Improved Regret Bounds for Gaussian Process Upper Confidence Bound in Bayesian Optimization

Shogo Iwazaki
LY Corporation
Tokyo, Japan
siwazaki@lycorp.co.jp

Abstract

This paper addresses the Bayesian optimization problem (also referred to as the Bayesian setting of the Gaussian process bandit), where the learner seeks to minimize the regret under a function drawn from a known Gaussian process (GP). Under a Matérn kernel with a certain degree of smoothness, we show that the Gaussian process upper confidence bound (GP-UCB) algorithm achieves $\widetilde{O}(\sqrt{T})$ cumulative regret with high probability. Furthermore, our analysis yields $O(\sqrt{T \ln^2 T})$ regret under a squared exponential kernel. These results fill the gap between the existing regret upper bound for GP-UCB and the best-known bound provided by Scarlett [46]. The key idea in our proof is to capture the concentration behavior of the input sequence realized by GP-UCB, enabling a more refined analysis of the GP's information gain.

1 Introduction

We study the Bayesian optimization (BO) problem, where the learner seeks to minimize the regret under a random function drawn from a known Gaussian process (GP) [18, 19]. Throughout this paper, we focus on the GP-UCB algorithm [51], which combines the posterior distribution of GP with the optimism principle. Due to its simple algorithm construction and general theoretical framework provided by Srinivas et al. [51], GP-UCB has played an important role in the advancement of the BO field. On the other hand, our theoretical understanding of the performance of GP-UCB has not been improved from [51] in the Bayesian setting, while its frequentist counterpart is studied in several existing works [11, 61]. Specifically, the current regret upper bound for GP-UCB, as provided by Srinivas et al. [51], is known to be worse than that of the algorithm in [46], which achieves state-of-the-art $O(\sqrt{T \ln T})$ cumulative regret. Then, the natural question is whether there is further room for improvement in the existing regret upper bound of GP-UCB. This paper provides an affirmative answer to this question by showing that GP-UCB achieves $\widetilde{O}(\sqrt{T})$ regret with high probability.

Contribution. We summarize our contributions as follows.

- We show that the GP-UCB proposed by Srinivas et al. [51] achieves $\widetilde{O}(\sqrt{T})$ regret with high probability under a Matérn kernel with a certain degree of smoothness (precise condition is provided in Theorem 3). Here, $\widetilde{O}(\cdot)$ is the order notation that hides polylogarithmic dependence. This result is comparable to state-of-the-art $O(\sqrt{T \ln T})$ regret provided by Scarlett [46] up to a polylogarithmic factor and strictly improves upon the existing $\widetilde{O}(T^{\frac{\nu+d}{2\nu+d}})$ upper bound of GP-UCB [51, 58]. Here, d and ν denote the dimension of the input domain and smoothness parameter, respectively.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

- Furthermore, for a squared exponential kernel, we establish $O\left(\sqrt{T \ln^2 T}\right)$ cumulative regret of GP-UCB. This improves the existing $O\left(\sqrt{T \ln^{d+2} T}\right)$ upper bound provided by Srinivas et al. [51] for any $d \geq 1$.
- The key idea behind our analysis is to refine the existing information gain bounds by leveraging algorithm-dependent behavior and sample path properties of the GP. We also discuss the applicability of this technique to other algorithms and settings in Section 4.

1.1 Related Works

BO has been extensively studied in the past few decades. Some of them are constructed so as to maximize the utility-based acquisition function defined through the GP posterior, including expected improvement [37], knowledge gradient [17], and the entropy-based algorithms [24]. The theoretical aspect of BO has also been actively studied through the lens of the bandit algorithms, such as GP-UCB [51], Thompson sampling [43], and information directed sampling [44]. In contrast to the noisy observation setting, which these algorithms focus on, algorithms for the noise-free setting form a separate line of research [14, 23, 32]. Extensions of these algorithms to more advanced settings have also been well-studied, e.g., contextual [34], parallel observation [15], high-dimensional [29], time-varying [6], and multi-fidelity setting [30]. Unlike the Bayesian assumption on the objective function adopted in this paper, existing works also extensively study the frequentist assumption of the function, which is also referred to as the frequentist setting of BO or GP bandits [7, 9, 11, 26, 35, 45, 47, 56, 59].

Among the existing studies, [46] is closely related to this paper, which propose a successive elimination-based algorithm and shows an $O(\sqrt{T \ln T})$ upper bound and an $\Omega(\sqrt{T})$ lower bound of the cumulative regret for a one-dimensional BO problem. The fundamental theoretical assumptions and the high-level idea of our analysis are built on the proof provided by Scarlett [46]. Following [46], Wang et al. [60] also proves similar regret guarantees under the one-dimensional Brownian motion.

In addition to [46], some parts of our analysis are inspired by the technique leveraged in [8, 28]. Firstly, Cai et al. [8] studies the GP-UCB algorithm through a relaxed version of regret, which is called *lenient regret*. In our analysis, the cumulative regret is decomposed into the lenient regret-based term, and we leverage their technique to analyze it. Secondly, Janz et al. [28] proposed the input partitioning-based algorithm for obtaining a superior regret in the frequentist setting. Roughly speaking, the high-level idea of their analysis is based on the fact that tighter information gain bounds can be obtained within a properly shrinking partition of the input. The key idea provided in Section 3.1 is motivated by this fact, while our analysis itself is substantially different from that in [28].

2 Preliminaries

Let $f : \mathcal{X} \to \mathbb{R}$ be a black-box objective function whose input domain $\mathcal{X}$ is $\mathcal{X} \coloneqq [0, r]^d$ with some $r > 0$. At each step $t \in \mathbb{N}_+$, the learner chooses a query point $\boldsymbol{x}_t \in \mathcal{X}$, and then receives a noisy observation $y_t = f(\boldsymbol{x}_t) + \epsilon_t$. Here, ϵ_t is a mean-zero noise random variable. We consider a Bayesian setting, where the objective function f and the noise sequence (ϵ_t) are drawn from a known zero-mean Gaussian process (GP) and a Gaussian distribution, respectively. We formally describe it using the following assumptions.

Assumption 1. *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the known positive definite kernel with $\forall \boldsymbol{x} \in \mathcal{X}, k(\boldsymbol{x}, \boldsymbol{x}) \leq 1$. Then, assume $f \sim \mathcal{GP}(0, k)$, where $\mathcal{GP}(0, k)$ denotes the mean-zero GP characterized by the covariance function k.*

Assumption 2. *The noise sequence $(\epsilon_t)_{t \in \mathbb{N}_+}$ is mutually independent. Furthermore, assume $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$, where $\sigma > 0$ is the known constant. Here, $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2.*

These are standard sets of assumptions in the existing theory of BO [43, 51]. Specifically, in Assumption 1, we focus on the following squared exponential (SE) kernel k_{SE} and Matérn kernel

Algorithm 1 Gaussian process upper confidence bound

Require: Kernel k, confidence width parameters $(\beta_t^{1/2})_{t\in\mathbb{N}_+}$.
1: **for** $t = 1, 2, \ldots$ **do**
2: $\quad \boldsymbol{x}_t \leftarrow \arg\max_{\boldsymbol{x}\in\mathcal{X}} \mu(\boldsymbol{x}; \mathbf{X}_{t-1}, \boldsymbol{y}_{t-1}) + \beta_t^{1/2}\sigma(\boldsymbol{x}; \mathbf{X}_{t-1})$.
3: $\quad$ Observe y_t and update the posterior mean and variance.
4: **end for**

$k_{\text{Matérn}}$:

$$k_{\text{SE}}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \exp\left(-\frac{\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|_2^2}{2\ell^2}\right), \quad k_{\text{Matérn}}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|_2}{\ell}\right)^{\nu} J_\nu\left(\frac{\sqrt{2\nu}\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|_2}{\ell}\right), \tag{1}$$

where $\ell > 0$ and $\nu > 0$ are the known lengthscale and smoothness parameters, respectively. In addition, $J_\nu(\cdot)$ and $\Gamma(\cdot)$ respectively denote modified Bessel and Gamma functions. Under Assumptions 1 and 2, the learner can infer the function f through the GP posterior distribution. Let $\mathcal{H}_t \coloneqq (\boldsymbol{x}_i, y_i)_{i\le t}$ be the history that the learner obtained up to the end of step t. Given $\mathcal{H}_t$, the posterior distribution of f is again GP, whose posterior mean and variance are respectively defined as

$$\mu(\boldsymbol{x}; \mathbf{X}_t, \boldsymbol{y}_t) = \boldsymbol{k}(\mathbf{X}_t, \boldsymbol{x})^\top(\mathbf{K}(\mathbf{X}_t, \mathbf{X}_t) + \sigma^2\boldsymbol{I}_t)^{-1}\boldsymbol{y}_t, \tag{2}$$

$$\sigma^2(\boldsymbol{x}; \mathbf{X}_t) = k(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}(\mathbf{X}_t, \boldsymbol{x})^\top(\mathbf{K}(\mathbf{X}_t, \mathbf{X}_t) + \sigma^2\boldsymbol{I}_t)^{-1}\boldsymbol{k}_t(\mathbf{X}_t, \boldsymbol{x}), \tag{3}$$

where $\boldsymbol{k}(\mathbf{X}_t, \boldsymbol{x}) \coloneqq [k(\boldsymbol{x}, \widetilde{\boldsymbol{x}})]_{\boldsymbol{x}\in\mathbf{X}_t}$ and $\boldsymbol{y}_t \coloneqq (y_1, \ldots, y_t)^\top$ are the t-dimensional kernel and output vectors, respectively. Here, we set $\mathbf{X}_t = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t)$. Furthermore, $\mathbf{K}(\mathbf{X}_t, \mathbf{X}_t) \coloneqq [k(\boldsymbol{x}, \widetilde{\boldsymbol{x}})]_{\boldsymbol{x},\widetilde{\boldsymbol{x}}\in\mathbf{X}_t}$ and $\boldsymbol{I}_t$ denote $t \times t$-gram matrix and $t \times t$-identity matrix, respectively.

Learner's goal. Under the total step size $T \in \mathbb{N}_+$, the learner's goal is to minimize the cumulative regret $R_T \coloneqq \sum_{t\in[T]} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)$, where $\boldsymbol{x}^* \in \operatorname{argmax}_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$ and $[T] = \{1, \ldots, T\}$.

Maximum information gain. To quantify the regret, the existing theory utilizes the following information-theoretic quantity $\gamma_T(\mathcal{X})$ arising from GP:

$$\gamma_T(\mathcal{X}) = \sup_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T\in\mathcal{X}} I(\mathbf{X}_T), \text{ where } I(\mathbf{X}_T) = \frac{1}{2}\ln\det(\boldsymbol{I}_T + \sigma^{-2}\mathbf{K}(\mathbf{X}_T, \mathbf{X}_T)). \tag{4}$$

The quantity $\gamma_T(\mathcal{X})$ is referred to as the *maximum information gain* (MIG) over $\mathcal{X}$ [51], since $I(\mathbf{X}_T)$ equals the mutual information between the function values $(f(\boldsymbol{x}_t))_{t\in[T]}$ and the outputs $(y_t)_{t\in[T]}$ under Assumptions 1 and 2, and the input sequence $\mathbf{X}_t = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t)$. MIG plays a vital role in the theoretical analysis of BO, and its increasing speed is analyzed in several commonly used kernels. For example, $\gamma_T(\mathcal{X}) = O(\ln^{d+1} T)$ as $T \to \infty$ under $k = k_{\text{SE}}$ [51]. For the notational convenience, we also define $\gamma_i(\mathcal{X}) = \gamma_{\lceil i\rceil}(\mathcal{X})$ for any non-integer $i > 0$.

Probabilistic property of GP sample path. The existing theory of GP-UCB under the Bayesian setting utilizes the regularity conditions of the realized sample path of GP. We summarize the existing known properties of the GP sample path in the following lemmas.

Lemma 1 (Lipchitz condition of sample path, e.g., [51])**.** *Suppose $k = k_{\text{SE}}$ or $k = k_{\text{Matérn}}$ with $\nu > 2$. Assume Assumption 1. Then, there exist the constants $a, b > 0$ such that*

$$\forall L > 0,\ \mathbb{P}\left(\forall \boldsymbol{x}, \widetilde{\boldsymbol{x}} \in \mathcal{X},\ |f(\boldsymbol{x}) - f(\widetilde{\boldsymbol{x}})| \le L\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|_1\right) \ge 1 - da\exp\left(-\frac{L^2}{b^2}\right). \tag{5}$$

Lemma 2 (Sample path condition for the global maximizer, e.g., [13, 14, 46])**.** *Suppose $k = k_{\text{SE}}$ or $k = k_{\text{Matérn}}$ with $\nu > 2$. Assume Assumption 1. Then, for any $\delta_{\text{GP}} \in (0, 1)$, there exist the strictly positive constants $c_{\text{gap}}, c_{\text{sup}}, c_{\text{lin}}, c_{\text{quad}}, \rho_{\text{lin}}, \rho_{\text{quad}} > 0$ such that the following statements simultaneously hold with probability at least $1 - \delta_{\text{GP}}$:*

1. *The function f has a unique maximizer $\boldsymbol{x}^* \in \mathcal{X}$ such that $f(\boldsymbol{x}^*) > f(\widetilde{\boldsymbol{x}}^*) + c_{\text{gap}}$ holds for any local maximizer $\widetilde{\boldsymbol{x}}^* \in \mathcal{X}$ of f.*

2. *The sup-norm of the sample path is bounded as* $\|f\|_\infty \leq c_{\sup}$.

3. *If the maximizer* $\boldsymbol{x}^*$ *is at the boundary of* $\mathcal{X}$*, the function* f *satisfies* $\forall \boldsymbol{x} \in \mathcal{B}_2(\rho_{\mathrm{lin}}; \boldsymbol{x}^*),\ f(\boldsymbol{x}^*) - c_{\mathrm{lin}}\|\boldsymbol{x}^* - \boldsymbol{x}\|_2 \geq f(\boldsymbol{x})$*, where* $\mathcal{B}_2(\rho_{\mathrm{lin}}; \boldsymbol{x}^*) \coloneqq \{\boldsymbol{x} \in \mathcal{X} \mid \|\boldsymbol{x}^* - \boldsymbol{x}\|_2 \leq \rho_{\mathrm{lin}}\}$ *is the L2-ball on* $\mathcal{X}$*, whose radius and center are* ρ_{lin} *and* $\boldsymbol{x}^*$*, respectively.*

4. *If the maximizer* $\boldsymbol{x}^*$ *is a internal point of* $\mathcal{X}$*, the function* f *satisfies* $\forall \boldsymbol{x} \in \mathcal{B}_2(\rho_{\mathrm{quad}}; \boldsymbol{x}^*),\ f(\boldsymbol{x}^*) - c_{\mathrm{quad}}\|\boldsymbol{x}^* - \boldsymbol{x}\|_2^2 \geq f(\boldsymbol{x})$.

Lemma 1 states that the sample path f of GP is a Lipschitz function with high probability. This property is leveraged in the theory of GP-UCB to control the discretization error arising from the confidence bound construction in the continuous input domain. As described in [51], Lemma 1 is a direct consequence of Theorem 5 in [21] under the existence of fourth-order mixed partial derivatives of the kernel, which are satisfied under $k = k_{\mathrm{SE}}$ and $k = k_{\mathrm{Matérn}}$ with $\nu > 2$[1]. Lemma 2 specifies the regularity condition of f related to the maximizer $\boldsymbol{x}^*$. Here, property 1 is implied from the fact that the GP-sample path has a unique maximizer almost surely under k_{SE} and $k_{\mathrm{Matérn}}$ [e.g., Lemma 2.6 in 33]. Property 2 is implied from, e.g., the compactness of $\mathcal{X}$ and the almost-sure continuity of the sample path under k_{SE} and $k_{\mathrm{Matérn}}$. Properties 3 and 4 also hold automatically under $k = k_{\mathrm{SE}}$ and $k = k_{\mathrm{Matérn}}$ with $\nu > 2$ and are used in existing works. See Theorem 5 in [13], Assumption 3 in [46], and the discussions provided by them for further details. Note that the properties in Lemma 2 are not used in the existing proof of GP-UCB in [51]. As described in the next section, we analyze the realized input sequence $\mathbf{X}_T$ of GP-UCB by relating it to conditions in Lemma 2.

Summary of existing analysis of GP-UCB. We briefly summarize the existing analysis of GP-UCB (Algorithm 1) provided by Srinivas et al. [51]. Based on Assumptions 1 and 2, we can construct the high-probability confidence bound of the underlying function value $f(\boldsymbol{x})$ for each $\boldsymbol{x}$ and $t \in \mathbb{N}_+$ through the posterior distribution of $f(\boldsymbol{x})$. Specifically, by choosing a properly designed finite representative input set $\mathcal{X}_t \subset \mathcal{X}$ and taking into account the discretization error with Lemma 1, Srinivas et al. [51] showed the following events hold simultaneously with probability at least $1 - \delta$:

1. **Confidence bound.** For any $t \in \mathbb{N}_+$, the function value at the queried point $\boldsymbol{x}_t$ satisfies $\mu(\boldsymbol{x}_t; \mathbf{X}_{t-1}, \boldsymbol{y}_{t-1}) - \beta_t^{1/2}\sigma(\boldsymbol{x}_t; \mathbf{X}_{t-1}) \leq f(\boldsymbol{x}_t)$. Furthermore, for any $t \in \mathbb{N}_+$, any function value $f(\boldsymbol{x})$ on $\mathcal{X}_t$ satisfies $f(\boldsymbol{x}) \leq \mu(\boldsymbol{x}; \mathbf{X}_{t-1}, \boldsymbol{y}_{t-1}) + \beta_t^{1/2}\sigma(\boldsymbol{x}; \mathbf{X}_{t-1})$.
2. **Discretization error.** The discretization error arising from $\mathcal{X}_t$ is at most $1/t^2$. Namely, $|f(\boldsymbol{x}) - f([\boldsymbol{x}]_t)| \leq 1/t^2$ holds for any $\boldsymbol{x} \in \mathcal{X}$ and $t \in \mathbb{N}_+$, where $[\boldsymbol{x}]_t$ denotes one of the closest points of $\boldsymbol{x}$ on $\mathcal{X}_t$.

In the above statements, $\beta_t^{1/2}$ is chosen based on the constants a, b in Lemma 1 and the length r of $\mathcal{X}$, and is defined as

$$\beta_t = 2\ln\frac{2t^2\pi^2}{3\delta} + 2d\ln\left(t^2 dbr\sqrt{\ln\frac{4da}{\delta}}\right). \tag{6}$$

The above two events and the UCB-selection rule for $\boldsymbol{x}_t$ imply

$$R_T = \sum_{t=1}^{T} f(\boldsymbol{x}^*) - f([\boldsymbol{x}^*]_t) + \sum_{t=1}^{T} f([\boldsymbol{x}^*]_t) - f(\boldsymbol{x}_t) \leq \frac{\pi^2}{6} + 2\beta_T^{1/2}\sum_{t=1}^{T}\sigma(\boldsymbol{x}_t; \mathbf{X}_{t-1}). \tag{7}$$

In the above expression, the upper bound $\sum_{t=1}^{T} f(\boldsymbol{x}^*) - f([\boldsymbol{x}^*]_t) \leq \sum_{t=1}^{T} 1/t^2 \leq \pi^2/6$ follows from the second event (discretization error). The inequality $\sum_{t=1}^{T} f([\boldsymbol{x}^*]_t) - f(\boldsymbol{x}_t) \leq 2\beta_T^{1/2}\sum_{t=1}^{T}\sigma(\boldsymbol{x}_t; \mathbf{X}_{t-1})$ also follows from the first event (confidence bound) and the definition of $\boldsymbol{x}_t$. See the proof of Theorem 2 in [51] for details. The above inequality suggests that the regret upper bound of GP-UCB depends on the sum of the posterior standard deviations $\sum_{t=1}^{T}\sigma(\boldsymbol{x}_t; \mathbf{X}_{t-1})$. Srinivas et al. [51] provides the upper bound of this term by leveraging the information gain $I(\mathbf{X}_T)$ as follows:

$$\sum_{t=1}^{T}\sigma(\boldsymbol{x}_t; \mathbf{X}_{t-1}) \leq \sqrt{CTI(\mathbf{X}_T)} \leq \sqrt{CT\gamma_T(\mathcal{X})}, \tag{8}$$

[1] Differentiability of $k_{\mathrm{Matérn}}$ is derived in the existing works, e.g., Chapter 2.7 in [52].

where $C = \frac{2}{\ln(1+\sigma^{-2})}$. From Eqs. (7) and (8), we conclude that the regret upper bound of GP-UCB is $O\left(\sqrt{\beta_T T \gamma_T(\mathcal{X})}\right)$ with probability at least $1-\delta$. By combining the explicit upper bound of $\gamma_T(\mathcal{X})$ [51, 58], we also obtain $O\left(\sqrt{T \ln^{d+2} T}\right)$ and $\widetilde{O}\left(T^{\frac{\nu+d}{2\nu+d}}\right)$ regret upper bounds for SE and Matérn kernels, respectively.

3 Improved Regret Bound for GP-UCB

The following theorem presents our main result: a new regret upper bound for GP-UCB.

Theorem 3 (Improved regret upper bound for GP-UCB)**.** *Suppose Assumptions 1 and 2 hold. Set $k = k_{\mathrm{SE}}$ or $k = k_{\text{Matérn}}$ with $\nu > 2$. Furthermore, assume that d, ν, ℓ, r , and σ^2 are fixed constants. Fix any $\delta_{\mathrm{GP}} \in (0,1)$, and set the confidence width parameter β_t of GP-UCB as defined in Eq.* (6) *with any fixed $\delta \in (0, 1-\delta_{\mathrm{GP}})$. Then, with probability at least $1-\delta_{\mathrm{GP}}-\delta$, the cumulative regret of GP-UCB (Algorithm 1) satisfies*

$$R_T = \begin{cases} \widetilde{O}\left(\sqrt{T}\right) & \text{if } k = k_{\text{Matérn}} \text{ with } 2\nu + d \leq \nu^2, \\ O\left(\sqrt{T \ln^2 T}\right) & \text{if } k = k_{\mathrm{SE}}. \end{cases} \tag{9}$$

The hidden constants in the above expressions may depend on $\ln(1/\delta), d, \nu, \ell, r, \sigma^2$, and the constants $c_{\mathrm{sup}}, c_{\mathrm{gap}}, \rho_{\mathrm{lin}}, \rho_{\mathrm{quad}}, c_{\mathrm{lin}}, c_{\mathrm{quad}}$ corresponding with δ_{GP}, which are guaranteed to exist by Lemma 2.

We would like to note the following three aspects of our results. First, the constants associated with the sample path properties defined in Lemma 2 are used solely for analyzing the regret. On the other hand, the existing algorithm provided by Scarlett [46], which shows the same $\widetilde{O}(\sqrt{T})$ regret as ours, requires prior information about these constants for the algorithm run. This is often unrealistic in practice. Secondly, our result does not imply the upper bound of Bayesian expected regret $\mathbb{E}[R_T]$. The main issue is that the dependence of the constants in Lemma 2 on δ_{GP} is not explicitly known. We leave future work to break this limitation; however, note that the same limitation exists in the algorithm provided by Scarlett [46]. Thirdly, our results in Theorem 3 only focus on the dependence of the total step size T in the regret. Therefore, we cannot claim any improvements of the regret on the dependence of the other parameters. For example, compared to the existing $R_T = O(\sqrt{T \ln^{d+2} T})$ regret under $k = k_{\mathrm{SE}}$, our regret upper bound $R_T = O(\sqrt{T \ln^2 T})$ indeed avoids the dependence of d in the logarithmic factor; however, under the joint limit of d and T ($d, T \to \infty$), it easily behaves super-linearly even under the slowly increasing d (e.g., $d = \Theta(\ln \ln T)$) due to the hidden constants in the regret.

3.1 Intuitive Explanation of our Analysis

Before we describe the proof, we provide an intuitive explanation of why GP-UCB achieves a tighter regret than the existing $O(\sqrt{\beta_T T \gamma_T(\mathcal{X})})$ upper bound. The motivation for our new analysis comes from the observation that the upper bound of the information gain: $I(\mathbf{X}_T) \leq \gamma_T(\mathcal{X})$ in Eq. (8) is not always tight depending on the specific realization of the input sequence $\mathbf{X}_T$. To see this, let us observe the following two simple extreme cases of $\mathbf{X}_T$ where the inequality $I(\mathbf{X}_T) \leq \gamma_T(\mathcal{X})$ is loose and tight:

- **Case I:** $I(\mathbf{X}_T) \leq \gamma_T(\mathcal{X})$ **is loose**: Let us assume all the input is equal to the unique maximizer $\boldsymbol{x}^*$ (namely, $\forall t \in [T], \boldsymbol{x}_t = \boldsymbol{x}^*$). Then, when the kernel function satisfies $\forall \boldsymbol{x} \in \mathcal{X}, k(\boldsymbol{x}, \boldsymbol{x}) = 1$ as with k_{SE} and $k_{\text{Matérn}}$, we have:

$$I(\mathbf{X}_T) = \frac{1}{2} \ln \det(\boldsymbol{I}_T + \sigma^{-2} \mathbf{K}(\mathbf{X}_T, \mathbf{X}_T)) = \frac{1}{2} \sum_{i=1}^{T} \ln(1 + \sigma^{-2} \lambda_i) = \frac{1}{2} \ln(1 + \sigma^{-2} T), \tag{10}$$

  where λ_i is the i-th eigenvalue of $\mathbf{K}(\mathbf{X}_T, \mathbf{X}_T) = \mathbf{1}\mathbf{1}^\top$ with $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^T$. The third equation uses the fact that $\mathbf{1}\mathbf{1}^\top$ is rank 1, and its unique non-zero eigenvalue is T.
- **Case II:** $I(\mathbf{X}_T) \leq \gamma_T(\mathcal{X})$ **is tight**: Let us assume that $(\boldsymbol{x}_t)$ is the same as the input sequence generated by the maximum variance reduction (MVR) algorithm (namely, $\forall t \in [T], \boldsymbol{x}_t \in$

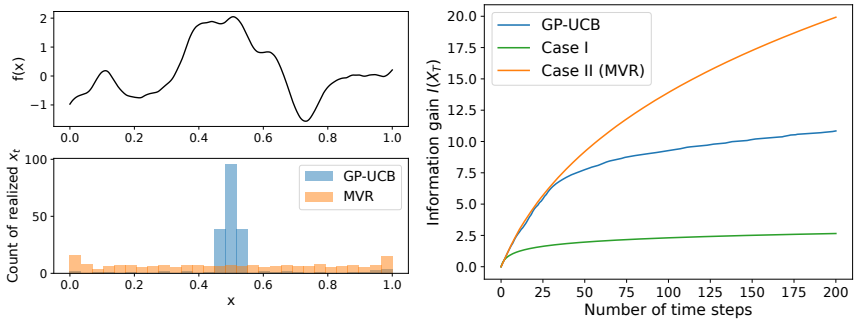

Figure 1: The behavior of the realized input sequence $\mathbf{X}_T$ (left) and the corresponding information gain $I(\mathbf{X}_T)$ (right) in the 1-dimensional BO problem with $\sigma^2 = 3$. The top left figure shows the objective function f realized by GP under $k = k_{\text{Matérn}}$ with $\nu = 5/2$. The bottom left figure shows the histogram of the realized inputs: $(\boldsymbol{x}_t)_{t \in [200]}$ with GP-UCB (blue) and MVR (orange) under f in the top left figure. Furthermore, the right plot shows the corresponding information gain $I(\mathbf{X}_t)$ under GP-UCB or MVR. We also plot $I(\mathbf{X}_t) \coloneqq 0.5 \ln(1 + \sigma^{-2} t)$, corresponding to Case I described in Section 3.1. We can observe that the inputs selected by GP-UCB are concentrated around the maximizer from the left figure. Then, from the right figure, we also observe that the corresponding information gain increases more slowly than that of MVR, and behaves similarly to Case I on $t \geq 30$. More comprehensive empirical results are also provided in Appendix D.

$\operatorname{argmax}_{\boldsymbol{x} \in \mathcal{X}} \sigma(\boldsymbol{x}; \mathbf{X}_{t-1})$) [51, 56]. Then, from the discussion in Sections 2 and 5 in [51], we already know that $\gamma_T(\mathcal{X}) \leq (1 - 1/e)^{-1} I(\mathbf{X}_T)$. This suggests that $I(\mathbf{X}_T) \leq \gamma_T(\mathcal{X})$ is tight up to a constant factor when $\mathbf{X}_T$ is realized by MVR.

From Case I, we observe that $I(\mathbf{X}_T)$ satisfies $\Theta(\ln T) \leq I(\mathbf{X}_T) \leq \gamma_T(\mathcal{X})$ depending on $\mathbf{X}_T$. Furthermore, by comparing the input sequences in cases I and II, we expect that $I(\mathbf{X}_T)$ becomes small if $\mathbf{X}_T$ concentrates around the neighborhood of $\boldsymbol{x}^*$, while $I(\mathbf{X}_T)$ becomes large if $\mathbf{X}_T$ spread over entire input domain $\mathcal{X}$. Then, from the fact that the worst-case regret of GP-UCB increases sub-linearly with the speed of $O(\sqrt{\beta_T T \gamma_T(\mathcal{X})})$, we can deduce that the input sequence $\mathbf{X}_T$ of GP-UCB will eventually concentrate around the maximizer $\boldsymbol{x}^*$ if $\boldsymbol{x}^*$ is unique and $\|f\|_\infty$ is not extremely small[2]. We provide an illustrative image in Figure 1. Our proof is designed so as to capture the above intuition that $I(\mathbf{X}_T)$ could be improved from $\gamma_T(\mathcal{X})$ to $\Theta(\ln T)$ under "favorable" sample path f.

3.2 Proof of Theorem 3

Let $\mathcal{A}$ be an event such that the two high-probability events of the original GP-UCB proof (described in the last paragraph in Section 2) and Lemma 2 with the confidence level δ_{GP} simultaneously hold. Note that event $\mathcal{A}$ occurs with probability at least $1 - \delta_{\text{GP}} - \delta$ from the union bound. Therefore, it is enough to prove our upper bound under $\mathcal{A}$. To encode the high-level idea in the previous section, we need to capture the concentration behavior of input sequence $\mathbf{X}_T$ around maximizer $\boldsymbol{x}^*$. From this motivation, given some constant $\varepsilon > 0$, we decompose the regret as $R_T = R_T^{(1)}(\varepsilon) + R_T^{(2)}(\varepsilon)$, where:

$$R_T^{(1)}(\varepsilon) = \sum_{t \in \mathcal{T}(\varepsilon)} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t), \; R_T^{(2)}(\varepsilon) = \sum_{t \in \mathcal{T}^c(\varepsilon)} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t). \tag{11}$$

We set $\mathcal{T}(\varepsilon) = \{t \in [T] \mid f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) > \varepsilon\}$ and $\mathcal{T}^c(\varepsilon) = [T] \setminus \mathcal{T}(\varepsilon)$ in the above definition. A key observation is that, if we set sufficiently small ε depending on the constants in Lemma 2, the inputs $(\boldsymbol{x}_t)$ in $R_T^{(2)}(\varepsilon)$ (namely, inputs $(\boldsymbol{x}_t)$ such that $f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) \leq \varepsilon$ holds) are on the locally

[2]Specifically, if $T\|f\|_\infty \leq O(\sqrt{\beta_T T \gamma_T(\mathcal{X})})$, we cannot make any claims about $\mathbf{X}_T$ based on the worst-case bound since any sequence $\mathbf{X}_T$ satisfies the worst-case bound without concentrating around maximizer. This is why our analysis technique does not improve the worst-case regret in the frequentist setting. Indeed, in the proof of the worst-case lower bound for the frequentist setting [47], the existence of the function f with $T\|f\|_\infty = O(\sqrt{\beta_T T \gamma_T(\mathcal{X})})$ is guaranteed.

linear or quadratic region around the maximizer $\boldsymbol{x}^*$ due to conditions 1, 3, and 4 in Lemma 2. The formal descriptions are provided in Lemma 20 in Appendix C. This fact is originally leveraged in [46] to analyze the successive elimination-based algorithm. In the analysis of GP-UCB, it enables us to analyze the behavior of the sub-input sequence $\{\boldsymbol{x}_t \mid f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) \leq \varepsilon\}$ through the regularity constants c_{lin} and c_{quad}. Below, we formally give the upper bound for $R_T^{(2)}(\varepsilon)$.

Lemma 4 (General upper bound of $R_T^{(2)}$)**.** *Suppose $(\boldsymbol{x}_t)_{t\in[T]}$ is the input query sequence realized by the GP-UCB algorithm. Furthermore, let $\overline{\gamma}_t$ is the upper bound of MIG $\gamma_t(\mathcal{X})$ such that $\overline{\gamma}_t/t$ is non-increasing on $[\overline{T}, \infty)$ with some $\overline{T} \in \mathbb{N}_+$[3]. Then, under event $\mathcal{A}$, we have*

$$R_T^{(2)}(\varepsilon) \leq 2c_{\mathrm{sup}}\overline{T} + \frac{\pi^2}{3}\left(\log_2 T + 1\right) + \frac{2\sqrt{2C\beta_T T}}{\sqrt{2}-1} \max_{i\in[\bar{i}]} \sqrt{\max\left\{\gamma_{(T/2^{i-1})}(\mathcal{B}_2(c_{\mathrm{lin}}^{-1}\eta_i; \boldsymbol{x}^*)), \gamma_{(T/2^{i-1})}\left(\mathcal{B}_2\left(\sqrt{c_{\mathrm{quad}}^{-1}\eta_i}; \boldsymbol{x}^*\right)\right)\right\}},$$

where $C = 2/\ln(1+\sigma^{-2})$, $\bar{i} = \lfloor \log_2 \frac{T}{\overline{T}} \rfloor + 1$, $\eta_i = \frac{2\left(2\sqrt{C\beta_T(T/2^{i-1})\overline{\gamma}_{T/2^{i-1}}} + \frac{\pi^2}{6}\right)}{(T/2^{i-1})}$, and $\varepsilon = \min\{\varepsilon_1, \varepsilon_2\} > 0$ with $\varepsilon_1 = \min\{c_{\mathrm{gap}}, c_{\mathrm{lin}}\rho_{\mathrm{lin}}\}$ and $\varepsilon_2 = \min\{c_{\mathrm{gap}}, c_{\mathrm{quad}}\rho_{\mathrm{quad}}^2\}$.

We give the full proof in Appendix A.1. Here, the dominant term in the above lemma is given as:

$$R_T^{(2)}(\varepsilon) = \widetilde{O}\left(\max_i \sqrt{T \max\left\{\gamma_{(T/2^{i-1})}(\mathcal{B}_2(\eta_i c_{\mathrm{lin}}^{-1}; \boldsymbol{x}^*)), \gamma_{(T/2^{i-1})}\left(\mathcal{B}_2\left(\sqrt{c_{\mathrm{quad}}^{-1}\eta_i}; \boldsymbol{x}^*\right)\right)\right\}}\right). \quad (12)$$

Note that η_i is decreasing as the time index $T/2^{i-1}$ of MIG increases. In other words, the input domains $\mathcal{B}_2(c_{\mathrm{lin}}^{-1}\eta_i; \boldsymbol{x}^*)$ and $\mathcal{B}_2\left(\sqrt{c_{\mathrm{quad}}^{-1}\eta_i}; \boldsymbol{x}^*\right)$ of MIG shrinks as the time index $T/2^{i-1}$ increases. This property is beneficial for obtaining a tighter upper bound than that from the existing technique. For example, under $k = k_{\mathrm{Matérn}}$ with $2\nu + d \leq \nu^2$, we can confirm that the dominant polynomial term in MIG is canceled out by the shrinking of the input domain in MIG. Namely, we can obtain the following result under $k = k_{\mathrm{Matérn}}$:

$$\max_i \max\left\{\gamma_{(T/2^{i-1})}(\mathcal{B}_2(\eta_i c_{\mathrm{lin}}^{-1}; \boldsymbol{x}^*)), \gamma_{(T/2^{i-1})}\left(\mathcal{B}_2\left(\sqrt{c_{\mathrm{quad}}^{-1}\eta_i}; \boldsymbol{x}^*\right)\right)\right\} = \widetilde{O}(1) \ \ (\text{as } T \to \infty), \quad (13)$$

which leads to $R_T^{(2)}(\varepsilon) = \widetilde{O}(\sqrt{T})$. This strictly improves the trivial upper bound $R_T^{(2)}(\varepsilon) = \widetilde{O}(\sqrt{T\gamma_T(\mathcal{X})})$ under $k = k_{\mathrm{Matérn}}$. The formal descriptions are given in the next lemma.

Lemma 5 (Upper bound of $R_T^{(2)}$ under k_{SE} and $k_{\mathrm{Matérn}}$)**.** *Suppose $(\boldsymbol{x}_t)_{t\in[T]}$ is the input sequence realized by the GP-UCB algorithm. Furthermore, ε is set as that in Lemma 4. Then, under event $\mathcal{A}$,*

$$R_T^{(2)}(\varepsilon) = \begin{cases} \widetilde{O}(\sqrt{T}) & \text{if } k = k_{\mathrm{Matérn}} \text{ with } 2\nu + d \leq \nu^2, \\ O\left(\sqrt{T\ln^2 T}\right) & \text{if } k = k_{\mathrm{SE}}. \end{cases} \quad (14)$$

The full proof is given in Appendix A.2. The remaining interest is the upper bound of $R_T^{(1)}(\varepsilon)$. The definition of $R_T^{(1)}(\varepsilon)$ is the same as the *lenient regret* [8], which is known to be smaller than the original regret R_T in GP-UCB. Although Cai et al. [8] studies the frequentist setting, their proof strategy is also applicable to the Bayesian setting as described in Section 3.4 in [8]. The following lemma provides the formal statement about the upper bound of $R_T^{(1)}(\varepsilon)$.

Lemma 6 (Upper bound of $R_T^{(1)}$, adaptation of the proof of Theorem 1 in [8])**.** *Fix any $\varepsilon > 0$. Suppose $k = k_{\mathrm{SE}}$ or $k = k_{\mathrm{Matérn}}$. Then, when running GP-UCB, $R_T^{(1)}(\varepsilon) = \widetilde{O}(1)$ holds under event $\mathcal{A}$.*

We provide the proof in Appendix A.3 for completeness. For both kernels, $R_T^{(1)}(\varepsilon)$ is dominated by the upper bound of $R_T^{(2)}(\varepsilon)$. Finally, we obtain the desired results by aggregating the inequalities in Lemmas 5 and 6.

[3]Namely, $\forall t \geq \overline{T}, \forall \epsilon \geq 0, \overline{\gamma}_t/t \geq \overline{\gamma}_{t+\epsilon}/(t+\epsilon)$ and $\forall t \geq \overline{T}, \gamma_t(\mathcal{X}) \leq \overline{\gamma}_t$ hold for some $\overline{T} \in \mathbb{N}_+$.

4 Discussions

Below, we discuss the limitations of our results and outline possible directions for future research.

- **Optimality.** Based on the $\Omega(\sqrt{T})$ lower bound on the expected regret provided by Scarlett [46], we conjecture that our $\widetilde{O}(\sqrt{T})$ high-probability regret bound for GP-UCB is near-optimal. However, it is not straightforward to extend the lower bound for the expected regret in [46] to a high probability result. Specifically, the lower bound in [46] is quantified by a mutual information term (Lemma 4 in [46]); however, to our knowledge, the technique used to handle this term appears to be specific to the expected regret setting. We believe that the rigorous optimality argument for the Bayesian high probability regret is an important direction for future research.
- **Smoothness condition.** In our result for the Matérn kernel, we require an additional smoothness constraint to obtain a $\widetilde{O}(\sqrt{T})$ regret bound[4] To overcome this issue in our proof, we believe that we need stronger regularity conditions on the sample path around the maximizer than those assumed in Lemma 2.
- **Extension to the expected regret.** Our regret bounds involve regularity constants that depend on the sample path. However, to our knowledge, there is no existing research that rigorously analyzes how these constants depend on the confidence level δ_{GP}. This makes it difficult to obtain the expected regret guarantees as with the original GP-UCB, whose expected regret bounds are established by properly decreasing the confidence level as a function of T (e.g., [40, 53]). To overcome this issue, further analysis for Lemma 2, or another idea to quantify the sample path regularities, is required.
- **Extension to other algorithms.** One limitation of our technique is its restricted applicability to other algorithms. To apply our proof, at least the algorithm should satisfy the following two conditions: (i) on any index subset, the sub-linear cumulative regret is obtained with high probability (Lemma 21), and (ii) the high probability lenient regret bound is provided (Lemma 6). The existing analysis of the other major algorithms in the Bayesian setting (e.g., Thompson sampling [43], information directed sampling [44]) does not provide these properties. Nevertheless, we believe that the high-level ideas in our proof (see Section 3.1) could be beneficial for future refined analyses of other algorithms.
- **Instance dependent analysis in the frequentist setting.** As described in the footnote in Section 3.1, we believe that our analysis does not improve the worst-case regret upper bound in the frequentist setting. On the other hand, our technique can be applied to the instance-dependent analysis [49] for GP-UCB. We expect that our proof strategy could yield a $\widetilde{O}(\sqrt{T})$ instance-dependent regret for GP-UCB by replacing the sample path conditions 3 and 4 in Lemma 2 with the *growth condition* (Definition 4 in [49]) of the function. It is an interesting direction for future research.

5 Conclusion

We provide a refined analysis of GP-UCB in the BO problem. For both SE and Matérn kernels, our results improve upon existing regret guarantees and fill the gap between the existing regret of GP-UCB and the current best upper bound in [46]. The core idea of our analysis is to capture the shrinking behavior of the input sequence by relating it to the worst-case upper bound and the sample path regularity conditions. Although our current analysis is limited to GP-UCB in the Bayesian setting, we believe it lays the foundation for several promising future research directions.

Acknowledgments

We thank Jonathan Scarlett and Shion Takeno for their valuable comments on revising the manuscript.

[4]For simplicity, in Theorem 3, we focus on the setting where the resulting regret becomes $\widetilde{O}(\sqrt{T})$ under $k_{\mathrm{Matérn}}$. This arises the requirement of the additional smoothness condition $2\nu + d \leq \nu^2$. On the other hand, we can also apply the same technique even under $2\nu + d > \nu^2$. In this case, resulting regret becomes strictly larger than $\widetilde{O}(\sqrt{T})$, while it is strictly smaller than $\widetilde{O}(T^{\frac{\nu+d}{2\nu+d}})$ of the original GP-UCB's analysis.

References

[1] Kendall Atkinson and Weimin Han. *Spherical harmonics and approximations on the unit sphere: an introduction*, volume 2044. Springer Science & Business Media, 2012.

[2] Douglas Azevedo and Valdir Antonio Menegatto. Sharp estimates for eigenvalues of integral operators generated by dot product kernels on the sphere. *Journal of Approximation Theory*, 2014.

[3] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 2017.

[4] Felix Berkenkamp, Angela P Schoellig, and Andreas Krause. No-regret Bayesian optimization with unknown hyperparameters. *Journal of Machine Learning Research*, 2019.

[5] Alberto Bietti and Francis Bach. Deep equals shallow for relu networks in kernel regimes. *International Conference on Learning Representations*, 2021.

[6] Ilija Bogunovic, Jonathan Scarlett, and Volkan Cevher. Time-varying Gaussian process bandit optimization. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

[7] Adam D Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 2011.

[8] Xu Cai, Selwyn Gomes, and Jonathan Scarlett. Lenient regret and good-action identification in Gaussian process bandits. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2021.

[9] Romain Camilleri, Kevin Jamieson, and Julian Katz-Samuels. High-dimensional experimental design and kernel bandits. In *Proc. International Conference on Machine Learning (ICML)*, 2021.

[10] Alexandre Capone, Armin Lederer, and Sandra Hirche. Gaussian process uniform error bounds with unknown hyperparameters for safety-critical applications. In *International Conference on Machine Learning*, 2022.

[11] Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *Proc. International Conference on Machine Learning (ICML)*, 2017.

[12] Andreas Christmann and Ingo Steinwart. Support vector machines. 2008.

[13] Nando de Freitas, Alex Smola, and Masrour Zoghi. Regret bounds for deterministic Gaussian process bandits. *arXiv preprint arXiv:1203.2177*, 2012.

[14] Nando De Freitas, Alex J. Smola, and Masrour Zoghi. Exponential regret bounds for Gaussian process bandits with deterministic observations. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, page 955–962. Omnipress, 2012.

[15] Thomas Desautels, Andreas Krause, and Joel W. Burdick. Parallelizing exploration-exploitation tradeoffs in Gaussian process bandit optimization. *Journal of Machine Learning Research*, 2014.

[16] Costas Efthimiou and Christopher Frye. *Spherical harmonics in p dimensions*. World Scientific, 2014.

[17] Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS journal on Computing*, 21(4):599–613, 2009.

[18] Peter I Frazier. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[19] Roman Garnett. *Bayesian optimization*. Cambridge University Press, 2023.

[20] Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Basri Ronen. On the similarity between the Laplace and neural tangent kernels. *Advances in Neural Information Processing Systems*, 2020.

[21] Subhashis Ghosal and Anindya Roy. Posterior consistency of Gaussian process prior for nonparametric binary regression. 2006.

[22] Andrew Gray and George Ballard Mathews. *A treatise on Bessel functions and their applications to physics*. Macmillan, 1895.

[23] Steffen Grünewälder, Jean-Yves Audibert, Manfred Opper, and John Shawe-Taylor. Regret bounds for Gaussian process bandit problems. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*. JMLR Workshop and Conference Proceedings, 2010.

[24] Philipp Hennig and Christian J Schuler. Entropy search for information-efficient global optimization. *The Journal of Machine Learning Research*, 13(1):1809–1837, 2012.

[25] Shogo Iwazaki and Shinya Suzumura. No-regret bandit exploration based on soft tree ensemble model. *Advances in Neural Information Processing Systems*, 2024.

[26] Shogo Iwazaki and Shion Takeno. Improved regret analysis in Gaussian process bandits: Optimality for noiseless reward, RKHS norm, and non-stationary variance. In *Proc. International Conference on Machine Learning (ICML)*, 2025.

[27] David Janz. *Sequential decision making with feature-linear models*. PhD thesis, 2022.

[28] David Janz, David Burt, and Javier Gonzalez. Bandit optimisation of functions in the Matérn kernel RKHS. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2486–2495. PMLR, 2020.

[29] Kirthevasan Kandasamy, Jeff Schneider, and Barnabas Poczos. High dimensional Bayesian optimisation and bandits via additive models. In *Proc. International Conference on Machine Learning (ICML)*, 2015.

[30] Kirthevasan Kandasamy, Gautam Dasarathy, Junier Oliva, Jeff Schneider, and Barnabas Poczos. Multi-fidelity Gaussian process bandit optimisation. *Journal of Artificial Intelligence Research*, 2019.

[31] Parnian Kassraie and Andreas Krause. Neural contextual bandits without regret. In *International Conference on Artificial Intelligence and Statistics*, pages 240–278. PMLR, 2022.

[32] Kenji Kawaguchi, Leslie P Kaelbling, and Tomás Lozano-Pérez. Bayesian optimization with exponential convergence. *Advances in neural information processing systems*, 28, 2015.

[33] Jeankyung Kim and David Pollard. Cube root asymptotics. *The Annals of Statistics*, pages 191–219, 1990.

[34] Andreas Krause and Cheng Ong. Contextual Gaussian process bandit optimization. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2011.

[35] Zihan Li and Jonathan Scarlett. Gaussian process bandit optimization with few batches. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

[36] Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer's theorem, feature maps, and smoothing. In *International Conference on Computational Learning Theory*. Springer, 2006.

[37] Jonas Močkus. On Bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974 6*, pages 400–404. Springer, 1975.

[38] Francis J Narcowich and Joseph D Ward. Scattered data interpolation on spheres: error estimates and locally supported basis functions. *SIAM Journal on Mathematical Analysis*, 33 (6):1393–1410, 2002.

[39] Francis J Narcowich, Xinping Sun, and Joseph D Ward. Approximation power of RBFs and their associated SBFs: a connection. *Advances in Computational Mathematics*, 2007.

[40] Biswajit Paria, Kirthevasan Kandasamy, and Barnabás Póczos. A flexible framework for multi-objective Bayesian optimization using random scalarizations. In *Uncertainty in Artificial Intelligence*, 2020.

[41] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

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

[43] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[44] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Advances in neural information processing systems*, 27, 2014.

[45] Sudeep Salgia, Sattar Vakili, and Qing Zhao. Random exploration in Bayesian optimization: Order-optimal regret and computational efficiency. In *Proc. International Conference on Machine Learning (ICML)*, 2024.

[46] Jonathan Scarlett. Tight regret bounds for Bayesian optimization in one dimension. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4500–4508. PMLR, 2018.

[47] Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. Lower bounds on regret for noisy Gaussian process bandit optimization. In *Proc. Conference on Learning Theory (COLT)*, 2017.

[48] Meyer Scetbon and Zaid Harchaoui. A spectral analysis of dot-product kernels. In *International conference on artificial intelligence and statistics*, 2021.

[49] Shubhanshu Shekhar and Tara Javidi. Instance dependent regret analysis of kernelized bandits. In *International Conference on Machine Learning*, 2022.

[50] Arno Solin and Simo Särkkä. Hilbert space methods for reduced-rank gaussian process regression. *Statistics and Computing*, 2020.

[51] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proc. International Conference on Machine Learning (ICML)*, 2010.

[52] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 1999.

[53] Shion Takeno, Yu Inatsu, and Masayuki Karasuyama. Randomized Gaussian process upper confidence bound with tighter Bayesian regret bounds. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 33490–33515. PMLR, 2023.

[54] Filip Tronarp, Toni Karvonen, and Simo Särkkä. Mixture representation of the matérn class with applications in state space approximations and Bayesian quadrature. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2018.

[55] Sattar Vakili and Julia Olkhovskaya. Kernelized reinforcement learning with order optimal regret bounds. *Advances in Neural Information Processing Systems*, 2023.

[56] Sattar Vakili, Nacime Bouziani, Sepehr Jalali, Alberto Bernacchia, and Da shan Shiu. Optimal order simple regret for Gaussian process bandits. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2021.

[57] Sattar Vakili, Michael Bromberg, Jezabel Garcia, Da-shan Shiu, and Alberto Bernacchia. Uniform generalization bounds for overparameterized neural networks. *arXiv preprint arXiv:2109.06099*, 2021.

[58] Sattar Vakili, Kia Khezeli, and Victor Picheny. On information gain and regret bounds in Gaussian process bandits. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

[59] Michal Valko, Nathan Korda, Rémi Munos, Ilias Flaounas, and Nello Cristianini. Finite-time analysis of kernelised contextual bandits. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 654–663. AUAI Press, 2013.

[60] Zexin Wang, Vincent YF Tan, and Jonathan Scarlett. Tight regret bounds for noisy optimization of a Brownian motion. *IEEE Transactions on Signal Processing*, 70:1072–1087, 2022.

[61] Justin Whitehouse, Zhiwei Steven Wu, and Aaditya Ramdas. Improved self-normalized concentration in Hilbert spaces: Sublinear regret for GP-UCB. *Proc. Neural Information Processing Systems (NeurIPS)*, 2023.

[62] Yun Yang, Anirban Bhattacharya, and Debdeep Pati. Frequentist coverage and sup-norm convergence rate in gaussian process regression. *arXiv preprint arXiv:1708.04753*, 2017.

[63] Fuzhen Zhang. *Matrix theory: basic results and techniques*. Springer Science & Business Media, 2011.

A Proofs in Section 3

A.1 Proof of Lemma 4

Proof. From Lemma 21, we have the following upper bound for any index set $\mathcal{T} \subset [T]$ under $\mathcal{A}$:

$$\sum_{t \in \mathcal{T}} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) \leq 2\sqrt{C\beta_T |\mathcal{T}| \overline{\gamma}_{|\mathcal{T}|}} + \frac{\pi^2}{6}. \tag{15}$$

Here, for any i such that $T/2^{i-1} \geq \overline{T}$, we set (η_i) as

$$\eta_i = \frac{2\left(2\sqrt{C\beta_T (T/2^{i-1}) \overline{\gamma}_{T/2^{i-1}}} + \frac{\pi^2}{6}\right)}{(T/2^{i-1})}. \tag{16}$$

As described in the proof below, these (η_i) are designed so that we can obtain the upper bound of $|\mathcal{T}(\eta_i)|$ in a dyadic manner. Here, we consider the upper bound of $|\mathcal{T}(\eta_i)|$ based on the worst-case upper bound in Eq. (15). From the definition of $\mathcal{T}(\eta)$ and Eq. (15) with $\mathcal{T} = [T]$, the condition $|\mathcal{T}(\eta_1)|\eta_1 \leq 2\sqrt{C\beta_T T \overline{\gamma}_T} + \pi^2/6$ must be satisfied; otherwise, we have $\sum_{t \in [T]} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) \geq \sum_{t \in \mathcal{T}(\eta_1)} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) \geq |\mathcal{T}(\eta_1)|\eta_1 > 2\sqrt{C\beta_T T \overline{\gamma}_T} + \pi^2/6$, which contradicts worst-case upper bound in Eq. (15). Therefore, we can obtain the following upper bound:

$$|\mathcal{T}(\eta_1)| \leq \max\left\{t \leq T \mid t\eta_1 \leq 2\sqrt{C\beta_T T \overline{\gamma}_T} + \frac{\pi^2}{6}\right\} = \frac{T}{2}. \tag{17}$$

Furthermore, since η_i is monotonic due to the condition about $\overline{\gamma}_t$, we have $\eta_1 \leq \eta_2$, which implies $\mathcal{T}(\eta_2) \subset \mathcal{T}(\eta_1)$. From Eq. (15) with $\mathcal{T} = \mathcal{T}(\eta_1)$, Eq. (17), and $\mathcal{T}(\eta_2) \subset \mathcal{T}(\eta_1)$, we further obtain

$$|\mathcal{T}(\eta_2)| \leq \max\left\{t \leq T/2 \mid t\eta_2 \leq 2\sqrt{C\beta_T (T/2) \overline{\gamma}_{(T/2)}} + \frac{\pi^2}{6}\right\} = \frac{T}{4}. \tag{18}$$

Similarly to $|\mathcal{T}(\eta_2)|$, we have $\mathcal{T}(\eta_3) \subset \mathcal{T}(\eta_2)$ and

$$|\mathcal{T}(\eta_3)| \leq \max\left\{t \leq T/4 \mid t\eta_3 \leq 2\sqrt{C\beta_T (T/4) \overline{\gamma}_{(T/4)}} + \frac{\pi^2}{6}\right\} = \frac{T}{8}. \tag{19}$$

By repeating this argument i times while $T/2^{i-1} \geq \overline{T}$ holds, we have the following inequality for any $i \leq \lfloor \log_2 \frac{T}{\overline{T}} \rfloor + 1$:

$$|\mathcal{T}(\eta_i)| \leq \max\left\{t \leq T/2^{i-1} \mid t\eta_i \leq \sqrt{C\beta_T (T/2^{i-1}) \overline{\gamma}_{(T/2^{i-1})}} + \frac{\pi^2}{6}\right\} = \frac{T}{2^i}. \tag{20}$$

Then, we have

$$\begin{aligned}
R_T^{(2)}(\varepsilon) &= \sum_{t \in \mathcal{T}^c(\varepsilon)} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) && (21)\\
&= \sum_{t \in \mathcal{T}^c(\varepsilon) \cap \mathcal{T}(\eta_1)} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) + \sum_{t \in \mathcal{T}^c(\varepsilon) \cap \mathcal{T}^c(\eta_1)} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) && (22)\\
&= \sum_{t \in \mathcal{T}^c(\varepsilon) \cap \mathcal{T}(\eta_1) \cap \mathcal{T}(\eta_2)} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) + \sum_{t \in \mathcal{T}^c(\varepsilon) \cap \mathcal{T}(\eta_1) \cap \mathcal{T}^c(\eta_2)} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) \\
&\quad + \sum_{t \in \mathcal{T}^c(\varepsilon) \cap \mathcal{T}^c(\eta_1)} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) && (23)\\
&= \sum_{t \in \mathcal{T}^c(\varepsilon) \cap \mathcal{T}(\eta_2)} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) + \sum_{i=1}^{2} \sum_{t \in \mathcal{T}^c(\varepsilon) \cap \mathcal{T}(\eta_{i-1}) \cap \mathcal{T}^c(\eta_i)} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t), && (24)
\end{aligned}$$

where the last line follows from $\mathcal{T}(\eta_2) \subset \mathcal{T}(\eta_1)$. In the above inequality, we define $\mathcal{T}(\eta_0)$ as $\mathcal{T}(\eta_0) = [T]$ for notational convenience. By repeatedly applying the above decomposition, we

obtain

$$\sum_{t\in\mathcal{T}^c(\varepsilon)\cap\mathcal{T}(\eta_2)} f(\boldsymbol{x}^*)-f(\boldsymbol{x}_t)+\sum_{i=1}^{2}\sum_{t\in\mathcal{T}^c(\varepsilon)\cap\mathcal{T}(\eta_{i-1})\cap\mathcal{T}^c(\eta_i)} f(\boldsymbol{x}^*)-f(\boldsymbol{x}_t) \tag{25}$$

$$=\sum_{t\in\mathcal{T}^c(\varepsilon)\cap\mathcal{T}(\eta_3)} f(\boldsymbol{x}^*)-f(\boldsymbol{x}_t)+\sum_{i=1}^{3}\sum_{t\in\mathcal{T}^c(\varepsilon)\cap\mathcal{T}(\eta_{i-1})\cap\mathcal{T}^c(\eta_i)} f(\boldsymbol{x}^*)-f(\boldsymbol{x}_t) \tag{26}$$

$$\vdots$$

$$=\sum_{t\in\mathcal{T}^c(\varepsilon)\cap\mathcal{T}(\eta_{\bar{i}})} f(\boldsymbol{x}^*)-f(\boldsymbol{x}_t)+\sum_{i=1}^{\bar{i}}\sum_{t\in\mathcal{T}^c(\varepsilon)\cap\mathcal{T}(\eta_{i-1})\cap\mathcal{T}^c(\eta_i)} f(\boldsymbol{x}^*)-f(\boldsymbol{x}_t), \tag{27}$$

where $\bar{i}=\lfloor\log_2\frac{T}{\overline{T}}\rfloor+1$. Regarding the first term in Eq. (27), we have

$$\sum_{t\in\mathcal{T}^c(\varepsilon)\cap\mathcal{T}(\eta_{\bar{i}})} f(\boldsymbol{x}^*)-f(\boldsymbol{x}_t)\le 2c_{\sup}|\mathcal{T}(\eta_{\bar{i}})|\le 2c_{\sup}\overline{T}, \tag{28}$$

where the last inequality follows from $|\mathcal{T}(\eta_{\bar{i}})|\le\overline{T}$, which is implied by $|\mathcal{T}(\eta_{\bar{i}})|\le T/2^{\bar{i}}$ from Eq. (20) and the definition of $\bar{i}$. Next, regarding the second term in Eq. (27), we first define $\mathcal{T}_i$ and $\mathcal{X}^{(i)}$ as $\mathcal{T}_i=\mathcal{T}^c(\varepsilon)\cap\mathcal{T}(\eta_{i-1})\cap\mathcal{T}^c(\eta_i)$ and $\mathcal{X}^{(i)}=\{\boldsymbol{x}_t\mid t\in\mathcal{T}_i\}$, respectively. Then, by applying Lemma 21 with $\mathcal{T}=\mathcal{T}_i$, we have

$$\sum_{t\in\mathcal{T}^c(\varepsilon)\cap\mathcal{T}(\eta_{i-1})\cap\mathcal{T}^c(\eta_i)} f(\boldsymbol{x}^*)-f(\boldsymbol{x}_t)=\sum_{t\in\mathcal{T}_i} f(\boldsymbol{x}^*)-f(\boldsymbol{x}_t) \tag{29}$$

$$\le 2\sqrt{C\beta_T|\mathcal{T}_i|I(\mathcal{X}^{(i)})}+\frac{\pi^2}{6} \tag{30}$$

$$\le 2\sqrt{C\beta_T|\mathcal{T}_i|\gamma_{|\mathcal{T}_i|}(\mathcal{X}^{(i)})}+\frac{\pi^2}{6} \tag{31}$$

$$\le 2\sqrt{C\beta_T|\mathcal{T}(\eta_{i-1})|\gamma_{|\mathcal{T}(\eta_{i-1})|}(\mathcal{X}^{(i)})}+\frac{\pi^2}{6} \tag{32}$$

$$\le 2\sqrt{C\beta_T(T/2^{i-1})\gamma_{(T/2^{i-1})}(\mathcal{X}^{(i)})}+\frac{\pi^2}{6}, \tag{33}$$

where the third inequality follows from $|\mathcal{T}_i|\le|\mathcal{T}(\eta_{i-1})|$, and the last inequality follows from Eq. (20). By aggregating Eqs. (27), (28), and (33), we obtain the following inequality under $\mathcal{A}$:

$$R_T^{(2)}(\varepsilon)\le 2c_{\sup}\overline{T}+2\sum_{i=1}^{\bar{i}}\left[\sqrt{C\beta_T(T/2^{i-1})\gamma_{(T/2^{i-1})}(\mathcal{X}^{(i)})}+\frac{\pi^2}{6}\right] \tag{34}$$

$$\le 2c_{\sup}\overline{T}+\frac{\pi^2}{3}\left(\log_2 T+1\right)+2\sqrt{C\beta_T T}\sum_{i=1}^{\bar{i}}\frac{1}{2^{(i-1)/2}}\sqrt{\gamma_{(T/2^{i-1})}(\mathcal{X}^{(i)})} \tag{35}$$

$$\le 2c_{\sup}\overline{T}+\frac{\pi^2}{3}\left(\log_2 T+1\right)+\frac{2\sqrt{2C\beta_T T}}{\sqrt{2}-1}\max_{i\in[\bar{i}]}\sqrt{\gamma_{(T/2^{i-1})}(\mathcal{X}^{(i)})}. \tag{36}$$

The last line follows from $\sum_{i=1}^{\bar{i}}\frac{1}{2^{(i-1)/2}}\le\sum_{i=1}^{\infty}\frac{1}{2^{(i-1)/2}}=\frac{1}{1-1/\sqrt{2}}=\frac{\sqrt{2}}{\sqrt{2}-1}$. The last part of the proof is to specify the radius of the ball $\mathcal{B}_2(\cdot;\boldsymbol{x}^*)$ such that $\mathcal{X}^{(i)}$ is included in it.

When $\boldsymbol{x}^*$ is on the boundary of $\mathcal{X}$. First, when $\boldsymbol{x}^*$ is on the boundary, condition 3 in Lemma 2 holds under $\mathcal{A}$. Then, from the definition of $\mathcal{T}^c(\varepsilon)$, ε, and Lemma 20, we have $\boldsymbol{x}\in\mathcal{B}_2(\rho_{\mathrm{lin}};\boldsymbol{x}^*)$ for any $\boldsymbol{x}\in\mathcal{X}^{(i)}$. This implies $\forall\boldsymbol{x}\in\mathcal{X}^{(i)}$, $f(\boldsymbol{x}^*)-f(\boldsymbol{x})\ge c_{\mathrm{lin}}\|\boldsymbol{x}-\boldsymbol{x}^*\|_2$ from condition 3 in Lemma 2. Since $\forall\boldsymbol{x}\in\mathcal{X}^{(i)}$, $f(\boldsymbol{x}^*)-f(\boldsymbol{x})\le\eta_i$ from $\mathcal{T}_i\subset\mathcal{T}^c(\eta_i)$, we have $\eta_i\ge c_{\mathrm{lin}}\|\boldsymbol{x}-\boldsymbol{x}^*\|_2\Leftrightarrow\eta_i c_{\mathrm{lin}}^{-1}\ge\|\boldsymbol{x}-\boldsymbol{x}^*\|_2$, which implies $\mathcal{X}^{(i)}\subset\mathcal{B}_2(\eta_i c_{\mathrm{lin}}^{-1};\boldsymbol{x}^*)$. Therefore, when $\boldsymbol{x}^*$ is on the boundary of $\mathcal{X}$, we have

$$\gamma_{(T/2^{i-1})}(\mathcal{X}^{(i)})\le\gamma_{(T/2^{i-1})}(\mathcal{B}_2(\eta_i c_{\mathrm{lin}}^{-1};\boldsymbol{x}^*)). \tag{37}$$

When x^* is on the interior of $\mathcal{X}$. Similarly to the boundary case, we consider the case where x^* is on the interior of $\mathcal{X}$. Then, condition 4 in Lemma 2 holds under $\mathcal{A}$. Furthermore, from the definition of $\mathcal{T}^c(\varepsilon)$, ε, and Lemma 20, we have $x \in \mathcal{B}_2(\rho_{\text{quad}}; x^*)$ for any $x \in \mathcal{X}^{(i)}$. This implies $\forall x \in \mathcal{X}^{(i)}, f(x^*) - f(x) \geq c_{\text{quad}}\|x - x^*\|_2^2$ from condition 4 in Lemma 2. Since $\forall x \in \mathcal{X}^{(i)}, f(x^*) - f(x) \leq \eta_i$ from $\mathcal{T}_i \subset \mathcal{T}^c(\eta_i)$, we have $\eta_i \geq c_{\text{quad}}\|x - x^*\|_2^2 \Leftrightarrow \sqrt{c_{\text{quad}}^{-1}\eta_i} \geq \|x - x^*\|_2$, which implies $\mathcal{X}^{(i)} \subset \mathcal{B}_2(\sqrt{c_{\text{quad}}^{-1}\eta_i}; x^*)$. Therefore, when x^* is on the interior of $\mathcal{X}$, we have

$$\gamma_{(T/2^{i-1})}(\mathcal{X}^{(i)}) \leq \gamma_{(T/2^{i-1})}\left(\mathcal{B}_2\left(\sqrt{c_{\text{quad}}^{-1}\eta_i}; x^*\right)\right). \tag{38}$$

Finally, combining Eq. (35) with Eqs. (37) and (38), we have

$$\begin{aligned} R_T^{(2)}(\varepsilon) \leq\ & 2c_{\text{sup}}\overline{T} + \frac{\pi^2}{3}\left(\log_2 T + 1\right) \\ & + \frac{2\sqrt{2C\beta_T T}}{\sqrt{2}-1}\max_{i\in[\tilde{t}]}\sqrt{\max\left\{\gamma_{(T/2^{i-1})}(\mathcal{B}_2(\eta_i c_{\text{lin}}^{-1}; x^*)), \gamma_{(T/2^{i-1})}\left(\mathcal{B}_2\left(\sqrt{c_{\text{quad}}^{-1}\eta_i}; x^*\right)\right)\right\}}. \end{aligned} \tag{39}$$

□

A.2 Proof of Lemma 5

To prove Lemma 5, we require the upper bound of MIG with the explicit dependence on the radius of the input domain. In Corollary 8 in Appendix B, we provide it with a full proof. Below, we establish the proof of Lemma 5 based on Corollary 8.

When $k = k_{\text{Matérn}}$. Set $C_{\text{Mat}} > 0$ as the constant such that the following inequalities hold:

$$\forall t \geq 2, \gamma_t(\mathcal{X}) \leq C_{\text{Mat}} t^{\frac{d}{2\nu+d}} \ln^{\frac{4\nu+d}{2\nu+d}} t, \tag{40}$$

$$\forall t \geq 2, \forall \eta > 0, \gamma_t\left(\{x \in \mathbb{R}^d \mid \|x\|_2 \leq \eta\}\right) \leq C_{\text{Mat}}\left(\eta^{\frac{2\nu d}{2\nu+d}} t^{\frac{d}{2\nu+d}} \ln^{\frac{4\nu+d}{2\nu+d}} t + \ln^2 t\right). \tag{41}$$

The existence of C_{Mat} is guaranteed by the upper bound of MIG established in Corollary 8[5]. Note that C_{Mat} is the constant that may depend on d, ℓ, ν, r, and σ^2. Furthermore, we set $\overline{\gamma}_t = C_{\text{Mat}} t^{\frac{d}{2\nu+d}} \ln^{\frac{4\nu+d}{2\nu+d}} t$. For function $g(t) \coloneqq \overline{\gamma}_t / t$, we have

$$g'(t) = -\frac{2\nu C_{\text{Mat}}}{2\nu+d} t^{-\frac{2\nu}{2\nu+d}-1} \ln^{\frac{4\nu+d}{2\nu+d}} t + C_{\text{Mat}} \frac{4\nu+d}{2\nu+d} t^{-\frac{2\nu}{2\nu+d}-1} \ln^{\frac{2\nu}{2\nu+d}} t \tag{42}$$

$$= \frac{C_{\text{Mat}}}{2\nu+d} t^{-\frac{2\nu}{2\nu+d}-1} (\ln^{\frac{2\nu}{2\nu+d}} t)\left(-2\nu \ln t + 4\nu + d\right). \tag{43}$$

From the above expression, if $2\nu \ln t \geq 4\nu + d \Leftrightarrow t \geq \exp(2 + d/(2\nu))$, $\overline{\gamma}_t / t$ is non-increasing. Therefore, we set $\overline{T} = \lceil \exp(2 + d/(2\nu)) \rceil$, which is independent of T. Here, for any $\eta > 0$ and $t \geq 2$, we have

$$\gamma_t\left(\mathcal{B}_2(\eta; x^*)\right) \leq \gamma_t\left(\{x \in \mathbb{R}^d \mid \|x - x^*\|_2 \leq \eta\}\right) \tag{44}$$

$$= \gamma_t\left(\{x \in \mathbb{R}^d \mid \|x\|_2 \leq \eta\}\right) \tag{45}$$

$$\leq C_{\text{Mat}}\left(\eta^{\frac{2\nu d}{2\nu+d}} t^{\frac{d}{2\nu+d}} \ln^{\frac{4\nu+d}{2\nu+d}} t + \ln^2 t\right), \tag{46}$$

where the second line follows from the fact that $k_{\text{Matérn}}$ is the stationary kernel (namely, $k_{\text{Matérn}}$ is transition invariant against any shift of inputs). Regarding η_i in Lemma 4, by setting T_i as

[5] If we rely on the result in [58], we can tighten the logarithmic term from $\ln^{\frac{4\nu+d}{2\nu+d}} t$ to $\ln^{\frac{2\nu}{2\nu+d}} t$; however, due to the technical issue of [58] described in Appendix B, we proceed our proof based on Corollary 8.

$T_i = T/2^{i-1}$, we have

$$\eta_i = \frac{2\left(2\sqrt{C\beta_T T_i \overline{\gamma}_{T_i}} + \frac{\pi^2}{6}\right)}{T_i} \tag{47}$$

$$= \frac{4\sqrt{C\beta_T T_i \overline{\gamma}_{T_i}}}{T_i} + \frac{\pi^2}{3T_i} \tag{48}$$

$$= \frac{4\sqrt{C\beta_T T_i C_{\mathrm{Mat}} T_i^{\frac{d}{2\nu+d}} \ln^{\frac{4\nu+d}{2\nu+d}} T_i}}{T_i} + \frac{\pi^2}{3T_i} \tag{49}$$

$$= 4\sqrt{CC_{\mathrm{Mat}}\beta_T}\left(T_i^{-\frac{\nu}{2\nu+d}} \ln^{\frac{4\nu+d}{4\nu+2d}} T_i\right) + \frac{\pi^2}{3T_i} \tag{50}$$

$$\leq \widetilde{C}_{\mathrm{Mat}}\sqrt{\beta_T}\left(T_i^{-\frac{\nu}{2\nu+d}} \ln^{\frac{4\nu+d}{4\nu+2d}} T_i\right), \tag{51}$$

where $\widetilde{C}_{\mathrm{Mat}} > 0$ is a sufficiently large constant such that $\widetilde{C}_{\mathrm{Mat}}\sqrt{\beta_T}\left(T_i^{-\frac{\nu}{2\nu+d}} \ln^{\frac{4\nu+d}{4\nu+2d}} T_i\right) \geq 4\sqrt{CC_{\mathrm{Mat}}\beta_T}\left(T_i^{-\frac{\nu}{2\nu+d}} \ln^{\frac{4\nu+d}{4\nu+2d}} T_i\right) + \frac{\pi^2}{3T_i}$ for any $T_i \geq 2$. Note that we can choose $\widetilde{C}_{\mathrm{Mat}} > 0$ without depending on T. From Eqs. (46) and (51), for any i, we have

$$\gamma_{T/2^{i-1}}\left(\mathcal{B}_2\left(c_{\mathrm{lin}}^{-1}\eta_i; \boldsymbol{x}^*\right)\right) \tag{52}$$

$$\leq C_{\mathrm{Mat}}\left(c_{\mathrm{lin}}^{-\frac{2\nu d}{2\nu+d}} \eta_i^{\frac{2\nu d}{2\nu+d}} T_i^{\frac{d}{2\nu+d}} \ln^{\frac{4\nu+d}{2\nu+d}} T_i + \ln^2 T_i\right) \tag{53}$$

$$\leq C_{\mathrm{Mat}}\left[c_{\mathrm{lin}}^{-\frac{2\nu d}{2\nu+d}} \widetilde{C}_{\mathrm{Mat}}^{\frac{2\nu d}{2\nu+d}} \beta_T^{\frac{\nu d}{2\nu+d}} \left(T_i^{-\frac{\nu}{2\nu+d}} \ln^{\frac{4\nu+d}{4\nu+2d}} T_i\right)^{\frac{2\nu d}{2\nu+d}} T_i^{\frac{d}{2\nu+d}} \ln^{\frac{4\nu+d}{2\nu+d}} T_i + \ln^2 T\right]. \tag{54}$$

We further consider the upper bound for the term depending on T_i: $\left(T_i^{-\frac{\nu}{2\nu+d}} \ln^{\frac{4\nu+d}{4\nu+2d}} T_i\right)^{\frac{2\nu d}{2\nu+d}} T_i^{\frac{d}{2\nu+d}} \ln^{\frac{4\nu+d}{2\nu+d}} T_i$. By aligning this term, we have

$$\left(T_i^{-\frac{\nu}{2\nu+d}} \ln^{\frac{4\nu+d}{4\nu+2d}} T_i\right)^{\frac{2\nu d}{2\nu+d}} T_i^{\frac{d}{2\nu+d}} \ln^{\frac{4\nu+d}{2\nu+d}} T_i = \widetilde{O}\left(T_i^{-\frac{2\nu^2 d}{(2\nu+d)^2} + \frac{d}{2\nu+d}}\right) \tag{55}$$

$$= \widetilde{O}\left(T_i^{\frac{d(2\nu+d)-2\nu^2 d}{(2\nu+d)^2}}\right) \tag{56}$$

$$= \widetilde{O}\left(T_i^{\frac{d(2\nu+d-2\nu^2)}{(2\nu+d)^2}}\right) \tag{57}$$

$$= o(1), \tag{58}$$

where the last line follows from $2\nu+d < 2\nu^2$, which is implied by the theorem's condition $2\nu+d \leq \nu^2$. Combining the above inequality with Eq. (54), we obtain $\gamma_{T/2^{i-1}}\left(\mathcal{B}_2\left(c_{\mathrm{lin}}^{-1}\eta_i; \boldsymbol{x}^*\right)\right) = \widetilde{O}(1)$. Similarly to the upper bound of $\gamma_{T/2^{i-1}}\left(\mathcal{B}_2\left(c_{\mathrm{lin}}^{-1}\eta_i; \boldsymbol{x}^*\right)\right)$, we have

$$\gamma_{T/2^{i-1}}\left(\mathcal{B}_2\left(\sqrt{c_{\mathrm{quad}}^{-1}\eta_i}; \boldsymbol{x}^*\right)\right) \tag{59}$$

$$\leq C_{\mathrm{Mat}}\left(c_{\mathrm{quad}}^{-\frac{\nu d}{2\nu+d}} \eta_i^{\frac{\nu d}{2\nu+d}} T_i^{\frac{d}{2\nu+d}} \ln^{\frac{4\nu+d}{2\nu+d}} T_i + \ln^2 T_i\right) \tag{60}$$

$$\leq C_{\mathrm{Mat}}\left[c_{\mathrm{quad}}^{-\frac{\nu d}{2\nu+d}} \widetilde{C}_{\mathrm{Mat}}^{\frac{\nu d}{2\nu+d}} \beta_T^{\frac{\nu d}{2(2\nu+d)}} \left(T_i^{-\frac{\nu}{2\nu+d}} \ln^{\frac{4\nu+d}{4\nu+2d}} T_i\right)^{\frac{\nu d}{2\nu+d}} T_i^{\frac{d}{2\nu+d}} \ln^{\frac{4\nu+d}{2\nu+d}} T_i + \ln^2 T\right]. \tag{61}$$

Furthermore, by noting condition $2\nu + d \leq \nu^2$, we have

$$\left(T_i^{-\frac{\nu}{2\nu+d}} \ln^{\frac{4\nu+d}{4\nu+2d}} T_i\right)^{\frac{\nu d}{2\nu+d}} T_i^{\frac{d}{2\nu+d}} \ln^{\frac{4\nu+d}{2\nu+d}} T_i = \widetilde{O}\left(T_i^{-\frac{\nu^2 d}{(2\nu+d)^2}+\frac{d}{2\nu+d}}\right) \tag{62}$$

$$= \widetilde{O}\left(T_i^{\frac{d(2\nu+d)-\nu^2 d}{(2\nu+d)^2}}\right) \tag{63}$$

$$= \widetilde{O}\left(T_i^{\frac{d(2\nu+d-\nu^2)}{(2\nu+d)^2}}\right) \tag{64}$$

$$= \widetilde{O}(1). \tag{65}$$

From the above inequalities, we have $\gamma_{T/2^{i-1}}\left(\mathcal{B}_2\left(\sqrt{c_{\mathrm{quad}}^{-1}}\eta_i; \boldsymbol{x}^*\right)\right) = \widetilde{O}(1)$. By noting $\gamma_{T/2^{i-1}}\left(\mathcal{B}_2\left(c_{\mathrm{lin}}^{-1}\eta_i; \boldsymbol{x}^*\right)\right) = \widetilde{O}(1)$ and $\gamma_{T/2^{i-1}}\left(\mathcal{B}_2\left(\sqrt{c_{\mathrm{quad}}^{-1}}\eta_i; \boldsymbol{x}^*\right)\right) = \widetilde{O}(1)$, we have the following inequality for any i.

$$\sqrt{\max\left\{\gamma_{(T/2^{i-1})}(\mathcal{B}_2(\eta_i c_{\mathrm{lin}}^{-1}; \boldsymbol{x}^*)), \gamma_{(T/2^{i-1})}\left(\mathcal{B}_2\left(\sqrt{c_{\mathrm{quad}}^{-1}}\eta_i; \boldsymbol{x}^*\right)\right)\right\}} = \widetilde{O}(1). \tag{66}$$

The above inequality and Lemma 4 implies

$$R_T^{(2)}(\varepsilon) \leq 2c_{\mathrm{sup}}\overline{T} + \frac{\pi^2}{3}\left(\log_2 T + 1\right) + \frac{2\sqrt{2C\beta_T T}}{\sqrt{2}-1} \times \widetilde{O}(1) \tag{67}$$

$$= \widetilde{O}(\sqrt{T}). \tag{68}$$

When $k = k_{\mathrm{SE}}$. The proof for $k = k_{\mathrm{SE}}$ is not straightforward as with the proof for $k = k_{\text{Matérn}}$. Specifically, we have to choose a proper $\overline{T}$ so as to obtain an $O(\ln T)$ upper bound of MIG. Let $C_{\mathrm{SE}} > 0$ be the constant such that the following inequalities hold:

$$\forall t \geq 2, \gamma_t(\mathcal{X}) \leq C_{\mathrm{SE}} \ln^{d+1} t, \tag{69}$$

$$\forall t \geq 2, \forall \eta \in (0, \sqrt{\frac{2\ell^2}{e^2 c_d}}), \gamma_t(\{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\|_2 \leq \eta\}) \leq C_{\mathrm{SE}}\left(\frac{\ln^{d+1} t}{\ln^d\left(\frac{2\ell^2}{\eta^2 e c_d}\right)} + \ln T\right). \tag{70}$$

The existence of such C_{SE} is guaranteed by Corollary 8. In the above inequalities, c_d is the constant defined in Corollary 8. We also set $\overline{\gamma}_t$ as $\overline{\gamma}_t = C_{\mathrm{SE}} \ln^{d+1} t$. We choose $\overline{T}$ later such that we can leverage the second statement in the above inequalities. Under $k = k_{\mathrm{SE}}$, we have

$$\eta_i = \frac{2\left(2\sqrt{C\beta_T T_i \overline{\gamma}_{T_i}} + \frac{\pi^2}{6}\right)}{T_i} \tag{71}$$

$$= \frac{4\sqrt{C\beta_T T_i \overline{\gamma}_{T_i}}}{T_i} + \frac{\pi^2}{3T_i} \tag{72}$$

$$= \frac{4\sqrt{C\beta_T T_i C_{\mathrm{SE}} \ln^{d+1} T_i}}{T_i} + \frac{\pi^2}{3T_i} \tag{73}$$

$$= 4\sqrt{CC_{\mathrm{SE}}\beta_T}\left(T_i^{-\frac{1}{2}} \ln^{\frac{d+1}{2}} T_i\right) + \frac{\pi^2}{3T_i} \tag{74}$$

$$\leq \widetilde{C}_{\mathrm{SE}}\sqrt{\beta_T}\left(T_i^{-\frac{1}{2}} \ln^{\frac{d+1}{2}} T_i\right), \tag{75}$$

where $\widetilde{C}_{\mathrm{SE}} > 0$ is a sufficiently large constant such that $\widetilde{C}_{\mathrm{SE}}\sqrt{\beta_T}\left(T_i^{-\frac{1}{2}} \ln^{\frac{d+1}{2}} T_i\right) \geq 4\sqrt{CC_{\mathrm{SE}}\beta_T}\left(T_i^{-\frac{1}{2}} \ln^{\frac{d+1}{2}} T_i\right) + \frac{\pi^2}{3T_i}$ for any $T_i \geq 2$. Hereafter, we define $\overline{\eta}_i \coloneqq \widetilde{C}_{\mathrm{SE}}\sqrt{\beta_T}\left(T_i^{-\frac{1}{2}} \ln^{\frac{d+1}{2}} T_i\right)$.

Then, to apply Eq. (70), we consider the lower bound of T_i such that $c_{\mathrm{lin}}^{-1}\overline{\eta}_i < \sqrt{2\ell^2/(e^2 c_d)}$ and $\sqrt{c_{\mathrm{quad}}^{-1}\overline{\eta}_i} < \sqrt{2\ell^2/(e^2 c_d)}$ hold. Regarding the condition $c_{\mathrm{lin}}^{-1}\overline{\eta}_i < \sqrt{2\ell^2/(e^2 c_d)}$, we have

$$c_{\mathrm{lin}}^{-1}\overline{\eta}_i < \sqrt{\frac{2\ell^2}{e^2 c_d}} \Leftrightarrow c_{\mathrm{lin}}^{-1}\sqrt{\frac{e^2 c_d}{2\ell^2}}\widetilde{C}_{\mathrm{SE}}\sqrt{\beta_T}\ln^{\frac{d+1}{2}} T_i < T_i^{\frac{1}{2}} \tag{76}$$

$$\Leftarrow c_{\mathrm{lin}}^{-1}\sqrt{\frac{e^2 c_d}{2\ell^2}}\widetilde{C}_{\mathrm{SE}}\sqrt{\beta_T}\ln^{\frac{d+1}{2}} T < T_i^{\frac{1}{2}} \tag{77}$$

$$\Leftrightarrow \frac{e^2 c_d \widetilde{C}_{\mathrm{SE}}^2}{2\ell^2 c_{\mathrm{lin}}^2}\beta_T \ln^{d+1} T < T_i. \tag{78}$$

Similarly, regarding the condition $\sqrt{c_{\mathrm{quad}}^{-1}\overline{\eta}_i} < \sqrt{2\ell^2/(e^2 c_d)}$, we have

$$\sqrt{c_{\mathrm{quad}}^{-1}\overline{\eta}_i} < \sqrt{\frac{2\ell^2}{e^2 c_d}} \Leftrightarrow c_{\mathrm{quad}}^{-1}\frac{e^2 c_d}{2\ell^2}\widetilde{C}_{\mathrm{SE}}\sqrt{\beta_T}\ln^{\frac{d+1}{2}} T_i < T_i^{\frac{1}{2}} \tag{79}$$

$$\Leftarrow c_{\mathrm{quad}}^{-1}\frac{e^2 c_d}{2\ell^2}\widetilde{C}_{\mathrm{SE}}\sqrt{\beta_T}\ln^{\frac{d+1}{2}} T < T_i^{\frac{1}{2}} \tag{80}$$

$$\Leftrightarrow \left(\frac{e^2 c_d \widetilde{C}_{\mathrm{SE}}}{2\ell^2 c_{\mathrm{quad}}}\right)^2 \beta_T \ln^{d+1} T < T_i. \tag{81}$$

From the above inequalities, we set $\overline{T}$ such that

$$\max\left\{\left(\frac{e^2 c_d \widetilde{C}_{\mathrm{SE}}}{2\ell^2 c_{\mathrm{quad}}}\right)^2, \frac{e^2 c_d \widetilde{C}_{\mathrm{SE}}^2}{2\ell^2 c_{\mathrm{lin}}^2}\right\}\beta_T \ln^{d+1} T < \overline{T}. \tag{82}$$

Then, from $T_i \geq \overline{T}$ and Eqs. (70), (76), and (82), we have

$$\gamma_{T/2^{i-1}}\left(\mathcal{B}_2\left(c_{\mathrm{lin}}^{-1}\eta_i; \boldsymbol{x}^*\right)\right) \leq \gamma_{T_i}\left(\{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\|_2 \leq c_{\mathrm{lin}}^{-1}\eta_i\}\right) \tag{83}$$

$$\leq C_{\mathrm{SE}}\left(\frac{\ln^{d+1} T_i}{\ln^d\left(\frac{2c_{\mathrm{lin}}^2\ell^2}{\overline{\eta}_i^2 e c_d}\right)} + \ln T_i\right) \tag{84}$$

$$\leq C_{\mathrm{SE}}\left(\frac{\ln^{d+1} T_i}{\ln^d\left(\frac{2c_{\mathrm{lin}}^2\ell^2}{\overline{\eta}_i^2 e c_d}\right)} + \ln T\right). \tag{85}$$

Furthermore, from $T_i \geq \overline{T}$ and Eqs. (70), (79), and (82),

$$\gamma_{T/2^{i-1}}\left(\mathcal{B}_2\left(\sqrt{c_{\mathrm{quad}}^{-1}\eta_i}; \boldsymbol{x}^*\right)\right) \leq \gamma_{T_i}\left(\left\{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\|_2 \leq \sqrt{c_{\mathrm{quad}}^{-1}\eta_i}\right\}\right) \tag{86}$$

$$\leq C_{\mathrm{SE}}\left(\frac{\ln^{d+1} T_i}{\ln^d\left(\frac{2c_{\mathrm{quad}}\ell^2}{\overline{\eta}_i e c_d}\right)} + \ln T\right). \tag{87}$$

Based on Eqs. (85) and (87), we further consider the lower bound of T_i such that

$$\frac{\ln^{d+1} T_i}{\ln^d\left(\frac{2c_{\mathrm{lin}}^2\ell^2}{\overline{\eta}_i^2 e c_d}\right)} = O(\ln T) \text{ and } \frac{\ln^{d+1} T_i}{\ln^d\left(\frac{2c_{\mathrm{quad}}\ell^2}{\overline{\eta}_i e c_d}\right)} = O(\ln T). \tag{88}$$

Regarding the first case in Eq. (88), we have

$$\frac{2c_{\text{lin}}^2\ell^2}{\overline{\eta}_i^2 ec_d} \geq \sqrt{T_i} \Leftrightarrow \frac{2c_{\text{lin}}^2\ell^2}{ec_d\widetilde{C}_{\text{SE}}^2\beta_T T_i^{-1}\ln^{d+1}T_i} \geq \sqrt{T_i} \tag{89}$$

$$\Leftrightarrow T_i^{1/2} \geq \frac{ec_d\widetilde{C}_{\text{SE}}^2\beta_T\ln^{d+1}T_i}{2c_{\text{lin}}^2\ell^2} \tag{90}$$

$$\Leftarrow T_i^{1/2} \geq \frac{ec_d\widetilde{C}_{\text{SE}}^2\beta_T\ln^{d+1}T}{2c_{\text{lin}}^2\ell^2} \tag{91}$$

$$\Leftrightarrow T_i \geq \left(\frac{ec_d\widetilde{C}_{\text{SE}}^2\beta_T\ln^{d+1}T}{2c_{\text{lin}}^2\ell^2}\right)^2. \tag{92}$$

Therefore, if $\overline{T} \geq \left(\frac{ec_d\widetilde{C}_{\text{SE}}^2\beta_T\ln^{d+1}T}{2c_{\text{lin}}^2\ell^2}\right)^2$, we have

$$C_{\text{SE}}\left(\frac{\ln^{d+1}T_i}{\ln^d\left(\frac{2c_{\text{lin}}^2\ell^2}{\overline{\eta}_i^2 ec_d}\right)}+\ln T\right) \leq C_{\text{SE}}\left(\frac{\ln^{d+1}T_i}{2^{-d}\ln^d T_i}+\ln T\right) \tag{93}$$

$$\leq C_{\text{SE}}\left(2^d\ln T+\ln T\right). \tag{94}$$

Regarding the second case in Eq. (88), we have

$$\frac{2c_{\text{quad}}\ell^2}{\overline{\eta}_i ec_d} \geq T_i^{1/4} \Leftrightarrow \frac{2c_{\text{quad}}\ell^2}{ec_d\widetilde{C}_{\text{SE}}\sqrt{\beta_T}T_i^{-1/2}\ln^{\frac{d+1}{2}}T_i} \geq T_i^{1/4} \tag{95}$$

$$\Leftrightarrow T_i^{1/4} \geq \frac{ec_d\widetilde{C}_{\text{SE}}\sqrt{\beta_T}\ln^{\frac{d+1}{2}}T_i}{2c_{\text{quad}}\ell^2} \tag{96}$$

$$\Leftarrow T_i^{1/4} \geq \frac{ec_d\widetilde{C}_{\text{SE}}\sqrt{\beta_T}\ln^{\frac{d+1}{2}}T}{2c_{\text{quad}}\ell^2} \tag{97}$$

$$\Leftrightarrow T_i \geq \left(\frac{ec_d\widetilde{C}_{\text{SE}}\sqrt{\beta_T}\ln^{\frac{d+1}{2}}T}{2c_{\text{quad}}\ell^2}\right)^4. \tag{98}$$

Hence, if $\overline{T} \geq \left(\frac{ec_d\widetilde{C}_{\text{SE}}\sqrt{\beta_T}\ln^{\frac{d+1}{2}}T}{2c_{\text{quad}}\ell^2}\right)^4$, we have

$$C_{\text{SE}}\left(\frac{\ln^{d+1}T_i}{\ln^d\left(\frac{2c_{\text{quad}}\ell^2}{\overline{\eta}_i ec_d}\right)}+\ln T\right) \leq C_{\text{SE}}\left(\frac{\ln^{d+1}T_i}{4^{-d}\ln^d T_i}+\ln T\right) \tag{99}$$

$$\leq C_{\text{SE}}\left(4^d\ln T+\ln T\right). \tag{100}$$

By aggregating the conditions (82), (92), and (98), we set $\overline{T}$ as the smallest natural number such that the following inequalities hold:

$$\overline{T} \geq \max\left\{\left(\frac{e^2c_d\widetilde{C}_{\text{SE}}}{2\ell^2c_{\text{quad}}}\right)^2, \frac{e^2c_d\widetilde{C}_{\text{SE}}^2}{2\ell^2c_{\text{lin}}^2}\right\}\beta_T\ln^{d+1}T, \text{ and} \tag{101}$$

$$\overline{T} \geq \max\left\{\left(\frac{ec_d}{2c_{\text{quad}}\ell^2}\right)^4, \left(\frac{ec_d}{2c_{\text{lin}}^2\ell^2}\right)^2\right\}\widetilde{C}_{\text{SE}}^4\beta_T^2\ln^{2(d+1)}T. \tag{102}$$

Then, from Eqs. (85), (87), (94), and (100), we have

$$\sqrt{\max\left\{\gamma_{(T/2^{i-1})}(\mathcal{B}_2(\eta_i c_{\text{lin}}^{-1};\boldsymbol{x}^*)), \gamma_{(T/2^{i-1})}\left(\mathcal{B}_2\left(\sqrt{c_{\text{quad}}^{-1}\eta_i};\boldsymbol{x}^*\right)\right)\right\}} = O(\sqrt{\ln T}). \tag{103}$$

Finally, by noting $\overline{T} = O(\ln^{2d+4} T)$, we obtain the following result from Lemma 4:

$$R_T^{(2)}(\varepsilon) = O\left(\ln^{2d+4} T + \sqrt{T \ln^2 T}\right). \tag{104}$$

Since d is a fixed constant, the above equation implies $R_T^{(2)}(\varepsilon) = O(\sqrt{T \ln^2 T})$. □

A.3 Proof of Lemma 6

Proof. From the upper bound of the discretization error in event $\mathcal{A}$, we have $\forall t \geq \sqrt{2/\varepsilon}, \forall \boldsymbol{x} \in \mathcal{X}, |f(\boldsymbol{x}) - f([\boldsymbol{x}]_t)| \leq \varepsilon/2$. Here, we set $\underline{\mathcal{T}}(\varepsilon) = \{t \in \mathbb{N}_+ \mid t \geq \sqrt{2/\varepsilon}\}$. By relying on the standard argument of MIG [51], we observe the following inequality for any realizations and $\varepsilon > 0$:

$$\min_{t \in \mathcal{T}(\varepsilon) \cap \underline{\mathcal{T}}(\varepsilon)} \sigma(\boldsymbol{x}_t; \mathbf{X}_{t-1}) \leq \sqrt{\frac{C \gamma_{|\mathcal{T}(\varepsilon) \cap \underline{\mathcal{T}}(\varepsilon)|}(\mathcal{X})}{|\mathcal{T}(\varepsilon) \cap \underline{\mathcal{T}}(\varepsilon)|}}, \tag{105}$$

where $\mathcal{T}(\varepsilon) = \{t \in [T] \mid f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) > \varepsilon\}$ and $C = 2/\ln(1 + \sigma^{-2})$. Under $\mathcal{A}$, we further have the following inequalities for any $\widetilde{t} \in \operatorname{argmin}_{t \in \mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)} \sigma(\boldsymbol{x}_t; \mathbf{X}_{t-1})$[6]:

$$\begin{aligned}
&\mu(\boldsymbol{x}_{\widetilde{t}}; \mathbf{X}_{\widetilde{t}-1}; \boldsymbol{y}_{\widetilde{t}-1}) + \beta_{\widetilde{t}}^{1/2} \sigma(\boldsymbol{x}_{\widetilde{t}}; \mathbf{X}_{\widetilde{t}-1}) && (106)\\
&= \mu(\boldsymbol{x}_{\widetilde{t}}; \mathbf{X}_{\widetilde{t}-1}; \boldsymbol{y}_{\widetilde{t}-1}) - \beta_{\widetilde{t}}^{1/2} \sigma(\boldsymbol{x}_{\widetilde{t}}; \mathbf{X}_{\widetilde{t}-1}) + 2\beta_{\widetilde{t}}^{1/2} \sigma(\boldsymbol{x}_{\widetilde{t}}; \mathbf{X}_{\widetilde{t}-1}) && (107)\\
&\leq f(\boldsymbol{x}_{\widetilde{t}}) + 2\beta_{\widetilde{t}}^{1/2} \sigma(\boldsymbol{x}_{\widetilde{t}}; \mathbf{X}_{\widetilde{t}-1}) && (108)\\
&< f(\boldsymbol{x}^*) - \varepsilon + 2\sqrt{\frac{C \beta_{\widetilde{t}} \gamma_{|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)|}(\mathcal{X})}{|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)|}} && (109)\\
&\leq |f(\boldsymbol{x}^*) - f([\boldsymbol{x}^*]_{\widetilde{t}})| + f([\boldsymbol{x}^*]_{\widetilde{t}}) - \varepsilon + 2\sqrt{\frac{C \beta_{\widetilde{t}} \gamma_{|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)|}(\mathcal{X})}{|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)|}} && (110)\\
&\leq \mu([\boldsymbol{x}^*]_{\widetilde{t}}; \mathbf{X}_{\widetilde{t}-1}; \boldsymbol{y}_{\widetilde{t}-1}) + \beta_{\widetilde{t}}^{1/2} \sigma([\boldsymbol{x}^*]_{\widetilde{t}}; \mathbf{X}_{\widetilde{t}-1}) - \frac{\varepsilon}{2} + 2\sqrt{\frac{C \beta_{\widetilde{t}} \gamma_{|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)|}(\mathcal{X})}{|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)|}}, && (111)
\end{aligned}$$

where the second inequality follows from the definition of $\mathcal{T}(\varepsilon)$, and the last inequality follows from $\widetilde{t} \in \underline{\mathcal{T}}(\varepsilon)$ and event $\mathcal{A}$. Therefore, under $\mathcal{A}$, the inequality $-\frac{\varepsilon}{2} + 2\sqrt{\frac{C \beta_{\widetilde{t}} \gamma_{|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)|}(\mathcal{X})}{|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)|}} \geq 0$ must hold; otherwise, $\mu(\boldsymbol{x}_{\widetilde{t}}; \mathbf{X}_{\widetilde{t}-1}; \boldsymbol{y}_{\widetilde{t}-1}) + \beta_{\widetilde{t}}^{1/2} \sigma(\boldsymbol{x}_{\widetilde{t}}; \mathbf{X}_{\widetilde{t}-1}) < \mu([\boldsymbol{x}^*]_{\widetilde{t}}; \mathbf{X}_{\widetilde{t}-1}; \boldsymbol{y}_{\widetilde{t}-1}) + \beta_{\widetilde{t}}^{1/2} \sigma([\boldsymbol{x}^*]_{\widetilde{t}}; \mathbf{X}_{\widetilde{t}-1})$, which contradicts $\boldsymbol{x}_{\widetilde{t}} \in \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{X}} \mu(\boldsymbol{x}; \mathbf{X}_{\widetilde{t}-1}; \boldsymbol{y}_{\widetilde{t}-1}) + \beta_{\widetilde{t}}^{1/2} \sigma(\boldsymbol{x}; \mathbf{X}_{\widetilde{t}-1})$. This further implies

$$|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)| \leq \frac{16 C \beta_{\widetilde{t}} \gamma_{|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)|}(\mathcal{X})}{\varepsilon^2} \leq \frac{16 C \beta_T \gamma_{|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)|}(\mathcal{X})}{\varepsilon^2} \tag{112}$$

for any $\varepsilon > 0$. Furthermore,

$$\begin{aligned}
R_T^{(1)}(\varepsilon) &= \sum_{t \in \mathcal{T}(\epsilon)} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) && (113)\\
&= 2c_{\sup}\sqrt{\frac{2}{\varepsilon}} + \sum_{t \in \mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) && (114)\\
&\leq 2c_{\sup}\sqrt{\frac{2}{\varepsilon}} + \frac{\pi^2}{6} + 2\sqrt{C \beta_T |\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)| \gamma_{|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)|}(\mathcal{X})} && (115)
\end{aligned}$$

for any $\varepsilon > 0$. In the above expressions, the last inequality follows from Lemma 21. The remaining part of the proof is to substitute the quantity $|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)|$ in Eq. (115) into its upper bound, which is deduced from Eq. (112) depending on the kernel.

[6] If $\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon) = \emptyset$, the theorem's statement clearly holds; therefore, we suppose $\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon) \neq \emptyset$ in this proof.

For $k = k_{\mathrm{SE}}$. Under $k = k_{\mathrm{SE}}$, we crudely take the upper bound of $|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)|$ as

$$|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)| \leq \frac{16C\beta_T\gamma_{|\mathcal{T}(\epsilon)\cap\underline{\mathcal{T}}(\varepsilon)|}(\mathcal{X})}{\varepsilon^2} \leq \frac{16C\beta_T\gamma_T(\mathcal{X})}{\varepsilon^2}. \tag{116}$$

The above upper bound implies $|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)| = O\left(\beta_T\gamma_T(\mathcal{X})\right)$. Since $\gamma_T(\mathcal{X}) = O(\ln^{d+1} T)$ under $k = k_{\mathrm{SE}}$, Eq. (115) implies

$$\begin{aligned}
R_T^{(1)}(\varepsilon) &\leq 2c_{\sup}\sqrt{\frac{2}{\varepsilon}} + \frac{\pi^2}{6} + O\left(\sqrt{\beta_T(\beta_T\gamma_T(\mathcal{X}))\ln^{d+1}(\beta_T\gamma_T(\mathcal{X}))}\right) && (117)\\
&= O\left(\beta_T\sqrt{(\ln^{d+1} T)\ln^{d+1}(\ln^{d+2} T)}\right) && (118)\\
&= O\left(\sqrt{(\ln T)^{d+3}(\ln\ln T)^{d+1}}\right) && (119)\\
&= \widetilde{O}(1). && (120)
\end{aligned}$$

For $k = k_{\text{Matérn}}$. Set $C_{\mathrm{Mat}} > 0$ as the constant such that the following inequality holds:

$$\forall t \geq 2, \gamma_t(\mathcal{X}) \leq C_{\mathrm{Mat}} t^{\frac{d}{2\nu+d}} \ln^{\frac{4\nu+d}{2\nu+d}} t. \tag{121}$$

The existence of C_{Mat} is guaranteed by the upper bound of MIG established in Corollary 8. Then, if $|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)| \geq 2$ holds, Eq. (112) implies

$$\begin{aligned}
&|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)| \leq \frac{16C\beta_T C_{\mathrm{Mat}}|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)|^{\frac{d}{2\nu+d}} \ln^{\frac{4\nu+d}{2\nu+d}} |\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)|}{\varepsilon^2} && (122)\\
&\Rightarrow |\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)| \leq \frac{16C\beta_T C_{\mathrm{Mat}}|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)|^{\frac{d}{2\nu+d}} \ln^{\frac{4\nu+d}{2\nu+d}} T}{\varepsilon^2} && (123)\\
&\Leftrightarrow |\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)|^{\frac{2\nu}{2\nu+d}} \leq \frac{16C\beta_T C_{\mathrm{Mat}} \ln^{\frac{4\nu+d}{2\nu+d}} T}{\varepsilon^2} && (124)\\
&\Leftrightarrow |\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)| \leq \left(\frac{16C\beta_T C_{\mathrm{Mat}} \ln^{\frac{4\nu+d}{2\nu+d}} T}{\varepsilon^2}\right)^{1+\frac{d}{2\nu}}. && (125)
\end{aligned}$$

Therefore, we have $|\mathcal{T}(\epsilon) \cap \underline{\mathcal{T}}(\varepsilon)| = \widetilde{O}(1)$ under fixed ε, d, and ν. Hence, from Eq. (115), we obtain $R_T^{(1)}(\varepsilon) = \widetilde{O}(1)$. □

B Information Gain Upper Bound

Our analysis requires the upper bound of MIG with explicit dependence on the radius of the input domain. Several existing works [4, 27, 28] established such a result by extending the proof in [51]. However, the proof strategy in [51] result in $\widetilde{O}(T^{\frac{d(d+1)}{2\nu+d(d+1)}})$ upper bound of MIG in Matérn kernel, which is strictly worse than the best achievable $\widetilde{O}(T^{\frac{d}{2\nu+d}})$ upper bound. Vakili et al. [58] shows $\widetilde{O}(T^{\frac{d}{2\nu+d}})$ upper bound of MIG with $\nu > 1/2$ under the uniform boundness assumption of the eigenfunctions. Furthermore, the following work [55] shows $\gamma_T(\{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\|_2 \leq \eta\}) = \widetilde{O}(\eta^{\frac{2\nu d}{2\nu+d}} T^{\frac{d}{2\nu+d}})$ for any radius $\eta > 0$ if there exist eigenfunctions uniformly bounded without depending on $\eta > 0$. Some of the related results supports the uniform boundness assumption under $d = 1$ [27, 62], or under the approximated version of the original Matérn kernel [42, 50]; however, to our knowledge, we are not aware of any literature that rigorously support uniform boundness assumption under the general compact input domain with $d \geq 2$ and $\nu > 1/2$. See Chapter 4.4 in [27] for the detailed discussion. Therefore, this section's goal is twofold: (i) prove $\widetilde{O}(T^{\frac{d}{2\nu+d}})$ upper bound as claimed in [58] without relying on the uniform boundness assumption, and (ii) clarify the explicit dependence on the input radius in the upper bound proved in (i).

Below, we formally describe our MIG upper bound.

Theorem 7. *Fix any $d \in \mathbb{N}_+$, $\lambda > 0$, and $T \in \mathbb{N}_+$. Let us assume $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\|_2 \leq 1\}$. Then,*

- *For $k = k_{\mathrm{SE}}$, $\gamma_T(\mathcal{X})$ satisfies*

$$\gamma_T(\mathcal{X}) \leq \frac{C_d^{(1)}}{\theta^d} \ln^{d+1}\left(1 + \frac{T}{\lambda^2}\right) + \ln\left(1 + \frac{T}{\lambda^2}\right) + C_d^{(2)} \exp\left(-\frac{2}{\theta} + \frac{1}{\theta^2}\right) \tag{126}$$

if $\theta \leq e^2 c_d$ and $T/(e-1) \geq \lambda^2$. Furthermore, for any $\theta > e^2 c_d$, we have

$$\gamma_T(\mathcal{X}) \leq \frac{C_d^{(3)}}{\ln^d\left(\frac{\theta}{e c_d}\right)} \ln^{d+1}\left(1 + \frac{T}{\lambda^2}\right) + C_d^{(4)} \ln\left(1 + \frac{T}{\lambda^2}\right) + C_d^{(5)}. \tag{127}$$

Here, we set $\theta = 2\ell^2$ and $c_d = \max\left\{1, \exp\left(\frac{1}{e}\left(\frac{d}{2} - 1\right)\right)\right\}$. Furthermore, $C_d^{(1)}, C_d^{(2)}, C_d^{(3)}, C_d^{(4)}, C_d^{(5)} > 0$ are the constants only depending on d.

- *For $k = k_{\text{Matérn}}$ with $\nu > 1/2$, $\gamma_T(\mathcal{X})$ satisfies*

$$\gamma_T(\mathcal{X}) \leq C(T, \nu, \lambda)\overline{\gamma}_T + C \tag{128}$$

where $C(T, \nu, \lambda) = \max\left\{1, \log_2\left(1 + \frac{\Gamma(\nu)}{C_\nu} \ln \frac{T^2}{\lambda^2}\right) + \frac{1}{\nu}\log_2\left(\frac{T^2}{\nu\Gamma(\nu)\lambda^2}\right) + 1\right\}$. Here, $C_\nu > 0$ and $C > 0$ are the constant that only depends on $\nu > 0$, and an absolute constant, respectively. Furthermore, $\overline{\gamma}_T$ is defined as

$$\overline{\gamma}_T = C_{d,\nu}^{(1)} \ln\left(1 + \frac{2T}{\lambda^2}\right) + C_{d,\nu}^{(2)}\left(\frac{T}{\lambda^2 \ell^{2\nu}}\right)^{\frac{d}{2\nu+d}} \ln^{\frac{2\nu}{2\nu+d}}\left(1 + \frac{2T}{\lambda^2}\right), \tag{129}$$

where $C_{d,\nu}^{(1)}, C_{d,\nu}^{(2)} > 0$ are the constants only depending on d and ν.

We also obtain the following corollary by adjusting the lengthscale parameter $\ell > 0$ based on the radius of the input domain.

Corollary 8. *Fix any $d \in \mathbb{N}_+$, $\lambda > 0$, $T \in \mathbb{N}_+$, $\eta > 0$. Let us assume $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\|_2 \leq \eta\}$. Then,*

- *For $k = k_{\mathrm{SE}}$, $\gamma_T(\mathcal{X})$ satisfies*

$$\gamma_T(\mathcal{X}) \leq \frac{C_d^{(3)}}{\ln^d\left(\frac{2\ell^2}{\eta^2 e c_d}\right)} \ln^{d+1}\left(1 + \frac{T}{\lambda^2}\right) + C_d^{(4)} \ln\left(1 + \frac{T}{\lambda^2}\right) + C_d^{(5)}. \tag{130}$$

if $2\ell^2/\eta^2 > e^2 c_d$.

- *For $k = k_{\text{Matérn}}$ with $\nu > 1/2$, $\gamma_T(\mathcal{X})$ satisfies Eq. (128), with*

$$\overline{\gamma}_T = C_{d,\nu}^{(1)} \ln\left(1 + \frac{2T}{\lambda^2}\right) + C_{d,\nu}^{(2)} \eta^{\frac{2\nu d}{2\nu+d}} \left(\frac{T}{\lambda^2 \ell^{2\nu}}\right)^{\frac{d}{2\nu+d}} \ln^{\frac{2\nu}{2\nu+d}}\left(1 + \frac{2T}{\lambda^2}\right). \tag{131}$$

The constants in the above statements are the same as those in Theorem 7.

While the above results ignore the explicit dependence on d and ν, all the other parameters are explicitly stated in our upper bound of MIG. We would like to emphasize that we do not rely on the uniform boundness assumption of the eigenfunctions to prove the above results. In the above results for $k = k_{\mathrm{SE}}$, we obtain the same $O(\ln^{d+1} T)$ upper bound as that in [58] except for the constant factor. For $k = k_{\text{Matérn}}$, we also obtain the same $\widetilde{O}(T^{\frac{d}{2\nu+d}})$ upper bound as that in [58], while the logarithmic dependence get worse from $O\left(\ln^{d/(2\nu+d)} T\right)$ to $O\left(\ln^{(4\nu+d)/(2\nu+d)} T\right)$. Furthermore, the above result reveals the explicit dependence of the radius η of the input domain. Regarding the case $k = k_{\text{Matérn}}$, our result suggests $\widetilde{O}(\eta^{\frac{2\nu d}{2\nu+d}} T^{\frac{d}{2\nu+d}})$ upper bound of MIG, which is consistent with that in [55] with uniform boundness assumption.

Proof overview. The basic proof strategy follows that in [58], which leverages the Mercer decomposition of the kernel. To bypass the uniform boundness assumption in the proof of [58], we must resort to other specific properties of the eigenfunction. However, except for some exceptional cases, the eigenfunction of the kernel on the general compact domain is difficult to specify in an analytical form and complex to analyze. To avoid this issue, instead of studying the original definition of the MIG on $\mathbb{R}^d$, we consider reducing the original MIG on $\mathcal{X} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\|_2 \leq 1\}$ to that on a hypersphere $\mathbb{S}^d \coloneqq \{\boldsymbol{x} \in \mathbb{R}^{d+1} \mid \|\boldsymbol{x}\|_2 = 1\}$ defined in $\mathbb{R}^{d+1}$. The eigensystems on $\mathbb{S}^d$ are one of the exceptional cases, whose eigenfunctions are specified as a special function on $\mathbb{S}^d$, called *spherical harmonics* [1, 16]. Indeed, by using the addition theorem of the spherical harmonics (Theorem 14), the existing works [25, 31, 57] already demonstrated that the upper bound of MIG on $\mathbb{S}^d$ is proved as with [58] without the uniform boundness assumption. We use their technique to show the upper bound of MIG under SE and Matérn kernels on $\mathbb{R}^d$, while the original motivation of these existing works is to study the MIG under the neural tangent kernel on a hypersphere. The remaining parts of this section are constructed as follows:

- In Section B.1, we show our core result (Lemma 9) that guarantees the MIG on $\{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\|_2 \leq 1\}$ is bounded from above by that on $\mathbb{S}^d$ up to logarithmic factor.
- In Section B.2, we summarize the basic known results about Mercer decomposition on $\mathbb{S}^d$, which is the foundation of the following subsections.
- In Section B.3, we provide the general upper bound of the information gain on $\mathbb{S}^d$ (Lemma 15), represented by the kernel function's eigenvalues. This subsection's result has no intrinsic change from those in [31, 57]; however, we provide details for completeness.
- In Section B.4, we provide the upper bound of the decaying rate of the eigenvalues in SE and Matérn kernels.
- In Section B.5, we establish the full proof of Theorem 7 based on the results in Sections B.1–B.4.

B.1 Reduction of the MIG on $\mathbb{R}^d$ to $\mathbb{S}^d$

Lemma 9 (Reduction to the hypersphere in $\mathbb{R}^{d+1}$)**.** *Fix any $d \in \mathbb{N}_+$, $\lambda > 0$, and $T \in \mathbb{N}_+$. Suppose $\mathcal{X} = \{(x_1, \ldots, x_d, 0)^\top \in \mathbb{R}^{d+1} \mid \sum_{i=1}^d x_i^2 \leq 1\}$, and define $\mathbb{S}^d$ as $\mathbb{S}^d = \{\boldsymbol{x} \in \mathbb{R}^{d+1} \mid \|\boldsymbol{x}\|_2 = 1\}$. Then,*

- *For $k = k_{\mathrm{SE}}$, we have*

$$\max_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \in \mathcal{X}} \ln \det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}(\mathbf{X}_T, \mathbf{X}_T)) \leq \max_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \in \mathbb{S}^d} \ln \det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}(\mathbf{X}_T, \mathbf{X}_T)). \quad (132)$$

- *For $k = k_{\text{Matérn}}$, we have*

$$\max_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \in \mathcal{X}} \ln \det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}(\mathbf{X}_T, \mathbf{X}_T)) \quad (133)$$

$$\leq C(T, \nu, \lambda) \max_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \in \mathbb{S}^d} \ln \det(\boldsymbol{I}_T + 2\lambda^{-2}\mathbf{K}(\mathbf{X}_T, \mathbf{X}_T)) + C, \quad (134)$$

where $C(T, \nu, \lambda) = \max\left\{0, \log_2\left(1 + \frac{\Gamma(\nu)}{C_\nu} \ln \frac{T^2}{\lambda^2}\right) + \frac{1}{\nu}\log_2\left(\frac{T^2}{\nu\Gamma(\nu)\lambda^2}\right) + 1\right\}$. Here, $C_\nu > 0$ is the constant that only dependes on $\nu > 0$. Furthermore, $C > 0$ is an absolute constant.

Proof. For any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \in \mathcal{X}$, we construct the new input sequence $\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_T$ on $\mathbb{S}^d$, where $\widetilde{\boldsymbol{x}}_i = \left(x_{i,1}, \ldots, x_{i,d}, \sqrt{1 - \sum_{j=1}^d x_{i,j}^2}\right)^\top$.

Under $k = k_{\mathrm{SE}}$. It is enough to show the following inequality:

$$\det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}(\mathbf{X}_T, \mathbf{X}_T)) \leq \det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}(\widetilde{\mathbf{X}}_T, \widetilde{\mathbf{X}}_T)), \quad (135)$$

where $\widetilde{\mathbf{X}}_T = (\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_T)$. From the definition of $\widetilde{\boldsymbol{x}}_i$, we rewrite R.H.S. in the above inequality as

$$\det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}(\widetilde{\mathbf{X}}_T, \widetilde{\mathbf{X}}_T)) = \det(\widetilde{\mathbf{K}} \odot (\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}(\mathbf{X}_T, \mathbf{X}_T)), \quad (136)$$

where $[\widetilde{\mathbf{K}}]_{i,j} = k(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_j)/k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Here, $A \odot B$ denotes the Hadamard product of the matrices A and B. Then, Oppenheim inequality (e.g., Theorem 7.27 in [63]) implies $\det(A \odot B) \geq \det(B) \prod_i A_{ii}$ for any positive semi-definite matrices A and B. Therefore, if $\widetilde{\mathbf{K}}$ is a positive semi-definite matrix, Eq. (136) immediately implies

$$\det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}(\widetilde{\mathbf{X}}_T, \widetilde{\mathbf{X}}_T)) \geq \det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}(\mathbf{X}_T, \mathbf{X}_T)) \prod_{i\in[T]} \widetilde{\mathbf{K}}_{ii} = \det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}(\mathbf{X}_T, \mathbf{X}_T)). \quad (137)$$

From the definition of k_{SE} and $\widetilde{\boldsymbol{x}}_i$, we have

$$\frac{k(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_j)}{k(\boldsymbol{x}_i, \boldsymbol{x}_j)} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 + \left(\sqrt{1 - \|\boldsymbol{x}_i\|_2^2} - \sqrt{1 - \|\boldsymbol{x}_j\|_2^2}\right)^2}{2\ell^2} + \frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\ell^2}\right) \quad (138)$$

$$= \exp\left(-\frac{\left(\sqrt{1 - \|\boldsymbol{x}_i\|_2^2} - \sqrt{1 - \|\boldsymbol{x}_j\|_2^2}\right)^2}{2\ell^2}\right). \quad (139)$$

The above equation suggests that $\widetilde{\mathbf{K}}$ is equal to the kernel matrix of the one-dimensional SE-kernel, whose inputs are transformed by $\sqrt{1 - \|\cdot\|_2^2}$. Since the SE kernel is positive definite, the matrix $\widetilde{\mathbf{K}}$ is also positive semi-definite, and we complete the proof for $k = k_{\mathrm{SE}}$.

Under $k = k_{\text{Matérn}}$. Similarly to the proof for $k = k_{\mathrm{SE}}$, we consider the application of Oppenheim inequality; however, the positive semi-definiteness of element-wise quotient matrix $\widetilde{\mathbf{K}}$ is unknown for $k = k_{\text{Matérn}}$. To avoid this problem, we leverage the following representation of $k_{\text{Matérn}}$, which is given as the form of the lengthscale mixture of the SE kernel [54]:

$$k(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \frac{1}{\Gamma(\nu)} \int_0^\infty z^{\nu-1} e^{-z} \exp\left(-\frac{\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_2^2}{2\ell^2 z \nu^{-1}}\right) \mathrm{d}z. \quad (140)$$

Based on the above representation, we decompose the original kernel function k into the following three components:

$$k(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = k_1(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) + k_2(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) + k_3(\boldsymbol{x}, \widetilde{\boldsymbol{x}}), \quad (141)$$

where:

$$k_1(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \frac{1}{\Gamma(\nu)} \int_0^{\eta_1} z^{\nu-1} e^{-z} \exp\left(-\frac{\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_2^2}{2\ell^2 z \nu^{-1}}\right) \mathrm{d}z, \quad (142)$$

$$k_2(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \frac{1}{\Gamma(\nu)} \int_{\eta_1}^{\eta_2} z^{\nu-1} e^{-z} \exp\left(-\frac{\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_2^2}{2\ell^2 z \nu^{-1}}\right) \mathrm{d}z, \quad (143)$$

$$k_3(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \frac{1}{\Gamma(\nu)} \int_{\eta_2}^{\infty} z^{\nu-1} e^{-z} \exp\left(-\frac{\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_2^2}{2\ell^2 z \nu^{-1}}\right) \mathrm{d}z \quad (144)$$

with some $\eta_2 > \eta_1 > 0$[7]. Then, as with the proof of Theorem 3 in [34], we have

$$\begin{aligned}&\ln\det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}(\mathbf{X}_T, \mathbf{X}_T)) \\ &\leq \underbrace{\ln\det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}_1(\mathbf{X}_T, \mathbf{X}_T))}_{(i)} + \underbrace{\ln\det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}_2(\mathbf{X}_T, \mathbf{X}_T))}_{(ii)} + \underbrace{\ln\det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}_3(\mathbf{X}_T, \mathbf{X}_T))}_{(iii)},\end{aligned} \quad (145)$$

where $\mathbf{K}_1(\mathbf{X}_T, \mathbf{X}_T), \mathbf{K}_2(\mathbf{X}_T, \mathbf{X}_T)$, and $\mathbf{K}_3(\mathbf{X}_T, \mathbf{X}_T)$ are the kernel matrix of k_1, k_2, and k_3, respectively. We set sufficiently small η_1 and large η_2 so that the crude upper bound of the first term (i)

[7] Note that the linear combination of the positive definite kernel with non-negative coefficients and its limit are also positive definite. Therefore, k_1, k_2, and k_3 are also positive definite as far as $\eta_2 > \eta_1$.

and the third term (iii) are sufficiently small. For this purpose, the following settings of η_1 and η_2 are sufficient (we confirm in the next paragraphs):

$$\eta_1 = \left(\frac{\nu\Gamma(\nu)\lambda^2}{T^2}\right)^{\frac{1}{\nu}}, \ \eta_2 = \max\left\{1, \frac{\Gamma(\nu)}{C_\nu}\ln\frac{T^2}{\lambda^2}\right\}, \tag{146}$$

where $C_\nu > 0$ is the constant such that $\forall z \geq 1,\ z^{\nu-1}e^{-z} \leq C_\nu e^{-z/2}$. Hereafter, we suppose that $\eta_1 < \eta_2$ holds with the above definition. The case $\eta_1 \geq \eta_2$ is considered in the last parts of the proof.

Upper bound for the first term (i). From the definition of k_1, we have

$$\begin{aligned}
|k_1(\boldsymbol{x}, \widetilde{\boldsymbol{x}})| &= \frac{1}{\Gamma(\nu)}\int_0^{\eta_1} z^{\nu-1}e^{-z}\exp\left(-\frac{\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|_2^2}{2\ell^2 z\nu^{-1}}\right)\mathrm{d}z && (147)\\
&\leq \frac{1}{\Gamma(\nu)}\int_0^{\eta_1} z^{\nu-1}e^{-z}\mathrm{d}z && (148)\\
&\leq \frac{1}{\Gamma(\nu)}\int_0^{\eta_1} z^{\nu-1}\mathrm{d}z && (149)\\
&= \frac{1}{\nu\Gamma(\nu)}[z^\nu]_0^{\eta_1} && (150)\\
&= \frac{1}{\nu\Gamma(\nu)}\eta_1^\nu. && (151)
\end{aligned}$$

Then, from the definition of η_1, we have

$$\frac{1}{\nu\Gamma(\nu)}\eta_1^\nu = \frac{1}{\nu\Gamma(\nu)}\left(\frac{\nu\Gamma(\nu)\lambda^2}{T^2}\right) = \frac{\lambda^2}{T^2}. \tag{152}$$

Therefore, by denoting the eigenvalues of $\mathbf{K}_1(\mathbf{X}_T, \mathbf{X}_T)$ with decreasing order as $(\lambda_i)_{i\in[T]}$[8], we have

$$\begin{aligned}
\ln\det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}_1(\mathbf{X}_T, \mathbf{X}_T)) &= \ln\prod_{i=1}^{T}(1+\lambda^{-2}\lambda_i) && (153)\\
&\leq \ln(1+\lambda^{-2}\lambda_1)^T && (154)\\
&\leq \ln(1+T^{-1})^T, && (155)
\end{aligned}$$

where the last inequality follows from $\lambda_1 \leq \sqrt{\sum_{i=1}^T \lambda_i^2} = \|\mathbf{K}_1(\mathbf{X}_T, \mathbf{X}_T)\|_F = \sqrt{\sum_{i,j} k_1(\boldsymbol{x}_i, \boldsymbol{x}_j)^2} \leq \lambda^2/T$. Since $\ln(1+T^{-1})^T \to 1$ as $T \to \infty$, there exists constant $C > 0$ such that $\ln\det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}_1(\mathbf{X}_T, \mathbf{X}_T)) \leq C$ for all $T \in \mathbb{N}_+$.

Upper bound for the third term (iii). From the definition of k_3, we have

$$\begin{aligned}
|k_3(\boldsymbol{x}, \widetilde{\boldsymbol{x}})| &= \frac{1}{\Gamma(\nu)}\int_{\eta_2}^{\infty} z^{\nu-1}e^{-z}\exp\left(-\frac{\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|_2^2}{2\ell^2 z\nu^{-1}}\right)\mathrm{d}z && (156)\\
&\leq \frac{1}{\Gamma(\nu)}\int_{\eta_2}^{\infty} z^{\nu-1}e^{-z}\mathrm{d}z && (157)\\
&\leq \frac{C_\nu}{\Gamma(\nu)}\int_{\eta_2}^{\infty} e^{-z/2}\mathrm{d}z && (158)\\
&= \frac{-2C_\nu}{\Gamma(\nu)}[e^{-z/2}]_{\eta_2}^{\infty} && (159)\\
&= \frac{2C_\nu}{\Gamma(\nu)}\exp\left(-\frac{\eta_2}{2}\right). && (160)
\end{aligned}$$

[8]Note that λ_i is non-negative from the positive semi-definiteness of $\mathbf{K}_1(\mathbf{X}_T, \mathbf{X}_T)$.

Then, from the definition of η_2, we have

$$\frac{2C_\nu}{\Gamma(\nu)}\exp\left(-\frac{\eta_2}{2}\right) \le \frac{2C_\nu}{\Gamma(\nu)}\exp\left(-\frac{\Gamma(\nu)}{2C_\nu}\ln\left(\frac{T^2}{\lambda^2}\right)\right) = \frac{\lambda^2}{T^2}. \tag{161}$$

By following the same arguments after Eq. (152) in the upper bound of the first term (i), we conclude that there exists constant $C > 0$ such that $\ln\det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}_3(\mathbf{X}_T, \mathbf{X}_T)) \le C$ for all $T \in \mathbb{N}_+$.

Upper bound for the second term (ii). We further divide k_2 with dyadic manner:

$$k_2(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \sum_{q=1}^{Q} k_2^{(q)}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}), \tag{162}$$

where:

$$k_2^{(q)}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \frac{1}{\Gamma(\nu)}\int_{\eta_1 2^{q-1}}^{\min\{\eta_1 2^q, \eta_2\}} z^{\nu-1}e^{-z}\exp\left(-\frac{\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|_2^2}{2\ell^2 z\nu^{-1}}\right)\mathrm{d}z. \tag{163}$$

Here, $Q \in \mathbb{N}_+$ is the minimum number such that $\eta_1 2^Q \ge \eta_2$ holds. Then, as with Eq. (145), we have

$$\ln\det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}_2(\mathbf{X}_T, \mathbf{X}_T)) \le \sum_{q=1}^{Q}\ln\det\left(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}_2^{(q)}(\mathbf{X}_T, \mathbf{X}_T)\right), \tag{164}$$

where $\mathbf{K}_2^{(q)}(\mathbf{X}_T, \mathbf{X}_T)$ is the kernel matrix of $k_2^{(q)}$. Next, for any q, we define new kernel function $\widetilde{k}^{(q)}(\boldsymbol{x}, \widetilde{\boldsymbol{x}})$ as

$$\widetilde{k}_2^{(q)}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = k_2^{(q)}(\boldsymbol{x}, \widetilde{\boldsymbol{x}})\exp\left(-\frac{\left(\sqrt{1-\|\boldsymbol{x}\|_2^2} - \sqrt{1-\|\widetilde{\boldsymbol{x}}\|_2^2}\right)^2}{2\ell^2\nu^{-1}\min\{\eta_1 2^q, \eta_2\}}\right). \tag{165}$$

We further denote the kernel matrix of $\widetilde{k}_2^{(q)}$ by $\widetilde{\mathbf{K}}_2^{(q)}(\mathbf{X}_T, \mathbf{X}_T)$. Then, from Oppenheim's inequality, we have

$$\ln\det\left(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}_2^{(q)}(\mathbf{X}_T, \mathbf{X}_T)\right) \le \ln\det\left(\boldsymbol{I}_T + \lambda^{-2}\widetilde{\mathbf{K}}_2^{(q)}(\mathbf{X}_T, \mathbf{X}_T)\right). \tag{166}$$

Furthermore, for any $z \in [\eta_1 2^{q-1}, \min\{\eta_1 2^q, \eta_2\}]$, the following kernel function $\widehat{k}_2^{(q)}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}; z)$ is positive definite (e.g., Lemma A.5 in [10]):

$$\widehat{k}_2^{(q)}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}; z) = 2\exp\left(-\frac{\left(\sqrt{1-\|\boldsymbol{x}\|_2^2} - \sqrt{1-\|\widetilde{\boldsymbol{x}}\|_2^2}\right)^2}{2\ell^2\nu^{-1}z}\right) - \exp\left(-\frac{\left(\sqrt{1-\|\boldsymbol{x}\|_2^2} - \sqrt{1-\|\widetilde{\boldsymbol{x}}\|_2^2}\right)^2}{2\ell^2\nu^{-1}\min\{\eta_1 2^q, \eta_2\}}\right). \tag{167}$$

Note that $2k_2^{(q)}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_j) - \widetilde{k}_2^{(q)}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is represented as

$$2k_2^{(q)}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_j) - \widetilde{k}_2^{(q)}(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{168}$$

$$= \frac{1}{\Gamma(\nu)}\int_{\eta_1 2^{q-1}}^{\min\{\eta_1 2^q, \eta_2\}} z^{\nu-1}e^{-z}\exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\ell^2 z\nu^{-1}}\right)\widehat{k}_2^{(q)}(\boldsymbol{x}_i, \boldsymbol{x}_j; z)\mathrm{d}z. \tag{169}$$

By noting that the product of two positive definite kernels is also positive definite, the above expression implies that $2\mathbf{K}_2^{(q)}(\widetilde{\mathbf{X}}_T, \widetilde{\mathbf{X}}_T) - \widetilde{\mathbf{K}}_2^{(q)}(\mathbf{X}_T, \mathbf{X}_T)$ is the positive semi-definite matrix. Therefore, we have[9]

$$\sum_{q=1}^{Q}\ln\det\left(\boldsymbol{I}_T + \lambda^{-2}\widetilde{\mathbf{K}}_2^{(q)}(\mathbf{X}_T, \mathbf{X}_T)\right) \le \sum_{q=1}^{Q}\ln\det\left(\boldsymbol{I}_T + 2\lambda^{-2}\mathbf{K}_2^{(q)}(\widetilde{\mathbf{X}}_T, \widetilde{\mathbf{X}}_T)\right) \tag{170}$$

$$\le Q\ln\det\left(\boldsymbol{I}_T + 2\lambda^{-2}\mathbf{K}(\widetilde{\mathbf{X}}_T, \widetilde{\mathbf{X}}_T)\right), \tag{171}$$

[9]For any positive semi-definite matrices A, B such that $A - B$ is positive semi-definite, we have $\lambda_i^{(A)} \ge \lambda_i^{(B)}$, where $(\lambda_i^{(A)})$ and $(\lambda_i^{(B)})$ is a non-negative eigenvalues of A and B with decreasing order. (This is a consequence of Courant–Fischer's min-max theorem.) Therefore, we have $\det(A) = \prod_i \lambda_i^{(A)} \ge \prod_i \lambda_i^{(B)} = \det(B)$ for such A and B.

where the second inequality follows from the fact that $\mathbf{K}(\widetilde{\mathbf{X}}_T, \widetilde{\mathbf{X}}_T) - \mathbf{K}_2^{(q)}(\widetilde{\mathbf{X}}_T, \widetilde{\mathbf{X}}_T)$ is positive semi-definite. From the definition of Q, we have

$$
\begin{aligned}
Q &\leq \log_2 \left(\frac{\eta_2}{\eta_1} \right) + 1 && (172)\\
&= \log_2 \eta_2 - \log_2 \eta_1 + 1 && (173)\\
&= \log_2 \max \left\{ 1, \frac{\Gamma(\nu)}{C_\nu} \ln \frac{T^2}{\lambda^2} \right\} - \log_2 \left(\frac{\nu \Gamma(\nu) \lambda^2}{T^2} \right)^{\frac{1}{\nu}} + 1 && (174)\\
&\leq \log_2 \left(1 + \frac{\Gamma(\nu)}{C_\nu} \ln \frac{T^2}{\lambda^2} \right) + \frac{1}{\nu} \log_2 \left(\frac{T^2}{\nu \Gamma(\nu) \lambda^2} \right) + 1. && (175)
\end{aligned}
$$

By combining Eqs. (164), (166), (171), and (175), we conclude

$$
\begin{aligned}
&\ln \det(\boldsymbol{I}_T + \lambda^{-2} \mathbf{K}_2(\mathbf{X}_T, \mathbf{X}_T)) && (176)\\
&\leq \left[\log_2 \left(1 + \frac{\Gamma(\nu)}{C_\nu} \ln \frac{T^2}{\lambda^2} \right) + \frac{1}{\nu} \log_2 \left(\frac{T^2}{\nu \Gamma(\nu) \lambda^2} \right) + 1 \right] \ln \det \left(\boldsymbol{I}_T + 2\lambda^{-2} \mathbf{K}(\widetilde{\mathbf{X}}_T, \widetilde{\mathbf{X}}_T) \right). && (177)
\end{aligned}
$$

By aggregating the upper bounds of (i), (ii), and (iii), we have the following inequality under $\eta_1 < \eta_2$:

$$
\ln \det(\boldsymbol{I}_T + \lambda^{-2} \mathbf{K}(\mathbf{X}_T, \mathbf{X}_T)) \leq C(T, \nu, \lambda) \ln \det \left(\boldsymbol{I}_T + 2\lambda^{-2} \mathbf{K}(\widetilde{\mathbf{X}}_T, \widetilde{\mathbf{X}}_T) \right) + 2C. \tag{178}
$$

Finally, if $\eta_1 \geq \eta_2$, we have

$$
\begin{aligned}
&\ln \det(\boldsymbol{I}_T + \lambda^{-2} \mathbf{K}(\mathbf{X}_T, \mathbf{X}_T)) && (179)\\
&\leq \ln \det(\boldsymbol{I}_T + \lambda^{-2} (\mathbf{K}_1(\mathbf{X}_T, \mathbf{X}_T) + \mathbf{K}_3(\mathbf{X}_T, \mathbf{X}_T))) && (180)\\
&\leq \ln \det(\boldsymbol{I}_T + \lambda^{-2} \mathbf{K}_1(\mathbf{X}_T, \mathbf{X}_T)) + \ln \det(\boldsymbol{I}_T + \lambda^{-2} \mathbf{K}_3(\mathbf{X}_T, \mathbf{X}_T)) && (181)\\
&\leq 2C && (182)\\
&\leq C(T, \nu, \lambda) \ln \det \left(\boldsymbol{I}_T + 2\lambda^{-2} \mathbf{K}(\widetilde{\mathbf{X}}_T, \widetilde{\mathbf{X}}_T) \right) + 2C, && (183)
\end{aligned}
$$

where the last inequality follows from $\ln \det \left(\boldsymbol{I}_T + 2\lambda^{-2} \mathbf{K}(\widetilde{\mathbf{X}}_T, \widetilde{\mathbf{X}}_T) \right) \geq 0$ and $C(T, \nu, \lambda) \geq 0$. The desired result is obtained by setting a new absolute constant C as $2C$ in the above inequality. □

B.2 Summary of Mercer Decomposition for Dot-Product Kernel on Sphere

In this subsection, we summarize the basic known results of the Mercer decomposition on $\mathbb{S}^d$. The content of this subsection is related to the analysis of the spherical harmonics. We refer to [1, 16] as the basic textbook. In the kernel method literature, the Mercer decomposition of the dot-product kernel and its eigendecay have been studied. See, e.g., [2, 36, 48]. Furthermore, the existing analysis of the neural tangent kernel also leverages the Mercer decomposition based on the spherical harmonics. We also refer to the appendix of [3, 5] as the related works of this subsection.

We first describe Mercer's theorem. Let $L^2(\mathcal{X}, \mu) \coloneqq \{ f : \mathcal{X} \to \mathbb{R} \mid \int_{\mathcal{X}} f^2(\boldsymbol{x}) \mu(\mathrm{d}\boldsymbol{x}) < \infty \}$ be the square-integrable functions on $\mathcal{X}$ under the measure μ. Furthermore, let us define the kernel integral operator $\mathcal{T}_k : L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu)$ of a square-integrable kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as $(\mathcal{T}_k f)(\cdot) = \int_{\mathcal{X}} k(\cdot, \boldsymbol{x}) f(\boldsymbol{x}) \mu(\mathrm{d}\boldsymbol{x})$. Then, Mercer's theorem guarantees that the positive kernel k is decomposed based on the eigenvalues and eigenfunctions sequence of $\mathcal{T}_k$ with absolute and uniform convergence on $\mathcal{X} \times \mathcal{X}$. We give the formal statement below.

Theorem 10 (Mercer's theorem, e.g., Theorem 4.49 in [12])**.** *Let $\mathcal{X}$ be a compact metric space, μ be a finite Borel measure whose support is $\mathcal{X}$, and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous and square integrable-positive definite kernel on $(\mathcal{X}, \mu)$. Suppose that $(\phi_i)_{i \in \mathbb{N}}$ and $(\lambda_i)_{i \in \mathbb{N}}$ are eigenfunctions and eigenvalues of the kernel integral operator $\mathcal{T}_k$, respectively. Namely, $(\phi_i)_{i \in \mathbb{N}}$ is an orthonormal bases of the eigenspace $\{ \mathcal{T}_k f \mid f \in L^2(\mathcal{X}, \mu) \}$ such that $\mathcal{T}_k \phi_i(\cdot) = \lambda_i \phi_i(\cdot)$ for all $i \in \mathbb{N}$. Then, we have*

$$
k(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \sum_{i \in \mathbb{N}} \lambda_i \phi_i(\boldsymbol{x}) \phi_i(\widetilde{\boldsymbol{x}}), \tag{184}
$$

where the convergence is absolute and uniform on $\mathcal{X} \times \mathcal{X}$.

Specifically, our interest is the Mercer decomposition of the kernel on $\mathbb{S}^d$. This is given as spherical harmonics on $\mathbb{S}^d$, which we define below.

Definition 1 (Spherical harmonics, e.g., Definition 2.7 in [1])**.** Fix any $d \geq 1$ and $m \in \mathbb{N}$. Let $\mathbb{Y}_m(\mathbb{R}^{d+1})$ be the all homogeneous polynomials of degree m in $\mathbb{R}^{d+1}$ that are also harmonic[10]. The space $\mathbb{Y}_m^{d+1} = \mathbb{Y}_m(\mathbb{R}^{d+1}) \mid_{\mathbb{S}^d}$ is called the spherical harmonic space of order m in $d+1$ dimensions. Any function in $\mathbb{Y}_m^{d+1}$ is called a spherical harmonic of order m in $d+1$ dimensions.

The following lemmas provide the properties of the spherical harmonics, which guarantee that the Mercer decomposition of the continuous dot-product kernel on $\mathbb{S}^d$ is defined based on spherical harmonics.

Lemma 11 (Dimension and completeness of sphererical harmonics, e.g., Chapter 2.1.3, Corollary 2.15, and Theorem 2.38 in [1])**.** *Fix any $d \geq 1$. Then, the following statements hold:*

- *For any $m \in \mathbb{N}$, we have $\dim(\mathbb{Y}_m^{d+1}) = N_{d+1,m}$ with $N_{d+1,m} = \frac{(2m+d-1)(m+d-2)!}{m!(d-1)!}$. Furthermore, For any $m, n \in \mathbb{N}$ with $m \neq n$, we have $\mathbb{Y}_m^{d+1} \perp \mathbb{Y}_n^{d+1}$[11].*
- *Let us define $(Y_{m,j})_{j \in [N_{d+1,m}]}$ be an orthonormal bases of $\mathbb{Y}_m^{d+1}$. Then, $\cup_{m \in \mathbb{N}} (Y_{m,j})_{j \in [N_{d+1,m}]}$ becomes an orthonormal bases of $L^2(\mathbb{S}^d, \sigma)$, where $\sigma(\cdot)$ is the induced Lebesgue measure on $\mathbb{S}^d$.*

Lemma 12 (Funk-Hecke Formula, e.g., Theorem 2.22 in [1] or Theorem 4.24 in [16])**.** *Fix any $d \geq 1$. Let $f : [-1, 1] \to \mathbb{R}$ be a continuous function. Define $|\mathbb{S}^{d-1}| := \frac{2\pi^{d/2}}{\Gamma(d/2)}$ as the surface area of $\mathbb{S}^{d-1}$. Then, for any $m \in \mathbb{N}$ and $Y_m \in \mathbb{Y}_m^{d+1}$, we have*

$$\int_{\mathbb{S}^d} f(\boldsymbol{z}^\top \boldsymbol{\eta}) Y_m(\boldsymbol{\eta}) \sigma(\mathrm{d}\boldsymbol{\eta}) = \lambda_m Y_m(\boldsymbol{z}), \tag{185}$$

where $\sigma(\cdot)$ is the induced Lebesgue measure on $\mathbb{S}^d$. Furthermore, λ_m is defined as

$$\lambda_m = |\mathbb{S}^{d-1}| \int_{-1}^{1} P_{m,d+1}(t) f(t) (1-t^2)^{\frac{d-2}{2}} \mathrm{d}t, \tag{186}$$

where $P_{m,d+1}(t)$ is the Legendre polynomial of degree m in $d+1$ dimensions, which is defined as

$$P_{m,d+1}(t) = m! \Gamma\left(\frac{d}{2}\right) \sum_{k=0}^{\lfloor m/2 \rfloor} (-1)^k \frac{(1-t^2)^k t^{m-2k}}{4^k k! (m-2k)! \Gamma\left(k + \frac{d}{2}\right)}. \tag{187}$$

Lemma 12 suggests that the spherical harmonics are eigenfunctions of the continuous dot-product kernel $k(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \widetilde{k}(\boldsymbol{x}^\top \widetilde{\boldsymbol{x}})$ on $\mathbb{S}^d$. Furthermore, Lemma 11 guarantees the $\cup_{m \in \mathbb{N}} (Y_{m,j})_{j \in [N_{d+1,m}]}$ forms an orthonormal bases of $L^2(\mathbb{S}^d, \sigma)$, which implies that they are the orthonormal bases of the eigenspace of $\mathcal{T}_k$. These facts give the following explicit form of Mercer decomposition for a continuous dot-product kernel.

Corollary 13. *Fix any $d \in \mathbb{N}_+$. Suppose $\mathcal{X} = \mathbb{S}^d$. Furthermore, assume the kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the positive definite kernel such that $\forall \boldsymbol{x}, \widetilde{\boldsymbol{x}} \in \mathcal{X}, k(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \widetilde{k}(\boldsymbol{x}^\top \widetilde{\boldsymbol{x}})$ with some continuous function $\widetilde{k} : [-1, 1] \to \mathbb{R}$. Then, we have the following Mercer decomposition of k:*

$$k(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \sum_{m=0}^{\infty} \lambda_m \sum_{j=1}^{N_{d+1,m}} Y_{m,j}(\boldsymbol{x}) Y_{m,j}(\widetilde{\boldsymbol{x}}), \tag{188}$$

where $(Y_{m,j}(\cdot))_{j \in [N_{d+1,m}]}$ denotes the spherical harmonics, which consist of orthonormal bases of $\mathbb{Y}_m^{d+1}$. Furthermore, $\lambda_m \geq 0$ is defined as

$$\lambda_m = |\mathbb{S}^{d-1}| \int_{-1}^{1} P_{m,d+1}(t) \widetilde{k}(t) (1-t^2)^{\frac{d-2}{2}} \mathrm{d}t. \tag{189}$$

[10] A polynomial $H(x_1, \ldots, x_{d+1})$ is called homogeneous of degree m if $H(tx_1, \ldots, tx_{d+1}) = t^n H(x_1, \ldots, x_{d+1})$. Furthermore, a polynomial $H(x_1, \ldots, x_{d+1})$ is called harmonic if $\Delta_{d+1} H = 0$, where Δ_{d+1} is the Laplace operator. See Chapter 4 in [16] or Chapter 2 in [1].

[11] Here, as with the second statement, we consider $L^2(\mathbb{S}^d, \sigma)$. Therefore, the inner product for any $f, g : \mathbb{S}^d \to \mathbb{R}$ is defined on $L^2(\mathbb{S}^d, \sigma)$ as $\int_{\mathbb{S}^d} f(\boldsymbol{x}) g(\boldsymbol{x}) \sigma(\mathrm{d}\boldsymbol{x})$.

Note that $\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_2 = \sqrt{2 - 2\boldsymbol{x}^\top \widetilde{\boldsymbol{x}}}$ for any $\boldsymbol{x}, \widetilde{\boldsymbol{x}} \in \mathbb{S}^d$. Therefore, we can represent k_{SE} and $k_{\mathrm{Matérn}}$ on $\mathbb{S}^d$ by Eq. (188). Finally, we describe the following addition theorem of the spherical harmonics, which plays a central role in avoiding the uniform boundness assumption in the existing proof of MIG on $\mathbb{S}^d$.

Lemma 14 (Addition theorem, e.g., Theorem 2.9 in [1] or Theorem 4.11 in [16])**.** *Fix any $d \geq 1$ and $m \in \mathbb{N}$. Let $(Y_{m,j})_{j \in [N_{d+1,m}]}$ be an orthonormal bases of $\mathbb{Y}_m^{d+1}$. Then, we have*

$$\forall \boldsymbol{x}, \widetilde{\boldsymbol{x}} \in \mathbb{S}^d, \quad \sum_{j=1}^{N_{d+1,m}} Y_{m,j}(\boldsymbol{x}) Y_{m,j}(\widetilde{\boldsymbol{x}}) = \frac{N_{d+1,m}}{|\mathbb{S}^{d+1}|} P_{m,d}(\boldsymbol{x}^\top \widetilde{\boldsymbol{x}}), \tag{190}$$

where $P_{m,d+1}(t)$ is the Legendre polynomial of degree m in $d+1$ dimensions, which is defined in Eq. (187)*.*

B.3 Upper Bound of MIG with Mercer Decomposition

By using Corollary 13 and Lemma 14 in the previous subsection, we can derive the following general form of the upper bound of MIG.

Lemma 15 (Adapted from [57])**.** *Suppose the kernel function k satisfies the condition in Corollary 13. Furthermore, assume $|k(\boldsymbol{x}, \widetilde{\boldsymbol{x}})| \leq 1$ for all $\boldsymbol{x}, \boldsymbol{x} \in \mathcal{X}$. Then, for any $M \in \mathbb{N}$, MIG on $\mathbb{S}^d$ satisfies*

$$\frac{1}{2} \max_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \in \mathbb{S}^d} \ln \det(\boldsymbol{I}_T + \lambda^{-2} \mathbf{K}(\mathbf{X}_T, \mathbf{X}_T)) \leq N_M \ln\left(1 + \frac{T}{\lambda^2}\right) + \frac{T}{|\mathbb{S}^{d+1}| \lambda^2} \sum_{m=M+1}^{\infty} \lambda_m N_{d+1,m}, \tag{191}$$

where $N_M = \sum_{m=0}^{M} N_{d+1,m}$.

The proof almost directly follows from [58], while a minor modification is required to deal with the unboundness of the eigenfunctions through the addition theorem. The same proof strategy is already provided in [31, 57] for analyzing the MIG of the neural tangent kernel on the sphere. Although our proof has no intrinsic change from their proof, we give the details below for completeness of our paper.

Proof. We first decompose the kernel matrix as $\mathbf{K}(\mathbf{X}_T, \mathbf{X}_T) = \mathbf{K}_{\mathrm{head}} + \mathbf{K}_{\mathrm{tail}}$, where $[\mathbf{K}_{\mathrm{head}}]_{i,l} = \sum_{m=0}^{M} \lambda_m \sum_{j=1}^{N_{d+1,m}} Y_{m,j}(\boldsymbol{x}_i) Y_{m,j}(\boldsymbol{x}_l)$ and $[\mathbf{K}_{\mathrm{tail}}]_{i,l} = \sum_{m=M+1}^{\infty} \lambda_m \sum_{j=1}^{N_{d+1,m}} Y_{m,j}(\boldsymbol{x}_i) Y_{m,j}(\boldsymbol{x}_l)$. Then, as with the proof in [58], the MIG is decomposed as

$$\frac{1}{2} \max_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \in \mathbb{S}^d} \ln \det(\boldsymbol{I}_T + \lambda^{-2} \mathbf{K}(\mathbf{X}_T, \mathbf{X}_T)) \tag{192}$$

$$= \frac{1}{2} \ln \det\left(\boldsymbol{I}_T + \frac{1}{\lambda^2} \mathbf{K}_{\mathrm{head}}\right) + \frac{1}{2} \ln \det\left(\boldsymbol{I}_T + \frac{1}{\lambda^2}\left(\boldsymbol{I}_T + \frac{1}{\lambda^2} \mathbf{K}_{\mathrm{head}}\right)^{-1} \mathbf{K}_{\mathrm{tail}}\right). \tag{193}$$

Based on the feature representation of the kernel, the first term is further bounded from above as follows (see [57, 58]):

$$\frac{1}{2} \ln \det\left(\boldsymbol{I}_T + \frac{1}{\lambda^2} \mathbf{K}_{\mathrm{head}}\right) \leq N_M \ln\left(1 + \frac{T}{\lambda^2 N_M}\right) \leq N_M \ln\left(1 + \frac{T}{\lambda^2}\right), \tag{194}$$

where the second inequality follows from $N_M \geq 1$. Regarding the second term, as with [58], we have

$$\frac{1}{2} \ln \det\left(\boldsymbol{I}_T + \frac{1}{\lambda^2}\left(\boldsymbol{I}_T + \frac{1}{\lambda^2} \mathbf{K}_{\mathrm{head}}\right)^{-1} \mathbf{K}_{\mathrm{tail}}\right) \tag{195}$$

$$\leq T \ln\left(T^{-1} \mathrm{Tr}\left(\boldsymbol{I}_T + \frac{1}{\lambda^2}\left(\boldsymbol{I}_T + \frac{1}{\lambda^2} \mathbf{K}_{\mathrm{head}}\right)^{-1} \mathbf{K}_{\mathrm{tail}}\right)\right) \tag{196}$$

$$\leq T \ln\left(T^{-1}\left(T + \frac{1}{\lambda^2} \mathrm{Tr}\left(\mathbf{K}_{\mathrm{tail}}\right)\right)\right), \tag{197}$$

where the first inequality follows from $\ln\det(A) \le T\ln(\mathrm{Tr}(A)/T)$ for any positive definite matrix $A \in \mathbb{R}^{T\times T}$ (e.g., Lemma 1 in [58]). Then, from addition theorem (Theorem 14), we have

$$\mathrm{Tr}\left(\mathbf{K}_{\text{tail}}\right) = \sum_{t=1}^{T}\sum_{m=M+1}^{\infty}\lambda_m \sum_{j=1}^{N_{d+1,m}} Y_{m,j}(\boldsymbol{x}_t)Y_{m,j}(\boldsymbol{x}_t) \tag{198}$$

$$= \sum_{t=1}^{T}\sum_{m=M+1}^{\infty}\lambda_m \frac{N_{d+1,m}}{|\mathbb{S}^{d+1}|}P_{m,d}(\boldsymbol{x}_t^\top \boldsymbol{x}_t) \tag{199}$$

$$= \frac{T}{|\mathbb{S}^{d+1}|}\sum_{m=M+1}^{\infty}\lambda_m N_{d+1,m}, \tag{200}$$

where the last line use $P_{m,d}(\boldsymbol{x}_t^\top \boldsymbol{x}_t) = P_{m,d}(1) = 1$. By combining the above equation with Eq. (197), we have

$$\frac{1}{2}\ln\det\left(\boldsymbol{I}_T + \frac{1}{\lambda^2}\left(\boldsymbol{I}_T + \frac{1}{\lambda^2}\mathbf{K}_{\text{head}}\right)^{-1}\mathbf{K}_{\text{tail}}\right) \le T\ln\left(1 + \frac{1}{\lambda^2|\mathbb{S}^{d+1}|}\sum_{m=M+1}^{\infty}\lambda_m N_{d+1,m}\right) \tag{201}$$

$$\le \frac{T}{\lambda^2|\mathbb{S}^{d+1}|}\sum_{m=M+1}^{\infty}\lambda_m N_{d+1,m}, \tag{202}$$

where the last line use $\forall z \in \mathbb{R}, \ln(1+z) \le z$. □

To obtain the explicit upper bound of Eq. (191), we introduce the following lemma.

Lemma 16 (Upper bound of $N_{d+1,m}$ and N_M)**.** *Fix any $d \in \mathbb{N}_+$. Then, for any $m \in \mathbb{N}_+$, we have*

$$N_{d+1,m} \le (d+1)e^{d-1}m^{d-1}. \tag{203}$$

Futhermore, for any $M \in \mathbb{N}$, we have

$$N_M \le 1 + (d+1)e^{d-1}M^d. \tag{204}$$

Proof. Recall $N_{d+1,m} = \frac{(2m+d-1)(m+d-2)!}{m!(d-1)!}$. Under $d = 1$, we have

$$N_{d+1,m} = \frac{(2m)(m-1)!}{m!} = 2 = (d+1)e^{d-1}m^{d-1} \tag{205}$$

for any $m \in \mathbb{N}_+$. Under $d \ge 2$, since $N_{d+1,m} = \frac{(2m+d-1)(m+d-2)!}{m!(d-1)!} = \frac{2m+d-1}{m}\binom{m+d-2}{d-1}$ and $\binom{m+d-2}{d-1} \le \left(\frac{(m+d-2)e}{d-1}\right)^{d-1}$, we have

$$N_{d+1,m} \le \frac{2m+d-1}{m}\left(\frac{(m+d-2)e}{d-1}\right)^{d-1} \tag{206}$$

$$\le (2+d-1)e^{d-1}\left(\frac{m+d-2}{d-1}\right)^{d-1} \tag{207}$$

$$\le (d+1)e^{d-1}m^{d-1}. \tag{208}$$

Finally, since $N_{d+1,0} = 1$, we have

$$N_M = 1 + \sum_{m=1}^{M} N_{d+1,m} \le 1 + (d+1)e^{d-1}\sum_{m=1}^{M} m^{d-1} \le 1 + (d+1)e^{d-1}M^{d-1}. \tag{209}$$

□

B.4 Eigendecay of SE and Matérn Kernel

To obtain the explicit upper bound of Eq. (191), we need the upper bound of the eigenvalue in Eq. (189) under SE and Matérn kernel. Regarding SE kernel, several existing works have already studied it [36, 39]. We formally provide the following lemma from [36].

Lemma 17 (Eigendecay for $k = k_{\mathrm{SE}}$ on $\mathbb{S}^d$, Theorem 2 in [36])**.** *Fix any $d \in \mathbb{N}_+$, $\theta > 0$, and define $\mathcal{X} = \mathbb{S}^d$. Suppose that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined as $k(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \exp\left(-\frac{\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|_2^2}{\theta}\right)$. Then, the eigenvalues $(\lambda_m)_{m\in\mathbb{N}_+}$ defined in* (189) *satisfy*

$$\lambda_m < |\mathbb{S}^d| \left(\frac{2e}{\theta}\right)^m \frac{(2e)^{\frac{d+1}{2}}\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{\pi}(2m+d-1)^{m+\frac{d}{2}}} \exp\left(-\frac{2}{\theta} + \frac{1}{\theta^2}\right). \tag{210}$$

Regarding Matérn kernel, we provide the upper bound of λ_m for $\nu > 1/2$ by extending the proof in [20], which studies λ_m for Laplace kernel (Matérn with $\nu = 1/2$). As with the proof in [20], we leverage the following lemma, which relates the spectral density of the kernel to λ_m.

Lemma 18 (Eigenvalues and spectral density, Theorem 4.1 in [38])**.** *Fix any $d \in \mathbb{N}_+$. Suppose that $k : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to \mathbb{R}$ is a positive definite, stationary, and isotropic kernel function on $\mathbb{R}^{d+1}$ such that $\forall \boldsymbol{x}, \widetilde{\boldsymbol{x}}, k(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \Phi(\boldsymbol{x} - \widetilde{\boldsymbol{x}})$ for some function $\Phi(\cdot)$. Furthermore, suppose $\Phi(\cdot)$ is represented as*

$$\Phi(\boldsymbol{x}) = \frac{1}{(2\pi)^{d+1}} \int_{\mathbb{R}^{d+1}} \widehat{\Phi}(\|\boldsymbol{\eta}\|_2) e^{i\boldsymbol{\eta}^\top \boldsymbol{x}} \mathrm{d}\boldsymbol{\eta}, \tag{211}$$

for some function $\widehat{\Phi}$ such that $\forall a \geq 0, \widehat{\Phi}(a) \geq 0$ and $\int_{\mathbb{R}^{d+1}} \widehat{\Phi}(\|\boldsymbol{\eta}\|_2)\mathrm{d}\boldsymbol{\eta} < \infty$. Then, there exists a function $\widetilde{k} : [-1, 1] \to \mathbb{R}$ such that $\forall \boldsymbol{x}, \widetilde{\boldsymbol{x}} \in \mathbb{S}^d, \widetilde{k}(\boldsymbol{x}^\top \widetilde{\boldsymbol{x}}) = k(\boldsymbol{x}, \widetilde{\boldsymbol{x}})$. Furthermore, λ_m in Eq. (189) *is given by*

$$\lambda_m = \int_0^\infty t\widehat{\Phi}(t) B^2_{m+\frac{d-1}{2}}(t)\mathrm{d}t, \tag{212}$$

where $B_{m+\frac{d-1}{2}}(\cdot)$ is the usual Bessel function of the first kind and of order $m + \frac{d-1}{2}$.

In the Matern kernel, the spectral density that satisfies the conditions in the lemma is defined when $\nu > 1/2$. Then, the explicit form of $\widehat{\Phi}(t)$ is given as:

$$\widehat{\Phi}(t) = \frac{C_{d,\nu}}{\ell^{2\nu}} \left(\frac{2\nu}{\ell^2} + t^2\right)^{-\left(\nu+\frac{d+1}{2}\right)}, \tag{213}$$

where

$$C_{d,\nu} = \frac{2^{d+1}\pi^{(d+1)/2}\Gamma\left(\nu + \frac{d+1}{2}\right)(2\nu)^\nu}{\Gamma(\nu)}. \tag{214}$$

See, Chapter 4.2 in [41]. By using Lemma 18, we obtain the following lemma.

Lemma 19 (Eigendecay for $k = k_{\text{Matérn}}$ on $\mathbb{S}^d$)**.** *Fix any $d \in \mathbb{N}_+$, $\ell > 0$, and define $\mathcal{X} = \mathbb{S}^d$. Suppose that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined as $k(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|_2}{\ell}\right) J_\nu\left(\frac{\sqrt{2\nu}\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|_2}{\ell}\right)$. Then, the eigenvalues $(\lambda_m)_{m\in\mathbb{N}_+}$ defined in* (189) *satisfies*

$$\lambda_m \leq \frac{\widetilde{C}_{d,\nu}}{\ell^{2\nu}} m^{-2\nu-d}. \tag{215}$$

if $m > 2\nu$ and $\nu > 1/2$. Here, $\widetilde{C}_{d,\nu}$ is defined as

$$\widetilde{C}_{d,\nu} = C_{d,\nu} \frac{\Gamma(2\nu + d)}{\Gamma^2\left(\nu + \frac{d+1}{2}\right)} \exp\left(2\nu + d + \frac{1}{6}\right). \tag{216}$$

Proof. From Lemma 18, we have

$$\lambda_m = \int_0^\infty t\widehat{\Phi}(t) B^2_{m+\frac{d-1}{2}}(t)\mathrm{d}t \tag{217}$$

$$= \frac{C_{d,\nu}}{\ell^{2\nu}} \int_0^\infty t\left(\frac{2\nu}{\ell^2} + t^2\right)^{-\left(\nu+\frac{d+1}{2}\right)} B^2_{m+\frac{d-1}{2}}(t)\mathrm{d}t \tag{218}$$

$$\leq \frac{C_{d,\nu}}{\ell^{2\nu}} \int_0^\infty t^{-2\nu-d} B^2_{m+\frac{d-1}{2}}(t)\mathrm{d}t. \tag{219}$$

As with the proof of Theorem 7 in [20], we evaluate the integral $\int_0^\infty t^{-2\nu-d} B^2_{m+\frac{d-1}{2}}(t)\mathrm{d}t$ by using the following identity (Chapter 13.4.1 in [22]):

$$\int_0^\infty \frac{B_p(at)B_q(at)}{t^z}\mathrm{d}t = \frac{\left(\frac{1}{2}a\right)^{z-1}\Gamma(z)\Gamma\left(\frac{1}{2}p+\frac{1}{2}q-\frac{1}{2}z+\frac{1}{2}\right)}{2\Gamma\left(\frac{1}{2}z+\frac{1}{2}q-\frac{1}{2}p+\frac{1}{2}\right)\Gamma\left(\frac{1}{2}z+\frac{1}{2}p+\frac{1}{2}q+\frac{1}{2}\right)\Gamma\left(\frac{1}{2}z+\frac{1}{2}p-\frac{1}{2}q+\frac{1}{2}\right)}, \tag{220}$$

where $p+q+1 > z > 0$. By setting $p = q = m + \frac{d-1}{2}$, $z = 2\nu + d$, and $a = 1$, we have $p+q+1 > z \Leftrightarrow m > \nu$. Hence, for any $m > \nu$, we have

$$\int_0^\infty t^{-2\nu-d} B^2_{m+\frac{d-1}{2}}(t)\mathrm{d}t = \left(\frac{1}{2}\right)^{2\nu+d-1} \frac{\Gamma(2\nu+d)\Gamma(m-\nu)}{2\Gamma^2\left(\nu+\frac{d+1}{2}\right)\Gamma(m+\nu+d)}. \tag{221}$$

Stirling's formula implies that there exists a constant $C > 0$ such that

$$\Gamma(m-\nu) \le C(m-\nu)^{m-\nu-\frac{1}{2}}\exp(-m+\nu)\exp\left(\frac{1}{12(m-\nu)}\right), \tag{222}$$

$$\Gamma(m+\nu+d) \ge C(m+\nu+d)^{m+\nu+d-\frac{1}{2}}\exp\left(-(m+\nu+d)\right). \tag{223}$$

Therefore, for any $m \ge 2\nu$ with $\nu > 1/2$, we have

$$\frac{\Gamma(m-\nu)}{\Gamma(m+\nu+d)} \le \frac{(m-\nu)^{m-\nu-\frac{1}{2}}\exp(-m+\nu)\exp\left(\frac{1}{12(m-\nu)}\right)}{(m+\nu+d)^{m+\nu+d-\frac{1}{2}}\exp\left(-(m+\nu+d)\right)} \tag{224}$$

$$\le \frac{(m-\nu)^{m-\nu-\frac{1}{2}}}{(m+\nu+d)^{m+\nu+d-\frac{1}{2}}}\exp\left(2\nu+d+\frac{1}{6}\right) \tag{225}$$

$$\le \frac{(m-\nu)^{m-\nu-\frac{1}{2}}}{(m-\nu)^{m+\nu+d-\frac{1}{2}}}\exp\left(2\nu+d+\frac{1}{6}\right) \tag{226}$$

$$= (m-\nu)^{-2\nu-d}\exp\left(2\nu+d+\frac{1}{6}\right) \tag{227}$$

$$\le 2^{2\nu+d}m^{-2\nu-d}\exp\left(2\nu+d+\frac{1}{6}\right), \tag{228}$$

where the second inequality follows from $m-\nu \ge 1/2$ due to $m \ge 2\nu \Leftrightarrow m-\nu \ge \nu$, the third inequality follows from $m+\nu+d \ge m-\nu$, and the last inequality follows from $m-\nu \ge m-m/2 \ge m/2$. By aggregating Eq. (219), (221), and (228), we have

$$\lambda_m \le \frac{C_{d,\nu}}{\ell^{2\nu}}\left(\frac{1}{2}\right)^{2\nu+d-1}\frac{\Gamma(2\nu+d)}{2\Gamma^2\left(\nu+\frac{d+1}{2}\right)}2^{2\nu+d}m^{-2\nu-d}\exp\left(2\nu+d+\frac{1}{6}\right) \tag{229}$$

$$= \frac{\widetilde{C}_{d,\nu}}{\ell^{2\nu}}m^{-2\nu-d}. \tag{230}$$

□

B.5 Proof of Theorem 7

Squared exponential kernel. From Lemma 17, we have

$$\lambda_m < |\mathbb{S}^d|\left(\frac{2e}{\theta}\right)^m \frac{(2e)^{\frac{d+1}{2}}\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{\pi}(2m+d-1)^{m+\frac{d}{2}}}\exp\left(-\frac{2}{\theta}+\frac{1}{\theta^2}\right) \tag{231}$$

$$\le |\mathbb{S}^d|\frac{(2e)^{\frac{d+1}{2}}\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{\pi}}\exp\left(-\frac{2}{\theta}+\frac{1}{\theta^2}\right)\left(\frac{2e}{\theta}\right)^m(2m)^{-m-\frac{d}{2}} \tag{232}$$

$$\le |\mathbb{S}^d|\frac{(2e)^{\frac{d+1}{2}}\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{\pi}}\exp\left(-\frac{2}{\theta}+\frac{1}{\theta^2}\right)\left(\frac{e}{\theta}\right)^m m^{-m-\frac{d}{2}}. \tag{233}$$

Here, we set $C_{d,\theta}$ as

$$C_{d,\theta} = \frac{|\mathbb{S}^d|}{|\mathbb{S}^{d+1}|} \frac{(2e)^{\frac{d+1}{2}} \Gamma\left(\frac{d+1}{2}\right)}{\sqrt{\pi}} \exp\left(-\frac{2}{\theta} + \frac{1}{\theta^2}\right). \tag{234}$$

Then,

$$\gamma_T(\mathcal{X}) \leq N_M \ln\left(1 + \frac{T}{\lambda^2}\right) + \frac{T}{|\mathbb{S}^{d+1}|\lambda^2} \sum_{m=M+1}^{\infty} \lambda_m N_{d+1,m} \tag{235}$$

$$\leq N_M \ln\left(1 + \frac{T}{\lambda^2}\right) + \frac{C_{d,\theta} T}{\lambda^2} \sum_{m=M+1}^{\infty} \left(\frac{e}{\theta}\right)^m m^{-m-\frac{d}{2}} N_{d+1,m} \tag{236}$$

$$\leq \left[1 + (d+1)e^{d-1}M^d\right] \ln\left(1 + \frac{T}{\lambda^2}\right) + \frac{C_{d,\theta} T}{\lambda^2}(d+1)e^{d-1} \sum_{m=M+1}^{\infty} \left(\frac{e}{\theta m}\right)^m m^{\frac{d}{2}-1}, \tag{237}$$

where the first and last inequalities follow from Lemmas 9, 15 and Lemma 16, respectively. Here, for any $d \in \mathbb{N}_+$ and $m \in \mathbb{N}_+$, we have $m^{\frac{d}{2}-1} \leq c_d^m$ with $c_d = \max\left\{1, \exp\left(\frac{1}{e}\left(\frac{d}{2} - 1\right)\right)\right\}$. Indeed, when $d \leq 2$, we have $m^{d/2-1} \leq 1 = c_d$. When $d \geq 3$, the function $g(m) = m^{\frac{1}{m}\left(\frac{d}{2}-1\right)}$ attains maximum at $m = e$ on $[1, \infty)$, which implies $g(m) \leq \exp\left(\frac{1}{e}\left(\frac{d}{2} - 1\right)\right) \Rightarrow m^{d/2-1} \leq \exp\left(\frac{1}{e}\left(\frac{d}{2} - 1\right)\right)^m \leq c_d^m$. Hence, we have

$$\gamma_T(\mathcal{X}) \leq \left[1 + (d+1)e^{d-1}M^d\right] \ln\left(1 + \frac{T}{\lambda^2}\right) + \frac{C_{d,\theta} T}{\lambda^2}(d+1)e^{d-1} \sum_{m=M+1}^{\infty} \left(\frac{ec_d}{\theta m}\right)^m. \tag{238}$$

In the remaining proof, we consider the upper bound of $\left(\frac{ec_d}{\theta m}\right)^m$ separately based on $\theta > 0$. If $\theta \leq e^2 c_d$, we have

$$\left(\frac{ec_d}{\theta m}\right)^m = \exp\left(-m \ln\left(\frac{\theta m}{ec_d}\right)\right) \leq \exp\left(-m\right) \tag{239}$$

for any m such that $\frac{\theta m}{ec_d} \geq e \Leftrightarrow m \geq \frac{e^2 c_d}{\theta}$. Then, by noting that the condition $T/(e-1) \geq \lambda^2$ implies $\ln\left(1 + \frac{T}{\lambda^2}\right) \geq 1$, we have the following inequalities by setting $M = \left\lfloor \frac{e^2 c_d}{\theta} \ln\left(1 + \frac{T}{\lambda^2}\right) \right\rfloor$:

$$\sum_{m=M+1}^{\infty} \left(\frac{ec_d}{\theta m}\right)^m \leq \sum_{m=M+1}^{\infty} \exp(-m) \tag{240}$$

$$\leq \int_M^{\infty} \exp(-m) \mathrm{d}m \tag{241}$$

$$\leq \exp(-M) \tag{242}$$

$$\leq \exp\left(-\frac{e^2 c_d}{\theta} \ln\left(1 + \frac{T}{\lambda^2}\right) + 1\right) \tag{243}$$

$$\leq e\left(1 + \frac{T}{\lambda^2}\right)^{-\frac{e^2 c_d}{\theta}} \tag{244}$$

$$\leq e\left(1 + \frac{T}{\lambda^2}\right)^{-1} \tag{245}$$

$$\leq e\frac{\lambda^2}{T}. \tag{246}$$

Therefore, for $\theta \leq e^2 c_d$, the following inequality holds from Eqs. (238) and (246), and the definition of M:

$$\gamma_T(\mathcal{X}) \leq \left[1 + (d+1)e^{d-1}\left(\frac{e^2 c_d}{\theta} \ln\left(1 + \frac{T}{\lambda^2}\right)\right)^d\right] \ln\left(1 + \frac{T}{\lambda^2}\right) + eC_{d,\theta}(d+1)e^{d-1}. \tag{247}$$

Next, if $\theta > e^2 c_d$, we have

$$\left(\frac{ec_d}{\theta m}\right)^m = \exp\left(-m \ln\left(\frac{\theta m}{ec_d}\right)\right) \leq \exp\left(-m \ln\left(\frac{\theta}{ec_d}\right)\right) \tag{248}$$

for any $m \in \mathbb{N}_+$. Then, similarly to the proof under $\theta \leq e^2 c_d$, we have the following inequalities for any $M \in \mathbb{N}$:

$$\begin{aligned}
\sum_{m=M+1}^{\infty} \left(\frac{ec_d}{\theta m}\right)^m &\leq \sum_{m=M+1}^{\infty} \exp\left(-m \ln\left(\frac{\theta}{ec_d}\right)\right) && (249)\\
&\leq \int_M^{\infty} \exp\left(-m \ln\left(\frac{\theta}{ec_d}\right)\right) \mathrm{d}m && (250)\\
&= \frac{1}{\ln\left(\frac{\theta}{ec_d}\right)} \exp\left(-M \ln\left(\frac{\theta}{ec_d}\right)\right) && (251)\\
&< \exp\left(-M \ln\left(\frac{\theta}{ec_d}\right)\right), && (252)
\end{aligned}$$

where the last inequality follows from $\theta/ec_d > e \Leftrightarrow \theta > e^2 c_d$. By setting $M = \left\lceil \frac{1}{\ln\left(\frac{\theta}{ec_d}\right)} \ln\left(1 + \frac{T}{\lambda^2}\right) \right\rceil$, we have

$$\sum_{m=M+1}^{\infty} \left(\frac{ec_d}{\theta m}\right)^m \leq \exp\left(-\ln\left(1 + \frac{T}{\lambda^2}\right)\right) = \left(1 + \frac{T}{\lambda^2}\right)^{-1} \leq \frac{\lambda^2}{T}. \tag{253}$$

Hence, for $\theta > e^2 c_d$, we have

$$\gamma_T(\mathcal{X}) \leq \left[1 + (d+1)e^{d-1}\left(\frac{1}{\ln\left(\frac{\theta}{ec_d}\right)} \ln\left(1 + \frac{T}{\lambda^2}\right) + 1\right)^d\right] \ln\left(1 + \frac{T}{\lambda^2}\right) + C_{d,\theta}(d+1)e^{d-1}. \tag{254}$$

Finally, aligning Eqs. (247) and (254) by focusing on the dependence on T, λ^2, and θ, we obtain the desired result.

Matérn kernel. Similarly to the proof for the SE kernel, for any $M \geq 2\nu$, we have

$$\begin{aligned}
&\frac{1}{2} \max_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T \in \mathbb{S}^d} \ln\det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}(\mathbf{X}_T, \mathbf{X}_T)) && (255)\\
&\leq N_M \ln\left(1 + \frac{T}{\lambda^2}\right) + \frac{T}{|\mathbb{S}^{d+1}|\lambda^2} \sum_{m=M+1}^{\infty} \lambda_m N_{d+1,m} && (256)\\
&\leq \left[1 + (d+1)e^{d-1}M^d\right] \ln\left(1 + \frac{T}{\lambda^2}\right) + \frac{T(d+1)e^{d-1}}{|\mathbb{S}^{d+1}|\lambda^2} \sum_{m=M+1}^{\infty} \lambda_m m^{d-1} && (257)\\
&\leq \left[1 + (d+1)e^{d-1}M^d\right] \ln\left(1 + \frac{T}{\lambda^2}\right) + \frac{T(d+1)\widetilde{C}_{d,\nu}e^{d-1}}{|\mathbb{S}^{d+1}|\lambda^2\ell^{2\nu}} \sum_{m=M+1}^{\infty} m^{-2\nu-1} && (258)\\
&= \left[1 + (d+1)e^{d-1}M^d\right] \ln\left(1 + \frac{T}{\lambda^2}\right) + \frac{T\overline{C}_{d,\nu}}{\lambda^2\ell^{2\nu}} \sum_{m=M+1}^{\infty} m^{-2\nu-1}, && (259)
\end{aligned}$$

where the second inequality follows from Lemma 16, and the third inequality follows from $M \geq 2\nu$ and Lemma 19. In the last equation, we set $\overline{C}_{d,\nu} = \frac{(d+1)\widetilde{C}_{d,\nu}e^{d-1}}{|\mathbb{S}^{d+1}|}$. Furthermore,

$$\sum_{m=M+1}^{\infty} m^{-2\nu-1} \leq \int_M^{\infty} m^{-2\nu-1}\mathrm{d}m = \frac{M^{-2\nu}}{2\nu}. \tag{260}$$

By balancing $M^d \ln(1 + T/\lambda^2)$ and $\frac{TM^{-2\nu}}{\lambda^2\ell^{2\nu}}$ under the condition $M \geq 2\nu$, we set $M = \left\lceil \max\left\{2\nu, \left\lceil \frac{T}{\lambda^2\ell^{2\nu}} \ln^{-1}\left(1 + \frac{T}{\lambda^2}\right)\right\rceil^{1/(2\nu+d)}\right\}\right\rceil$. Then,

$$\frac{1}{2} \max_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T \in \mathbb{S}^d} \ln\det(\boldsymbol{I}_T + \lambda^{-2}\mathbf{K}(\mathbf{X}_T, \mathbf{X}_T)) \tag{261}$$

$$\leq \left[1 + (d+1)e^{d-1}M^d\right] \ln\left(1 + \frac{T}{\lambda^2}\right) + \frac{T\overline{C}_{d,\nu}}{\lambda^2\ell^{2\nu}} \frac{M^{-2\nu}}{2\nu} \tag{262}$$

$$\leq \left[1 + (d+1)e^{d-1}M^d\right] \ln\left(1 + \frac{T}{\lambda^2}\right) + \frac{\overline{C}_{d,\nu}}{2\nu} M^d \ln\left(1 + \frac{T}{\lambda^2}\right) \tag{263}$$

$$= \ln\left(1 + \frac{T}{\lambda^2}\right) + \left[\frac{\overline{C}_{d,\nu}}{2\nu} + (d+1)e^{d-1}\right] M^d \ln\left(1 + \frac{T}{\lambda^2}\right) \tag{264}$$

$$= \ln\left(1 + \frac{T}{\lambda^2}\right) + C'_{d,\nu}\left((2\nu)^d + \left(\frac{T}{\lambda^2\ell^{2\nu}}\right)^{\frac{d}{2\nu+d}} \ln^{-\frac{d}{2\nu+d}}\left(1 + \frac{T}{\lambda^2}\right)\right) \ln\left(1 + \frac{T}{\lambda^2}\right) \tag{265}$$

$$= \left(C'_{d,\nu}(2\nu)^d + 1\right) \ln\left(1 + \frac{T}{\lambda^2}\right) + C'_{d,\nu}\left(\frac{T}{\lambda^2\ell^{2\nu}}\right)^{\frac{d}{2\nu+d}} \ln^{\frac{2\nu}{2\nu+d}}\left(1 + \frac{T}{\lambda^2}\right), \tag{266}$$

where we set $C'_{d,\nu} = \frac{\overline{C}_{d,\nu}}{2\nu} + (d+1)e^{d-1}$. In the above equations, the second inequality follows from

$$M^d \ln(1 + T/\lambda^2) \geq \frac{TM^{-2\nu}}{\lambda^2\ell^{2\nu}} \Leftrightarrow \frac{\lambda^2\ell^{2\nu}}{T} \ln(1 + T/\lambda^2) \geq M^{-2\nu-d} \tag{267}$$

$$\Leftrightarrow \left[\frac{\lambda^2\ell^{2\nu}}{T} \ln(1 + T/\lambda^2)\right]^{-\frac{1}{2\nu+d}} \leq M \tag{268}$$

$$\Leftrightarrow \left[\frac{T}{\lambda^2\ell^{2\nu}} \ln^{-1}(1 + T/\lambda^2)\right]^{\frac{1}{2\nu+d}} \leq M. \tag{269}$$

Finally, combining Eq. (266) with Lemma 9, we obtain the desired result [12]. □

C Auxiliary Lemmas

Lemma 20 (Sub-optimality gap and the neighborhood around the maximizer)**.** *Suppose f is continuous. Then, the following two statements hold:*

- *Under conditions 1 and 3 in Lemma 2, $\boldsymbol{x} \in \mathcal{B}_2(\rho_{\mathrm{lin}}; \boldsymbol{x}^*)$ holds for any $\boldsymbol{x} \in \mathcal{X}$ such that $f(\boldsymbol{x}^*) - f(\boldsymbol{x}) \leq \varepsilon_1$ with $\varepsilon_1 = \min\{c_{\mathrm{gap}}, c_{\mathrm{lin}}\rho_{\mathrm{lin}}\}$.*
- *Under conditions 1 and 4 in Lemma 2, $\boldsymbol{x} \in \mathcal{B}_2(\rho_{\mathrm{quad}}; \boldsymbol{x}^*)$ holds for any $\boldsymbol{x} \in \mathcal{X}$ such that $f(\boldsymbol{x}^*) - f(\boldsymbol{x}) \leq \varepsilon_2$ with $\varepsilon_2 = \min\{c_{\mathrm{gap}}, c_{\mathrm{quad}}\rho_{\mathrm{quad}}^2\}$.*

Proof. We only prove the first statement since the proof of the second statement is almost the same. When $\mathcal{B}_2(\rho_{\mathrm{lin}}; \boldsymbol{x}^*) = \mathcal{X}$, the statement is trivial. Hereafter, we assume $\mathcal{B}_2(\rho_{\mathrm{lin}}; \boldsymbol{x}^*) \neq \mathcal{X}$. Here , note that $f(\boldsymbol{x}^*) - f(\widetilde{\boldsymbol{x}}) \geq c_{\mathrm{lin}}\rho_{\mathrm{lin}}$ holds for any $\widetilde{\boldsymbol{x}} \in \mathcal{B}_2^b(\rho_{\mathrm{lin}}; \boldsymbol{x}^*)$ from condition 3 in Lemma 2, where $\mathcal{B}_2^b(\rho_{\mathrm{lin}}; \boldsymbol{x}^*) = \{\boldsymbol{x} \in \mathcal{X} \mid \|\boldsymbol{x} - \boldsymbol{x}^*\|_2 = \rho_{\mathrm{lin}}\}$. Furthermore, from the continuity of f and the compactness of $(\mathcal{X} \setminus \mathcal{B}_2(\rho_{\mathrm{lin}}; \boldsymbol{x}^*)) \cup \mathcal{B}_2^b(\rho_{\mathrm{lin}}; \boldsymbol{x}^*)$, there exists $\widetilde{\boldsymbol{x}}_* \in \operatorname{argmax}_{\boldsymbol{x} \in (\mathcal{X} \setminus \mathcal{B}_2(\rho_{\mathrm{lin}}; \boldsymbol{x}^*)) \cup \mathcal{B}_2^b(\rho_{\mathrm{lin}}; \boldsymbol{x}^*)} f(\boldsymbol{x})$. Then, we consider the following two cases separately.

- When $c_{\mathrm{gap}} \geq c_{\mathrm{lin}}\rho_{\mathrm{lin}}$, $\varepsilon_1 = c_{\mathrm{lin}}\rho_{\mathrm{lin}}$ holds. If there exists $\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{B}_2(\rho_{\mathrm{lin}}; \boldsymbol{x}^*)$ such that $f(\boldsymbol{x}^*) - f(\boldsymbol{x}) \leq \varepsilon_1 = c_{\mathrm{lin}}\rho_{\mathrm{lin}}$, we can choose $\widetilde{\boldsymbol{x}}_*$ such that $\widetilde{\boldsymbol{x}}_* \in \mathcal{X} \setminus \mathcal{B}_2(\rho_{\mathrm{lin}}; \boldsymbol{x}^*)$ since $f(\boldsymbol{x}^*) - f(\widetilde{\boldsymbol{x}}) \geq c_{\mathrm{lin}}\rho_{\mathrm{lin}}$ holds for any $\widetilde{\boldsymbol{x}} \in \mathcal{B}_2^b(\rho_{\mathrm{lin}}; \boldsymbol{x}^*)$. Furthermore, such $\widetilde{\boldsymbol{x}}_*$ is the local maximizer on $\mathcal{X}$, which satisfies $f(\boldsymbol{x}^*) - f(\widetilde{\boldsymbol{x}}_*) \leq f(\boldsymbol{x}^*) - f(\boldsymbol{x}) \leq \varepsilon_1 \leq c_{\mathrm{gap}}$. This contradicts condition 1 in Lemma 2.

[12] Note that we need to adjust the noise variance parameter by a factor $1/\sqrt{2}$ from Lemma 9.

- When $c_{\mathrm{gap}} < c_{\mathrm{lin}}\rho_{\mathrm{lin}}$, $\varepsilon_1 = c_{\mathrm{gap}}$ holds. If there exists $\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{B}_2(\rho_{\mathrm{lin}}; \boldsymbol{x}^*)$ such that $f(\boldsymbol{x}^*) - f(\boldsymbol{x}) \leq \varepsilon_1 = c_{\mathrm{gap}}$, we can choose $\widetilde{\boldsymbol{x}}_*$ such that $\widetilde{\boldsymbol{x}}_* \in \mathcal{X} \setminus \mathcal{B}_2(\rho_{\mathrm{lin}}; \boldsymbol{x}^*)$ since $f(\boldsymbol{x}^*) - f(\widetilde{\boldsymbol{x}}) \geq c_{\mathrm{lin}}\rho_{\mathrm{lin}} > c_{\mathrm{gap}}$ holds for any $\widetilde{\boldsymbol{x}} \in \mathcal{B}_2^b(\rho_{\mathrm{lin}}; \boldsymbol{x}^*)$. Furthermore, such $\widetilde{\boldsymbol{x}}_*$ is the local maximizer on $\mathcal{X}$, which satisfies $f(\boldsymbol{x}^*) - f(\widetilde{\boldsymbol{x}}_*) \leq f(\boldsymbol{x}^*) - f(\boldsymbol{x}) \leq \varepsilon_1 = c_{\mathrm{gap}}$. This contradicts condition 1 in Lemma 2.

□

Lemma 21 (Upper bound of regret of GP-UCB for any index subset)**.** *Fix any index set $\mathcal{T} \subset [T]$. Then, when running GP-UCB, we have the following inequality under $\mathcal{A}$:*

$$\sum_{t \in \mathcal{T}} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) \leq 2\sqrt{C\beta_T |\mathcal{T}| I(\mathbf{X}_{\mathcal{T}})} + \frac{\pi^2}{6} \leq 2\sqrt{C\beta_T |\mathcal{T}| \gamma_{|\mathcal{T}|}(\mathcal{X})} + \frac{\pi^2}{6}, \tag{270}$$

where $C = 2/\ln(1 + \sigma^{-2})$ and $\mathbf{X}_{\mathcal{T}} = (\boldsymbol{x}_t)_{t \in \mathcal{T}}$.

Proof. By following the proof strategy of GP-UCB, we have

$$\sum_{t \in \mathcal{T}} f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) \leq \sum_{t \in \mathcal{T}} \frac{1}{t^2} + \sum_{t \in \mathcal{T}} f([\boldsymbol{x}^*]_t) - f(\boldsymbol{x}_t) \leq 2\beta_T^{1/2} \sum_{t \in \mathcal{T}} \sigma(\boldsymbol{x}_t; \mathbf{X}_{t-1}) + \frac{\pi^2}{6} \tag{271}$$

due to event $\mathcal{A}$. Here, we define a new input sequence $(\widetilde{\boldsymbol{x}}_t)_{t \leq |\mathcal{T}|}$ as $\widetilde{\boldsymbol{x}}_t = \boldsymbol{x}_{j_t}$, where j_t is the t-th element in $\mathcal{T}$. Furthermore, we define $\widetilde{\mathbf{X}}_t = (\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_t)$. Then, from $\widetilde{\mathbf{X}}_t \subset \mathbf{X}_{j_t}$ and the monotonicity of the posterior variance against the input data, we have

$$\begin{aligned}
\sum_{t \in \mathcal{T}} \sigma(\boldsymbol{x}_t; \mathbf{X}_{t-1}) &= \sum_{t=1}^{|\mathcal{T}|} \sigma(\widetilde{\boldsymbol{x}}_t; \mathbf{X}_{j_t - 1}) && (272)\\
&\leq \sum_{t=1}^{|\mathcal{T}|} \sigma(\widetilde{\boldsymbol{x}}_t; \widetilde{\mathbf{X}}_{t-1}) && (273)\\
&\leq \sqrt{C|\mathcal{T}| I(\widetilde{\mathbf{X}}_{|\mathcal{T}|})} && (274)\\
&= \sqrt{C|\mathcal{T}| I(\mathbf{X}_{\mathcal{T}})} && (275)\\
&\leq \sqrt{C|\mathcal{T}| \gamma_{|\mathcal{T}|}(\mathcal{X})}, && (276)
\end{aligned}$$

where the second inequality follows from Theorems 5.3 and 5.4 in [51]. □

D Numerical Simulation for Information Gain

In addition to the simple example provided in Figure 1, we empirically confirm the gap between the worst-case (MVR) and GP-UCB's information gain under the Bayesian assumption. In Figures 2 and 3, we report the average and the quantile of realized information gain with GP-UCB, over 20 different sample paths, generated by changing the random seed, respectively. We conduct experiments under the same settings as in Figure 1 of the main text. We also report the information gain corresponding to the sequence of maximum variance reduction (MVR), following the same setup as Figure 1. In all cases, consistent with Figure 1 in the main text, we observe a noticeable gap in information gain between GP-UCB and MVR.

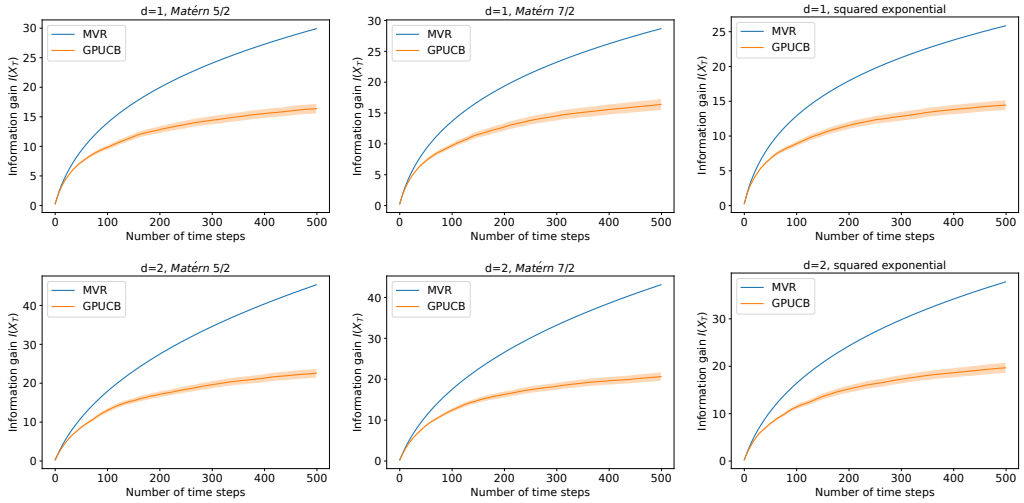

Figure 2: Average results of information gain under 20 different sample paths. The top row shows the results for the input space $\mathcal{X} = [0, 1]$ with lengthscale parameter $\ell = 0.1$. The bottom row corresponds to the input space $\mathcal{X} = [0, 1]^2$ with lengthscale $\ell = 0.25$, and from left to right, we use the SE kernel, the Matérn 5/2 kernel, and the Matérn 7/2 kernel. The shaded regions indicate one standard error.

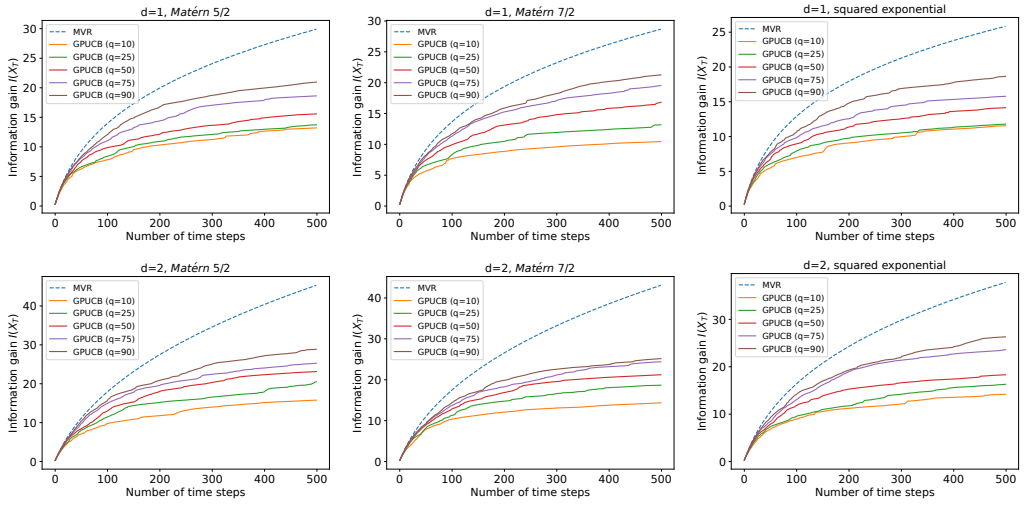

Figure 3: Quantiles of information gain over 20 different sample paths. We report the 10%, 25%, 50%, 75%, and 90% quantiles of the information gain of GP-UCB.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: "Contribution" paragraph in Section 1 reflects the paper's contributions.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: In Section 4, we discuss the limitations of our results and possible future directions.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: We provide complete assumptions in Section 2. The complete proofs of theoretical results in Section 3 are provided in the appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Detailed experimental settings are provided in Appendix D.

   Guidelines:

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

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

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

Answer: [No]

Justification: The main contribution of this paper is on the theoretical side, and open-access code is not essential.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The detailed experimental settings in Appendix D are provided.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The experiments in Appendix D provide one standard error bar.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [No]

   Justification: The experiments in Section D are minimal, and the computing resources are not significant.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We reviewed the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [NA]

    Justification: This paper only focuses on the theoretical aspect, does not propose a new algorithm, and does not relate to specific applications.

    Guidelines:

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

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

11. **Safeguards**

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

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [NA]

    Justification: This paper does not use existing assets.

    Guidelines:

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

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

13. **New assets**

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

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.