrt{(\psi_x^\top \mathbf{Z}^{-1}\psi_x)} \cdot c \\
&\leq \sqrt{(\psi_x^\top \hat{\mathbf{Z}}^{-1}\psi_x)} \cdot \sqrt{(\psi_x^\top \hat{\mathbf{Z}}^{-1}\psi_x)} \cdot c \\
&\leq c \cdot (\psi_x^\top \hat{\mathbf{Z}}^{-1}\psi_x) \\
\implies \psi_x \hat{\mathbf{Z}}^{-1}\psi_x &\leq \big(\psi_x \mathbf{Z}^{-1}\psi_x\big) \cdot \frac{1}{1-c},
\end{aligned}$$

as claimed. Thus, it suffices to show that $\|\mathbf{M}\|_{\text{op}}$ is small.

To that effect, note that we can write the operator $\mathbf{M}$ as $\mathbf{M} = \hat{\mathbf{Z}}^{-1/2}\mathbf{Z}^{1/2} - \hat{\mathbf{Z}}^{1/2}\mathbf{Z}^{-1/2} = \mathbf{C}^{-1} - \mathbf{C}^\top$ where, $\mathbf{C} := \mathbf{Z}^{-1/2}\hat{\mathbf{Z}}^{1/2}$. Consequently, using the definition of operator norm yields us,

$$\begin{aligned}
\|\mathbf{M}\|_{\text{op}}^2 = \|\mathbf{M}^\top\mathbf{M}\|_2 &= \|((\mathbf{C}^\top)^{-1} - \mathbf{C})(\mathbf{C}^{-1} - \mathbf{C}^\top)\|_2 \\
&= \|(\mathbf{C}\mathbf{C}^\top)^{-1} - \text{Id} + \mathbf{C}\mathbf{C}^\top - \text{Id}\|_2 \\
&\leq \|(\mathbf{C}\mathbf{C}^\top)^{-1} - \text{Id}\|_2 + \|\mathbf{C}\mathbf{C}^\top - \text{Id}\|_2. \qquad (26)
\end{aligned}$$

From the definition of $\mathbf{C}$, we have $\|\mathbf{C}\mathbf{C}^\top - \text{Id}\|_2 = \|\mathbf{Z}^{-\frac{1}{2}}\hat{\mathbf{Z}}\mathbf{Z}^{-\frac{1}{2}} - \mathbf{Id}\|_2 \leq b$, from the given statement in the Lemma. Note that if $\|\mathbf{C}\mathbf{C}^\top - \text{Id}\|_2 \leq b$ for some $b \in (0, 1/3)$, then all eigenvalues of $\mathbf{C}\mathbf{C}^\top$ lie in the interval $[1-b, 1+b]$. This implies that all the eigenvalues of $(\mathbf{C}\mathbf{C}^\top)^{-1}$ lie in the interval $[(1+b)^{-1}, (1-b)^{-1}]$. Hence, $\|(\mathbf{C}\mathbf{C}^\top)^{-1} - \text{Id}\|_2 \leq b/(1-b)$. On combining this with Eqn. (26), we can conclude that if $\|\mathbf{C}\mathbf{C}^\top - \text{Id}\|_2 \leq b$, then $\|\mathbf{M}\|_{\text{op}} \leq \sqrt{2b/(1-b)} < 1$. On combining this with the previous claim that relates $\psi_x^\top \hat{\mathbf{Z}}^{-1}\psi_x$ to $\psi_x^\top \mathbf{Z}^{-1}\psi_x$ through $\|\mathbf{M}\|_{\text{op}}$, we arrive at the result.

A.3. Proof of Lemma 3.4

Similar to the analysis in Appendix A.1, we fix an $R \in \mathcal{R}(n, \tau, \delta)$ and define projection matrices $\mathbf{P}$ and $\mathbf{Q}$ using the value of R as defined in Appendix A.1. We define the projection of the kernel operator $k(\cdot,\cdot)$ on the subspaces spanned by $\mathbf{P}$ and $\mathbf{Q}$ as follows:

$$k^{(\mathbf{P})}(x,y) = \sum_{j=1}^{R} \lambda_j \varphi_j(x)\varphi_j(y); \quad k^{(\mathbf{Q})}(x,y) = k(x,y) - k^{(\mathbf{P})}(x,y).$$

Recall that $\tilde{\gamma}_{X_n,\tau}$ denotes the information gain corresponding to the randomly drawn set of points $X_n = \{x_1, x_2, \ldots, x_n\}$. Similar to K_{X_n,X_n}, we also define $K^{(\mathbf{P})}_{X_n,X_n}$ and $K^{(\mathbf{Q})}_{X_n,X_n}$ as $K^{(\mathbf{P})}_{X_n,X_n} = [k^{(\mathbf{P})}(x_i,x_j)]_{i,j=1}^n$ and $K^{(\mathbf{Q})}_{X_n,X_n} = [k^{(\mathbf{Q})}(x_i,x_j)]_{i,j=1}^n$. It is straightforward to note that $K_{X_n,X_n} = K^{(\mathbf{P})}_{X_n,X_n} + K^{(\mathbf{Q})}_{X_n,X_n}$.

We first derive some auxiliary results on the spectrum of $K^{(\mathbf{P})}_{X_n,X_n}$ and $K^{(\mathbf{Q})}_{X_n,X_n}$ which will be useful in the analysis later. Recall that we defined $\mathbf{\Psi}_n := [\psi_{x_1}, \psi_{x_2}, \ldots, \psi_{x_n}]$. We can also rewrite K_{X_n,X_n}, $K^{(\mathbf{P})}_{X_n,X_n}$ and $K^{(\mathbf{Q})}_{X_n,X_n}$ in terms of $\mathbf{\Psi}_n$ as: $K_{X_n,X_n} = \mathbf{\Psi}_n^\top\mathbf{\Psi}_n$, $K^{(\mathbf{P})}_{X_n,X_n} = \mathbf{\Psi}_n^\top P\mathbf{\Psi}_n$ and $K^{(\mathbf{Q})}_{X_n,X_n} = \mathbf{\Psi}_n^\top\mathbf{Q}\mathbf{\Psi}_n$. Using this relation, note that the singular values of $K^{(\mathbf{P})}_{X_n,X_n} = (P\mathbf{\Psi}_n)^\top(P\mathbf{\Psi}_n)$ and $K^{(\mathbf{Q})}_{X_n,X_n} = \mathbf{\Psi}_n^\top\mathbf{Q}\mathbf{\Psi}_n$ are the same as that of $(P\mathbf{\Psi}_n)(P\mathbf{\Psi}_n)^\top = P\mathbf{\Psi}_n\mathbf{\Psi}_n^\top P$ and $(\mathbf{Q}\mathbf{\Psi}_n)(\mathbf{Q}\mathbf{\Psi}_n)^\top = \mathbf{Q}\mathbf{\Psi}_n\mathbf{\Psi}_n^\top\mathbf{Q}$ respectively.

For the spectrum of $K^{(\mathbf{P})}_{X_n,X_n}$, note that

$$\begin{aligned}
K^{(\mathbf{P})}_{X_n,X_n} &= (\mathbf{P}\mathbf{\Psi}_n)^\top(\mathbf{P}\mathbf{\Psi}_n) = ((n\mathbf{\Lambda})^{-1/2}\mathbf{P}\mathbf{\Psi}_n)^\top(n\mathbf{\Lambda})((n\mathbf{\Lambda})^{-1/2}\mathbf{P}\mathbf{\Psi}_n) \\
&= (\mathbf{P}(n\mathbf{\Lambda})^{-1/2}\mathbf{\Psi}_n)^\top\mathbf{P}(n\mathbf{\Lambda})\mathbf{P}(\mathbf{P}(n\mathbf{\Lambda})^{-1/2}\mathbf{\Psi}_n).
\end{aligned}$$

If $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_R$ denote the eigenvalues of $K^{(\mathbf{P})}_{X_n,X_n}$, then using Ostrowski's Theorem (Ostrowski, 1959), we can conclude that $\tilde{\lambda}_j = \theta_j n\lambda_j$ for all $j = 1, 2, \ldots, R$, where $\{n\lambda_j\}_{j=1}^R$ correspond to the eigenvalues of $n\mathbf{P}\mathbf{\Lambda}\mathbf{P}$ and θ_j lie between the smallest and largest eigenvalues of the matrix $n^{-1}(\mathbf{P}\mathbf{\Lambda}^{-1/2}\mathbf{\Psi}_n)^\top(\mathbf{P}\mathbf{\Lambda}^{-1/2}\mathbf{\Psi}_n)$. Note

that the singular values (in this case, also eigenvalues) of $n^{-1}(\mathbf{P}\mathbf{\Lambda}^{-1/2}\mathbf{\Psi}_n)^\top(\mathbf{P}\mathbf{\Lambda}^{-1/2}\mathbf{\Psi}_n)$ are the same as that of $n^{-1}(\mathbf{P}\mathbf{\Lambda}^{-1/2}\mathbf{\Psi}_n)(\mathbf{P}\mathbf{\Lambda}^{-1/2}\mathbf{\Psi}_n)^\top = n^{-1}\sum_{i=1}^n(\mathbf{P}\omega_i)(\mathbf{P}\omega_i)^\top$, where $\omega_i = \mathbf{\Lambda}^{-1/2}\psi_{x_i}$, as defined in Appendix A.1. Using Eqn. (13) and that $R \in \mathcal{R}(n,\tau,\delta)$ and $n \geq \overline{N}$, we can conclude that the following relation is true with probability $1-\delta/6$:

$$\left\| \left(\frac{1}{n}\sum_{i=1}^n (\mathbf{P}\omega_i)(\mathbf{P}\omega_i)^\top - \mathbf{P} \right) \right\|_2 \leq \frac{1}{27}.$$

Thus, we can conclude that eigenvalues of $n^{-1}(\mathbf{P}\mathbf{\Lambda}^{-1/2}\mathbf{\Psi}_n)^\top(\mathbf{P}\mathbf{\Lambda}^{-1/2}\mathbf{\Psi}_n)$ lie in the range $[26/27, 28/27]$ and consequently, $\tilde{\lambda}_j \geq 26n\lambda_j/27$.

As mentioned earlier, the singular values of $K^{(\mathbf{Q})}_{X_n,X_n}$ are the same as those of $\mathbf{Q}\mathbf{\Psi}_n\mathbf{\Psi}_n^\top\mathbf{Q}$. For the analysis, it suffices to have an upper bound on $\|K^{(\mathbf{Q})}_{X_n,X_n}\|_2$, or equivalently, $\|\mathbf{Q}\mathbf{\Psi}_n\mathbf{\Psi}_n^\top\mathbf{Q}\|_2$. Using the result from Moeller & Ullrich (2021, Proposition 3.8), we know that the following relation holds with probability $1-\delta/6$:

$$\|\mathbf{Q}\mathbf{\Psi}_n\mathbf{\Psi}_n^\top\mathbf{Q}\|_2 \leq 2\left\{42\log\left(\frac{12}{\delta}\right)T(R), n\lambda_{R+1}\right\}.$$

Since $R \in \mathcal{R}_{n,\tau}$, we can conclude that $\|K^{(\mathbf{Q})}_{X_n,X_n}\|_2 = \|\mathbf{Q}\mathbf{\Psi}_n\mathbf{\Psi}_n^\top\mathbf{Q}\|_2 \leq 2\tau/27$. We are now ready to prove the lemma.

Using the relation $K_{X_n,X_n} = K^{(\mathbf{P})}_{X_n,X_n} + K^{(\mathbf{Q})}_{X_n,X_n}$, we can decompose the information gain of X_n as follows:

$$\begin{aligned}
\tilde{\gamma}_{X_n,\tau} &= \frac{1}{2}\log\left(\det(I_n + \tau^{-1}K_{X_n,X_n})\right) \\
&= \frac{1}{2}\log\left(\det(I_n + \tau^{-1}K^{(\mathbf{P})}_{X_n,X_n} + \tau^{-1}K^{(\mathbf{Q})}_{X_n,X_n})\right) \\
&= \frac{1}{2}\log\left(\det((I_n + \tau^{-1}K^{(\mathbf{Q})}_{X_n,X_n})(I_n + \tau^{-1}(I_n + \tau^{-1}K^{(\mathbf{Q})}_{X_n,X_n})^{-1}K^{(\mathbf{P})}_{X_n,X_n}))\right) \\
&= \frac{1}{2}\underbrace{\log(\det(I + \tau^{-1}K^{(\mathbf{Q})}_{X_n,X_n}))}_{:=G_1} + \frac{1}{2}\underbrace{\log(\det(I + \tau^{-1}(I + \tau^{-1}K^{(\mathbf{Q})}_{X_n,X_n})^{-1}K^{(\mathbf{P})}_{X_n,X_n}))}_{:=G_2}.
\end{aligned}$$

This decomposition is similar to that derived in Vakili et al. (2021b, App. A, Eqn. 8) with the roles of $K^{(\mathbf{P})}_{X_n,X_n}$ and $K^{(\mathbf{Q})}_{X_n,X_n}$ interchanged.

We begin with G_1. Since $\|K^{(\mathbf{Q})}_{X_n,X_n}\|_2 \leq 2\tau/27$, all eigenvalues of $\tau^{-1}K^{(\mathbf{Q})}_{X_n,X_n}$ are less than 1. Using the relation $\log(1+x) \geq x/2$, which holds for all $x \in [0,1]$, we can lower bound G_1 as follows:

$$G_1 = \log(\det(I + \tau^{-1}K^{(\mathbf{Q})}_{X_n,X_n})) \geq \frac{1}{2\tau}\mathrm{trace}(K^{(\mathbf{Q})}_{X_n,X_n}).$$

Note $k^{(\mathbf{Q})}(X_i,X_i)$ are i.i.d. random variables with $\mathbb{E}[k^{(\mathbf{Q})}(X_i,X_i)] = \sum_{r=R+1}^\infty \lambda_r$ and $|k^{(\mathbf{Q})}(X_i,X_i)| \leq T(R)$. We can thus use Hoeffding inequality to obtain the following bound on $\mathrm{trace}(K^{(\mathbf{Q})}_{X_n,X_n})$ which holds with probability at least $1-\delta/6$:

$$\begin{aligned}
G_1 &\geq \frac{1}{2\tau}\mathrm{trace}(K^{(O)}_{X_n,X_n}) \\
&\geq \frac{1}{2\tau}\left[n\sum_{r=R+1}^\infty \lambda_r - T(R)\sqrt{n\log(12/\delta)}\right] \\
&\geq \frac{nT(R)}{2\tau F^2}\left(1 - F^2\sqrt{\frac{\log(12/\delta)}{n}}\right) \\
&\geq \frac{13nT(R)}{27\tau F^2}
\end{aligned}$$

In the third line, we used the fact that $T(R) \leq F^2\sum_{r=R+1}^\infty \lambda_r$ since $\|\varphi_j\|_\infty \leq F$ for all $j \in \mathbb{N}$ (Assumption 2.3). The fourth line uses the condition that $n \geq \overline{N}$.

To bound G_2, first note that using the condition on the spectrum on $\tau^{-1}K^{(\mathbf{Q})}_{X_n,X_n}$, we can conclude that all the eigenvalues of $(I+\tau^{-1}K^{(\mathbf{Q})}_{X_n,X_n})$ lie in the range $[1,2]$. Moreover, note that the spectrum of $(I+\tau^{-1}K^{(\mathbf{Q})}_{X_n,X_n})^{-1}K^{(\mathbf{P})}_{X_n,X_n}$ is the same as that of $(I+\tau^{-1}K^{(\mathbf{Q})}_{X_n,X_n})^{-1/2}K^{(\mathbf{P})}_{X_n,X_n}(I+\tau^{-1}K^{(\mathbf{Q})}_{X_n,X_n})^{-1/2}$. On using Ostrowski's Theorem (Ostrowski, 1959) along with range of eigenvalues of $(I+\tau^{-1}K^{(\mathbf{Q})}_{X_n,X_n})$, we can conclude that

$$G_2 = \log(\det(I+\tau^{-1}(I+\tau^{-1}K^{(\mathbf{Q})}_{X_n,X_n})^{-1}K^{(\mathbf{P})}_{X_n,X_n})) \geq \log(\det(I+(2\tau)^{-1}K^{(\mathbf{P})}_{X_n,X_n})).$$

Using the relation for the eigenvalues of $K^{(\mathbf{P})}_{X_n,X_n}$ derived earlier, we can further G_2 as follows:

$$\begin{aligned}
G_2 &\geq \log(\det(I+(2\tau)^{-1}K^{(\mathbf{P})}_{X_n,X_n})) \\
&\geq \sum_{j=1}^{R}\log(1+(2\tau)^{-1}\tilde{\lambda}_j) \\
&\geq \sum_{j=1}^{R}\log\left(1+\frac{13n\lambda_j}{27\tau}\right) \\
&\geq \sum_{j=1}^{R}\frac{13n\lambda_j}{13n\lambda_j+27\tau} \\
&\geq \frac{13n}{27F^2}\sup_{x\in\mathcal{X}}\sum_{j=1}^{R}\frac{\lambda_j}{n\lambda_j+\tau}\varphi_j^2(x).
\end{aligned}$$

In the fourth line, we used the relation $\log(1+x) \geq \frac{x}{x+1}$, which holds for all $x \geq 0$.

On combining the bounds for G_1 and G_2, we obtain,

$$\begin{aligned}
\tilde{\gamma}_{X_n,\tau} &= \frac{1}{2}(G_1+G_2) \\
&\geq \frac{13nT(R)}{54\tau F^2} + \frac{13n}{54F^2}\sup_{x\in\mathcal{X}}\sum_{j=1}^{R}\frac{\lambda_j}{n\lambda_j+\tau}\varphi_j^2(x) \\
&\geq \frac{13n}{54F^2}\left(\sup_{x\in\mathcal{X}}\sum_{j=1}^{R}\frac{\lambda_j}{n\lambda_j+\tau}\varphi^2(x) + \frac{T(R)}{\tau}\right) \\
&\geq \frac{13n}{54F^2}\sup_{x\in\mathcal{X}}\left(\sum_{j=1}^{R}\frac{\lambda_j}{n\lambda_j+\tau}\varphi_j^2(x) + \sum_{j=R+1}^{\infty}\frac{\lambda_j}{n\lambda_j+\tau}\varphi_j^2(x)\right) \\
&\geq \frac{13n}{54F^2}\sup_{x\in\mathcal{X}}\psi_x^\top\mathbf{Z}^{-1}\psi_x,
\end{aligned}$$

as required. Since each of the bounds on G_1 and the eigenvalues of $K^{(\mathbf{P})}_{X_n,X_n}$ and $K^{(\mathbf{Q})}_{X_n,X_n}$, holds with probability at least $1-\delta/6$, the overall bound holds with probability at least $1-\delta/2$.

B. Proof of Theorems 4.3 and 4.5

The proof of both the theorems is based along the lines of the proof of the Batched Pure Exploration (BPE) algorithm (Li & Scarlett, 2022). We first begin with a brief discussion about Assumption 4.2 and the choice of constants $\tilde{N}_{\varepsilon_0}$ and $\tilde{N}'_{\varepsilon_0}$ and then move on to the proof.

Definition B.1. Let $\Gamma : \mathcal{X} \to \mathcal{X}'$ be a map between two sets $\mathcal{X}, \mathcal{X}' \subset \mathbb{R}^d$. We call Γ to be a bi-Lipschitz map if the inverse map, Γ^{-1}, exists and the following relations hold for some $L, L' > 0$:

$$\|\Gamma(x)-\Gamma(y)\|_2 \leq L\|x-y\|_2 \;\; \forall x,y \in \mathcal{X}$$
$$\|\Gamma^{-1}(x)-\Gamma^{-1}(y)\|_2 \leq L'\|x-y\|_2 \;\; \forall x,y \in \mathcal{X}'.$$

We refer to (L, L') the Lipschitz constant pair of Γ. We also define normalized Lipschitz constant pair of Γ to be the pair $(\tilde{L}, \tilde{L}') = \left(L \left(\frac{\text{vol}(\mathcal{X})}{\text{vol}(\mathcal{X}')} \right)^{1/d}, L' \left(\frac{\text{vol}(\mathcal{X}')}{\text{vol}(\mathcal{X})} \right)^{1/d} \right)$.

The normalized Lipschitz constant pair quantifies solely the change due to structure and discounts for the change in size between $\mathcal{X}$ and $\mathcal{X}'$. The following is a restatement of Assumption 4.2.

Assumption B.2. Let $\mathcal{L}_\eta^f = \{x \in \mathcal{X} | f(x) \geq \eta\}$ denote the level set of f for $\eta \in [-B, B]$. Let $\mathcal{X}'$ be a subset of $\mathcal{L}_\eta^f$ with a finite number of connected components. Let $\text{UCB}_t(x; \mathcal{X}')$ denote the upper confidence bound on the function f at any point $x \in \mathcal{X}'$, constructed using t points sampled uniformly at random from each component of $\mathcal{X}'$. For any $\eta' \geq \eta$, we define $\mathcal{L}_{\eta'}^{\text{UCB}_t} = \{x \in \mathcal{X}' | \text{UCB}_t(x; \mathcal{X}') \geq \eta'\}$ to be the level set of the upper confidence bound at level η'. Let $\eta_0 := \sup f - \varepsilon_0$ for some fixed $\varepsilon_0 > 0$. We assume the following for all $\mathcal{X}'$ and $t \geq \overline{N}$.

1. For any $\eta \geq \eta_0$, the number of connected components in $\mathcal{L}_\eta^f$ are at most M_f.
2. For any $\eta' \geq \eta \geq \eta_0$, the number of connected components of $\mathcal{L}_{\eta'}^{\text{UCB}_t}$ are at most M more than those of $\{x \in \mathcal{X}' : f(x) \geq \eta'\}$ with probability $1 - \delta/(2 \log_2 T)$, where the probability is taken over the randomness in query points and noise (if any).
3. Let $\mathcal{L}_{\eta'}^{\text{UCB}_t, i}$ denote the ith such connected component of $\mathcal{L}_{\eta'}^{\text{UCB}_t}$. We assume that there exists a bi-Lipschitzian map $\Gamma_{\eta', i} : \mathcal{X} \to \mathcal{L}_{\eta'}^{\text{UCB}_t, i}$ with normalized Lipschitz constant pair $\tilde{L}_{\eta', i}, \tilde{L}'_{\eta', i} > 0$ for all η', i. Let $L_f = \sup_{\eta', i} \tilde{L}_{\eta', i}$ and $L'_f = \sup_{\eta', i} \tilde{L}'_{\eta', i}$. We assume that $L_f, L'_f < \infty$.

Assumption 4.2 is an assumption on the regularity of the level sets of the function f. The term M_f can be thought of as the number of local maximas of f and hence finiteness of M_f is a mild assumption on f satisfied by functions encountered in practice. Moreover, the knowledge of M_f is only required for analysis and not for the algorithm to run. Furthermore, since we only require the number of connected components to be finite for values close to the maximum value, we allow the function f have a large number of local maximas with small value. The last condition is to ensure that the these connected components are topologically regular enough and to avoid certain pathological cases. In particular, the existence of a bi-Lipschitzian map between two sets implies topological similarity between the two sets. Intuitively, this assumption ensures that the shape of the level-sets is not "too arbitrary". Note that such an assumption on the level sets of UCB is relatively mild as the RKHS endows smoothness properties to the UCB which translate to a degree of topological regularity of level sets (Alberti et al., 2011; Lee, 2010). Please refer to Appendix D for additional discussion on Assumption 4.2.

The constants $\tilde{N}_{\varepsilon_0}$ and $\tilde{N}'_{\varepsilon_0}$ are chosen to ensure that the conditions in Assumption 4.2 are satisfied. Specifically, we choose the constants to ensure that the algorithm encounters level sets corresponding to $\eta \geq f(x^*) - \varepsilon_0$. Thus, we set

$$\tilde{N}_{\varepsilon_0} := \min \left\{ n \in \mathbb{N} : 4BC_1 n^{1-\beta} (\log(4n/\delta))^\beta \leq \varepsilon_0 \right\}$$
$$\tilde{N}'_{\varepsilon_0} := \min \left\{ n \in \mathbb{N} : 4\alpha_\tau(\delta/6) \cdot C_0 \cdot \left(\frac{n}{\tau} \right)^{\frac{1}{\beta} - 1} + \frac{2B}{T} + R \sqrt{\frac{2}{T\tau} \log \left(\frac{4T}{\delta'} \right)} \leq \varepsilon_0 \right\},$$

where C_0 and C_1 are the constants in Eqns. (9) and (10) respectively. On invoking the results from Lemmas B.4, B.6 and Theorem 3.1, we can conclude that the choices of $\tilde{N}_{\varepsilon_0}$ and $\tilde{N}'_{\varepsilon_0}$ ensure that the condition in Assumption 4.2 is satisfied with probability $1 - \delta/4$ and $1 - \delta/6$ respectively.

B.1. Proof of Theorem 4.3

At a high level, the bound on regret is obtained by first separately bounding the regret during every epoch r and then summing it across all epochs. During any epoch r, since REDS chooses points uniformly at random from the current domain $\mathcal{X}_r$, we simply bound the regret incurred at each point queried during this epoch by the worst case scenario, i.e., $\varsigma_r := f(x^*) - \inf_{x \in \mathcal{X}_r} f(x)$. This leads to an upper bound of $\varsigma_r N_r M'$ on the regret incurred during epoch r, where M' denotes the bound on the number of connected components in $\mathcal{X}_r$. Since poorly performing regions of the domain are eliminated as the algorithm proceeds, $\inf_{x \in \mathcal{X}_r} f(x)$ gets closer to $f(x^*)$, reducing the regret in each epoch as the algorithm proceeds.

The following two lemmas ensure the correctness of the algorithm and help bound the regret incurred during each epoch.

Lemma B.3. *$x^* \in \mathcal{X}_r$ for all $r \geq 1$.*

Lemma B.4. *For all epochs r, we have,*

$$\varsigma_r \leq \begin{cases} 2B & \text{if } r = 1, \\ 4B \sup_{x \in \mathcal{X}_{r-1}} \sigma_{r-1}(x) & \text{if } r \geq 2. \end{cases}$$

Lemma B.5. *For all epochs r, with probability at least $1 - \delta/2$, the number of connected components in $\mathcal{X}_r$ are at most $(M_f + M)(1 + \log_2(T/N_1))$.*

We defer the proof of these lemmas to Appendix B.3. Equipped with these lemmas, we move on to the proof of Theorem 4.3. The regret incurred by REDS can be bounded as

$$\begin{aligned} R(T) = \sum_{t=1}^{T} f(x^*) - f(x_t) &\leq \sum_{r=1}^{S} \varsigma_r N_r (M_f + M)(1 + \log_2(T/N_1)) \\ &\leq 2BN_1 + 4B(M_f + M)(1 + \log_2(T/N_1)) \sum_{r=2}^{S} \left[N_r \cdot \sup_{x \in \mathcal{X}_{r-1}} \sigma_{r-1}(x) \right]. \end{aligned}$$

In the above expression, S denotes the total number of epochs that begin during a run of REDS algorithm before reaching a total of T queries. Since the epoch lengths double every epoch, we have $S \leq 1 + \log_2(T/N_1)$. We can further bound $R(T)$ using Lemma 4.6 (which in turn is based on Theorem 3.1) to bound the worst-case posterior standard deviation in the above equation. Since $\mathcal{X}_{r-1}$ is compact ($\mathcal{X}_{r-1}$ is closed by definition and $\mathcal{X}_{r-1}$ is bounded because $\mathcal{X}_{r-1} \subseteq \mathcal{X}$) and $N_{r-1} \geq N_1 \geq C_{L_f, L'_f} \overline{N}$, we can invoke Lemma 4.6 to conclude

$$R(T) \leq 2BN_1 + 4BC_2 C'_{L_f, L'_f} (M_f + M)(1 + \log_2(T/N_1)) \sum_{r=2}^{S} N_r \cdot N_{r-1}^{(1-\beta)/2} (\log(n/\delta'))^{\beta/2}, \tag{27}$$

where $\delta' = \delta / 4 \log_2 T$, $C_2 = \sqrt{C_1}$ and $C_{L_f, L'_f}, C'_{L_f, L'_f}$ are the constants from Lemma 4.6 and depend only on L_f, L'_f. For simplicity, we define $C_f := C'_{L_f, L'_f}(M_f + M)$, as a constant that depends only on the function f. On plugging in the values of N_r, Eqn. (27) simplifies to

$$\begin{aligned} R(T) &\leq 2BN_1 + 4BC_2 C_f (1 + \log_2 T) \sum_{r=2}^{S} N_r \cdot N_{r-1}^{(1-\beta)/2} (\log(n/\delta'))^{\beta/2} \\ &\leq 2BN_1 + 4BC_2 C_f (1 + \log_2 T) N_1^{(3-\beta)/2} \sum_{r=2}^{S} 2^{r-1} \cdot 2^{(r-2)(1-\beta)/2} \left(\log \left(\frac{N_1}{\delta'} \cdot 2^{r-2} \right) \right)^{\beta/2} \\ &\leq 2BN_1 + 8BC_2 C_f (1 + \log_2 T) N_1^{(3-\beta)/2} \sum_{r=0}^{S-2} 2^{r(3-\beta)/2} \left(\log \left(\frac{N_1}{\delta'} \cdot 2^{r} \right) \right)^{\beta/2}. \end{aligned} \tag{28}$$

We consider three separate cases based on the value of β:

- $\beta < 3$: Under this case, Eqn. (28) can be simplified as follows:

$$\begin{aligned} R(T) &\leq 2BN_1 + 8BC_2 C_f (1 + \log_2 T) N_1^{(3-\beta)/2} \sum_{r=0}^{S-2} 2^{r(3-\beta)/2} \left(\log \left(\frac{N_1}{\delta'} \cdot 2^{r} \right) \right)^{\beta/2} \\ &\leq 2BN_1 + 8BC_2 C_f (1 + \log_2 T) N_1^{(3-\beta)/2} \left(\log \left(\frac{T}{\delta'} \right) \right)^{3/2} \sum_{r=0}^{S-2} 2^{r(3-\beta)/2} \\ &\leq 2BN_1 + 8BC_2 C_f (1 + \log_2 T) N_1^{(3-\beta)/2} \left(\log \left(\frac{T}{\delta'} \right) \right)^{3/2} \frac{2^{(S-1)(3-\beta)/2} - 1}{2^{(3-\beta)/2} - 1} \\ &\leq 2BN_1 + \frac{8BC_2 C_f}{2^{(3-\beta)/2} - 1} T^{(3-\beta)/2} \left(\log \left(\frac{T}{\delta'} \right) \right)^{3/2} (1 + \log_2 T). \end{aligned}$$

- $\beta = 3$: For this value of β, Eqn. (28) can be simplified as follows:

$$\begin{aligned}
R(T) &\leq 2BN_1 + 8BC_2C_f(1+\log_2 T)N_1^{(3-\beta)/2}\sum_{r=0}^{S-2} 2^{r(3-\beta)/2}\left(\log\left(\frac{N_1}{\delta'}\cdot 2^r\right)\right)^{\beta/2} \\
&\leq 2BN_1 + 8BC_2C_f(1+\log_2 T)\cdot\left(\log\left(\frac{T}{\delta'}\right)\right)^{3/2}\cdot\sum_{r=0}^{S-2} 1 \\
&\leq 2BN_1 + 8BC_2C_f\cdot\left(\log\left(\frac{T}{\delta'}\right)\right)^{3/2}\cdot\log\left(\frac{T}{N_1}\right)(1+\log_2 T).
\end{aligned}$$

- $\beta > 3$: For this range, we have,

$$\begin{aligned}
R(T) &\leq 2BN_1 + 8BC_2C_f(1+\log_2 T)N_1^{(3-\beta)/2}\sum_{r=0}^{S-2} 2^{r(3-\beta)/2}\left(\log\left(\frac{N_1}{\delta'}\cdot 2^r\right)\right)^{\beta/2} \\
&\leq 2BN_1 + 8BC_2C_f(1+\log_2 T)\cdot\left(\log\left(\frac{T}{\delta'}\right)\right)^{3/2}\cdot\sum_{r=0}^{S-2} 2^{r(3-\beta)/4}\left[\frac{\log(N_1\cdot 2^r)+\log(1/\delta')}{N_1\cdot 2^{r/2}}\right]^{(\beta-3)/2} \\
&\leq 2BN_1 + 8BC_2C_f(1+\log_2 T)\cdot\left(\log\left(\frac{T}{\delta'}\right)\right)^{3/2}\cdot\left[\frac{\log(N_1/\delta')}{N_1}\right]^{(\beta-3)/2}\cdot\sum_{r=0}^{S-2} 2^{r(3-\beta)/4} \\
&\leq 2BN_1 + 8BC_2C_f(1+\log_2 T)\cdot\left(\log\left(\frac{T}{\delta'}\right)\right)^{3/2}\cdot\left[\frac{\log(N_1/\delta')}{N_1}\right]^{(\beta-3)/2}\cdot\sum_{r=0}^{\infty} 2^{r(3-\beta)/4} \\
&\leq 2BN_1 + \frac{8BC_2C_f}{1-2^{(3-\beta)/4}}\cdot\left(\log\left(\frac{T}{\delta'}\right)\right)^{3/2}(1+\log_2 T).
\end{aligned}$$

  In the third step, we used the fact that $\dfrac{\log(N_1\cdot 2^r/\delta')}{N_1\cdot 2^{r/2}}$ is a decreasing function of r for all $r \geq 0$ and in the fifth step we used the fact that $N_1 \geq \log(N_1/\delta')$ since $N_1 \geq \overline{N}(\delta')$.

On combining all the cases, we arrive at the result. Lastly, note that the relation on $\tilde{N}_{\varepsilon_0}$ guarantees that Assumption 4.2 holds with probability $1-\delta/4$, Lemma B.5 holds with probability $1-\delta/2$ and the relation in Eqn. (27) holds for all epochs simultaneously with probability $1-\delta/4$. Thus, the overall theorem holds with probability $1-\delta$. The statement in Corollary 4.4 follows immediately from the above proof by plugging in $\beta = 1+2\nu/d$.

B.2. Proof of Theorem 4.5

The proof of Theorem 4.5 is almost identical to that of Theorem 4.3. The following lemma is a counterpart to Lemma B.4 for the noisy case.

Lemma B.6. *For all epochs r, the following relation holds with probability at least $1-\delta/6$:*

$$\varsigma_r \leq \begin{cases} 2B & \text{if } r = 1, \\ 4\alpha_\tau(\delta')\left[\sup_{x\in\mathcal{X}_{r-1}}\sigma_{r-1,\tau}(x)\right] + \frac{2B}{T} + R\sqrt{\frac{2}{T\tau}\log\left(\frac{4T}{\delta'}\right)} & \text{if } r \geq 2. \end{cases}$$

The proof of this lemma is identical to that of Lemma B.4 with the definitions of UCB and LCB changed according to the noisy setup (See (Vakili et al., 2021a) for an exact derivation). On using Lemma 4.6 (for the noisy case) along with Lemma B.6, we can rewrite Eqn. (27) as

$$\begin{aligned}
R(T) &\leq 2BN_1 + (M_f+M)(1+\log_2 T)\sum_{r=2}^{S} N_r\cdot\left[4\sqrt{C_\tau}C'_{L_f,L'_f}\alpha_\tau(\delta'/2)\sqrt{\frac{\gamma_{N_{r-1},\tau}}{N_{r-1}}} + \frac{2B}{T} + R\sqrt{\frac{2}{T\tau}\log\left(\frac{4T}{\delta'}\right)}\right] \\
&\leq 2BN_1 + (M_f+M)(1+\log_2 T)\sum_{r=2}^{S} N_r\cdot\left[4\sqrt{C_\tau}C'_{L_f,L'_f}\alpha_\tau(\delta'/2)\sqrt{\frac{\gamma_{T,\tau}}{N_{r-1}}} + \frac{2B}{T} + R\sqrt{\frac{2}{T\tau}\log\left(\frac{4T}{\delta'}\right)}\right],
\end{aligned} \tag{29}$$

where second line follows using monotonicity of $\gamma_{n,\tau}$ i.e., $\gamma_{n_1,\tau} \leq \gamma_{n_2,\tau}$ for all $n_1 \leq n_2$ and C_τ is the leading constant in Eqn. (9). On plugging in the values of N_r in Eqn. (29), we obtain,

$$
\begin{aligned}
R(T) &\leq 2BN_1 + \sum_{r=2}^{S} N_r \cdot \Bigg[4\sqrt{C_\tau} C'_{L_f, L'_f}(M_f + M)\alpha_\tau(\delta'/2)\sqrt{\frac{\gamma_{T,\tau}}{N_{r-1}}} + \frac{2B(M_f+M)}{T} \\
&\qquad + R(M_f + M)\sqrt{\frac{2}{T\tau}\log\left(\frac{4T}{\delta'}\right)}\Bigg](1 + \log_2 T) \\
&\leq 2BN_1 + \sum_{r=2}^{S} \Bigg[4\sqrt{N_1 C_\tau} C_f \alpha_\tau(\delta'/2)\sqrt{\gamma_{T,\tau}} \cdot 2^{r-1} \cdot 2^{-(r-2)/2} + (M_f + M) \cdot \frac{2BN_1}{T} \cdot 2^{r-1} \\
&\qquad + (M_f + M) \cdot RN_1 \sqrt{\frac{2}{T\tau}\log\left(\frac{4T}{\delta'}\right)} \cdot 2^{r-1}\Bigg](1 + \log_2 T) \\
&\leq 2BN_1 + 16\sqrt{N_1 C_\tau} C_f \alpha_\tau(\delta'/2)\sqrt{\gamma_{T,\tau}} \left(\sum_{r=0}^{S-2} 2^{r/2}\right) \log_2 T \\
&\qquad + 2(M_f + M)\log_2 T \cdot \left(\frac{4B}{T} + 2R\sqrt{\frac{2}{T\tau}\log\left(\frac{4T}{\delta'}\right)}\right) N_1 \left(\sum_{r=0}^{S-2} 2^r\right) \\
&\leq 2BN_1 + \frac{16}{\sqrt{2}-1}\sqrt{N_1 C_\tau} C_f \alpha_\tau(\delta'/2)\sqrt{\gamma_{T,\tau}} \cdot \sqrt{\frac{T}{N_1}} \cdot \log_2 T \\
&\qquad + 2(M_f + M)\log_2 T \cdot \left(\frac{4B}{T} + 2R\sqrt{\frac{2}{T\tau}\log\left(\frac{4T}{\delta'}\right)}\right) \cdot N_1 \cdot \frac{T}{N_1} \\
&\leq 2BN_1 + \frac{16}{\sqrt{2}-1}\sqrt{C_\tau} C_f \alpha_\tau(\delta'/2)\sqrt{T\gamma_{T,\tau}} \cdot \log_2 T \\
&\qquad + 8B(M_f + M)\log_2 T + 4R(M_f + M)\sqrt{\frac{2T}{\tau}\log\left(\frac{4T}{\delta'}\right)}\log_2 T,
\end{aligned}
$$

where $C_f = C'_{L_f, L'_f}(M_f + M)$ as before. Hence, $R(T)$ satisfies $\tilde{\mathcal{O}}(\sqrt{T\gamma_{T,\tau}})$, as required. Lastly, the relation on $\tilde{N}'_{\varepsilon_0}$ guarantees that Assumption 4.2 holds with probability $1 - \delta/6$, Lemma B.5 holds with probability $1 - \delta/2$, Lemma B.6 holds with probability $1 - \delta/6$ and the relation in Eqn. (29) (consequence of Lemma 4.6) holds for all epochs simultaneously with probability $1 - \delta/6$. Thus, the overall theorem holds with probability $1 - \delta$.

B.3. Proof of Auxiliary Lemmas

B.3.1. Proof of Lemma B.3

The main ingredient in the proof is the relation: $|f(x) - \mu_{r-1}(x)| \leq B\sigma_{r-1}(x)$, which holds for all $x \in \mathcal{X}_{r-1}$ and across all epochs r. This is a well-known relation in the literature (Vakili et al., 2021a; Lyu et al., 2020) that bounds the predictive performance of the posterior mean in terms of posterior variance.

We use induction to prove the lemma. Since $\mathcal{X}_1 = \mathcal{X}$ and $x^* \in \mathcal{X}$ holds by definition, $x^* \in \mathcal{X}_1$. Assume that $x^* \in \mathcal{X}_{r-1}$. Using the relation $|f(x) - \mu_{r-1}(x)| \leq B\sigma_{r-1}(x)$, we can conclude,

$$
\sup_{x' \in \mathcal{X}_{r-1}} \mathrm{LCB}_{r-1}(x') = \sup_{x' \in \mathcal{X}_{r-1}} (\mu_{r-1}(x') - B\sigma_{r-1}(x')) \leq \sup_{x' \in \mathcal{X}_{r-1}} f(x') = f(x^*) \leq \mathrm{UCB}_{r-1}(x^*),
$$

where we used the inductive hypothesis to establish $\sup_{x' \in \mathcal{X}_{r-1}} f(x') = f(x^*)$. This implies that $x^* \in \mathcal{X}_r$, as required.

B.3.2. Proof of Lemma B.4

We separately show the bounds for $r = 1$ and $r \geq 2$. For the first epoch, we have,

$$
\varsigma_1 = f(x^*) - \inf_{x \in \mathcal{X}_1} f(x) = f(x^*) - \inf_{x \in \mathcal{X}} f(x) \leq 2 \sup_{x \in \mathcal{X}} f(x) \leq 2B.
$$

We used the fact that $\sup_{x\in\mathcal{X}} f(x) = \sup_{x\in\mathcal{X}} f^\top \psi_x \leq \sup_{x\in\mathcal{X}} \|f\|_{\mathcal{H}_k}\|\psi_x\|_{\mathcal{H}_k} \leq B$. Consider any epoch $r \geq 2$. For the analysis, we define

$$\mathcal{X}'_r := \{x \in \mathcal{X}_{r-1} : f(x) + 2B\sigma_{r-1}(x) \geq \sup_{x'\in\mathcal{X}_{r-1}} f(x') - 2B\sigma_{r-1}(x')\}.$$

The region $\mathcal{X}'_r$ satisfies $\mathcal{X}_r \subseteq \mathcal{X}'_r$. To establish this, we once again employ the relation $|f(x) - \mu_{r-1}(x)| \leq B\sigma_{r-1}(x)$. Using the relation, we can conclude that

$$\begin{aligned}
\mathrm{UCB}_{r-1}(x) &= \mu_{r-1}(x) + B\sigma_{r-1}(x) \leq (f(x) + B\sigma_{r-1}(x)) + B\sigma_{r-1}(x) = f(x) + 2B\sigma_{r-1}(x)\\
\mathrm{LCB}_{r-1}(x) &= \mu_{r-1}(x) - B\sigma_{r-1}(x) \geq (f(x) - B\sigma_{r-1}(x)) - B\sigma_{r-1}(x) = f(x) - 2B\sigma_{r-1}(x).
\end{aligned}$$

The inclusion $\mathcal{X}_r \subseteq \mathcal{X}'_r$ follows immediately from the definition of $\mathcal{X}_r$ and $\mathcal{X}'_r$ and the above expressions.

Consider the following relation which holds for any $x \in \mathcal{X}'_r$.

$$\begin{aligned}
f(x) + 2B\sigma_{r-1}(x) &\geq \sup_{x'\in\mathcal{X}_{r-1}} f(x') - 2B\sigma_{r-1}(x')\\
\implies f(x) &\geq \sup_{x'\in\mathcal{X}_{r-1}} [f(x') - 2B\sigma_{r-1}(x')] - \sup_{x''\in\mathcal{X}_{r-1}} [2B\sigma_{r-1}(x'')]\\
&\geq \sup_{x'\in\mathcal{X}_{r-1}} f(x') - \sup_{x''\in\mathcal{X}_{r-1}} [4B\sigma_{r-1}(x'')]\\
&\geq f(x^*) - \sup_{x''\in\mathcal{X}_{r-1}} [4B\sigma_{r-1}(x'')]. \qquad (30)
\end{aligned}$$

In the last line, we used Lemma B.3 to conclude $\sup_{x'\in\mathcal{X}_{r-1}} f(x') = f(x^*)$. Since $\mathcal{X}_r \subset \mathcal{X}'_r$, we can use Eqn. (30) to obtain an upper bound on ς_r as follows:

$$\begin{aligned}
\varsigma_r &= f(x^*) - \inf_{x\in\mathcal{X}_r} f(x)\\
&\leq f(x^*) - \inf_{x\in\mathcal{X}'_r} f(x)\\
&\leq f(x^*) - \left[f(x^*) - \sup_{x'\in\mathcal{X}_{r-1}} 4B\sigma_{r-1}(x')\right]\\
&\leq 4B \sup_{x'\in\mathcal{X}_{r-1}} \sigma_{r-1}(x').
\end{aligned}$$

B.3.3. Proof of Lemma 4.6

We begin with the noiseless case. For brevity, we drop the subscript 0 from the posterior variance corresponding to the noiseless case. Consider a kernel k and let $\mathcal{H}$ and $\mathcal{H}'$ denote the RKHS induced by k on $\mathcal{X}$ and $\mathcal{X}'$. Since $\mathcal{X}' \subset \mathcal{X}$, it is straightforward to note that $\mathcal{H}' \subseteq \mathcal{H}$. Using the result from Wendland (2004, Theorem 10.46), we know that for every $f \in \mathcal{H}'$ there exists a natural extension $\mathscr{E}f \in \mathcal{H}$ such that $\|\mathscr{E}f\|_{\mathcal{H}} = \|f\|_{\mathcal{H}'}$. Consequently, we can conclude $\{f : \|f\|_{\mathcal{H}'} \leq 1\} \subseteq \{f : \|f\|_{\mathcal{H}} \leq 1\}$. Lastly, note that $\mathcal{H}'$ is same as the RKHS of the kernel $k'(x, y) = k(\Gamma(x), \Gamma(y))$ over the domain $\mathcal{X}$. Here Γ denotes the bi-Lipschitian map $\Gamma : \mathcal{X} \to \mathcal{X}'$ as given by Assumption 4.2.

Let $X \subset \mathcal{X}$ be any set of distinct points and $\sigma'_X(x)$ and $\sigma_X(x)$ denote the posterior standard deviation at any point x computed using the kernels k' and k. Using the dual formulation of posterior variance, we have the following relation:

$$\sigma'_X(x) = \sup_{\substack{f\in\mathcal{H}'\\ \|f\|_{\mathcal{H}'}\leq 1\\ f(X)=\{0\}}} f(x) \leq \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}\leq 1\\ f(X)=\{0\}}} f(x) = \sigma_X(x).$$

In the above relation, we used the fact that $\mathcal{H}' \subset \mathcal{H}$ and the unit ball in $\mathcal{H}'$ is contained in the unit ball in $\mathcal{H}$. This implies that the prediction made using the kernel k' has a smaller error than the prediction made by using kernel k. If we set $X = \Gamma^{-1}(X')$[4], then the above is equivalent to saying that the prediction error using kernel k corresponding to set of points $X' \in \mathcal{X}'$ is smaller than the prediction error using kernel k corresponding to set of points $X \in \mathcal{X}$.

[4]For any operator Γ and $X = \{x_1, x_2, \ldots, x_n\}$, we use the shorthand $\Gamma(X)$ for the set $\{\Gamma(x_1), \Gamma(x_2), \ldots, \Gamma(x_n)\}$.

Since the points X' are distributed uniformly in $\mathcal{X}'$, the points $X = \Gamma^{-1}(X')$ are distributed according to density $\vartheta(x) = \frac{\det(\nabla\Gamma(x))}{\mathrm{vol}(\mathcal{X}')}$ for all $x \in \mathcal{X}$, where $\det(A)$ denotes the determinant of a matrix A and $\nabla\Gamma$ denotes the Jacobian of Γ. Note that $\nabla\Gamma$ (and hence the density ϑ) is well-defined almost everywhere (a.e.) as a consequence of Rademacher's theorem (Rudin, 1987, Chp. 7) and Lipschitz continuity of Γ.

Let ϱ_{unif} denote the uniform distribution on $\mathcal{X}$ (i.e., the Lebesgue measure). We construct a (random) subset of X, denoted by Y, as follows. Each point x_i for $i \in \{1, 2, \ldots, n\}$ is added into Y independently of others with probability $c_\vartheta \frac{\varrho_{\mathrm{unif}}(x_i)}{\vartheta(x_i)}$, where $c_\vartheta = \inf_x \frac{\vartheta(x)}{\varrho_{\mathrm{unif}}(x)}$ (where the infimum is taken over where ϑ is well defined). It is straightforward to note that the samples in Y are distributed according to ϱ_{unif}. Using the Bernstein inequality for sum of Bernoulli random variables, we can conclude that $|Y|$, the number of points in Y satisfies the relation $|Y| \geq \frac{c_\vartheta n}{2C_\vartheta}$ with probability $1-\delta$ as long as $\frac{3c_\vartheta n}{16C_\vartheta} \geq \log(2/\delta)$. Here $C_\vartheta = \sup_x \frac{\vartheta(x)}{\varrho_{\mathrm{unif}}(x)}$. Since $Y \subseteq X$, the prediction based on the values of X is no worse than the prediction based on the values of Y. Thus,

$$\sup_{x' \in \mathcal{X}'} \sigma^2_{X'}(x') \leq \sup_{x \in \mathcal{X}} \sigma^2_X(x) \leq \sup_{x \in \mathcal{X}} \sigma^2_Y(x)$$

An identical result holds for the noisy case using an identical series of arguments using the kernel $k_\tau(x, x') = k(x, x) + \tau\delta_{x=x'}$ (Kanagawa et al., 2018), where $\delta_{x=x'}$ denotes the dirac delta function. We can invoke the result from Theorem 3.1 for uniform samples on $\mathcal{X}$ to bound $\sigma^2_Y(x)$ under both the noisy and noiseless settings to obtain the following relations

$$\sup_{x' \in \mathcal{X}'} \sigma^2_{X',\tau}(x') \leq \sup_{x \in \mathcal{X}} \sigma^2_{Y,\tau}(x) \leq \frac{C_\vartheta}{c_\vartheta} \cdot \frac{216}{13} \cdot F^2\tau \cdot \frac{\gamma_{n,\tau}}{n},$$
$$\sup_{x' \in \mathcal{X}'} \sigma^2_{X',0}(x') \leq \sup_{x \in \mathcal{X}} \sigma^2_{Y,0}(x) \leq \frac{C_\vartheta}{c_\vartheta} \cdot \frac{216}{13} \cdot F^2 \cdot n^{1-\beta}.$$

We only need to obtain a bound the ratio C_ϑ/c_ϑ that is independent of n to complete the proof. Using the Lipschitzness of Γ and Γ^{-1}, we can conclude that

$$L_f'^{-d} \leq |\det(\nabla\Gamma)| \leq L_f^d.$$

Using the definition of c_ϑ, we have,

$$c_\vartheta = \inf_x \frac{\vartheta(x)}{\varrho_{\mathrm{unif}}(x)} = \inf_x \frac{\det(\nabla\Gamma(x))\mathrm{vol}(\mathcal{X})}{\mathrm{vol}(\mathcal{X}')} \geq \frac{\mathrm{vol}(\mathcal{X})}{L_f'^d \mathrm{vol}(\mathcal{X}')} = \tilde{L_f}'^{-d}.$$

Similarly,

$$C_\vartheta = \sup_x \frac{\vartheta(x)}{\varrho_{\mathrm{unif}}(x)} = \sup_x \frac{\det(\nabla\Gamma(x))\mathrm{vol}(\mathcal{X})}{\mathrm{vol}(\mathcal{X}')} \leq \frac{L_f^d \mathrm{vol}(\mathcal{X})}{\mathrm{vol}(\mathcal{X}')} = \tilde{L_f}^d.$$

Hence, $C_\vartheta/c_\vartheta \leq (\tilde{L}_f/\tilde{L}'_f)^d$ depends only on $(\tilde{L}_f, \tilde{L}'_f)$ and is independent of n, as required.

B.4. Proof of Lemma B.5

In order to establish this bound, we claim that it is sufficient to show that the number of connected components increase by at most $M_f + M$ in each epoch. If the number of connected components increase by at most $M_f + M$ in each epoch, then the total number of connected components in any epoch is bounded by $(M_f + M)(1 + \log_2(T/N_1))$. This follows from the fact that there are at most $1 + \log_2(T/N_1)$ epochs as the epoch lengths double every time. Thus, for the rest of the proof we focus on establishing the relation that the number of connected components increase by at most $M_f + M$ in each epoch.

To obtain this relation, we bound the number of connected components in $\mathcal{X}_{r+1} = \mathcal{L}^{\mathrm{UCB}_{N_{r+1}}(\cdot;\mathcal{X}_r)}_{\eta_{r+1}}$ where η_{r+1} is the threshold corresponding to the algorithm for a general epoch r. By Assumption 4.2, these are M more than those in $\mathcal{X}_r \cap \mathcal{L}^f_{\eta_{r+1}}$. Since $\mathcal{X}_r$ is the level set of upper confidence bound, which is always greater than f, it contains the level sets of f at the same threshold. Thus, every connected component of $\mathcal{L}^f_{\eta_{r+1}}$ completely belongs to a connected component of $\mathcal{X}_r$. Consequently, number of connected components in $\mathcal{X}_r \cap \mathcal{L}^f_{\eta_{r+1}}$ is at most sum of the number of connected components in $\mathcal{X}_r$ and $\mathcal{L}^f_{\eta_{r+1}}$. Thus, at every iteration, the number of connected components increases at most by $M_f + M$. Lastly, based on Assumption 4.2, this relation holds for each epoch with probability $1 - \delta/(2\log_2 T)$. Using a union bound argument, we can conclude that this holds for all epochs r with probability at least $1 - \delta/2$.

C. Additional Details on the Experimental Setup

We compare the regret performance and the running time of the BPE (Li & Scarlett, 2022), GP-ThreDS (Salgia et al., 2021) and REDS algorithm (Algorithm 1) for three commonly used benchmark functions in Bayesian Optimization, namely, Branin (Azimi et al., 2012; Picheny et al., 2013), Hartmann-4D (Picheny et al., 2013) and Hartmann-6D (Picheny et al., 2013). The analytical expressions for the three benchmark functions are given as follows:

- Branin function, denoted by $B(x_1, x_2)$, is defined over $\mathcal{X} = [0, 1]^2$.

$$B(x_1, x_2) = -\frac{1}{51.95}\left(\left(v - \frac{5.1u^2}{4\pi^2} + \frac{5u}{\pi} - 6\right)^2 + \left(10 - \frac{10}{8\pi}\right)\cos(u) - 44.81\right),$$

  where $u = 15x_1 - 5$ and $v = 15x_2$.

- Hartmann-4D function, denoted by $H_4(x_1, x_2, x_3, x_4)$, is defined over $\mathcal{X} = [0, 1]^4$.

$$H_4(x_1, x_2, x_3, x_4) = \sum_{i=1}^{4} w_i \exp\left(-\sum_{j=1}^{4} A_{ij}(x_j - C_{ij})^2\right).$$

- Hartmann-6D function, denoted by $H_6(x_1, x_2, x_3, x_4, x_5, x_6)$, is defined over $\mathcal{X} = [0, 1]^6$.

$$H_6(x_1, x_2, x_3, x_4, x_5, x_6) = \sum_{i=1}^{4} w_i \exp\left(-\sum_{j=1}^{6} A_{ij}(x_j - C_{ij})^2\right).$$

In the definitions above, A_{ij} and C_{ij} refer to the $(i, j)^{\text{th}}$ element of the matrices A and C and w_i denotes the ith element of the vector w, defined below:

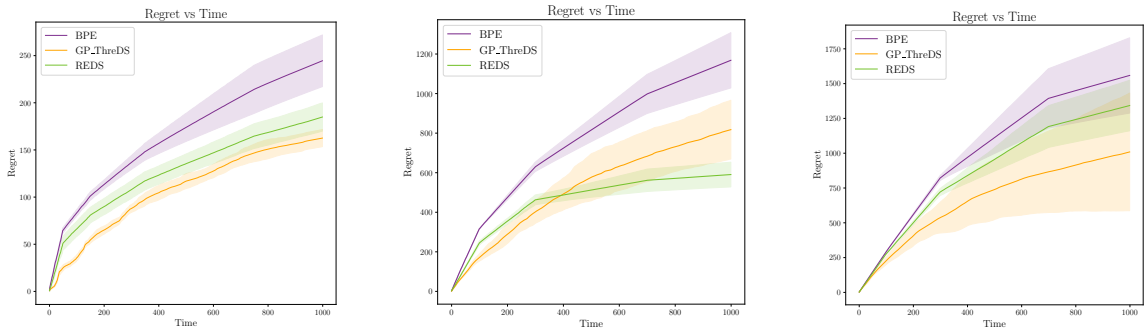
$$w = \begin{pmatrix} 1.0 & 1.2 & 3.0 & 3.2 \end{pmatrix}^\top$$

$$A = \begin{pmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{pmatrix}$$

$$C = 10^{-4} \cdot \begin{pmatrix} 1312 & 1696 & 5569 & 124 & 8283 & 5886 \\ 2329 & 4135 & 8307 & 3736 & 1004 & 9991 \\ 2348 & 1451 & 3522 & 2883 & 3047 & 6650 \\ 4047 & 8828 & 8732 & 5743 & 1091 & 381 \end{pmatrix}$$

For BPE and REDS, we consider a discretized version of the domain consisting of 2000, 7000 and 20000 points chosen uniformly at random from the domain for the Branin, Hartmann-4D and Hartmann-6D functions respectively. We use the exponentially growing epoch schedule for both BPE and REDS as described in (Algorithm 1) for a fair comparison. We implement GP-ThreDS as described in (Salgia et al., 2021). For each node in the tree, we consider a discretization, chosen uniformly at random, of size 100, 200 and 500 for the Branin, Hartmann-4D and Hartmann-6D functions respectively. The values of (a, b) (the lower and upper bound on $f(x^*)$) are set to $(0.5, 1.2)$, $(0, 3.8)$ and $(0, 3.5)$ for Branin, Hartmann-4D and Hartmann-6D respectively. We set $\tau = 0.2$ for all experiments. The value of α_τ is set to 1 across all experiments, except for BPE with Hartmann-4D and Hartmann-6D for which we set it to 0.75. These values are obtained using a grid search over $[0.25, 2]$ in steps of 0.25. The parameter N_1 in REDS and BPE was set to 50 for Branin and 100 for Hartmann-4D and Hartmann-6D functions. For the implementation of REDS, we sample uniformly directly from the entire domain $\mathcal{X}_r$. We do not separately consider each connected component as it is not possible to define them for discrete domain.

For all the experiments, we used the Square exponential kernel. The length scale was set to 0.2 for Branin and 1 for Hartmann-4D and Hartmann-6D functions. We corrupted the observations with a zero mean Gaussian noise to the with a standard deviation of 0.2. All the algorithms were run for $T = 1000$ time steps. We recorded the cumulative regret and time taken by different algorithms for 10 Monte Carlo runs for each benchmark function.



(a) Branin (b) Hartmann-4D (c) Hartmann-6D

Figure 2: Cumulative regret averaged over 10 Monte Carlo runs for all algorithms across different benchmark functions. The shaded region represents the error bars upto one standard deviation. As evident from the plots, the regret of REDS is comparable to that of BPE and GP-ThreDS.

	BPE	GP-ThreDS	REDS
Branin	29.84 ± 6.13	4.37 ± 0.28	$\mathbf{0.32} \pm 0.08$
Hartmann-4D	38.45 ± 3.93	7.59 ± 0.54	$\mathbf{0.47} \pm 0.11$
Hartmann-6D	119.71 ± 23.75	19.33 ± 0.54	$\mathbf{1.19} \pm 0.08$

Table 2: Time taken (in seconds) by different algorithms across the different benchmark functions.

The regret for the algorithms over different functions is plotted in Figure 2. The shaded region represents the error bars upto standard deviation on either side. The running times, with an error bar of one standard deviation, are tabulated in Table 2. As evident from the plots in Figure 2, the regret incurred by REDS is comparable to that of other algorithms for all benchmark functions. At the same time, REDS offers about a $15\times$ and $100\times$ speedup in terms of runtime over the GP-ThreDS and BPE (See Table 2), demonstrating the practical benefits of our proposed methodology of random sampling.

D. Additional Discussion on Assumption 4.2

In this section, we discuss some additional details about Assumption 4.2. At a high level, the assumption is required to guarantee the topological regularity of level sets at each iteration, which allows for analytical convenience in the proof of Theorems 4.3 and 4.5. Note that during any epoch r, the active domain $\mathcal{X}_r$ is a collection of closed, connected components and hence is compact. This implies that we can directly invoke Theorem 3.1 to ensure that the posterior variance corresponding to N randomly sampled points from this domain decays as $\sqrt{\gamma_N/N}$. However, the leading constant in Theorem 3.1 implicitly depends on the geometry of the domain. Thus, a direct application of Theorem 3.1 results in appearance of leading constants that are difficult to characterize and may potentially affect the order of regret with T. Assumption 4.2 allows us to circumvent this issue by providing a means to precisely characterize this constant and consequently the order of regret in terms of number of samples. By considering each connected component individually, we can ensure that each component is explored irrespective of their volume. While the adaptive sampling approaches that use an acquisition function implicitly account for this in the step where the acquisition function is maximized, we need to explicitly incorporate this into algorithm design as small regions are less likely to ne sampled than the larger ones under our non-adaptive scheme. Moreover, the topological regularity of each connected component ensures that the leading constant in Theorem 4.2 cannot be much worse than that corresponding to the original domain $\mathcal{X}$. See Lemma 4.6 and its proof in Sec. B.3.3 for more details. We believe that this is a minor technical requirement in our analysis and can potentially be improved with a combination of different algorithm design and refined analytical tools.

While it is not possible to completely eliminate the need for Assumption 4.2 in our work, we propose below two variants of the assumption that are potentially more intuitive or less restrictive.

D.1. Replacing UCB with posterior mean

Instead of assuming regularity of level sets of the Upper Confidence Bound (UCB) of f, it is sufficient to assume the regularity of level sets of the posterior mean of f constructed using the randomly sampled points. Since the posterior mean represents the underlying f more closely than the UCB, especially in the noise-free case, it is more intuitive to understand the regularity of level sets of f being endowed to that of posterior mean when compared to being endowed to that of UCB.

Under this variant of Assumption 4.2, it is sufficient to modify the update rule in line 10 of Alg. 1 to the following for noise-free observations:

$$\mathcal{X}_{r+1} = \left\{ x \in \mathcal{X}_r \middle| \mu_r(x) \geq \sup_{x' \in \mathcal{X}_r} \mu_r(x') - 2BC\sqrt{\frac{\gamma_{N_r}}{N_r}} \right\},$$

where C is the leading constant in Lemma 4.6. For the noisy case, the update rule can be modified to

$$\mathcal{X}_{r+1} = \left\{ x \in \mathcal{X}_r \middle| \mu_{r,\tau}(x) \geq \sup_{x' \in \mathcal{X}_r} \mu_{r,\tau}(x') - 2\alpha_{\tau,\delta} C\sqrt{\frac{\gamma_{N_r}}{N_r}} - 2c_{T,\tau,\delta} \right\}.$$

It is straightforward to note that even after this modification, Lemma B.4 continues to hold by replacing the term $\sup_{x \in \mathcal{X}_{r-1}} \sigma_{r-1}(x)$ with its upper bound from Lemma 4.6. Thus, the analysis goes through almost as is with this modification, retaining the original order of the regret.

D.2. Replacing the value of η_0

In the current version of Assumption 4.2, $\eta_0 = f(x^*) - \varepsilon_0$ is a non-adaptively chosen constant value. Such a bound is useful if ε_0 is tuned or known to be smaller than the range of the function. Often, it might be not be possible to obtain to know such a bound ahead of time. In such a case, we replace the non-adaptive bound with an adaptive one. In particular, we can set $\eta_0 := c \sup f + (1-c) \inf f$ for some $c \in (0,1)$. Such a choice gives us the advantage of being adaptive to the range thereby making the algorithm more robust to the knowledge of ε_0.

To ensure the condition that we only encounter level sets for $\eta \geq cf(x^*) + (1-c)f(x_*)$ during the algorithm, i.e., after the first epoch, the termination criteria of the first epoch needs to be updated to an adaptive one as opposed to the current non-adaptive strategy of having a fixed bound. Here $x_* \in \arg\min_{x \in \mathcal{X}} f$. Specifically, after taking $C_{L_f, L'_f} \overline{N}(\delta/\log_2 T)$ samples, we continue to run the first epoch until the following condition is met:

$$c \cdot \sup \text{UCB} + (1-c) \inf \text{UCB} \leq \sup \text{LCB} - 2B \sup_{x \in \mathcal{X}} \sigma(x).$$

For the remaining epochs, we continue to run the algorithm as earlier.

We claim that the above termination condition ensures that the algorithm only encounters level sets corresponding to $\eta \geq \eta_0 = c \sup f + (1-c) \inf f$. The proof of the claim is straightforward. Note that the following relation holds by definition of UCB:

$$c \cdot \sup \text{UCB} + (1-c) \inf \text{UCB} \geq cf(x^*) + (1-c)f(x_*).$$

Thus, after the termination condition is met, we will always have

$$\{x : \text{UCB}(x) \geq \sup \text{LCB}(x)\} \subseteq \{x : f(x) \geq \sup \text{LCB}(x) - 2B \sup_{x \in \mathcal{X}} \sigma(x)\} \subseteq \{x : f(x) \geq cf(x^*) + (1-c)f(x_*)\},$$

as required.

Moreover, we also claim that this modified version of REDS also achieves the same regret guarantees as the original version. In order to establish the claim, we only need to bound the regret in the first epoch as the regret analysis for the remaining epochs carries through as is. Moreover, WLOG we assume this happens after $C_{L_f, L'_f} \overline{N}(\delta/\log_2 T)$ samples have been taken. Otherwise, anyway the current analysis goes through. To bound the regret in the first epoch, we first obtain a bound on the number of samples taken in the first epoch. To this effect, note that the termination condition would have been definitely satisfied if

$$cf(x^*) + (1-c)f(x_*) + 2B \sup_{x \in \mathcal{X}} \sigma(x) \leq \sup \text{LCB} - 2B \sup_{x \in \mathcal{X}} \sigma(x) \leq f(x^*) - 4B \sup_{x \in \mathcal{X}} \sigma(x),$$

or equivalently,

$$6B \sup_{x \in \mathcal{X}} \sigma(x) \leq cf(x^*) + (1-c)f(x_*).$$

Using Theorem 3.1 to bound the left hand side of the above relation, we can conclude that $N_1 = \mathcal{O}((f(x^*) - f(x_*))^{\frac{2}{1-\beta}})$. Since the instantaneous regret is trivially bounded by $f(x^*) - f(x_*)$, the regret in the first epoch is bounded by $C \cdot \min\{(f(x^*) - f(x_*))^{\frac{3-\beta}{1-\beta}}, T \cdot (f(x^*) - f(x_*))\} = \mathcal{O}(T^{\frac{3-\beta}{2}})$, matching the overall regret bound of REDS.