Noname manuscript No.
(will be inserted by the editor)

Lower Bounds on the Noiseless Worst-Case Complexity of Efficient Global Optimization

Wenjie Xu · Yuning Jiang*[*] ·
Emilio T. Maddalena · Colin N. Jones

Received: date / Accepted: date

Abstract Efficient global optimization is a widely used method for optimizing expensive black-box functions. In this paper, we study the worst-case oracle complexity of the efficient global optimization problem. In contrast to existing kernel-specific results, we derive a unified lower bound for the oracle complexity of efficient global optimization in terms of the metric entropy of a ball in its corresponding reproducing kernel Hilbert space (RKHS). Moreover, we show that this lower bound nearly matches the upper bound attained by non-adaptive search algorithms, for the commonly used squared exponential kernel and the Matérn kernel with a large smoothness parameter ν. This matching is up to a replacement of $d/2$ by d and a logarithmic term $\log \frac{R}{\epsilon}$, where d is the dimension of input space, R is the upper bound for the norm of the unknown black-box function, and ϵ is the desired accuracy. That is to say, our lower bound is nearly optimal for these kernels.

Keywords Efficient Global Optimization · Worst-Case Complexity · Reproducing Kernel Hilbert Space

Communicated by Paul I. Barton.

1 Introduction

Black-box optimization by sequentially evaluating different candidate solutions without access to gradient information is a pervasive problem. For ex-

[*] Corresponding author.
E-mail: yuning.jiang@ieee.org

W. Xu, Y. Jiang, E. Maddalena, C. Jones
Automatic Control Laboratory
EPFL, Switzerland

W. Xu is also with Swiss Federal Laboratories
for Materials Science and Technology (EMPA).

ample, tuning the hyperparameters of machine learning models [3, 30], optimizing control system performance [2, 40] and discovering drugs or designing materials [10, 21], etc., can all be formulated as a black-box optimization problem without explicit gradient information. Therefore, efficient global optimization [13,29], as a sample-efficient method to solve the expensive black-box optimization problem without explicit gradient information, has recently been receiving much attention. Efficient global optimization is based on the idea of constructing a surrogate function using Gaussian process regression or kernel ridge regression to guide the search of optimal solution [13].

In many applications, e.g., tuning the hyperparameters of a deep neural network (where the objective function in discrete variables, such as number of layers, can be regarded as a restriction of continuous function), each sample can take significant resources such as time and computation. For such problems, understanding the sample complexity of efficient global optimization is of great theoretical interest and practical relevance.

There is a large body of literature on the convergence rates of particular efficient global optimization algorithms [7, 26, 31, 33, 34, 37]. Two typical analysis set-ups are the Bayesian and non-Bayesian settings[1]. In the Bayesian setting, the black-box function is assumed to be sampled from a Gaussian process, whereas in the non-Bayesian setting, the black-box function is assumed to be regular in the sense of having a bounded norm in the corresponding reproducing kernel Hilbert space.

As a complement to convergence analysis of different algorithms, complexity analysis tries to understand the inherent hardness of a problem. Specifically, we are interested in answering the question: *for a class of optimization problems, how many queries to an oracle, which returns some information about the function, are necessary to guarantee the identification of a solution with objective value at most ϵ worse than the optimal value* [22]? Without a complexity analysis, we cannot tell whether existing algorithms can be improved further in terms of convergence rate. This problem is well studied for convex optimization (e.g., in [22]), but less well understood for efficient global optimization.

Intuitively, the complexity of efficient global optimization largely depends on the richness or complexity of the functions inside the corresponding reproducing kernel Hilbert space (RKHS). Indeed, selecting the proper RKHS or the kernel function k is an important research question in the literature [14, 15]. Intuitively, the choice of the kernel functions captures the prior knowledge on the black-box function to optimize. As an extreme example, if we know the ground-truth black-box function is linear, we can adopt the linear kernel. Then after a finite number of noiseless function evaluations, we can uniquely determine the ground-truth function and hence the optimal solution. However, agnostically selecting simple kernels may lead to a surrogate function that is not expressive enough. For example, when the black-box function is nonlinear, using an RKHS with a linear kernel can not learn the ground-truth function

[1] The Bayesian setting is typically referred to as Bayesian optimization.

well. For such a function, it is more reasonable to select a more expressive kernel such as squared exponential kernel. To measure the complexity of a set of functions, *metric entropy* [16] is widely used in learning theory. However, as far as we know, the explicit connection between a complexity measure such as metric entropy for a function set and the problem complexity of efficient global optimization has not been established.

This paper focuses on the complexity analysis of efficient global optimization with general kernel functions in the non-Bayesian and noiseless setting. Although the noisy setting is more realistic from the practical point of view, it is also critical to consider the noiseless setting from the complexity-theoretic point of view. The rationale is that the noise may introduce additional statistical complexity to the problem and corrupts the inherent complexity analysis of the efficient global optimization. In addition, noiseless setting is not a simple extension of the noisy setting. Existing analysis under noisy setting (e.g., [5, 25, 27, 28]) typically relies on strictly positive noise variance. Simply setting noise variance to zero makes the analysis and results diminish. For example, the noisy bound for Squared Exponential (SE) kernel in [28] is $\Omega(\frac{\sigma^2}{\epsilon^2}\left(\log\frac{R}{\epsilon})^{\frac{d}{2}}\right)$, which is dominated by $\frac{\sigma^2}{\epsilon^2}$, where σ^2 is the noise variance, ϵ is the desired accuracy,[2] and R is the function norm upper bound. Simply setting $\sigma = 0$ gives a meaningless $\Omega(0)$ bound. Without the analysis under noiseless setting, it is unclear whether this dominant $\frac{\sigma^2}{\epsilon^2}$ term is due to noise or the inherent complexity of the RKHS.

Works	**Noise**	**SE kernel**	**Matérn kernel**	**General kernel**
[4]	No	N/A	$\Omega\left((\frac{1}{\epsilon})^{\frac{d}{\nu}}\right)$	N/A
[28]	Yes	$\Omega\left(\frac{\sigma^2}{\epsilon^2}\left(\log\frac{R}{\epsilon}\right)^{d/2}\right)$	$\Omega\left(\frac{\sigma^2}{\epsilon^2}\left(\frac{R}{\epsilon}\right)^{d/\nu}\right)$	N/A
Ours	No	$\Omega\left(\left(\log\frac{R}{\epsilon}\right)^{d/2-1}\right)$	$\Omega\left(\frac{\left(\frac{R}{\epsilon}\right)^{\frac{d}{\nu+d/2}}}{\log\frac{R}{\epsilon}}\right)$	$\Omega\left(\frac{\log\mathcal{N}(S(\mathcal{X}),4\epsilon,\|\cdot\|_\infty)}{\log\left(\frac{R}{\epsilon}\right)}\right)$

Table 1 A summary of the state-of-the-art complexity result for efficient global optimization. σ^2 is the noise variance. R is the function norm upper bound. d is the dimension of input space. ν is the smoothness parameter of Matérn kernel. N/A means 'not applicable'. $S(\mathcal{X})$ is the ball in the corresponding reproducing kernel Hilbert space, with input set $\mathcal{X}$. $\mathcal{N}(\cdot,\cdot,\cdot)$ is the standard covering number to be formally defined in Sec. 4.

To highlight our originality and contribution, a comparison of our results with the state-of-the-art complexity analysis is given in Tab. 1. As far as we know, our work is the first to give a unified general lower bound in terms of metric entropy. Interestingly, we also notice that the commonly seen $\Theta(1/\epsilon^2)$ term in the noisy setting disappears in the noiseless setting, which matches our intuition that estimating a point with Gaussian noise typically takes $\Theta(1/\epsilon^2)$ sample complexity. Specifically, our contributions include:

– We introduce a new set of analysis techniques and derive a *general* unified lower bound for the deterministic oracle complexity of efficient global op-

[2] We will use ϵ to denote the desired accuracy throughout the paper.

timization in terms of the *metric entropy* of the function space ball in the corresponding reproducing kernel Hilbert space, providing a unified and intuitive understanding for the complexity of efficient global optimization.
- Our *general* lower bound allows us to leverage existing estimates of the covering number of the function space ball in the RKHS to derive kernel-specific lower bounds for the commonly used squared exponential kernel and Matérn kernel with a large smoothness parameter ν, without the commonly seen $1/\epsilon^2$ term for the noisy setting interestingly. Furthermore, the lower bound for squared exponential kernel under noiseless setting is derived for the first time, to the best of our knowledge.
- We further show that these kernel-specific lower bounds nearly match the upper bounds attained by some non-adaptive search algorithms, where the upper bound for the squared exponential kernel is newly derived in this paper. Hence, our general lower bound is close to optimal for these specific kernels.

2 Related work

There has been a large body of literature on analyzing the complexity and the convergence properties of efficient global optimization. We first summarize the relevant literature area by area. We then highlight the position and the original contribution of our paper.

Algorithm-dependent Convergence Analysis. One line of research analyzes the property of particular types of algorithms. For example, some papers [9, 17] analyze the consistency of efficient global optimization algorithms. [34, 37] analyze the convergence property of the expected improvement algorithm. [33] proposes a maximum variance reduction algorithm that achieves optimal order simple regret for particular kernel functions. Under the assumption of Hölder continuity of the covariance function, lower and upper bounds are derived for the Bayesian setting in [12]. Among this set of literature, the works on information-theoretic upper bounds are more relevant to our metric-entropy lower bound. [31] derives an information-theoretic upper bound for the cumulative regret of the upper confidence bound algorithm. [26] gives an information-theoretic analysis of Thompson sampling. However, there is no existing work that provides a complementary information-theoretic lower bound.

Kernel-specific Lower Bound Analysis. As for lower bounds or complexity analysis, [4] derives a lower bound of simple regret for Matérn kernel in a noise-free setting. [28] provides lower bounds of both simple regret and cumulative regret for the squared exponential and Matérn kernels. With the Matérn kernel, a tight regret bound has been provided for Bayesian optimization in one dimension in [27]. With heavy-tailed noise in the non-Bayesian setting, a cumulative regret lower bound has been provided for the Matérn and squared exponential kernels in [25]. More recently, [5] provides lower bounds for both standard and robust Gaussian process bandit optimization.

However, unlike the information-theoretic upper bound shown in [31], the existing lower bound results are mostly (if not all) restricted to specific kernel functions (mostly squared exponential and Matérn). The explicit connection between the optimization lower bound and the complexity of the RKHS has not been established so far in the existing literature. In this paper, we establish such a connection by constructing a lower bound in terms of *metric entropy*.

Covering Number Estimate in RKHS. Another area of research relevant to this paper is the estimate of covering number or metric entropy in function spaces. Some of the classical results are used in this paper. In [8, Sec. 3.3], the covering number for the function space ball in a Besov space is estimated. A technique to derive a lower estimate of the covering number for a stationary kernel is developed in [42], and as an application, a lower bound of a function space ball's covering number for the squared exponential kernel is derived.

General Information-based Complexity Analysis. Our focus is efficient global optimization in this paper, due to its increasing popularity and lack of a unified and intuitive understanding for its complexity. Nevertheless, there have also been many classical works in the general area of information-based complexity analysis. For example, it is shown that the optimal convergence rates of global optimization are equivalent to those of approximation in the sup-norm [23]. However, approximation in the sup-norm itself is another hard problem with its complexity to be understood. There is also another set of results that try to connect the finite rank approximation, which is more general than sample based interpolation, with metric entropy [8, 18, 32]. However, they can not be directly applied to our efficient global optimization problem, due to the general finite rank approximation definitions that are inconsistent with our sample based efficient global optimization setting.

Minimax Rates for Kernel Regression. In learning theory, there are well-established results on covering number bound of learning errors. Many existing works [6, 24] derive covering number bounds for the generalization error of learning problems with RKHS or more general hypothetical spaces. However, in a typical learning setting, the sample points and corresponding observations are assumed to be identically and independently distributed, with observations corrupted by noise. To the contrast, the setting we consider in this paper is an essentially different global optimization problem. Specifically, our goal is to identify a solution with desired level of optimality and the sample point can be adaptively selected.

Position and Originality of Our Work. Despite the rich literature summarized above, we notice two major limitations of the state-of-the-art complexity bounds. Firstly, existing analysis (see, e.g., in [4,5]) is typically restricted to a specific group of kernels (most commonly, the Squared Exponential kernel and the Matérn kernel). A unified understanding of the optimization complexity is lacking. Our work addresses this limitation by providing a unified general lower bound in terms of metric entropy, which recovers (close-to) state-of-the-art lower bounds when restricted to specific kernels. Secondly, the lower bounds with noise can be dominated by a $\Theta\left(\frac{1}{\epsilon^2}\right)$ term (e.g., in [28] for squared exponential kernel), which may corrupts the understanding for the complexity of efficient global optimization. Our work addresses this limitation by proving bounds in the noiseless regime.

3 Problem Statement

We consider efficient global optimization in a non-Bayesian setting [31]. Specifically, we optimize a deterministic function f from a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with input space $\mathbb{R}^d$, where d is the dimension. $\mathcal{H}$ is equipped with the reproducing kernel $k(\cdot,\cdot):\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$. Let $\mathcal{X}\subset\mathbb{R}^d$ be the known feasible set (e.g., a hyperbox) of the optimization problem. In the following, we will use $[n]$ to denote the set $\{1,2,\cdots,n\}$. We assume that

Assumption 3.1 *$\mathcal{X}$ is compact and nonempty.*

Assumption 3.1 is reasonable because in many applications (e.g., continuous hyperparameter tuning) of efficient global optimization, we are able to restrict the optimization into certain ranges based on domain knowledge. Regarding the black-box function $f\in\mathcal{H}$ that we aim to optimize, we assume that,

Assumption 3.2 *$\|f\|_{\mathcal{H}}\leq R$, where R is a positive real number and $\|\cdot\|_{\mathcal{H}}$ is the norm induced by the inner product associated with $\mathcal{H}$.*

Assumption 3.2 requires that the function to be optimized is regular in the sense that it has bounded norm in the RKHS, which is a common assumption (e.g., [4,28]) for complexity and convergence analysis.

Assumption 3.3 *$k(x_1,x_2)\leq 1,\forall x_1,x_2\in\mathcal{X}$ and $k(x_1,x_2)$ is continuous on $\mathbb{R}^d\times\mathbb{R}^d$.*

Assumption 3.3 is a common assumption for analyzing the convergence and complexity of efficient global optimization. It holds for a large class of commonly used kernel functions (e.g., Matérn kernel and squared exponential kernel) after normalization.

Our problem[3] is formulated as

$$\min_{x\in\mathcal{X}} \quad f(x). \tag{1}$$

We know that

$$f(x_1)-f(x_2)=\langle f,k(x_1,\cdot)-k(x_2,\cdot)\rangle\leq\|f\|_{\mathcal{H}}\left\|k(x_1,\cdot)-k(x_2,\cdot)\right\|_{\mathcal{H}}.$$

Hence, it can be shown under Assumptions 3.2 and 3.3, that f is continuous and thus (1) has an optimal solution on the compact set $\mathcal{X}$. As in standard efficient global optimization, we restrict ourselves to the zero-order oracle case. That is, our algorithm can only query the function value $f(x)$ but not higher-order information at a point x in each step. Based on the function evaluations

[3] In the Gaussian process bandit literature, the maximizition formulation is usually adopted, while in the global optimization literature, the minimization formulation is usually adopted. Here, we adopt the latter.

before the current step, the algorithm sequentially decides the next point to sample. In this paper, we only consider oracle query (namely, function evaluation) complexity without considering the complexity of solving auxiliary optimization problems in typical efficient global optimization algorithms (e.g., maximizing the expected improvement).

In this paper, we focus on the performance metric of *simple regret* $r_{(t)}$.

Definition 3.1 (Simple regret) After t function evaluations, simple regret $r_{(t)} \coloneqq \min_{\tau\in[t]} f(x_\tau) - \min_{x\in\mathcal{X}} f(x)$, where $[t] \coloneqq \{1, 2, \cdots, t\}$.

Note that in some of the literature, simple regret is also defined as $f(\hat{x}_t) - \min_{x\in\mathcal{X}} f(x)$, where $\hat{x}_t$ is one additional point reported after t steps. Since we can always pay one more function evaluation for the reported point, this definition difference will not impact our convergence or complexity analysis.

4 Preliminary

To analyze the problem complexity of efficient global optimization, we need a metric to measure the complexity of the RKHS. As an extreme example, if we choose a linear kernel, the underlying function to be optimized is a linear function. Hence, we can reconstruct it after a finite number of steps and compute the optimum without any error. The covering number is such a widely used metric to measure the complexity of an RKHS [41]. To facilitate our discussion, we introduce some concepts about the complexity of function sets.

Given a normed vector space $(V, \|\cdot\|)$ and a subset $G \subset V$, for $\epsilon > 0$, we make the following complexity related definitions [39].

Definition 4.1 (ϵ-covering) $\{v_1, \cdots, v_N\}$ is an ϵ-covering of G if

$$G \subset \cup_{i\in[N]} B_{\|\cdot\|}(v_i, \epsilon),$$

where $B_{\|\cdot\|}(v_i, \epsilon)$ is the ball in V centered at v_i with radius ϵ with respect to the norm $\|\cdot\|$.

Definition 4.2 (ϵ-packing) $\{v_1, \cdots, v_N\} \subset G$ is an ϵ-packing of G if

$$\min_{i\neq j} \|v_i - v_j\| > \epsilon.$$

Definition 4.3 (Covering number) The covering number $\mathcal{N}(G, \epsilon, \|\cdot\|)$ is defined to be $\min\{n \mid \exists\epsilon\text{-covering } \{v_1, \cdots, v_n\} \text{ with cardinality } n\}$.

Definition 4.4 (Packing number) The packing number $\mathcal{M}(G, \epsilon, \|\cdot\|)$ is defined to be $\max\{n \mid \exists\epsilon\text{-packing } \{v_1, \cdots, v_n\} \text{ with cardinality } n\}$.

Definition 4.5 (Metric entropy) The metric entropy of $(G, \|\cdot\|)$ is defined to be $\log \mathcal{N}(G, \epsilon, \|\cdot\|)$, where $\mathcal{N}$ is the covering number.

It can be verified that,

Proposition 4.1 (Thm. IV, [16]) $\mathcal{N}(G,\epsilon,\|\cdot\|)\leq\mathcal{M}(G,\epsilon,\|\cdot\|)\leq\mathcal{N}(G,\frac{\epsilon}{2},\|\cdot\|)$.

To facilitate the subsequent complexity analysis, we use $x_1,x_2,\cdots,x_t$ to denote the sequence of evaluated points up to step t. We now formalize the concept of deterministic algorithm for solving the efficient global optimization problem.

Definition 4.6 (Deterministic algorithm) A deterministic algorithm $\mathcal{A}$ for solving the optimization problem in (1) is a sequence of mappings $(\pi_t)_{t=1}^{\infty}$, where $\pi_t:(\mathcal{X}\times\mathbb{R})^{t-1}\to\mathcal{X}, t\geq 2$ and $\pi_1:\{\emptyset\}\to\mathcal{X}$. When running the algorithm $\mathcal{A}$, the sample at step t is $x_t=\pi_t((x_\tau,f(x_\tau))_{\tau=1}^{t-1}), t\geq 2$ and $x_1=\pi_1(\emptyset)$.

Note that deterministic algorithms include most of the popular acquisition functions based efficient global optimization algorithms (e.g., lower/upper confidence bound [31] and expected improvement [13]).

We assume that the first sample point x_1 is deterministic, either given before running the algorithm or chosen by the algorithm. Now, if we suppose that f is such that the algorithm observes a sequence of 0's for every function evaluation $f(x_\tau)$, it will generate a deterministic sample trajectory. We will see in our main result that this trajectory can be used to construct adversarial functions to derive the lower bound. We formally define it below.

Definition 4.7 (Zero sequence) Given a deterministic algorithm $\mathcal{A}=(\pi_t)_{t=1}^{\infty}$. We set $x_1^0=\pi_1(\emptyset)$. Applying the recurrence relationship $x_t^0=\pi_t((x_\tau^0,0)_{\tau=1}^{t-1})$, we get a deterministic sequence $x_1^0,x_2^0,\cdots,x_t^0,\cdots$, which only depends on the algorithm $\mathcal{A}$. We call this sequence the zero sequence of the algorithm $\mathcal{A}$.

5 Main Results

Our strategy to derive the lower bound is decomposing the RKHS into two orthogonal subspaces with one of them expanding as more samples are obtained, as shown in Fig. 1. Then, we can project the function space ball into these two subspaces. We will show that as the number of sampled points grows, the covering number of the ball's projection into one subspace increases and the other decreases. We derive the lower bound on the number of optimization steps by bounding the increase/decrease rate. All the proofs of the lemmas and theorems are attached in the Appendix, except Lem. 5.4 and Thm. 5.1. Before proceeding, we introduce some notations.

Notations For $f\in\mathcal{H}$, $f|_{\mathcal{X}}:\mathcal{X}\to\mathbb{R}$ is defined as $f|_{\mathcal{X}}(x)=f(x),\forall x\in\mathcal{X}$. For $Q\subset\mathcal{H}$, we use $Q(\mathcal{X})$ to denote the set $\{f|_{\mathcal{X}}|f\in Q\}$, which is a subset of $C(\mathcal{X},\|\cdot\|_\infty)$, the continuous function space over $\mathcal{X}$. $Q(\mathcal{X})$ is considered as a subset of $C(\mathcal{X},\|\cdot\|_\infty)$ in $\mathcal{N}(Q(\mathcal{X}),\epsilon,\|\cdot\|_\infty)$ and $\mathcal{M}(Q(\mathcal{X}),\epsilon,\|\cdot\|_\infty)$.

We first decompose the RKHS into two orthogonal subspaces.

Definition 5.1 $\mathcal{H}_t^{\|}:=\{\sum_{i\in[t]}\alpha_i k(x_i,\cdot)|\alpha_i\in\mathbb{R}\}$, $\mathcal{H}_t^{\perp}:=\{f\in\mathcal{H}|f(x_i)=0,\forall i\in[t]\}$.

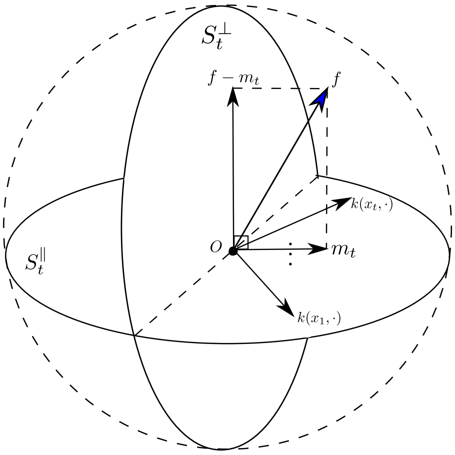

Fig. 1 The function space view of our proof strategy.

Notice that $\mathcal{H}_t^{||}$ expands when we have more and more function evaluation data. In parallel, $\mathcal{H}_t^{\perp}$ shrinks. We then consider the intersection of the function space ball S with $\mathcal{H}_t^{||}$ and $\mathcal{H}_t^{\perp}$.

Definition 5.2 $S \coloneqq \{f|f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq R\}, S_t^{||} \coloneqq \mathcal{H}_t^{||} \cap S, S_t^{\perp} \coloneqq \mathcal{H}_t^{\perp} \cap S.$

With these definitions, we can show that any function in S can be decomposed into two functions in $S_t^{||}$ and $S_t^{\perp}$, respectively.

Lemma 5.1 *$\forall f \in S$, there exists $m_t \in S_t^{||}$, such that $f - m_t \in S_t^{\perp}$.*

Remark 5.1 *When the matrix $K = (k(x_i, x_j))_{i,j\in[t]}$ is invertible, we can check that $m_t(x) = f_X^T K^{-1} K_{Xx}$, where $f_X = [f(x_1), f(x_2), \cdots, f(x_t)]^T$ and $K_{Xx} = [k(x_1, x), k(x_2, x), \cdots, k(x_t, x)]^T$, satisfies $m_t \in S_t^{||}$ and $f - m_t \in S_t^{\perp}$. The function $m_t(x)$ is exactly the posterior mean function in Gaussian process regression.*

Intuitively, we can add some function from $S_t^{\perp}$ to f without changing the historical evaluations at $x_1, \cdots, x_t$. If we have some way of lower bounding the complexity of $S_t^{\perp}$, we may be able to find a perturbing function from $S_t^{\perp}$ that leads to sub-optimality. We will try to lower bound the complexity of $S_t^{\perp}$ through Lem. 5.2 and Lem. 5.3.

Since $S_t^{||}$ and $S_t^{\perp}$ are orthogonal to each other in the RKHS, it is intuitive that the complexity of S can be decomposed into the complexity of $S_t^{\perp}$ and $S_t^{||}$. Formally, we have Lem. 5.2.

Lemma 5.2 *For any $\epsilon_t^{||} > 0, \epsilon_t^{\perp} > 0$, we have*

$$\mathcal{M}(S_t^{\perp}(\mathcal{X}), \epsilon_t^{\perp}, \|\cdot\|_\infty) \geq \frac{\mathcal{N}(S(\mathcal{X}), \epsilon_t, \|\cdot\|_\infty)}{\mathcal{N}(S_t^{||}(\mathcal{X}), \epsilon_t^{||}, \|\cdot\|_\infty)},$$

where $\epsilon_t = \epsilon_t^{||} + \epsilon_t^{\perp}$.

Lem. 5.2 is proved based on Lem. 5.1. With Lem. 5.2, we can lower bound $\mathcal{M}(S_t^{\perp}(\mathcal{X}), \epsilon_t^{\perp}, \|\cdot\|_\infty)$ if we are able to upper bound $\mathcal{N}(S_t^{||}(\mathcal{X}), \epsilon_t^{||}, \|\cdot\|_\infty)$.

Since $S_t^{||}$ is inside a finite dimensional space $\mathcal{H}_t^{||}$, we can show that,

Lemma 5.3 *If* $0 < \epsilon < \frac{R}{4}$, *we have* $\log \mathcal{N}(S_t^{||}(\mathcal{X}), \epsilon, \|\cdot\|_\infty) \leq 2t \log\left(\frac{R}{\epsilon}\right)$.

We then give the following key lemma.

Lemma 5.4 *For* $0 < \epsilon < \epsilon_0$, *if* $t \leq \frac{\log \mathcal{N}(S(\mathcal{X}), 4\epsilon, \|\cdot\|_\infty)}{4 \log\left(\frac{R}{\epsilon}\right)}$, *then for any sample sequence* $x_1, \cdots, x_t$, *we have,*

$$\frac{\mathcal{N}(S(\mathcal{X}), 4\epsilon, \|\cdot\|_\infty)}{\mathcal{N}(S_t^{||}(\mathcal{X}), \epsilon, \|\cdot\|_\infty)} \geq 2,$$

where $\epsilon_0 = \sup\{\delta | \delta > 0, \log \mathcal{N}(S(\mathcal{X}), 4\delta, \|\cdot\|_\infty) > 2\log 2\}$.

Proof By assumption that $t \leq \frac{\log \mathcal{N}(S(\mathcal{X}), 4\epsilon, \|\cdot\|_\infty)}{4 \log\left(\frac{R}{\epsilon}\right)}$, we have

$$2t \log\left(\frac{R}{\epsilon}\right) \leq \frac{1}{2} \log \mathcal{N}(S(\mathcal{X}), 4\epsilon, \|\cdot\|_\infty).$$

By $\epsilon < \epsilon_0$ and the definition of ϵ_0, $\frac{1}{2}\log \mathcal{N}(S(\mathcal{X}), 4\epsilon, \|\cdot\|_\infty) - \log 2 > 0$. We also notice that $\log \mathcal{N}(S(\mathcal{X}), R, \|\cdot\|_\infty) = 0 < 2\log 2$ and thus, $\epsilon_0 \leq \frac{R}{4}$. We then can apply Lem. 5.3 to derive,

$$\begin{aligned} \log \mathcal{N}(S_t^{||}(\mathcal{X}), \epsilon, \|\cdot\|_\infty) \leq & \ 2t \log\left(\frac{R}{\epsilon}\right) \\ \leq \frac{1}{2} \log \mathcal{N}(S(\mathcal{X}), 4\epsilon, \|\cdot\|_\infty) + & \underbrace{\frac{1}{2} \log \mathcal{N}(S(\mathcal{X}), 4\epsilon, \|\cdot\|_\infty) - \log 2}_{\text{positive}} \\ = \ \log \mathcal{N}(S(\mathcal{X}), 4\epsilon, \|\cdot\|_\infty) - \log 2, & \end{aligned}$$

where the first inequality follows by Lem. 5.3 and the second by assumption on t. So $\frac{\mathcal{N}(S(\mathcal{X}), 4\epsilon, \|\cdot\|_\infty)}{\mathcal{N}(S_t^{||}(\mathcal{X}), \epsilon, \|\cdot\|_\infty)} \geq 2$. □

We are now ready to give our main result in Thm. 5.1.

Theorem 5.1 *If there exists a deterministic algorithm that achieves simple regret* $r_{(T)} \leq \epsilon$ *for any function* $f \in S$ *in* T *function evaluations for our problem* (1), *it is necessary that,*

$$T = \Omega\left(\frac{\log \mathcal{N}(S(\mathcal{X}), 4\epsilon, \|\cdot\|_\infty)}{\log(\frac{R}{\epsilon})}\right). \tag{2}$$

Before we prove Thm. 5.1, we give a sketch of the proof. For any deterministic algorithm and any number of optimization steps t, we consider the corresponding deterministic zero sequence $x_1^0, x_2^0, \cdots, x_t^0$ as defined in Def. 4.7. We try to construct an adversarial function inside the corresponding $S_t^\perp$ with 0 function value at the points $x_i^0, i \in [t]$ and low function values at some point that is not sampled. The possible minimal value of such an adversarial function links to the covering number of the set $S_t^\perp(\mathcal{X})$, which can be lower bounded by combining Lem. 5.2 and Lem. 5.3.

*Proof (**Proof of Thm. 5.1**)* Given an deterministic algorithm $\mathcal{A} = (\pi_t)_{t=1}^{+\infty}$, if it always gets the evaluations 0, then the sample trajectory satisfies,

$$x_t^0 = \pi_t\left((x_\tau^0, 0)_{\tau=1}^{t-1}\right), t \geq 2,$$

which is exactly the zero sequence of the algorithm. Note that the zero sequence x_t^0 only depends on the deterministic algorithm $\mathcal{A}$. Once we fix the algorithm, the zero sequence is fixed.

We want to check the feasibility of the problem (3),

$$\min_{s \in \mathcal{S}, x \in \mathcal{X}} \quad 1 \quad \text{s.t.} \quad \begin{cases} s\left(x_n^0\right) = 0, \ \forall n = 1, \ldots, t, \\ s(x) < -\epsilon. \end{cases} \tag{3}$$

Any feasible solution of (3) has some ‘adversarial’ property against the algorithm $\mathcal{A}$. In fact, suppose that $(\bar{s}, \bar{x})$ is a feasible solution for problem (3), when we run the algorithm $\mathcal{A}$ over $\bar{s}$, the sample sequence up to step t is exactly the zero sequence truncated at step t and $r_{(t)} = \min_{\tau \in [t]} \bar{s}(x_\tau^0) - \min_{x \in \mathcal{X}} \bar{s}(x) > \epsilon$. Now the question is under what condition, the problem (3) is feasible. Since we are analyzing the asymptotic rate, we restrict to the case $\epsilon < \epsilon_0$, where ϵ_0 is given in Lem. 5.4. By Lem. 5.4 and Lem. 5.2, if $t \leq \frac{\log \mathcal{N}(S(\mathcal{X}), 4\epsilon, \|\cdot\|_\infty)}{4 \log\left(\frac{R}{\epsilon}\right)}$, for the sample sequence $x_1^0, \cdots, x_t^0$ corresponding to any given algorithm, we have,

$$\mathcal{M}(S_t^\perp(\mathcal{X}), 3\epsilon, \|\cdot\|_\infty) \geq \frac{\mathcal{N}(S(\mathcal{X}), 4\epsilon, \|\cdot\|_\infty)}{\mathcal{N}(S_t^\parallel(\mathcal{X}), \epsilon, \|\cdot\|_\infty)} \geq 2.$$

Therefore, there exists functions $f_1, f_2 \in S_t^\perp$, such that, $\|f_1|_\mathcal{X} - f_2|_\mathcal{X}\|_\infty \geq 3\epsilon$. So $\|f_1|_\mathcal{X}\|_\infty + \|f_2|_\mathcal{X}\|_\infty \geq \|f_1|_\mathcal{X} - f_2|_\mathcal{X}\|_\infty \geq 3\epsilon$ and at least one of f_1 and f_2 has L_∞ norm over the set $\mathcal{X}$ at least $\frac{3\epsilon}{2}$. Without loss of generality, we assume $\|f_1|_\mathcal{X}\|_\infty \geq \frac{3\epsilon}{2}$. Since for $\forall g \in S_t^\perp$, $-g \in S_t^\perp$, there exists $\hat{f} \in S_t^\perp$ (either f_1 or $-f_1$), such that,

$$\inf_{x \in \mathcal{X}} \hat{f}(x) \leq -\frac{3\epsilon}{2}.$$

When applying the given algorithm to $\hat{f}$, if $t \leq \frac{\log \mathcal{N}(S(\mathcal{X}), 4\epsilon, \|\cdot\|_\infty)}{4 \log\left(\frac{R}{\epsilon}\right)}$, the suboptimality gap or the simple regret $r_{(t)}$ is at least $\frac{3}{2}\epsilon$. Therefore, to reduce the simple regret $r_{(T)} \leq \epsilon$ for all the functions in S within T steps, it is necessary that,

$$T = \Omega\left(\frac{\log \mathcal{N}(S(\mathcal{X}), 4\epsilon, \|\cdot\|_\infty)}{\log(\frac{R}{\epsilon})}\right). \qquad \square$$

To verify the effectiveness of Thm. 5.1, we apply it to a simple case in Ex. 5.1.

Example 5.1 *For the quadratic kernel $k(x, y) = (x^T y)^2$, the corresponding RKHS is finite dimensional and is given as [20],*

$$\mathcal{H} = \left\{ f_A(x) = x^T A x | A \in \mathcal{S}^{d\times d} \right\}, \tag{4}$$

where $\mathcal{S}^{d\times d}$ is the set of symmetric matrices of size $d \times d$. We know that,

$$\langle f_{A_1}, f_{A_2} \rangle_{\mathcal{H}} = \langle A_1, A_2 \rangle_{\mathrm{F}}, \tag{5}$$

where $\langle \cdot, \cdot \rangle_{\mathrm{F}}$ is the Frobenius inner product. Since $\mathcal{S}^{d\times d}$ can be embedded into $\mathbb{R}^{\frac{d\times(d+1)}{2}}$ and the metric entropy for compact set in Euclidean space is $\Theta\left(\log \frac{1}{\epsilon}\right)$ as discussed in [39], the lower bound in Thm. 5.1 reduces to a constant. By applying a grid search algorithm for the quadratic kernel, we can identify the ground truth function after a finite number of steps and determine the optimal solution without any error. Therefore, the lower bound is tight in ϵ for the quadratic kernel.

5.1 Comparison with upper bounds for commonly used kernels

Ex. 5.1 demonstrates the validity of Thm. 5.1 for simple quadratic kernel functions. In this section, we will derive kernel-specific lower bounds for the squared exponential kernel and the Matérn kernels by using Thm. 5.1 and existing estimates of the covering numbers for their RKHS's. We compare our lower bounds with derived/existing upper bounds and show that they nearly match.

5.1.1 Squared Exponential kernel

One widely used kernel in efficient global optimization is the squared exponential (SE) kernel given by

$$k(x, y) = \exp\left\{ -\frac{\|x - y\|^2}{\sigma^2} \right\}. \tag{6}$$

In this case, we restrict to $\mathcal{X} = [0, 1]^d$. By applying Thm. 5.1, we have,

Theorem 5.2 *With $\mathcal{X} = [0, 1]^d$ and using the squared exponential kernel, if there exists a deterministic algorithm that achieves simple regret $r_{(T)} \leq \epsilon$ for any function $f \in S$ in T function evaluations for our problem (1), it is necessary that,*

$$T = \Omega\left(\left(\log \frac{R}{\epsilon} \right)^{d/2-1} \right). \tag{7}$$

Furthermore, there exists a deterministic algorithm and T satisfying

$$T = \mathcal{O}\left(\left(\log \frac{R}{\epsilon}\right)^{d}\right)$$

such that the algorithm achieves $r_{(T)} \leq \epsilon$ in T function evaluations for any $f \in S$.

The upper bound part is obtained through sampling non-adaptively to reduce the posterior variance to a uniform low level in $\mathcal{X}$. In this theorem, we focus on the asymptotic analysis of efficient global optimization and hide the coefficients that may depend on the dimension. We notice that the upper bound and lower bound are both polynomial in $\log \frac{1}{\epsilon}$ and nearly match, up to a replacement of $d/2$ by d in the order and one additional logarithmic term $\log \frac{R}{\epsilon}$.

5.1.2 Matérn kernel

In this section, we consider the Matérn kernel,

$$k(x,y) = C_\nu(\|x-y\|) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{\|x-y\|}{\rho}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\frac{\|x-y\|}{\rho}\right), \quad (8)$$

where ρ and ν are positive parameters of the kernel function, Γ is the gamma function, and K_ν is the modified Bessel function of the second kind.

Theorem 5.3 *With $\mathcal{X} = [0,1]^d$ and the Matérn kernel, if there exists a deterministic algorithm that achieves simple regret $r_{(T)} \leq \epsilon$ for any function $f \in S$ in T function evaluations for our problem* (1)*, it is necessary that,*

$$T = \Omega\left(\left(\frac{R}{\epsilon}\right)^{\frac{d}{\nu+d/2}} \left(\log \frac{R}{\epsilon}\right)^{-1}\right). \quad (9)$$

Furthermore, there exists a deterministic algorithm and T satisfying,

$$T = \mathcal{O}\left(\left(\frac{R}{\epsilon}\right)^{\frac{d}{\nu}}\right), \quad (10)$$

such that the algorithm achieves $r_{(T)} \leq \epsilon$ in T function evaluations for any $f \in S$.

Remark 5.2 *The upper bound part of Thm. 5.3 is proved by Thm. 1 of [4]. We also notice that [4] provides a lower bound of the same order as the upper bound in Eq.* (10)*, which means that the upper bound order is also the optimal lower bound order.*

Remark 5.3 *When $\nu \geq \frac{1}{2}d$, our lower bound can further imply the lower bound of $\Omega\left(\left(\frac{R}{\epsilon}\right)^{\frac{d}{2\nu}} \left(\log \frac{R}{\epsilon}\right)^{-1}\right)$, which nearly matches the upper bound, up to* *a replacement of $d/2$ by d and a logarithmic term $\log \frac{R}{\epsilon}$. However, when $\frac{\nu}{d}$ is small, there is still a significant gap between the lower bound implied by our general lower bound and the optimal lower bound.*

Remark 5.4 *There are two possible reasons why the bound is not tight. One potential reason is that we apply a conservative lower estimate for the metric entropy corresponding to the Matérn kernel. The other is that our metric-entropy approach is limited in the regime of small smoothness parameter ν. Filling this gap is left as future work.*

6 Experiments

In this section, we will first give a demonstration of adversarial functions, on which two common algorithms, the lower confidence bound (LCB) [31] and the expected improvement (EI) [13], perform poorly and achieve the optimization lower bound. Both algorithms model the unknown black-box function as sampled from a Gaussian process. The idea of LCB algorithm is minimizing the lower confidence bound, which is defined to be posterior mean minus a coefficient times posterior standard deviation, to get the next sample point in each step. The EI algorithm maximizes the expected improvement with respect to the best observed value so far to get the next sample point. Then we run the two algorithms on a set of randomly sampled functions and compare the average performance and the adversarial performance in terms of simple regret. The algorithms are implemented based on GPy [11] and CasADi [1]. All the auxiliary optimization problems in the algorithms are solved using the solver IPOPT [35] with multiple different starting points. Our experiments take about 15 hours on a device with `AMD Ryzen Threadripper 3990X 64-Core Processor` and 251 GB RAM.

6.1 Demonstration of adversarial functions

In our proof of Thm. 5.1, we use a particular set of adversarial functions, which reveal value 0 to the algorithm and have low values somewhere else. In this section, we demonstrate such adversarial functions for two popular algorithms, expected improvement and lower confidence bound.

We use the Matérn kernel in one dimension with $\nu = \frac{5}{2}, \rho = 1, \sigma^2 = 1$. We set the compact set to $\mathcal{X} = [-10, 10]$ and assume that the RKHS norm upper bound is $R = 1$. We apply both lower confidence bound algorithm with the constant weight 1 for the posterior standard deviation and the expected improvement algorithm. We manually assign $x_1 = 0$ as the first sampled point and derive the adversarial function by solving Prob. (11).

$$\min_{x\in\mathcal{X}} \min_{s\in\mathcal{H}} \; s(x) \quad \text{s.t.} \begin{cases} s\left(x_n^0\right) = 0, \;\; \forall n = 1, \ldots, t, \\ \|s\|_{\mathcal{H}} \leq R \end{cases} \tag{11}$$

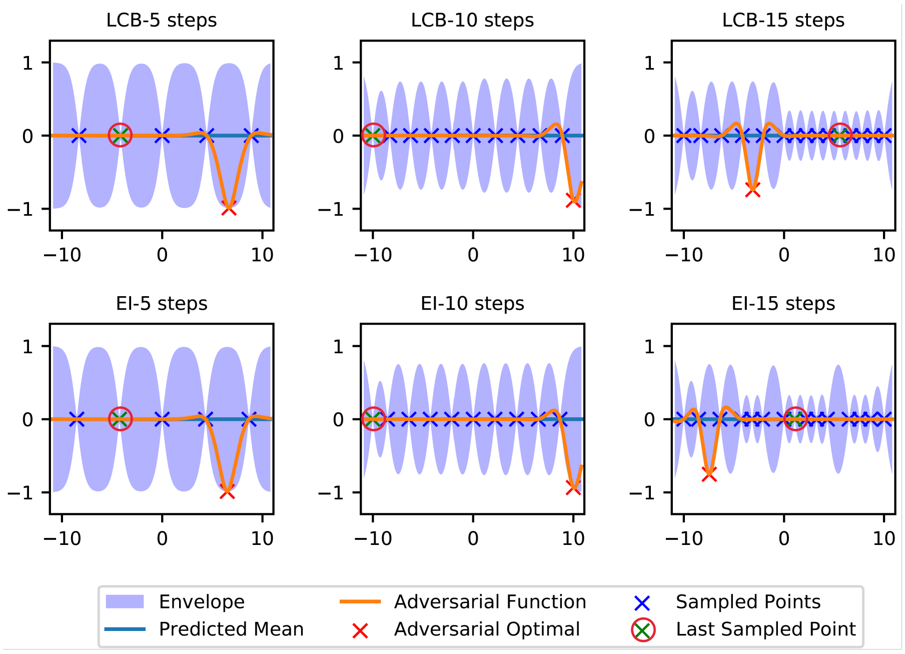

Fig. 2 Demonstrations of adversarial functions in dimension one.

Thanks to the optimal recovery property [38, Thm 13.2], the optimal value for the inner problem of (11) can be analytically derived as

$$-R\sqrt{k(x,x)-k(x,X)^T K^{-1}k(X,x)}.$$

Fig. 2 demonstrates the adversarial functions inside the corresponding RKHS with bounded norm of 1, which have value 0 at all the sampled points but have low global optimal value somewhere else. We notice that the envelope formed by the functions inside the ball with consistent evaluation data shrinks as more and more data becomes available. Intuitively, any algorithm needs to sample sufficiently densely globally in the adversarial case in order to find a close-to-optimal solution.

6.2 Average vs. adversarial performance

The proofs of Thm. 5.2 and Thm. 5.3 indicate that a non-adaptive sampling algorithm can achieve a close-to-optimal worst-case convergence rate. However, in practice, adaptive algorithms (e.g., lower confidence bound and expected improvement) are usually adopted and perform better. There could potentially be a gap between average-case convergence and worst-case convergence. To perform such a comparison, we randomly sample a set of functions from the RKHS to run the algorithms over. Specifically, we first uniformly sam-

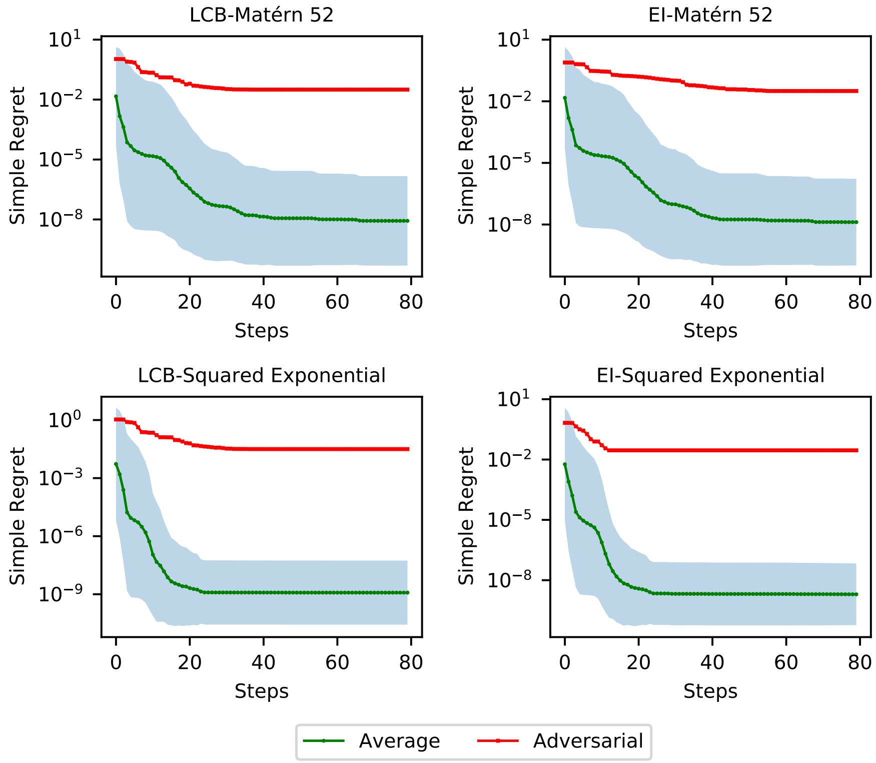

Fig. 3 Comparison of average performance (±standard deviation shown as shaded area, over 100 instances) and adversarial performance. Adversarial simple regret is defined as opposite of the optimal value of Prob. (11), namely the simple regret of the adversarial function at different optimization steps. Since the simple regret is defined as the best sampled function value minus the global optimal value (see the definition 3.1), this plot can also be seen as the convergence rate plot if the algorithm reports the best sampled point.

ple a finite set of knots $X \subset \mathcal{X}$ and then sample the function values f_X on the knots from the marginal distribution of the Gaussian process, which is a finite-dimensional Gaussian distribution. We then construct the minimal norm interpolant of the knots as the sampled function. To be consistent with the bounded norm assumption, we reject the functions with a norm value larger than R.

We use simple regret, which is defined to be $\min_{\tau \in [t]} f(x_\tau) - \min_{x \in \mathcal{X}} f(x)$, to measure the performance of different algorithms. We set $\mathcal{X} = [0,1]^3 \subset \mathbb{R}^3$ and set the length scales and variances of both the Matérn kernel function (with $\nu = \frac{5}{2}$) and the squared exponential kernel. Fig. 3 shows the comparison of average simple regret and adversarial simple regret. We observe that the average performance is much better than the performance on adversarial functions in terms of simple regret. Intuitively, adversarial functions are only a subset of *needle-in-haystack* functions, with most region flat and somewhere very small, when t becomes large. For those adversarial functions such as shown in Fig. 2, it can be difficult for the efficient global optimization algorithms to “see” the trend of the function. For common functions inside the function space ball, however, the algorithms are still able to detect the trend of the function value and find a near-to-optimal solution quickly.

7 Conclusions

In this paper, we provide a general lower bound on the worst-case suboptimality or simple regret for noiseless efficient global optimization in a non-Bayesian setting in terms of the metric entropy of the corresponding reproducing kernel Hilbert space (RKHS). We apply the general lower bounds to commonly used specific kernel functions, including the squared exponential kernel and the Matérn kernel. We further derive upper bounds and compare them to the lower bounds and find that they nearly match, except for the case for the Matérn kernel when $\frac{\nu}{d}$ is small. Two interesting future research directions are deriving an upper bound on the worst-case convergence rate in terms of metric entropy and characterizing the average-case convergence rate. We also conjecture that introducing randomness into the existing algorithms can improve the worst-case performance. An expected analysis challenge is that our current approach is sensitive to randomness. We also leave the extension of our analysis to the noisy case as future work.

Acknowledgements This work was supported by the Swiss National Science Foundation under the NCCR Automation project, grant agreement 51NF40_180545 and the RISK project (Risk Aware Data-Driven Demand Response, grant number 200021_175627).

Appendix

A Proof of Lemma 5.1

Consider the optimization problem below,

$$\min_{s\in\mathcal{H}} \quad \|s\|_{\mathcal{H}}^2 \quad \text{s.t.} \quad s\left(x_n\right)=f(x_n), \quad \forall n=1,\ldots,t. \tag{12}$$

Based on the representer theorem [36, Theorem 1.3.1], the optimal solution of (12) has the form $\sum_{i=1}^t \alpha_i k(x_i,\cdot)$. By using the constraint $s(x_n)=f(x_n)$, we can derive $K\alpha=f_X$, where $f_X=[f(x_1),f(x_2),\cdots,f(x_n)]^T$ and $K=(k(x_i,x_j))_{i\in[n],j\in[n]}$. With this restriction, we transform the problem in (12) to the problem in (13).

$$\min_{\alpha\in\mathbb{R}^t} \quad \alpha^T K\alpha \quad \text{s.t.} \quad K\alpha=f_X \tag{13}$$

We take α^* as the solution to the problem in (13), whose feasibility is guaranteed by representer theorem [36] and the non-emptiness of the feasible set (f is feasible for (12)). Therefore, $m_t(x)=(\alpha^*)^T K_{Xx}$ is the optimal solution to (12). Since f is a feasible solution for the problem (12), $\|m_t\|_{\mathcal{H}}\leq\|f\|_{\mathcal{H}}\leq R$. In addition, $(f-m_t)(x_i)=f(x_i)-m_t(x_i)=0,\forall i\in[t]$. And $\|f-m_t\|_{\mathcal{H}}^2=\|f\|_{\mathcal{H}}^2+\|m_t\|_{\mathcal{H}}^2-2\langle f,m_t\rangle=\|f\|_{\mathcal{H}}^2-\|m_t\|_{\mathcal{H}}^2\leq R^2$. So $m_t\in S_t^{\|}$ and $f-m_t\in S_t^{\perp}$.

B Proof of Lemma 5.2

Let $(p_1, p_2, \cdots, p_m)$ be an $\epsilon_t^{||}$-covering of $S_t^{||}(\mathcal{X})$ and $(q_1, q_2, \cdots, q_n)$ an $\epsilon_t^{\perp}$-covering of $S_t^{\perp}(\mathcal{X})$. Then $\forall f \in S$, by Lem. 5.1, $f = m_t + (f - m_t)$, where $m_t \in S_t^{||}$ and $f - m_t \in S_t^{\perp}$. By the definition of covering, $\exists p_i$, such that $\|m_t|_{\mathcal{X}} - p_i\|_\infty \leq \epsilon_t^{||}$ and $\exists q_j$, such that $\|(f - m_t)|_{\mathcal{X}} - q_j\|_\infty \leq \epsilon_t^{\perp}$. So

$$\|f|_{\mathcal{X}} - (p_i + q_j)\|_\infty \leq \|m_t|_{\mathcal{X}} - p_i\|_\infty + \|(f - m_t)|_{\mathcal{X}} - q_j\|_\infty \leq \epsilon_t^{||} + \epsilon_t^{\perp} = \epsilon_t.$$

So the set $\{p_i + q_j | i \in [m], j \in [n]\}$ is an ϵ_t-covering of $S(\mathcal{X})$ and we have the cardinality

$$|\{p_i + q_j | i \in [m], j \in [n]\}| = \mathcal{N}(S_t^{\perp}(\mathcal{X}), \epsilon_t^{\perp}, \|\cdot\|_\infty)\mathcal{N}(S_t^{||}(\mathcal{X}), \epsilon_t^{||}, \|\cdot\|_\infty) \geq \mathcal{N}(S(\mathcal{X}), \epsilon_t, \|\cdot\|_\infty).$$

So $\mathcal{M}(S_t^{\perp}(\mathcal{X}), \epsilon_t^{\perp}, \|\cdot\|_\infty) \geq \mathcal{N}(S_t^{\perp}(\mathcal{X}), \epsilon_t^{\perp}, \|\cdot\|_\infty) \geq \frac{\mathcal{N}(S(\mathcal{X}), \epsilon_t, \|\cdot\|_\infty)}{\mathcal{N}(S_t^{||}(\mathcal{X}), \epsilon_t^{||}, \|\cdot\|_\infty)}$.

C Proof of Lemma 5.3

We first introduce the set, $\mathcal{E}_t = \{\alpha \in \mathbb{R}^t | \alpha^T K_t \alpha \leq R^2\}$, where $K_t = (k(x_i, x_j))_{i,j \in [t]}$. Without loss of generality, we assume that K_t has full rank in the following analysis. Notice that if this condition does not hold, we only need to restrict to the subspace spanned by the eigenvectors of K_t with strictly positive eigenvalues and consider the intersection of $\mathcal{E}_t$ with the subspace. Since the restriction only reduces the essential dimension, the upper bound still holds. We introduce the norm $\|\alpha\|_{K_t} = \sqrt{\alpha^T K_t \alpha}$. We then have, $\forall f(x) = \alpha^T k(X, x) \in S_t^{||}(\mathcal{X}), g(x) = \beta^T k(X, x) \in S_t^{||}(\mathcal{X})$, we have

$$\|f - g\|_\infty = \sup_{x \in \mathcal{X}} |(\alpha - \beta)^T k(X, x)| = \sup_{x \in \mathcal{X}} |\langle \sum_{i \in [t]} (\alpha_i - \beta_i) k(x_i, \cdot), k(x, \cdot) \rangle| \tag{14}$$

$$\leq \sup_{x \in \mathcal{X}} \left\| \sum_{i \in [t]} (\alpha_i - \beta_i) k(x_i, \cdot) \right\|_{\mathcal{H}} \|k(x, \cdot)\|_{\mathcal{H}} \leq \sup_{x \in \mathcal{X}} \|\alpha - \beta\|_{K_t} \sqrt{k(x, x)} \leq \|\alpha - \beta\|_{K_t}.$$

Therefore, we have $\mathcal{N}(S_t^{||}(\mathcal{X}), \epsilon, \|\cdot\|_\infty) \leq \mathcal{N}(\mathcal{E}_t, \epsilon, \|\cdot\|_{K_t})$. We further have,

$$\mathcal{N}(\mathcal{E}_t, \epsilon, \|\cdot\|_{K_t}) \leq \mathcal{M}(\mathcal{E}_t, \epsilon, \|\cdot\|_{K_t}) \leq \frac{\mathrm{Vol}\left(B_{\|\cdot\|_{K_t}}\left(0, R + \frac{\epsilon}{2}\right)\right)}{\mathrm{Vol}\left(B_{\|\cdot\|_{K_t}}\left(0, \frac{\epsilon}{2}\right)\right)} \tag{15a}$$

$$= \left(\frac{2R}{\epsilon} + 1\right)^t \leq \left(\frac{R^2}{2\epsilon^2} + \frac{R^2}{2\epsilon^2}\right)^t = \left(\frac{R}{\epsilon}\right)^{2t}. \tag{15b}$$

The second inequality in (15) follows by that if $\alpha_1, \alpha_2, \cdots, \alpha_M$ is an ϵ-packing of the set $\mathcal{E}_t$, then $\cup_{i \in [M]} B_{\|\cdot\|_{K_t}}\left(\alpha_i, \frac{\epsilon}{2}\right) \subset B_{\|\cdot\|_{K_t}}\left(0, R + \frac{\epsilon}{2}\right)$ and $B_{\|\cdot\|_{K_t}}\left(\alpha_i, \frac{\epsilon}{2}\right) \cap B_{\|\cdot\|_{K_t}}\left(\alpha_j, \frac{\epsilon}{2}\right) = \emptyset, \forall i \neq j$ by the definition of packing. The third inequality in (15) follows by the assumption of $0 < \epsilon < \frac{R}{4}$. So $\log \mathcal{N}(\mathcal{E}_t, \epsilon, \|\cdot\|_{K_t}) \leq 2t \log\left(\frac{R}{\epsilon}\right)$. Therefore, $\log \mathcal{N}(S_t^{||}(\mathcal{X}), \epsilon, \|\cdot\|_\infty) \leq 2t \log\left(\frac{R}{\epsilon}\right)$.

D Proof of Theorem 5.2

By [42, Example 1], the covering number satisfies,

$$\log \mathcal{N}(S(\mathcal{X}), 4\epsilon, \|\cdot\|_\infty) = \Omega\left(\log\left(\frac{R}{\epsilon}\right)^{\frac{d}{2}}\right). \tag{16}$$

Therefore, Thm. 5.1 implies that,

$$T = \Omega\left(\left(\log\frac{R}{\epsilon}\right)^{\frac{d}{2}-1}\right). \tag{17}$$

We now focus on proving the upper bound part. To facilitate the following proof, we define,

$$\underline{f}_t(x) = m_t(x) - \sigma_t(x)\sqrt{R^2 - f_X^T K^{-1} f_X}, \tag{18}$$

$$\bar{f}_t(x) = m_t(x) + \sigma_t(x)\sqrt{R^2 - f_X^T K^{-1} f_X}, \tag{19}$$

where $m_t(x) = f_X^T K^{-1} K_{Xx}$ and $\sigma_t(x) = \sqrt{k(x,x) - K_{xX} K^{-1} K_{Xx}}$. Note that with squared exponential kernel and the sampled points set X to be used in this proof, the invertibility of the matrix K is guaranteed. As implied by [19, Prop. 1],

$$f(x) \in [\underline{f}_t(x), \bar{f}_t(x)]. \tag{20}$$

We consider the algorithm that evaluates the grid points

$$X = \left\{\left(\frac{k_1}{N}, \frac{k_2}{N}, \cdots, \frac{k_d}{N}\right)\middle| k_i \in \{0, 1, \cdots, N-1\}\right\}$$

without adaptation, and evaluate the point $\tilde{x}_t$ before termination after $t = N^d$ function evaluations on the grid points, where $\tilde{x}_t$ is given as,

$$\tilde{x}_t = \arg\min_{x \in \mathcal{X}} \underline{f}_t(x). \tag{21}$$

Let x^* denote the ground truth optimal solution. We can bound the suboptimality,

$$\begin{aligned}
f(\tilde{x}_t) - \min_{x\in\mathcal{X}} f(x) &\leq \bar{f}_t(\tilde{x}_t) - f(x^*) && \text{(22a)}\\
&\leq \bar{f}_t(\tilde{x}_t) - \underline{f}_t(x^*) && \text{(22b)}\\
&\leq \bar{f}_t(\tilde{x}_t) - \underline{f}_t(\tilde{x}_t) && \text{(22c)}\\
&= 2\sigma_t(\tilde{x}_t)\sqrt{R^2 - f_X^T K^{-1} f_X} && \text{(22d)}\\
&\leq 2R\sigma_t(\tilde{x}_t), && \text{(22e)}
\end{aligned}$$

where the inequalities in (22a) and (22b) follow by (20) and the inequality (22c) follows by the definition of $\tilde{x}_t$ in (21). We now try to upper bound $\sigma_t(\tilde{x}_t)$. We first introduce a set of Lagrangian interpolation functions,

$$w_{\alpha,N}(x) = \prod_{i\in[N]} \prod_{j\in[N], j\neq\alpha_i} \frac{x_i - j/N}{\alpha_i/N - j/N}, x \in [0,1]^d, \alpha \in [N]^d.$$

Let $w_N(x) = \big(w_{\alpha,N}(x)\big)_{\alpha\in[N]^d}$ and $\beta_N(x) = K^{-1}K_{Xx} \in \mathbb{R}^{N^d}$, we have

$$(\sigma_t(x))^2 = k(x,x) - K_{xX}K^{-1}K_{Xx} \tag{23a}$$

$$= k(x,x) - 2K_{xX}\beta_N(x) + \beta_N(x)^T K \beta_N(x) \tag{23b}$$

$$= \min_{y\in\mathbb{R}^{N^d}} (k(x,x) - 2K_{xX}y + y^T K y) \tag{23c}$$

$$\leq k(x,x) - 2K_{xX}w_N(x) + w_N(x)^T K w_N(x). \tag{23d}$$

Let $k_0(x)$ denote the function $k(0,x)$ and $\hat{k}_0$ its corresponding Fourier transformation. By inverse Fourier transformation, we have,

$$k(x,x) - 2K_{xX}w_N(x) + w_N(x)^T K w_N(x) \tag{24a}$$

$$= (2\pi)^{-d}\int_{\mathbb{R}^d} \hat{k}_0(\xi)\left(1 - 2\sum_{\alpha\in[N]^d} w_{\alpha,N}(x)e^{i\xi\cdot\left(x-\frac{\alpha}{N}\right)}\right. \tag{24b}$$

$$\left. + \sum_{\alpha\in[N]^d,\beta\in[N]^d} w_{\alpha,N}(x)e^{i\xi\cdot\left(\frac{\beta}{N}-\frac{\alpha}{N}\right)}w_{\beta,N}(x)\right)\mathrm{d}\xi \tag{24c}$$

$$= (2\pi)^{-d}\int_{\mathbb{R}^d} \hat{k}_0(\xi)\left|1 - \sum_{\alpha\in[N]^d} w_{\alpha,N}(x)e^{i\xi\cdot\left(x-\frac{\alpha}{N}\right)}\right|^2 \mathrm{d}\xi \tag{24d}$$

$$= (2\pi)^{-d}\int_{\mathbb{R}^d} \hat{k}_0(\xi)\left|e^{-i\frac{\xi}{N}\cdot Nx} - \sum_{\alpha\in[N]^d} w_{\alpha,N}(x)e^{-i\frac{\xi}{N}\cdot\alpha}\right|^2 \mathrm{d}\xi \tag{24e}$$

$$= (2\pi)^{-d}\int_{\mathbb{R}^d} \hat{k}_0(\xi)\left|e^{-i\frac{\xi}{N}\cdot Nx} - \sum_{\alpha\in[N]^d} w_{\alpha,N}(x)e^{-i\frac{\xi}{N}\cdot\alpha}\right|^2 \mathrm{d}\xi \tag{24f}$$

$$= (2\pi)^{-d}\int_{\xi\in[-\frac{N}{2},\frac{N}{2}]^d} \hat{k}_0(\xi)\left|e^{-i\frac{\xi}{N}\cdot Nx} - \sum_{\alpha\in[N]^d} w_{\alpha,N}(x)e^{-i\frac{\xi}{N}\cdot\alpha}\right|^2 \mathrm{d}\xi \tag{24g}$$

$$+ (2\pi)^{-d}\int_{\xi\notin[-\frac{N}{2},\frac{N}{2}]^d} \hat{k}_0(\xi)\left|e^{-i\frac{\xi}{N}\cdot Nx} - \sum_{\alpha\in[N]^d} w_{\alpha,N}(x)e^{-i\frac{\xi}{N}\cdot\alpha}\right|^2 \mathrm{d}\xi. \tag{24h}$$

To proceed, we need to use the Lem. D.1 [41].

Lemma D.1 (Lemma 4.1, [41]) *Let $x \in [0,1]^d$ and $N \in \mathbb{N}$. Then $\sum_{\alpha\in[N]^d}\left|w_{\alpha,N}(x)\right| \leq \left(N2^N\right)^d$ and for $\theta \in \left[-\frac{1}{2},\frac{1}{2}\right]^d$, there holds*

$$\left|e^{-i\theta\cdot Nx} - \sum_{\alpha\in[N]^d} w_{\alpha,N}(x)e^{-i\theta\cdot\alpha}\right| \leq d\left(1+\frac{1}{2^N}\right)^{d-1}\left(\max_{1\leq j\leq d}|\theta_j|\right)^N.$$

We apply the bounds in Lem. $D.1$ to Eq. (25) and have,

$$k(x,x) - 2K_{xX}w_N(x) + w_N(x)^T K w_N(x) \tag{25a}$$

$$= (2\pi)^{-d}\int_{\xi\in[-\frac{N}{2},\frac{N}{2}]^d} \hat{k}_0(\xi)\left|e^{-i\frac{\xi}{N}\cdot Nx} - \sum_{\alpha\in[N]^d} w_{\alpha,N}(x)e^{-i\frac{\xi}{N}\cdot\alpha}\right|^2 \mathrm{d}\xi \tag{25b}$$

$$+ (2\pi)^{-d}\int_{\xi\notin[-\frac{N}{2},\frac{N}{2}]^d} \hat{k}_0(\xi)\left|e^{-i\frac{\xi}{N}\cdot Nx} - \sum_{\alpha\in[N]^d} w_{\alpha,N}(x)e^{-i\frac{\xi}{N}\cdot\alpha}\right|^2 \mathrm{d}\xi \tag{25c}$$

$$
\leq d \left(1+\frac{1}{2^N}\right)^{d-1} \max_{1\leq j\leq d} \left\{ (2\pi)^{-d} \int_{\xi\in[-\frac{N}{2},\frac{N}{2}]^d} \hat{k}_0(\xi) \left(\frac{|\xi_j|}{N}\right)^N \mathrm{d}\xi \right\} \tag{25d}
$$

$$
+ \frac{\left(1+\left(N2^N\right)^d\right)^2}{(2\pi)^d} \int_{\xi\notin[-\frac{N}{2},\frac{N}{2}]^d} \hat{k}_0(\xi)\,\mathrm{d}\xi. \tag{25e}
$$

We know that $\hat{k}_0(\xi) = (\sigma\sqrt{\pi})^d e^{-\frac{\sigma^2|\xi|^2}{4}}$. Similar to the analysis in the proof of Example 4 of [41], we first try to bound the first term in the upper bound derived in Eq. (25).

$$
(2\pi)^{-d} \int_{\xi\in[-\frac{N}{2},\frac{N}{2}]^d} (\sigma\sqrt{\pi})^d e^{-\frac{\sigma^2|\xi|^2}{4}} \left(\frac{|\xi_j|}{N}\right)^N \mathrm{d}\xi \tag{26a}
$$

$$
= \frac{\sigma\sqrt{\pi}}{2\pi} \int_{-N/2}^{N/2} e^{-\frac{\sigma^2\xi_j^2}{4}} \left(\frac{|\xi_j|}{N}\right)^N \left(\prod_{k\neq j} \int_{-N/2}^{N/2} \frac{\sigma\sqrt{\pi}}{2\pi} e^{-\frac{\sigma^2\xi_k^2}{4}}\,\mathrm{d}\xi_k \right) \mathrm{d}\xi_j \tag{26b}
$$

$$
\leq \frac{\sigma\sqrt{\pi}}{2\pi} \int_{-N/2}^{N/2} e^{-\frac{\sigma^2|\xi_j|^2}{4}} \left(\frac{|\xi_j|}{N}\right)^N \mathrm{d}\xi_j \leq \frac{2}{\sqrt{\pi}} \left(\frac{2}{\sigma N}\right)^N \Gamma\left(\frac{N+1}{2}\right), \tag{26c}
$$

where $\Gamma(\cdot)$ is the Gamma function. The first inequality in (26) follows by that

$$
\int_{-N/2}^{N/2} \frac{\sigma\sqrt{\pi}}{2\pi} e^{-\frac{\sigma^2\xi_k^2}{4}}\,\mathrm{d}\xi_k \leq \int_{-\infty}^{+\infty} \frac{\sigma\sqrt{\pi}}{2\pi} e^{-\frac{\sigma^2\xi_k^2}{4}}\,\mathrm{d}\xi_k = 1
$$

and the second inequality in (26) follows by that

$$
\begin{aligned}
\int_{-N/2}^{N/2} e^{-\frac{\sigma^2|\xi_j|^2}{4}} \left(\frac{|\xi_j|}{N}\right)^N \mathrm{d}\xi_j =& 2\int_{0}^{N/2} e^{-\frac{\sigma^2|\xi_j|^2}{4}} \left(\frac{|\xi_j|}{N}\right)^N \mathrm{d}\xi_j \\
\leq& 2\int_{0}^{+\infty} e^{-\frac{\sigma^2|\xi_j|^2}{4}} \left(\frac{|\xi_j|}{N}\right)^N \mathrm{d}\xi_j
\end{aligned}
$$

and the definition of Gamma function. Applying Stirling's formula yields

$$
(2\pi)^{-d} \int_{\xi\in[-\frac{N}{2},\frac{N}{2}]^d} (\sigma\sqrt{\pi})^d e^{-\frac{\sigma^2|\xi|^2}{4}} \left(\frac{|\xi_j|}{N}\right)^N \mathrm{d}\xi \tag{27a}
$$

$$
\leq \frac{2}{\sqrt{\pi}} \left(\frac{2}{\sigma N}\right)^N \Gamma\left(\frac{N+1}{2}\right) \leq 2\left(\frac{2}{\sigma N}\right)^N \left(\frac{N+1}{2e}\right)^{\frac{N+1}{2}} \frac{1}{\sqrt{N+1}} e^{\frac{1}{6(N+1)}} \tag{27b}
$$

$$
= \sqrt{\frac{2}{e}} e^{\frac{1}{6(N+1)}} \left(\frac{2\sqrt{\frac{N+1}{2}}}{\sigma\sqrt{e}N} \right)^N \leq \sqrt{2e} \left(\frac{2}{\sigma\sqrt{eN}}\right)^N, \tag{27c}
$$

where the second inequality in (27) follows by the Stirling's formula that

$$
\Gamma(u) \leq \sqrt{2\pi} u^{u-\frac{1}{2}} e^{-u} e^{\frac{1}{12u}}, \quad u > 0
$$

and the last inequality follows by $e^{\frac{1}{6(N+1)}} \leq e$ and $\frac{N+1}{2} \leq N$. We are now to bound the second term in Eq. (25) as follows,

$$
(2\pi)^{-d} \int_{\xi\notin[-\frac{N}{2},\frac{N}{2}]^d} (\sigma\sqrt{\pi})^d e^{-\frac{\sigma^2|\xi|^2}{4}}\,\mathrm{d}\xi = \left(\frac{\sigma\sqrt{\pi}}{2\pi} \int_{\xi_j\notin[-\frac{N}{2},\frac{N}{2}]} e^{-\frac{\sigma^2|\xi_j|^2}{4}}\,\mathrm{d}\xi_j \right)^d
$$

$$= \left(\frac{\sigma\sqrt{\pi}}{\pi}\int_{N/2}^{+\infty} e^{-\frac{\sigma^2}{4}(t^2-t/2)}e^{-\frac{\sigma^2}{4}\cdot\frac{t}{2}}\,\mathrm{d}t\right)^d \leq \left(\frac{\sigma\sqrt{\pi}}{\pi}\int_{N/2}^{+\infty} e^{-\frac{\sigma^2}{4}\left((\frac{N}{2})^2-(\frac{N}{4})\right)}e^{-\frac{\sigma^2}{4}\cdot\frac{t}{2}}\,\mathrm{d}t\right)^d$$

$$= \left(\frac{\sigma}{\sqrt{\pi}}e^{-\frac{\sigma^2 N(N-1)}{16}}\frac{8}{\sigma^2}e^{-\frac{\sigma^2 N}{16}}\right)^d = \left(\frac{8}{\sigma\sqrt{\pi}}\right)^d e^{-\frac{\sigma^2}{16}dN^2}. \tag{28}$$

Combining (25), (27) and (28) yields

$$k(x,x) - 2K_{xX}w_N(x) + w_N(x)^T K w_N(x) \tag{29a}$$

$$\leq \sqrt{2e}d\left(1+\frac{1}{2^N}\right)^{d-1}\left(\frac{2}{\sigma\sqrt{eN}}\right)^N + \left(1+\left(N2^N\right)^d\right)^2\left(\frac{8}{\sigma\sqrt{\pi}}\right)^d e^{-\frac{\sigma^2}{16}dN^2} \tag{29b}$$

$$\leq \sqrt{2e}d2^{d-1}\left(\frac{2}{\sigma\sqrt{eN}}\right)^N + 4\left(N2^N\right)^{2d}\left(\frac{8}{\sigma\sqrt{\pi}}\right)^d e^{-\frac{\sigma^2}{16}dN^2} \tag{29c}$$

$$= \sqrt{2e}d2^{d-1}\left(\frac{2}{\sigma\sqrt{eN}}\right)^N + 4\left(\frac{8}{\sigma\sqrt{\pi}}\right)^d e^{-\frac{\sigma^2}{16}dN^2+2d(N\log 2+\log N)} \tag{29d}$$

$$\leq \sqrt{2e}d2^{d-1}\left(\frac{2}{\sigma\sqrt{eN}}\right)^N + 4\left(\frac{8}{\sigma\sqrt{\pi}}\right)^d e^{-\frac{\sigma^2}{16}dN^2+2d(\log 2+1)N}, \tag{29e}$$

where the first inequality follows by combining (25), (27) and (28), the second inequality follows by that $1+\frac{1}{2^N}\leq 2$ and $1+(N2^N)^d \leq 2(N2^N)^d$, and the last inequality follows by that $\log N \leq N$. Let $N \geq \max\{\frac{32(\log 2+2)}{\sigma^2}, \frac{4e^{2d-1}}{\sigma^2}\}$, we have

$$k(x,x) - 2K_{xX}w_N(x) + w_N(x)^T K w_N(x) \tag{30a}$$

$$\leq \sqrt{2e}d2^{d-1}\left(\frac{2}{\sigma\sqrt{eN}}\right)^N + 4\left(\frac{8}{\sigma\sqrt{\pi}}\right)^d e^{-\frac{\sigma^2}{16}dN^2+2d(\log 2+1)N} \tag{30b}$$

$$\leq \sqrt{2e}d2^{d-1}\left(e^{-d}\right)^N + 4\left(\frac{8}{\sigma\sqrt{\pi}}\right)^d e^{-dN} \tag{30c}$$

$$= \left(\sqrt{2e}d2^{d-1} + 4\left(\frac{8}{\sigma\sqrt{\pi}}\right)^d\right)e^{-dN}. \tag{30d}$$

Combining (23), (30) and that $N^d = t$, we have

$$\sigma_t(x) \leq \left(\sqrt{2e}d2^{d-1} + 4\left(\frac{8}{\sigma\sqrt{\pi}}\right)^d\right)^{\frac{1}{2}} e^{-\frac{1}{2}dt^{1/d}}, \forall x \in [0,1]^d. \tag{31}$$

Combining that $f(\tilde{x}_t) - \min_{x\in\mathcal{X}} f(x) \leq 2R\sigma_t(\tilde{x}_t)$ in (22) and (31), we have

$$f(\tilde{x}_t) - \min_{x\in\mathcal{X}} f(x) \leq 2R\left(\sqrt{2e}d2^{d-1} + 4\left(\frac{8}{\sigma\sqrt{\pi}}\right)^d\right)^{\frac{1}{2}} e^{-\frac{1}{2}dt^{1/d}}. \tag{32}$$

Setting the right hand side to be smaller than ϵ, we observe that the number of steps t only needs to be $\mathcal{O}\left(\left(\log\frac{R}{\epsilon}\right)^d\right)$. This completes the proof.

E Proof of Theorem 5.3

By Lem. 3 in [4], the RKHS on $[0,1]^d$ is equivalent to Sobolev Hilbert space $H^{\nu+\frac{d}{2}}((0,1)^d)$. Implied by [8, Thm. 1, Sec. 3.3.3], the covering number of the function space ball in

$H^{\nu+\frac{d}{2}}((0,1)^d)$ is lower bounded by $\Omega\left(\left(\frac{R}{\epsilon}\right)^{\frac{d}{\nu+d/2}}\right)$. Therefore,

$$\log\mathcal{N}(S(\mathcal{X}),4\epsilon,\|\cdot\|_\infty)=\Omega\left(\left(\frac{R}{\epsilon}\right)^{\frac{d}{\nu+d/2}}\right). \tag{33}$$

We then apply Thm. 5.1 such that we can get the lower bound

$$T=\Omega\left(\left(\frac{R}{\epsilon}\right)^{\frac{d}{\nu+d/2}}\left(\log\left(\frac{R}{\epsilon}\right)\right)^{-1}\right). \tag{34}$$

The upper bound is implied by Thm. 1 in [4].

References

1. Andersson, J.A., Gillis, J., Horn, G., Rawlings, J.B., Diehl, M.: CasADi: a software framework for nonlinear optimization and optimal control. Mathematical Programming Computation **11**(1), 1–36 (2019)
2. Bansal, S., Calandra, R., Xiao, T., Levine, S., Tomlin, C.J.: Goal-driven dynamics learning via Bayesian optimization. In: 2017 IEEE 56th Annual Conference on Decision and Control (CDC), pp. 5168–5173. IEEE (2017)
3. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. Journal of Machine Learning Research **13**(2) (2012)
4. Bull, A.D.: Convergence rates of efficient global optimization algorithms. Journal of Machine Learning Research **12**(10) (2011)
5. Cai, X., Scarlett, J.: On lower bounds for standard and robust Gaussian process bandit optimization. In: International Conference on Machine Learning, pp. 1216–1226. PMLR (2021)
6. Cucker, F., Smale, S.: On the mathematical foundations of learning. Bulletin of the American Mathematical Society **39**(1), 1–49 (2002)
7. De Freitas, N., Smola, A.J., Zoghi, M.: Exponential regret bounds for Gaussian process bandits with deterministic observations. In: Proceedings of the 29th International Conference on Machine Learning, pp. 955–962 (2012)
8. Edmunds, D.E., Triebel, H.: Function Spaces, Entropy Numbers, Differential Operators, vol. 120. Cambridge University Press (1996)
9. Frazier, P.I., Powell, W.B.: Consistency of sequential Bayesian sampling policies. SIAM Journal on Control and Optimization **49**(2), 712–731 (2011)
10. Frazier, P.I., Wang, J.: Bayesian optimization for materials design. In: Information Science for Materials Discovery and Design, pp. 45–75. Springer (2016)
11. GPy: GPy: A Gaussian process framework in python. `http://github.com/SheffieldML/GPy` (since 2012)
12. Grünewälder, S., Audibert, J.Y., Opper, M., Shawe-Taylor, J.: Regret bounds for Gaussian process bandit problems. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, pp. 273–280. JMLR Workshop and Conference Proceedings (2010)
13. Jones, D.R., Schonlau, M., Welch, W.J.: Efficient global optimization of expensive black-box functions. Journal of Global Optimization **13**(4), 455–492 (1998)
14. Khemchandani, R., Jayadeva, Chandra, S.: Optimal kernel selection in twin support vector machines. Optimization Letters **3**, 77–88 (2009)
15. Kim, S.J., Magnani, A., Boyd, S.: Optimal kernel selection in kernel fisher discriminant analysis. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 465–472 (2006)
16. Kolmogorov, A.N., Tikhomirov, V.M.: ε-entropy and ε-capacity of sets in function spaces. Uspekhi Matematicheskikh Nauk **14**(2), 3–86 (1959)
17. Locatelli, M.: Bayesian algorithms for one-dimensional global optimization. Journal of Global Optimization **10**(1), 57–76 (1997)

18. Lorentz, G.: Metric entropy and approximation. Bulletin of the American Mathematical Society **72**(6), 903–937 (1966)
19. Maddalena, E.T., Scharnhorst, P., Jones, C.N.: Deterministic error bounds for kernel-based learning techniques under bounded noise. Automatica **134**, 109,896 (2021)
20. Mairal, J., Vert, J.P.: Machine learning with kernel methods. Lecture Notes, January **10** (2018)
21. Negoescu, D.M., Frazier, P.I., Powell, W.B.: The knowledge-gradient algorithm for sequencing experiments in drug discovery. INFORMS Journal on Computing **23**(3), 346–363 (2011)
22. Nemirovskij, A.S., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. Wiley-Interscience (1983)
23. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems: Standard information for functionals, vol. 2. European Mathematical Society (2008)
24. Raskutti, G., J Wainwright, M., Yu, B.: Minimax-optimal rates for sparse additive models over kernel classes via convex programming. Journal of Machine Learning Research **13**(2) (2012)
25. Ray Chowdhury, S., Gopalan, A.: Bayesian optimization under heavy-tailed payoffs. Advances in Neural Information Processing Systems **32** (2019)
26. Russo, D., Van Roy, B.: An information-theoretic analysis of thompson sampling. Journal of Machine Learning Research **17**(1), 2442–2471 (2016)
27. Scarlett, J.: Tight regret bounds for Bayesian optimization in one dimension. In: International Conference on Machine Learning, pp. 4500–4508. PMLR (2018)
28. Scarlett, J., Bogunovic, I., Cevher, V.: Lower bounds on regret for noisy Gaussian process bandit optimization. In: Conference on Learning Theory, pp. 1723–1742. PMLR (2017)
29. Shahriari, B., Swersky, K., Wang, Z., Adams, R.P., De Freitas, N.: Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE **104**(1), 148–175 (2015)
30. Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M., Prabhat, M., Adams, R.: Scalable Bayesian optimization using deep neural networks. In: International Conference on Machine Learning, pp. 2171–2180. PMLR (2015)
31. Srinivas, N., Krause, A., Kakade, S.M., Seeger, M.W.: Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. IEEE Transactions on Information Theory **58**(5), 3250–3265 (2012)
32. Steinwart, I.: A short note on the comparison of interpolation widths, entropy numbers, and Kolmogorov widths. Journal of Approximation Theory **215**, 13–27 (2017)
33. Vakili, S., Bouziani, N., Jalali, S., Bernacchia, A., Shiu, D.s.: Optimal order simple regret for Gaussian process bandits. Advances in Neural Information Processing Systems **34** (2021)
34. Vazquez, E., Bect, J.: Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. Journal of Statistical Planning and Inference **140**(11), 3088–3095 (2010)
35. Wächter, A., Biegler, L.T.: On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Mathematical Programming **106**(1), 25–57 (2006)
36. Wahba, G.: Spline Models for Observational Data. SIAM (1990)
37. Wang, Z., de Freitas, N.: Theoretical analysis of Bayesian optimisation with unknown Gaussian process hyper-parameters. arXiv preprint arXiv:1406.7758 (2014)
38. Wendland, H.: Scattered Data Approximation, vol. 17. Cambridge University Press (2004)
39. Wu, Y.: Lecture notes on information-theoretic methods for high-dimensional statistics. Lecture Notes for ECE598YW (UIUC) **16** (2017)
40. Xu, W., Jones, C.N., Svetozarevic, B., Laughman, C.R., Chakrabarty, A.: VABO: Violation-Aware Bayesian Optimization for closed-loop control performance optimization with unmodeled constraints. arXiv preprint arXiv:2110.07479 (2021)
41. Zhou, D.X.: The covering number in learning theory. Journal of Complexity **18**(3), 739–767 (2002)
42. Zhou, D.X.: Capacity of reproducing kernel spaces in learning theory. IEEE Transactions on Information Theory **49**(7), 1743–1752 (2003)