Uncertainty Quantification for Bayesian Optimization

Rui Tuo*
Wm Michael Barnes '64
Department of Industrial and Systems Engineering
Texas A&M University

Wenjia Wang*
Data Science and Analytics Thrust
The Hong Kong University of Science
and Technology (Guangzhou)
and Department of Mathematics
The Hong Kong University of Science and Technology

Abstract

Bayesian optimization is a class of global optimization techniques. In Bayesian optimization, the underlying objective function is modeled as a realization of a Gaussian process. Although the Gaussian process assumption implies a random distribution of the Bayesian optimization outputs, quantification of this uncertainty is rarely studied in the literature. In this work, we propose a novel approach to assess the output uncertainty of Bayesian optimization algorithms, which proceeds by constructing confidence regions of the maximum point (or value) of the objective function. These regions can be computed efficiently, and their confidence levels are guaranteed by the uniform error bounds for sequential Gaussian process regression newly developed in the present work. Our theory provides a unified uncertainty quantification framework for all existing sequential sampling policies and stopping criteria.

1 INTRODUCTION

The empirical and data-driven nature of data science field makes uncertainty quantification one of the central questions that need to be addressed in order to guide and safeguard decision makings. In this work, we focus on Bayesian optimization, which is effective in solving global optimization problems for complex blackbox functions. Our objective is to quantify the uncertainty of Bayesian optimization outputs. Such uncertainty comes from the Gaussian process prior, random input and stopping time. Closed-form solution of the output uncertainty is usually intractable because of the complicated sampling scheme and stopping criterion.

Let f be an underlying continuous function over Ω, a compact subset of $\mathbb{R}^p$. The goal of global optimization is to find the maximum of f, denoted by $\max_{x\in\Omega} f(x)$, or the point x_{max} which satisfies $f(x_{max}) = \max_{x\in\Omega} f(x)$. In many scenarios, objective functions can be expensive to evaluate. For example, f defined by a complex computer model may take a long time to run. Bayesian optimization is a powerful technique to deal with this type of problems, and has been widely used in areas including designing engineering systems (Forrester et al., 2008; Jones et al., 1998), materials and drug design (Frazier and Wang, 2016; Negoescu et al., 2011; Solomou et al., 2018), chemistry (Häse et al., 2018), deep neural networks (Diaz et al., 2017; Klein et al., 2017), and reinforcement learning (Marco et al., 2017; Wilson et al., 2014).

In Bayesian optimization, f is treated as a realization of a stochastic process, denoted by Z. Usually, people assume that Z is a Gaussian process. Every Bayesian optimization algorithm defines a sequential sampling procedure, which successively generates new input points, based on the acquired function evaluations over all previous input points. Usually, the next input point is determined by maximizing an acquisition function. Examples of acquisition functions include probability of improvement (Kushner, 1964), expected improvement (Huang et al., 2006; Jones et al., 1998; Mockus et al., 1978; Picheny et al., 2013a), Gaussian process upper confidence bound (Azimi et al., 2010; Contal et al., 2013; Desautels et al., 2014; Srinivas et al., 2010), entropy search (Hennig and Schuler, 2012), predictive entropy search (Hernández-Lobato

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

* These authors contributed equally to this work.

et al., 2014), entropy search portfolio (Shahriari et al., 2014), knowledge gradient (Scott et al., 2011; Wu and Frazier, 2016; Wu et al., 2017), etc. We refer to Frazier (2018); Shahriari et al. (2016) for an introduction to popular Bayesian optimization methods.

Although Bayesian optimization has received considerable attention and numerous techniques have emerged in recent years, how to quantify the uncertainty of the outputs from a Bayesian optimization algorithm is rarely discussed in the literature. Since we assume that f is a *random realization* of Z, x_{max} and $f(x_{max})$ should also be random. However, the highly nontrivial distributions of x_{max} and $f(x_{max})$ make uncertainty quantification rather challenging.

In this work, we develop efficient methods to construct confidence regions of x_{max} and $f(x_{max})$ for Bayesian optimization algorithms, where function f is a realization of Gaussian process Z. Our uncertainty quantification method *does not* rely on the specific formulae or strategies, and can be applied to all existing methods in an abstract sense. We show that by using the collected data of any instance algorithm of Bayesian optimization, Algorithm 3 gives a confidence upper limit with theoretical guarantees of their confidence level in Corollary 3. To the best of our knowledge, this is the *first* theoretical result of the uncertainty quantification on the maximum estimator for Bayesian optimization, under the assumption that f is a realization of a Gaussian process. Compared with the traditional point-wise predictive standard deviation of Gaussian process regression, denoted by $\sigma(x)$, our bound is only inflated by a factor proportional to $\sqrt{\log(e\sigma/\sigma(x))}$, where σ is the prior standard deviation.

It is worth noting that uncertainty quantification typically differs from convergence analysis of algorithms. In Bayesian optimization, the latter topic has been studied more often. See, for instance, Bect et al. (2019); Calvin (1997, 2005); Ryzhov (2016); Vazquez and Bect (2010); Yarotsky (2013). These analyses provide theoretical guarantee on the convergence of Bayesian optimization algorithms, but do not directly lead to techniques for uncertainty quantification. Recall that in this work, we assume that the underlying function f is a realization of a Gaussian process, and therefore, the sample path properties of f, such as the smoothness, should be governed by the covariance function of the Gaussian process. This Gaussian process assumption differs from those in some existing works in the analysis of Bayesian optimization, e.g., Bull (2011); Astudillo and Frazier (2019); Yarotsky (2013), where the underlying function f is assumed to be a *deterministic* function satisfying pre-specified smoothness conditions.

The rest of this paper is structured as follows. In Section 2, we present some preliminaries, including an introduction to Gaussian process regression and Bayesian optimization. Section 3 presents uncertainty quantification results under fixed designs. Section 4 introduces our methods and main theoretical results. How to calibrate the constant in our method is introduced in Section 5. Numerical results are presented in Section 6. Conclusions and discussion are made in Section 7. Technical details are given in the Appendix.

2 PRELIMINARIES

In this section, we provide a brief introduction to Gaussian process regression and review some existing methods in Bayesian optimization.

2.1 Gaussian Process Regression

Recall that in Bayesian optimization, the objective function f is assumed to be a realization of a Gaussian process Z. In this work, we suppose that Z is stationary and has mean zero, variance σ^2 and correlation function Ψ, i.e., $\mathrm{Cov}(Z(x), Z(x')) = \sigma^2\Psi(x-x')$ with $\Psi(0) = 1$. Under certain regularity conditions, Bochner's theorem (Wendland, 2004) suggests that the Fourier transform (with a specific choice of the constant factor) of Ψ, denoted by $\tilde{\Psi}$, is a probability density function and satisfies the inversion formula $\Psi(x) = \int_{\mathbb{R}^p} \cos(\omega^T x)\tilde{\Psi}(\omega)d\omega$. We call $\tilde{\Psi}$ the *spectral density* of Ψ. Some popular choices of correlation functions and their spectral densities are discussed in Section 3.3. Throughout this work, we further assume Ψ satisfies the following condition. For a vector $\omega = (\omega_1, \dots, \omega_p)^T$, define its l_1-norm as $\|\omega\|_1 = |\omega_1| + \dots + |\omega_p|$.

Condition 1. *The correlation function Ψ has a spectral density, denoted by $\tilde{\Psi}$, and*

$$A_0 = \int_{\mathbb{R}^p} \|\omega\|_1 \tilde{\Psi}(\omega) d\omega < +\infty. \tag{1}$$

Remark 1. *The l_1-norm in* (1) *can be replaced by the usual Euclidean norm. However, we use the former here because they usually have explicit expressions. See Section 3.3 for details.*

Remark 2. *Condition 1 imposes a smoothness condition on the correlation function Ψ, which is equivalent to the mean squared differentiability (Stein, 1999) of the Gaussian process Z. Note that the mean squared differentiability differs from the sample path differentiability. We refer to Driscoll (1973); Steinwart (2019) for results on the relationship between the sample path smoothness of Z (thus f) and the smoothness of correlation function Ψ.*

Suppose the set of points $X = (x_1, \ldots, x_n)$ is given. Then f can be reconstructed via Gaussian process regression. Let $Y = (Z(x_1), \ldots, Z(x_n))^T$ be the vector of evaluations of the Gaussian process at points $x_1, ..., x_n$. The following results are well-known and can be found in Rasmussen and Williams (2006). For any untried point x, conditional on Y, $Z(x)$ follows a normal distribution. The conditional mean and variance of $Z(x)$ are

$$\mu(x) := \mathbb{E}[Z(x)|Y] = r^T(x)K^{-1}Y, \tag{2}$$

$$\sigma^2(x) := \mathrm{Var}[Z(x)|Y] = \sigma^2(1 - r^T(x)K^{-1}r(x)), \tag{3}$$

where $r(x) = (\Psi(x - x_1), \ldots, \Psi(x - x_n))^T, K = (\Psi(x_j - x_k))_{jk}$. Since we assume that f is a realization of Z, $\mu(x)$ can serve as a reconstruction of f.

2.2 Bayesian Optimization

In Bayesian optimization, we evaluate f over a set of input points, denoted by $x_1, \ldots, x_n$. We call them the *design points*, because these points can be chosen according to our will. There are two categories of strategies to choose design points. We can choose all the design points before we evaluate f at any of them. Such a design set is call a *fixed design*. An alternative strategy is called *sequential sampling*, in which the design points are not fully determined at the beginning. Instead, points are added sequentially, guided by the information from the previous input points and the corresponding acquired function values. An instance algorithm defines a sequential sampling scheme which determines the next input point x_{n+1} by maximizing an *acquisition function* $a(x; X_n, Y_n)$, where $X_n = (x_1, \ldots, x_n)$ consists of previous input points, and $Y_n = (f(x_1), \ldots, f(x_n))^T$ consists of corresponding outputs. The acquisition function can be either deterministic or random given X_n, Y_n. A general Bayesian optimization procedure under sequential sampling scheme is shown in Algorithm 1.

Algorithm 1 Bayesian optimization (described in Shahriari et al. (2016))

1: **Input:** A Gaussian process prior of f, initial observation data X_1, Y_1.
2: **for** $n = 1, 2 \ldots,$ **do**
3: Find $x_{n+1} = \mathrm{argmax}_{x \in \Omega}\, a(x; X_n, Y_n)$, evaluate $f(x_{n+1})$, update data and the posterior probability distribution on f.
4: **Output:** The point evaluated with the largest $f(x)$.

A number of acquisition functions are proposed in the literature, for example:

1. Expected improvement (EI) (Jones et al., 1998; Mockus et al., 1978), with the acquisition function $a_{\mathrm{EI}}(x; X_n, Y_n) := \mathbb{E}((Z(x) - y_n^*)\mathbf{1}(Z(x) - y_n^*)|X_n, Y_n)$, where $\mathbf{1}(\cdot)$ is the indicator function, and $y_n^* = \max_{1 \le i \le n} f(x_i)$.

2. Gaussian process upper confidence bound (Srinivas et al., 2010), with the acquisition function $a_{\mathrm{UCB}}(x; X_n, Y_n) := \mu_n(x) + \beta_n \sigma_n(x)$, where β_n is a parameter, and $\mu_n(x)$ and $\sigma_n(x)$ are posterior mean and variance of f after nth iteration, respectively.

3. Predictive entropy search (Hernández-Lobato et al., 2014), with the acquisition function $a_{\mathrm{PES}}(x; X_n, Y_n) := \mathcal{H}(y|X_n, Y_n, x) - \mathbb{E}_{p(x_{max}|X_n, Y_n)}\mathcal{H}(y|X_n, Y_n, x, x_{max})$, where $\mathcal{H}(y|X_n, Y_n, x)$ and $\mathcal{H}(y|X_n, Y_n, x, x_{max})$ are the differential entropy of the posterior distribution $p(y|X_n, Y_n, x)$ and $p(y|X_n, Y_n, x, x_{max})$, respectively. The expectation can be approximated via Thompson samples. Another entropy search acquisition function is introduced by Hennig and Schuler (2012), who also provide an efficient way to approximate the distribution of x_{max} based on the Gaussian process prior.

Among the above acquisition functions, a_{EI} and a_{UCB} are deterministic functions of (x, X_n, Y_n), whereas a_{PES} is random in practice because Thompson sampling depends on a random sample from the posterior Gaussian process. We refer to Shahriari et al. (2016) for general discussions and popular methods in Bayesian optimization.

In practice, one also needs to determine when to stop Algorithm 1. Usually, decisions are made in consideration of the budget and the accuracy requirement. For instance, practitioners can stop Algorithm 1 after finishing a fixed number of iterations (Frazier, 2018) or no further significant improvement of function values can be made (Acerbi and Ji, 2017). Although stopping criteria plays no role in the analysis of the algorithms' asymptotic behaviors, they can greatly affect the output uncertainty.

3 OPTIMIZATION WITH GAUSSIAN PROCESS REGRESSION UNDER FIXED DESIGNS

Before investigating the more important sequential sampling schemes, we shall first consider fixed designs in this section, because the latter situation is simpler and will serve as an important intermediate step to the general problem in Section 4.

3.1 Uniform Error Bound

Although the conditional distribution of $Z(x)$ is simple as shown in (2)-(3), those for x_{max} and $Z(x_{max})$ are highly non-trivial because they are nonlinear functionals of Z. In this work, we construct confidence regions for the maximum points and values using a uniform error bound for Gaussian process regression, given in Theorem 1. We will use the notion $a \vee b := \max(a, b)$. Also, we shall use the convention $0/0 = 0$ in all statements in this article related to error bounds.

Theorem 1. *Suppose Condition 1 holds. Let* $M = \sup_{x\in\Omega} \frac{Z(x)-\mu(x)}{\sigma(x)\log^{1/2}(e\sigma/\sigma(x))}$, *where* $\mu(x)$ *and* $\sigma(x)$ *are given in (2) and (3), respectively. Then the followings are true.*

1. $\mathbb{E}M \leq C_0\sqrt{p(1 \vee \log(A_0 D_\Omega))}$, *where* C_0 *is a universal constant,* A_0 *is as in Condition 1, and* $D_\Omega = diam(\Omega)$ *is the Euclidean diameter of* Ω.

2. *For any* $t > 0$, $\mathbb{P}(M - \mathbb{E}M > t) \leq e^{-t^2/2}$.

In practice, Part 2 of Theorem 1 is hard to use directly because $\mathbb{E}M$ is difficult to calculate accurately. Instead, we can replace $\mathbb{E}M$ by its upper bound in Part 1 of Theorem 1. We state such a result in Corollary 1. Its proof is trivial.

Corollary 1. *Under the conditions and notation of Theorem 1, for any constant* C *such that* $\mathbb{E}M \leq C\sqrt{p(1 \vee \log(A_0 D_\Omega))}$, *we have that for any* $t > 0$,

$$\mathbb{P}(M - C\sqrt{p(1 \vee \log(A_0 D_\Omega))} > t) \leq e^{-t^2/2}.$$

To use Corollary 1, we need to determine the universal constant C and the moment of the spectral density A_0. According to our numerical simulations in Section 5 and Section F of the Supplementary material, we recommend using $C = 1$ in practice. We shall discuss the calculation of A_0 in Section 3.3.

3.2 Uncertainty Quantification

In light of Corollary 1, we can construct a confidence upper limit of f. Algorithm 2 describes how to compute the confidence upper limit of f at a given untried x. For notational simplicity, we regard the dimension p, the variance σ^2, the moment A_0 and the universal constant C as global variables so that Algorithm 2 has access to all of them.

Based on the UCL function in Algorithm 3, we can construct a confidence region for x_{max} and a confidence interval for $f(x_{max})$. These regions do not have explicit expressions. However, they can be approximated by calling UCL with many different x's. Let

Algorithm 2 Uniform confidence upper limit at a given point: $\mathrm{UCL}(x, t, X, Y)$

1: **Input**: Untried point x, significance parameter t, data $X = (x_1, \ldots, x_n)^T, Y$.
2: Set $r = (\Psi(x - x_1), \ldots, \Psi(x - x_n))^T, K = (\Psi(x_j - x_k))_{jk}$. Calculate $\mu = r^T K^{-1} Y, s = \sqrt{\sigma^2(1 - r^T K^{-1} r)}$.
3: **Output**: $\mu + s\sqrt{\log(e\sigma/s)}(C\sqrt{p(1 \vee \log(A_0 D_\Omega))} + t)$.

$Y = (f(x_1), \ldots, f(x_n))^T$. The confidence region for x_{max} is defined as

$$CR_t := \left\{x \in \Omega : \mathrm{UCL}(x, t, X, Y) \geq \max_{1\leq i\leq n} f(x_i)\right\}. \quad (4)$$

The confidence interval for $f(x_{max})$ is defined as

$$CI_t := \left[\max_{1\leq i\leq n} f(x_i), \max_{x\in\Omega} \mathrm{UCL}(x, t, X, Y)\right]. \quad (5)$$

It is worth noting that the probability in Corollary 1 is *not* a posterior probability. Therefore, the regions given by (4) and (5) should be regarded as frequentist confidence regions under the Gaussian process model, rather than Bayesian credible regions. Such a frequentist nature has an alternative interpretation, shown in Corollary 2. Corollary 2 simply translates Corollary 1 from the language of stochastic processes to a deterministic function approximation setting, which fits the Bayesian optimization framework better. It shows that CR_t in (4) and CI_t in (5) are confidence region of x_{max} and $f(x_{max})$ with confidence level $1 - e^{-t^2/2}$, respectively. In particular, to obtain a 95% confidence region, we use $t = 2.448$.

Corollary 2. *Let* $C(\Omega)$ *be the space of continuous functions on* Ω *and* $\mathbb{P}_Z$ *be the law of* Z. *Then there exists a set* $B \subset C(\Omega)$ *so that* $\mathbb{P}_Z(B) \geq 1 - e^{-t^2/2}$ *and for any* $f \in B$, *its maximum point* x_{max} *is contained in* CR_t *defined in (4), and* $f(x_{max})$ *is contained in* CI_t *defined in (5).*

In practice, the shape of CR_t can be highly irregular and representing the region of CR_t can be challenging. If Ω is of one or two dimensions, we can choose a fine mesh over Ω and call $\mathrm{UCL}(x, t, X, Y)$ for each mesh grid point x. In a general situation, we suggest calling $\mathrm{UCL}(x, t, X, Y)$ with randomly chosen x's and using the k-nearest neighbors algorithm to represent CR_t.

3.3 Calculating A_0

For an arbitrary Ψ, calculation of A_0 in (1) can be challenging. Fortunately, for two most popular correlation functions in one dimension, namely the Gaussian and the Matérn correlation functions (Rasmussen and Williams, 2006), A_0 can be calculated in closed form. The results are summarized in Table 1. In Table 1, $\Gamma(\cdot)$ is the Gamma function and $K_\nu(\cdot)$ is the modified Bessel function of the second kind.

For multi-dimensional problems, a common practice is to use *product* correlation functions. Specifically, suppose $\Psi_1, \ldots, \Psi_p$ are one-dimensional correlation functions. Then their product $\Psi(x) = \prod_{i=1}^p \Psi(x_i)$ forms a p-dimensional correlation function, where $x = (x_1, \ldots, x_p)^T$. If a product correlation function is used, the calculation of A_0 is easy. It follows from the elementary properties of Fourier transform that $\tilde{\Psi}(x) = \prod_{i=1}^p \tilde{\Psi}_i(x_i)$. Let X_i be a random variable with probability density function Ψ_i. Then $A_0 = \sum_{i=1}^p \mathbb{E}|X_i|$, i.e., the value of A_0 corresponding to a product correlation function is the sum of those given by the marginal correlation functions. If each Ψ_i is either a Gaussian or Matérn correlation function, then $\mathbb{E}|X_i|$'s can be read from Table 1.

4 UNCERTAINTY QUANTIFICATION FOR SEQUENTIAL SAMPLINGS

In Bayesian optimization, sequential samplings are more popular, because such approaches can utilize the information from the previous responses and choose new design points in the area which is more likely to contain the maximum points. In this section, we present our uncertainty quantification methodology for sequential samplings, as well as the theoretical guarantees.

4.1 Methodology

In this subsection, we construct confidence regions for the maximum points and values under sequential sampling scheme, as presented in Algorithm 3. In the rest of this work, let T be the number of iterations when an instance of Algorithm 1 stops and D_Ω be the diameter of Ω. It is worth noting that under sequential sampling scheme, T can be a random variable, which introduces additional randomness of the confidence interval, and complicates the theoretical analysis. Given n, we denote $X_{1:n} = (x_1, \ldots, x_{m_n})$, where each x_i is corresponding to one data point and m_n is the number of sampled points after n iterations of the algorithm, and $Y_{1:n} = (f(x_1), \ldots, f(x_{m_n}))^T$. In this work, we allow $m_n \geq 1$, which means that we can sample one point or a batch of points at a time in each iteration.

Let $g(x) = \mathrm{UCL}(x, t, X_{1:T}, Y_{1:T})$,

$$CR_t^{\mathbf{seq}} := \left\{ x \in \Omega : g(x) \geq \max_{1 \leq i \leq m_T} f(x_i) \right\}, \tag{6}$$

$$CI_t^{\mathbf{seq}} := \left[\max_{1 \leq i \leq m_T} f(x_i), \max_{x \in \Omega} g(x) \right]. \tag{7}$$

Algorithm 3 Confidence regions for $x_{\max}$ and $f(x_{\max})$

1: **Input:** Significance parameter t, data $X_{1:T}, Y_{1:T}$ collected from an instance of Bayesian optimization algorithm.

2: For point $x \in \Omega$, set $r(x) = (\Psi(x - x_1), \ldots, \Psi(x - x_{m_T}))^T, K = (\Psi(x_j - x_k))_{jk}$. Calculate

$$\mu_T(x) = r(x)^T K^{-1} Y_{1:T}, \tag{8}$$

$$s_T(x) = \sqrt{\sigma^2(1 - r(x)^T K^{-1} r(x))}. \tag{9}$$

3: Compute $\mathrm{UCL}(x, t, X_{1:T}, Y_{1:T})$ via Algorithm 2.

4: Let $g(x) = \mathrm{UCL}(x, t, X_{1:T}, Y_{1:T})$. Calculate $CR_t^{\mathbf{seq}}$ and $CI_t^{\mathbf{seq}}$ via (6) and (7), respectively.

5: **Output:** The confidence region $CR_t^{\mathbf{seq}}$ for x_{max} and the confidence interval $CI_t^{\mathbf{seq}}$ for $f(x_{max})$.

In Section 4.2, we will show that under the condition that f is a realization of Z, $CR_t^{\mathbf{seq}}$ and $CI_t^{\mathbf{seq}}$ are confidence regions of x_{max} and $f(x_{max})$, respectively, with a simultaneous confidence level at least $1 - e^{-t^2/2}$, respectively. In particular, to obtain a 95% confidence region, we use $t = 2.448$. The calculation of A_0 follows the discussion in Section 3.3, and we recommend using $C = 1$ in practice, as in Algorithm 2.

4.2 Theory

To facilitate our mathematical analysis, we first state the general Bayesian optimization framework in a rigorous manner. Recall that we assume that f is a realization of a Gaussian process Z with correlation function Ψ. From this Bayesian point of view, we shall not differentiate f and Z in this section.

Denote the vectors of input and output points in the nth iteration as X_n and Y_n, respectively. Let $X_{1:n}$ and $Y_{1:n}$ be as in Section 4.1. Because $X_{1:n}$ and $Y_{1:n}$ are random, the data $(X_{1:n}, Y_{1:n}^T)$ is associated with the σ-algebra $\mathcal{F}_n$, defined as the σ-algebra generated by $(X_{1:n}, Y_{1:n}^T)$. When the algorithm just starts, the data is an empty set, which is associated with the trivial σ-algebra $\mathcal{F}_0$. In each sampling-evaluation iteration, a sequential sampling strategy, which determines the next sample point or a batch of points based on the current data, is applied. Clearly, such strategy should not depend on unobserved data. After each sampling-evaluation iteration, a stopping criterion is checked

Correlation family	Gaussian	Matérn
Correlation function	$\exp\{-(x/\theta)^2\}$	$\frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{2\sqrt{\nu}\lvert x\rvert}{\theta}\right)^{\nu} K_\nu\left(\frac{2\sqrt{\nu}\lvert x\rvert}{\theta}\right)$
Spectral density	$\frac{\theta}{2\sqrt{\pi}}\exp\{-\omega^2\theta^2/4\}$	$\frac{\Gamma(\nu+1/2)}{\Gamma(\nu)\sqrt{\pi}}\left(\frac{4\nu}{\theta^2}\right)^{\nu}\left(\omega^2+\frac{4\nu}{\theta^2}\right)^{-(\nu+1/2)}$
A_0	$\frac{2}{\sqrt{\pi}\theta}$	$\frac{4\sqrt{\nu}\Gamma(\nu+1/2)}{\sqrt{\pi}(2\nu-1)\theta\Gamma(\nu)}$ for $\nu > 1/2$

Table 1: Gaussian And Matérn Correlation Families.

and to determine whether to terminate the algorithm. A stopping decision should depend only on the current data and/or prespecified values such as computational budget, and should not depend on unobserved data either. Let T be the number of iterations when the algorithm stops. Then a Bayesian optimization algorithm must satisfy the following conditions.

1. Conditional on $\mathcal{F}_{n-1}$, X_n and Z are mutually independent for $n = 1, 2, \ldots$.

2. T is a stopping time with respect to the filtration $\{\mathcal{F}_n\}_{n=0}^{\infty}$. We further require $1 \leq T < +\infty, a.s.$, to ensure a meaningful Bayesian optimization procedure.

We shall establish a generic theory that bounds the uniform prediction error, which can be applied to any instance algorithms of Bayesian optimization. It is worth noting that several literature, including Sniekers and van der Vaart (2015); Yoo et al. (2016); Yang et al. (2017); Kuriki et al. (2019); Azzimonti et al. (2019); Azaïs et al. (2010), investigate uncertainty quantification methods which are not within the Bayesian optimization or sequential sampling scheme, and cannot be directly applied to quantify the uncertainties of outputs of Bayesian optimization.

In Bayesian optimization, sequential samplings are more popular, because such approaches can utilize the information from the previous responses and choose new design points in the area which is more likely to contain the maximum points. Similar to Section 3, we first quantify the uncertainty of $Z(\cdot) - \mu_T(\cdot)$. Note that $Z(\cdot) - \mu_T(\cdot)$ is generally *not* a Gaussian process, because in the sequential samplings situation, the stopping time T is random. Nonetheless, an error bound similar to that in Theorem 1 is still valid. In the following theorem, we define

$$\mu_n(x) := r_n^T(x) K_n^{-1} Y_{1:n}, \tag{10}$$
$$\sigma_n^2(x) := \sigma^2(1 - r_n^T(x) K_n^{-1} r_n(x)), \tag{11}$$

where $r_n(x) = (\Psi(x - x_1), \ldots, \Psi(x - x_{m_n}))^T, K_n = (\Psi(x_j - x_k))_{jk}$.

Theorem 2. (Uncertainty quantification for sequential samplings) *Suppose Condition 1 holds. Given an instance of Bayesian optimization algorithm, let*

$$M_n = \sup_{x\in\Omega} \frac{Z(x) - \mu_n(x)}{\sigma_n(x)\log^{1/2}(e\sigma/\sigma_n(x))},$$

where $\mu_n(x)$ and $\sigma_n(x)$ are given in (10) *and* (11)*, respectively. Then for any $t > 0$,*

$$\mathbb{P}(M_T - C\sqrt{p(1 \vee \log(A_0 D_\Omega))} > t) \leq e^{-t^2/2}, \tag{12}$$

where C, A_0, D_Ω are the same as in Corollary 1.

The proof of Theorem 2 can be found in Appendix D. The major difficulty of proving Theorem 2 is that the stopping time T is random, which introduces extra uncertainties of the output of a Bayesian optimization algorithm. The probability bound (12) has a major advantage: the constant C is independent of the specific Bayesian optimization algorithm, and it can be chosen the same as that for fixed designs. This suggests that when calibrating C via numerical simulations (see Section 5 and Appendix F), we only need to simulate for fixed-design problems, and the resulting constant C can be used for the uncertainty quantification of all past and possible future Bayesian optimization algorithms. The dimension p does not influence our uncertainty quantification a lot, in the sense that only $\sqrt{p}$ appears in (12). However, the dimension will strongly influence the performance of Bayesian optimization (Bull, 2011), which could lead to a large confidence region.

Analogous to Corollary 2, we can restate Theorem 2 under a deterministic setting in terms of Corollary 3. In this situation, we have to restrict ourselves to *deterministic* instances of Bayesian optimization algorithms, in the sense that the sequential sampling strategy is a deterministic map, such as the first two examples in Section 2.2.

Corollary 3. *Let $C(\Omega)$ be the space of continuous functions on Ω and $\mathbb{P}_Z$ be the law of Z. Given a deterministic instance of Bayesian optimization algorithm, there exists a set $B \subset C(\Omega)$ so that $\mathbb{P}_Z(B) \geq 1 - e^{-t^2/2}$ and for any $f \in B$, its maximum point x_{max} is contained in $CR_t^{\mathbf{seq}}$ defined in* (6)*, and $f(x_{max})$ is contained in $CI_t^{\mathbf{seq}}$ defined in* (7).

5 CALIBRATING C VIA SIMULATION STUDIES

To construct confidence regions (4) and (6), and confidence intervals (5) and (7), we need to specify the constant C. An upper bound of the constant C in Theorem 1 can be obtained by examine the proof of Lemma A.1 and Theorem A.1. However, this theoretical upper bound can be too large for practical use. In this work, we consider estimating C via numerical simulation. The details are presented in Appendix F. Here we outline the main conclusions of our simulation studies.

Our main conclusions are: 1) $C = 1$ is a robust choice for most of the cases; 2) for the cases with Gaussian correlation functions or small $A_0 D_\Omega$, choosing $C = 1$ may lead to very conservative confidence regions. We suggest practitioners first consider $C = 1$ to obtain robust confidence regions. When users believe that this robust confidence region is too conservative, they can use the value in Table F.1 or F.2 corresponding to their specific setting, or run similar numerical studies as in Appendix F to calibrate their own C.

6 NUMERICAL EXPERIMENTS

In this section, we conduct several numerical studies to compare the performance between the proposed confidence interval $CI_t^{\mathbf{seq}}$ as in (7) and the naive bound of Gaussian process. The nominal confidence levels are 95% for both methods. The naive 95% confidence upper bound, denoted by CI_G, is defined as the usual pointwise upper bound of Gaussian process, i.e.,

$$CI_G := \left[\max_{1\leq i\leq m_T} f(x_i), \max_{x\in\Omega} \mu_T(x) + q_{0.05}\sigma_T(x)\right], \tag{13}$$

where $q_{0.05}$ is the 0.95 quantile of the standard normal distribution, $\mu_T(x)$ and $\sigma_T(x)$ are given in (8) and (9), respectively. As suggested in Section 5, we use $C = 1$ and $t = 2.448$ in $CI_t^{\mathbf{seq}}$.

6.1 Well-Specified Gaussian Process

We first consider that the underlying truth is a Gaussian process with known covariance function. We consider the Matérn correlation functions (see Table 1) with $\nu = 1.5, 2.5, 3.5$, and $A_0 D_\Omega = 25$. We simulate Gaussian processes on $\Omega = [0, 1]^2$ for each ν. We use optimal Latin hypercube designs (Stocki, 2005) to generate five initial points. We employs a_{UCB} (see Section 2.2) as the acquisition function, where the parameter β_n is chosen as the theoretically optimal parameter, suggested by Srinivas et al. (2010).

We repeat the above procedure 100 times to estimate the coverage rate by calculating the relative frequency of the event $f(x_{\max}) \in CI_t^{\mathbf{seq}}$ or $f(x_{\max}) \in CI_G$. We also compare $CI_t^{\mathbf{seq}}$ and CI_G with the "optimal upper bound" in the sense that we choose a constant a_ν and the confidence upper bound

$$CI_a := \left[\max_{1\leq i\leq m_T} f(x_i), \max_{x\in\Omega} \mu_T(x) + a_\nu\sigma_T(x)\right],$$

such that the relative frequency of the event $f(x_{\max}) \in CI_a$ is exactly 95%, where a_ν only depends on ν. Then we plot the coverage rate of $CI_t^{\mathbf{seq}}$ and CI_G, and the width of $CI_t^{\mathbf{seq}}$, CI_G, and CI_a under 5, 10, 15, 20, 25, 30 iterations, respectively.

Panels 1 and 2 in Figure 1 shows the coverage rates and the width of the confidence intervals under different smoothness with $\nu = 1.5, 2.5, 3.5$. From the Panel 1 in Figure 1, we find that the coverage rate of $CI_t^{\mathbf{seq}}$ is almost 100% for all the experiments, while CI_G has a lower coverage rate no more than 75%. Thus the proposed method is conservative while the naive one is permissive. Such a result shows that using the naive method may be risky in practice, because the naive one underestimates the uncertainties. The coverage results support our theory and conclusions made in Section 4.2. As shown by the Panel 2 in Figure 1, the widths of $CI_t^{\mathbf{seq}}$ are about five times of CI_G, and about 2-2.5 times of CI_a. The ratio decreases as the number of iterations increases. The inflation in the width of confidence intervals is the cost of gaining confidence.

6.2 Misspecified Gaussian Process

Although our theory does not cover the case when the Gaussian process is misspecified, we conduct numerical studies to study the influence of model misspecification. The misspecification is in the sense that the correlation function is misspecified. Specifically, we consider the Matérn correlation functions with $\nu_0 = 1.5, 2.5, 3.5$, and $A_0 D_\Omega = 25$. The rest of the settings are the same as those in Section 6.1. The only difference is that, we use Matérn correlation functions with $\nu = 1.5, 2.5, 3.5$ to construct predictors and confidence intervals for each ν_0. The coverage rates of $CI_t^{\mathbf{seq}}$ and CI_G and the width of $CI_t^{\mathbf{seq}}$, CI_G, and CI_a are shown in Panels 1-6 in Figure G.1 in Appendix G. We find similar patterns as discussed in Section 6.1, namely, the proposed confidence interval is conservative while the naive one is permissive. We also find that for each ν_0, the width of confidence intervals does not change a lot for different ν. These findings indicate that our theory may work for the misspecified case, while the theoretical development needs further study.

We also consider another case of misspecification,

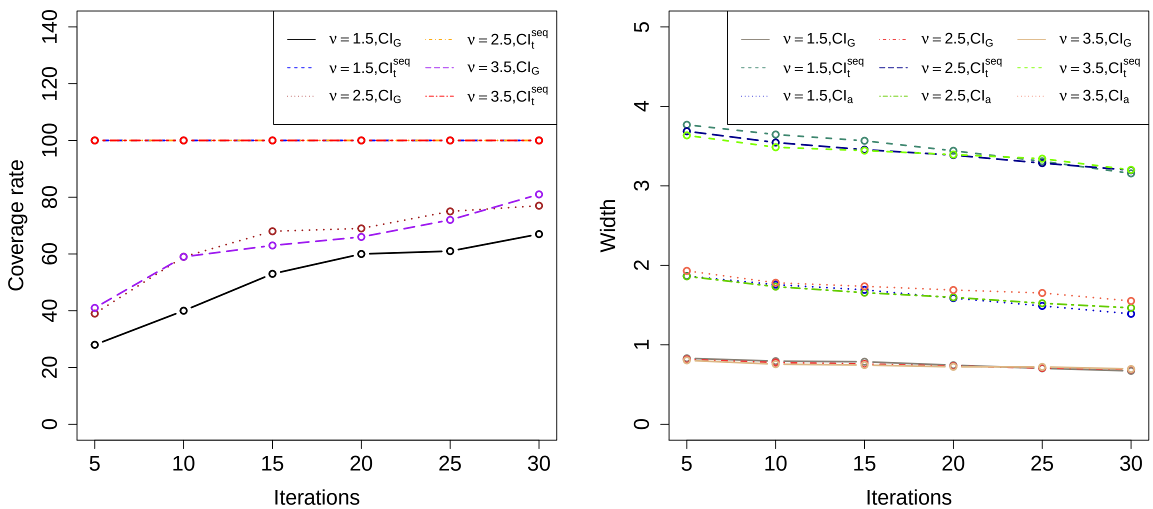

Figure 1: Coverage rates of $CI_t^{\mathbf{seq}}$ and CI_G (Panel 1) and widths of $CI_t^{\mathbf{seq}}$, CI_G, and CI_a (Panel 2) under different scenarios. The nominal confidence level is 95%. The Gaussian process is well specified. More figures of misspecified cases are available in Appendix G.

where the correlation function used in the prediction is a rational quadratic kernel (Rasmussen and Williams, 2006) defined as

$$\Psi_{QK}(x, x') = \left(1 + \frac{\|x - x'\|^2}{2\alpha\phi}\right)^{-\alpha},$$

where $\alpha, \phi > 0$ are parameters. Here we consider $\alpha = 1$ as in Hennig and Schuler (2012). We choose $\phi = 10$ and 20. The results are in Figure G.2 in Appendix G. From Figure G.2 it can be seen that both $CI_t^{\mathbf{seq}}$ and CI_G do not achieve the nominal level, but $CI_t^{\mathbf{seq}}$ is much closer to the nominal level than the naive one. This indicates that if the misspecification is severe, then it has a strong influence of uncertainty quantification, while our method is more robust than the naive one.

6.3 Deterministic Functions

In this subsection, we consider three deterministic functions. Because a deterministic function is no longer random (thus not a Gaussian process), a model misspecification occurs. Furthermore, there is no definition of “confidence interval” for a deterministic function. Therefore, we evaluate the confidence intervals $CI_t^{\mathbf{seq}}$ and CI_G by checking whether they cover the optimal point after certain number of iterations.

In both numerical examples in this subsection, we use a_{UCB} (defined in Section 2.2) as the acquisition function. The iteration numbers we consider are 5, 10, 15, 20, 25, 30. There are three parameters needed to be specified in the Gaussian process regression: the smoothness parameter ν, the constant A_0 in Condition 1, and the variance σ^2. Following the usual approaches in Gaussian process regression (Santner et al., 2003), we impose a specific ν, and estimate A_0 and σ^2 via maximum likelihood estimation based on the initial evaluations of the function values on the initial points. These estimated parameters are used for constructing the prior distribution of underlying functions and evaluation of a_{UCB} in Algorithm 1, and constructing confidence intervals $CI_t^{\mathbf{seq}}$ (by Algorithm 3) and CI_G (by (13)). For the conciseness, we put all the details and numerical results in this subsection to Appendix H, and only list the test functions we used in this section. Here we select three test functions from the Optimization category of Virtual Library of Simulation Experiemnts: Test function and Datasets (http://www.sfu.ca/ ssurjano/optimization.html).

- Modified test function in Higdon (2002):

$$f_1(x) = 1.5\sin(2\pi x/2) - 0.2\sin(2\pi x/2.5) - (x-1)^2/120,$$

where $x \in [0, 8]$. The modification is made because the original function is quite easy to be optimized.

- The test function in Keane et al. (2008):

$$f_3(x) = -(6x-2)^2\sin(12x-4), x \in [0, 1].$$

- The rescaled form of the Branin-Hoo function

(Picheny et al., 2013b):

$$f_4(x) = -\frac{1}{51.95}\Bigg(\left(\bar{x}_2 - \frac{5.1\bar{x}_1^2}{4\pi^2} + \frac{5\bar{x}_1}{\pi} - 6\right)^2 + \left(10 - \frac{10}{8\pi}\right)\cos(\bar{x}_1) - 44.82\Bigg),$$

where $\bar{x}_1 = 15x_1 - 5, \bar{x}_2 = 15x_2$, and $x = (x_1, x_2)^T \in [0,1]^2$. This function has mean zero and variance one.

From Tables H.1-H.3 in Appendix H, we can see that our proposed confidence interval is more robust than the naive confidence interval. Unlike misspecified Gaussian processes, the smoothness plays a more important role in the effectiveness of the confidence intervals. This suggests that when the underlying truth is a deterministic function, this kind of model misspecification has a strong impact on the quality of uncertainty quantification. Robust uncertainty quantification methodologies for deterministic functions will be pursued in the future.

7 CONCLUSIONS AND DISCUSSION

In this work, we propose a novel methodology to construct confidence regions for the outputs given by any Bayesian optimization algorithm with theoretical guarantees. To the best of our knowledge, this is the *first* result of this kind. As a cost of its high flexibility, the confidence regions may be somewhat conservative, because they are constructed based on generic probability inequalities that may not be tight enough. Nevertheless, given the fact that naive methods may be highly permissive, the proposed method can be useful when a conservative approach is preferred, such as in reliability assessments. To improve the power of the proposed method, one needs to seek for more accurate inequalities in a future work. One might also need to derive better error bounds tailored to specific acquisition functions and specify the constants in the upper bounds, and find robust confidence intervals, or other uncertainty quantification methods, which can mitigate the impact of model misspecification. Other possible future extensions include considering more surrogate models, such as Decision trees or Tree Parzen estimators.

Acknowledgements

The authors are grateful to all reviewers for their helpful comments and suggestions. Tuo's research is supported by NSF DMS-1914636 and CCF-1934904.

References

Acerbi, L. and Ji, W. (2017). Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. In *Advances in Neural Information Processing Systems*, pages 1836–1846.

Adler, R. J. and Taylor, J. E. (2009). *Random Fields and Geometry*. Springer Science & Business Media.

Astudillo, R. and Frazier, P. I. (2019). Bayesian optimization of composite functions. *arXiv preprint arXiv:1906.01537*.

Azaïs, J.-M., Bercu, S., Fort, J.-C., Lagnoux, A., and Lé, P. (2010). Simultaneous confidence bands in curve prediction applied to load curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(5):889–904.

Azimi, J., Fern, A., and Fern, X. Z. (2010). Batch Bayesian optimization via simulation matching. In *Advances in Neural Information Processing Systems*, pages 109–117.

Azzimonti, D., Ginsbourger, D., Rohmer, J., and Idier, D. (2019). Profile extrema for visualizing and quantifying uncertainties on excursion regions: application to coastal flooding. *Technometrics*, 61(4):474–493.

Bect, J., Bachoc, F., and Ginsbourger, D. (2019). A supermartingale approach to Gaussian process based sequential design of experiments. *Bernoulli*. To appear.

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

Calvin, J. (2005). One-dimensional global optimization for observations with noise. *Computers & Mathematics with Applications*, 50(1-2):157–169.

Calvin, J. M. (1997). Average performance of a class of adaptive algorithms for global optimization. *The Annals of Applied Probability*, 7(3):711–730.

Contal, E., Buffoni, D., Robicquet, A., and Vayatis, N. (2013). Parallel Gaussian process optimization with upper confidence bound and pure exploration. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 225–240. Springer.

Desautels, T., Krause, A., and Burdick, J. W. (2014). Parallelizing exploration-exploitation tradeoffs in Gaussian process bandit optimization. *Journal of Machine Learning Research*, 15(1):3873–3923.

Diaz, G. I., Fokoue-Nkoutche, A., Nannicini, G., and Samulowitz, H. (2017). An effective algorithm for hyperparameter optimization of neural networks. *IBM Journal of Research and Development*, 61(4/5):9–1.

Driscoll, M. F. (1973). The reproducing kernel Hilbert space structure of the sample paths of a Gaussian process. *Probability Theory and Related Fields*, 26(4):309–316.

Forrester, A., Sobester, A., and Keane, A. (2008). *Engineering Design via Surrogate Modelling: A Practical Guide*. John Wiley & Sons.

Frazier, P. I. (2018). A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*.

Frazier, P. I. and Wang, J. (2016). Bayesian optimization for materials design. In *Information Science for Materials Discovery and Design*, pages 45–75.

Häse, F., Roch, L. M., Kreisbeck, C., and Aspuru-Guzik, A. (2018). Phoenics: A Bayesian optimizer for chemistry. *ACS Central Science*, 4(9):1134–1145.

Hennig, P. and Schuler, C. J. (2012). Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(6).

Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. (2014). Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems*, pages 918–926.

Higdon, D. (2002). Space and space-time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues*, pages 37–56. Springer.

Huang, D., Allen, T. T., Notz, W. I., and Zeng, N. (2006). Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of Global Optimization*, 34(3):441–466.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492.

Keane, A., Forrester, A., and Sobester, A. (2008). *Engineering design via surrogate modelling: a practical guide*. American Institute of Aeronautics and Astronautics, Inc.

Klein, A., Falkner, S., Bartels, S., Hennig, P., and Hutter, F. (2017). Fast Bayesian optimization of machine learning hyperparameters on large datasets. In *Artificial Intelligence and Statistics*, pages 528–536.

Kuriki, S., Wynn, H. P., et al. (2019). Optimal experimental design that minimizes the width of simultaneous confidence bands. *Electronic Journal of Statistics*, 13(1):1099–1134.

Kushner, H. J. (1964). A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering*, 86(1):97–106.

Marco, A., Berkenkamp, F., Hennig, P., Schoellig, A. P., Krause, A., Schaal, S., and Trimpe, S. (2017). Virtual vs. real: Trading off simulations and physical experiments in reinforcement learning with Bayesian optimization. In *2017 IEEE International Conference on Robotics and Automation*, pages 1557–1563.

Mockus, J., Tiesis, V., and Zilinskas, A. (1978). The application of Bayesian methods for seeking the extremum. *Towards Global Optimization*, 2(117-129).

Negoescu, D. M., Frazier, P. I., and Powell, W. B. (2011). The knowledge-gradient algorithm for sequencing experiments in drug discovery. *INFORMS Journal on Computing*, 23(3):346–363.

Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*, volume 63. SIAM.

Picheny, V., Ginsbourger, D., Richet, Y., and Caplin, G. (2013a). Quantile-based optimization of noisy computer experiments with tunable precision. *Technometrics*, 55(1):2–13.

Picheny, V., Wagner, T., and Ginsbourger, D. (2013b). A benchmark of kriging-based infill criteria for noisy optimization. *Structural and Multidisciplinary Optimization*, 48(3):607–626.

Pollard, D. (1990). Empirical processes: Theory and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages i–86. JSTOR.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Ryzhov, I. O. (2016). On the convergence rates of expected improvement methods. *Operations Research*, 64(6):1515–1528.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer Science & Business Media.

Scott, W., Frazier, P., and Powell, W. (2011). The correlated knowledge gradient for simulation optimization of continuous parameters using Gaussian process regression. *SIAM Journal on Optimization*, 21(3):996–1026.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175.

Shahriari, B., Wang, Z., Hoffman, M. W., Bouchard-Côté, A., and de Freitas, N. (2014). An entropy search portfolio for Bayesian optimization. *Stat*, 1050:18.

Sniekers, S. and van der Vaart, A. (2015). Credible sets in the fixed design model with Brownian motion prior. *Journal of Statistical Planning and Inference*, 166:78–86.

Solomou, A., Zhao, G., Boluki, S., Joy, J. K., Qian, X., Karaman, I., Arróyave, R., and Lagoudas, D. C. (2018). Multi-objective Bayesian materials discovery: Application on the discovery of precipitation strengthened niti shape memory alloys through micromechanical modeling. *Materials & Design*, 160:810–827.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *the 27th International Conference on Machine Learning*.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media.

Steinwart, I. (2019). Convergence types and rates in generic Karhunen-Loève expansions with applications to sample path properties. *Potential Analysis*, 51(3):361–395.

Stocki, R. (2005). A method to improve design reliability using optimal latin hypercube sampling. *Computer Assisted Mechanics and Engineering Sciences*, 12(4):393.

Talagrand, M. (1996). Majorizing measures: the generic chaining. *The Annals of Probability*, 24(3):1049–1103.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

Vazquez, E. and Bect, J. (2010). Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *Journal of Statistical Planning and inference*, 140(11):3088–3095.

Wang, W., Tuo, R., and Wu, C. F. J. (2020). On prediction properties of kriging: Uniform error bounds and robustness. *Journal of the American Statistical Association*, 115(530):920–930.

Wendland, H. (2004). *Scattered Data Approximation*, volume 17. Cambridge University Press.

Wilson, A., Fern, A., and Tadepalli, P. (2014). Using trajectory data to improve Bayesian optimization for reinforcement learning. *Journal of Machine Learning Research*, 15(1):253–282.

Wu, J. and Frazier, P. (2016). The parallel knowledge gradient method for batch Bayesian optimization. In *Advances in Neural Information Processing Systems*, pages 3126–3134.

Wu, J., Poloczek, M., Wilson, A. G., and Frazier, P. (2017). Bayesian optimization with gradients. In *Advances in Neural Information Processing Systems*, pages 5267–5278.

Yang, Y., Shang, Z., and Cheng, G. (2017). Non-asymptotic theory for nonparametric testing. *arXiv preprint arXiv:1702.01330*.

Yarotsky, D. (2013). Univariate interpolation by exponential functions and Gaussian RBFs for generic sets of nodes. *Journal of Approximation Theory*, 166:163–175.

Yoo, W. W., Ghosal, S., et al. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *The Annals of Statistics*, 44(3):1069–1102.

Supplementary Materials: Uncertainty Quantification for Bayesian Optimization

A INEQUALITIES FOR GAUSSIAN PROCESSES

In this section, we review some inequalities on the maximum of a Gaussian process. Let $G(x)$ be a separable zero-mean Gaussian process with $x \in \Gamma$. Define the metric on Γ by

$$\mathfrak{d}_g(G(x_1), G(x_2)) = \sqrt{\mathbb{E}(G(x_1) - G(x_2))^2}.$$

The ϵ-covering number of the metric space $(\Gamma, \mathfrak{d}_g)$, denoted as $N(\epsilon, \Gamma, \mathfrak{d}_g)$, is the minimum integer N so that there exist N distinct balls in $(\Gamma, \mathfrak{d}_g)$ with radius ϵ, and the union of these balls covers Γ. Let D be the diameter of Γ with respect to the metric $\mathfrak{d}_g$. The supremum of a Gaussian process is closely tied to a quantity called the *entropy integral*, defined as

$$\int_0^{D/2} \sqrt{\log N(\epsilon, \Gamma, \mathfrak{d}_g)} d\epsilon. \tag{A.1}$$

For detailed discussion of entropy integral, we refer to Adler and Taylor (2009).

Lemma A.1 provides an upper bound on the expectation of the maximum value of a Gaussian process, which is Theorem 1.3.3 of Adler and Taylor (2009). Note that the right-hand side of (A.2) is an upper bound of Talagrand's majorizing measure (Talagrand, 1996). Talagrand's bound, albeit sharper, is not numerically tractable in the current context, because the covariance function of the GP regression posterior is highly nonstationary and complicated. Therefore, we apply Lemma A.1 in our proofs.

Lemma A.1. *Let $G(x)$ be a separable zero-mean Gaussian process with x lying in a $\mathfrak{d}_g$-compact set Γ, where $\mathfrak{d}_g$ is the metric. Let N be the ϵ-covering number. Then there exists a universal constant η such that*

$$\mathbb{E}\left(\sup_{x\in\Gamma} G(x)\right) \leq \eta \int_0^{D/2} \sqrt{\log N(\epsilon, \Gamma, \mathfrak{d}_g)} d\epsilon. \tag{A.2}$$

Lemma A.2, which is Theorem 2.1.1 of Adler and Taylor (2009), presents a concentration inequality.

Lemma A.2. *Let G be a separable Gaussian process on a $\mathfrak{d}_g$-compact Γ with mean zero, then for all $u > 0$,*

$$\mathbb{P}\left(\sup_{x\in\Gamma} G(x) - \mathbb{E}(\sup_{x\in\Gamma} G(x)) > u\right) \leq e^{-u^2/2\sigma_\Gamma^2}, \tag{A.3}$$

where $\sigma_\Gamma^2 = \sup_{x\in\Gamma} \mathbb{E}G(x)^2$.

Theorem A.1 is a slightly strengthened version of Theorem 1 of Wang et al. (2020). Its proof, in Section E, is based on Lemmas A.1-A.2 and some machinery from scattered data approximation Wendland (2004).

Theorem A.1. *Suppose Condition 1 holds. Let $\mu(x)$ and $\sigma(x)$ be as in (2) and (3), respectively, and $D_\Omega = diam(\Omega)$ be the Euclidean diameter of Ω. Then for any $u > 0$, and any closed deterministic subset $A \subset \Omega$, with probability at least $1 - \exp\{-u^2/(2\sigma_A^2)\}$, the kriging prediction error has the upper bound*

$$\sup_{x\in A} Z(x) - \mu(x) \leq \eta_1 \sigma_A \sqrt{p(1 \vee \log(A_0 D_\Omega))} \sqrt{\log(e\sigma/\sigma_A)} + u, \tag{A.4}$$

where A_0 is defined in Condition 1, η_1 is a universal constant, and $\sigma_A = \sup_{x\in A} \sigma(x)$.

B PROOF OF THEOREM 1

We proof Theorem 1 by partitioning Ω into subregions, and applying Theorem A.1 on each of them. Let $\Omega_i = \{x \in \Omega | \sigma e^{-i} \leqslant \sigma(x) \leqslant \sigma e^{-i+1}\}$, for $i = 1, \ldots$. Let $\sigma_i = \sup_{x\in\Omega_i} \sigma(x)$.

Take $\eta_2 = \eta_1\sqrt{2}e$. By Theorem A.1, we have

$$
\begin{aligned}
& P\left(\sup_{x\in\Omega} \frac{Z(x)-\mu(x)}{\sigma(x)\log^{1/2}(e\sigma/\sigma(x))} > \eta_2\sqrt{p(1\vee\log(A_0D_\Omega))} + u\right) \\
\leqslant & \sum_{i=1}^{\infty} P\left(\sup_{x\in\Omega_i} \frac{Z(x)-\mu(x)}{\sigma(x)\log^{1/2}(e\sigma/\sigma(x))} > \eta_2\sqrt{p(1\vee\log(A_0D_\Omega))} + u\right) \\
\leqslant & \sum_{i=1}^{\infty} P\left(\sup_{x\in\Omega_i} Z(x)-\mu(x) > (\eta_2\sqrt{p(1\vee\log(A_0D_\Omega))} + u)\sigma e^{-i}\sqrt{i}\right) \\
\leqslant & \sum_{i=1}^{\infty} P\left(\sup_{x\in\Omega_i} Z(x)-\mu(x) > (\eta_2\sqrt{p(1\vee\log(A_0D_\Omega))} + u)\sigma_i\log^{1/2}(e\sigma/\sigma_i)/(\sqrt{2}e)\right) \\
\leqslant & \sum_{i=1}^{\infty} \exp\left\{-u^2\log(e\sigma/\sigma_i)/(4e^2)\right\} \\
\leqslant & \sum_{i=1}^{\infty} \exp\left\{-iu^2/(4e^2)\right\} = \frac{\exp\left\{-u^2/(4e^2)\right\}}{1-\exp\left\{-u^2/(4e^2)\right\}},
\end{aligned}
$$

which, together with the fact that $M \geq 0$, implies the following upper bound of $\mathbb{E}M$

$$
\begin{aligned}
\mathbb{E}M &= \int_0^\infty \mathbb{P}(M > x)dx \\
&\leq \left(\int_0^{\eta_2\sqrt{p(1\vee\log(A_0D_\Omega))}+1} + \int_{\eta_2\sqrt{p(1\vee\log(A_0D_\Omega))}+1}^\infty\right)\mathbb{P}(M > x)dx \\
&\leq \eta_2\sqrt{p(1\vee\log(A_0D_\Omega))} + 1 + \int_1^\infty \frac{2\exp\left\{-x^2/(4e^2)\right\}}{1-\exp\left\{-x^2/(4e^2)\right\}}dx \\
&\leq C_0\sqrt{p(1\vee\log(A_0D_\Omega))}.
\end{aligned}
$$

To access the tail probability, we note that $M - \mathbb{E}M$ is also a Gaussian process with mean zero. Thus by Lemma A.2, we have

$$\mathbb{P}(M - \mathbb{E}M > t) \leq e^{-t^2/2\sigma_M^2},$$

where

$$\sigma_M^2 = \sup_{x\in\Omega} \mathbb{E}\frac{(Z(x)-\mu(x))^2}{\sigma(x)^2\log(e\sigma/\sigma(x))} \leq 1.$$

Hence, we complete the proof.

C INDEPENDENCE IN SEQUENTIAL GAUSSIAN PROCESS MODELING

The proof of Theorem 2 relies on certain independence properties of sequential Gaussian process modeling shown in Lemmas C.1-C.2. First we introduce some notation. For an arbitrary function f, and $X = (x_1, \ldots, x_n)$, define $f(X) = (f(x_1), \ldots, f(x_n))^T$, and

$$\mathcal{I}_{\Psi,X}f(x) = r^T(x)K^{-1}f(X), \tag{C.1}$$

where $r = (\Psi(x - x_1), \ldots, \Psi(x - x_n))^T, K = (\Psi(x_j - x_k))_{jk}$. For notational convenience, we define $\mathcal{I}_{\Psi,\varnothing}f = 0$.

Lemma C.1. *Let Z be a stationary Gaussian process with mean zero and correlation function Ψ. For two sets of scattered points $X' \subset X = (x_1, \ldots, x_n)$, we have*

$$Z - \mathcal{I}_{\Psi,X'}Z = (Z - \mathcal{I}_{\Psi,X}Z) + \mathcal{I}_{\Psi,X}(Z - \mathcal{I}_{\Psi,X'}Z). \tag{C.2}$$

In addition, if X and X' are deterministic sets, then the residual $Z - \mathcal{I}_{\Psi,X}Z$ and the vector of observed data $(Z(x_1), \ldots, Z(x_n))^T$ are mutually independent Gaussian process and vector, respectively.

Proof. It is easily seen that $\mathcal{I}_{\Psi,X}$ and $\mathcal{I}_{\Psi,X'}$ are linear operators and $\mathcal{I}_{\Psi,X'}\mathcal{I}_{\Psi,X} = \mathcal{I}_{\Psi,X}$, which implies (C.2).

The residual $Z - \mathcal{I}_{\Psi,X}Z$ is a Gaussian process because $\mathcal{I}_{\Psi,X}$ is linear. The independence between the Gaussian process and the vector can be proven by calculation the covariance

$$\begin{aligned}
& \operatorname{Cov}(Z(x') - \mathcal{I}_{\Psi,X'}Z(x'), Z(X)) \\
= \; & \operatorname{Cov}(Z(x') - r^T(x')K^{-1}Z(X), Z(X)) \\
= \; & r(x') - r(x') = 0,
\end{aligned}$$

which completes the proof. □

Lemma C.2. *For any instance algorithm of Bayesian optimization, the following statements are true.*

1. *Conditional on $\mathcal{F}_{n-1}$ and X_n, the residual process $Z(\cdot) - \mu_n(\cdot)$ is independent of $\mathcal{F}_n$.*
2. *Conditional on $\mathcal{F}_n$, the residual process $Z(\cdot) - \mu_n(\cdot)$ is a Gaussian process with same law as $Z'(\cdot) - \mathcal{I}_{\Psi,X_{1:n}}Z'(\cdot)$, where Z' is an independent copy of Z.*

Proof. We use induction on n. For $n = 1$, the desired results are direct consequences of Lemma C.1, because the design set is suppressed conditional on $\mathcal{F}_0$.

Now suppose that we have proven already the assertion for n and want to conclude it for $n+1$. First, we invoke the decomposition given by Lemma C.1 to have

$$Z' - \mathcal{I}_{\Psi,X_{1:n}}Z' = (Z' - \mathcal{I}_{\Psi,X_{1:(n+1)}}Z') + \mathcal{I}_{\Psi,X_{1:(n+1)}}(Z' - \mathcal{I}_{\Psi,X_{1:n}}Z'). \tag{C.3}$$

Because $\mu_n = \mathcal{I}_{\Psi,X_{1:n}}Z$, we also have

$$Z - \mu_n = (Z - \mu_{n+1}) + \mathcal{I}_{\Psi,X_{1:(n+1)}}(Z - \mu_n). \tag{C.4}$$

By the inductive hypothesis, $Z - \mu_n$ has the same law as $Z' - \mathcal{I}_{\Psi,X_{1:n}}Z'$ conditional on $\mathcal{F}_n$, denoted by $Z - \mu_n \overset{\mathrm{d}}{=} Z' - \mathcal{I}_{\Psi,X_{1:n}}Z' | \mathcal{F}_n$. Our assumption that X_{n+1} is independent of (Z, Z') conditional on $\mathcal{F}_n$ implies that X_{n+1} is independent of $(Z - \mu_n, Z' - \mathcal{I}_{\Psi,X_{1:n}}Z')$ as well. Thus,

$$Z - \mu_n \overset{\mathrm{d}}{=} Z' - \mathcal{I}_{\Psi,X_{1:n}}Z' | \mathcal{F}_n, X_{n+1}.$$

Clearly, this equality in distribution is preserved by acting $\mathcal{I}_{\Psi,X_{1:(n+1)}}$ on both sides, which implies

$$\left(Z - \mu_n, \mathcal{I}_{\Psi,X_{1:(n+1)}}(Z - \mu_n)\right) \overset{\mathrm{d}}{=} \left(Z' - \mathcal{I}_{\Psi,X_{1:n}}Z', \mathcal{I}_{\Psi,X_{1:(n+1)}}(Z' - \mathcal{I}_{\Psi,X_{1:n}}Z')\right) | \mathcal{F}_n, X_{n+1}.$$

Incorporating the above equation with (C.3) and (C.4) yields

$$(Z - \mu_{n+1}, Z - \mu_n)) \overset{\mathrm{d}}{=} \left(Z' - \mathcal{I}_{\Psi,X_{1:(n+1)}}Z', Z' - \mathcal{I}_{\Psi,X_{1:n}}Z'\right) | \mathcal{F}_n, X_{n+1}. \tag{C.5}$$

Now we consider the vectors $V := Z(X_{n+1}) - \mu_n(X_{n+1})$ and $V' = Z'(X_{n+1}) - \mathcal{I}_{\Psi,X_{1:n}}Z'(X_{n+1})$. Then (C.5) implies

$$(Z - \mu_{n+1}, V)) \overset{\mathrm{d}}{=} \left(Z' - \mathcal{I}_{\Psi,X_{1:(n+1)}}Z', V'\right) | \mathcal{F}_n, X_{n+1}. \tag{C.6}$$

Because V' consists of observed data, we can apply Lemma C.1 to obtain that, conditional on $\mathcal{F}_n$ and X_{n+1}, $Z' - \mathcal{I}_{\Psi,X_{1:(n+1)}}Z'$ is independent of V', which, together with (C.6), implies that $Z - \mu_{n+1}$ and V are independent conditional on $\mathcal{F}_n$ and X_{n+1}. Because $\mu_n(X_{n+1})$ is measurable with respect to the σ-algebra generated by $\mathcal{F}_n$ and X_{n+1}, we obtain that $Z - \mu_{n+1}$ is independent of $Z(X_{n+1})$ conditional on $\mathcal{F}_n$ and X_{n+1}, which proves Statement 1. Combining Statement 1 and (C.5) yields Statement 2. □

D PROOF OF THEOREM 2

The law of total probability implies

$$\begin{aligned}
& \mathbb{P}(M_T - C\sqrt{p(1\vee\log(A_0D_\Omega))} > t) \\
= & \sum_{i=n}^{\infty}\mathbb{P}(M_T - C\sqrt{p(1\vee\log(A_0D_\Omega))} > t|T=n)\mathbb{P}(T=n) \\
= & \sum_{n=1}^{\infty}\mathbb{P}(M_n - C\sqrt{p(1\vee\log(A_0D_\Omega))} > t|T=n)\mathbb{P}(T=n) \\
= & \sum_{n=1}^{\infty}\mathbb{E}\left\{\mathbb{P}(M_n - C\sqrt{p(1\vee\log(A_0D_\Omega))} > t|\mathcal{F}_n)\Big|T=n\right\}\mathbb{P}(T=n),
\end{aligned}$$

where the last equality follows from the fact that $\{T=n\}\in\mathcal{F}_n$, namely, T is a stopping time. Clearly, the desired results are proven if we can show $\mathbb{P}(M_n - C\sqrt{p(1\vee\log(A_0D_\Omega))} > t|\mathcal{F}_n) < e^{-t^2/2}$. Now we resort to part 2 of Lemma C.2, which states that conditional on $\mathcal{F}_n$, $Z(\cdot)-\mu_n(\cdot)$ is identical in law to its independent copy $Z'(\cdot)-\mathcal{I}_{\Psi,X_{1:n}}Z'(\cdot)$. Although the event $\{M_n - C\sqrt{p(1\vee\log(A_0D_\Omega))} > t\}$ looks complicated, it is measurable with respect to $Z(\cdot)-\mu_n(\cdot)$. Thus, we arrive at

$$\begin{aligned}
& \mathbb{P}(M_n - C\sqrt{p(1\vee\log(A_0D_\Omega))} > t|\mathcal{F}_n) \\
= & \mathbb{P}\left(\sup_{x\in\Omega}\frac{Z'(x)-\mathcal{I}_{\Phi,X_{1:n}}Z'(x)}{\sigma_n(x)\log^{1/2}(e\sigma/\sigma_n(x))} - C\sqrt{p(1\vee\log(A_0D_\Omega))} > t|\mathcal{F}_n\right).
\end{aligned} \tag{D.1}$$

Because Z' is independent of Z, the part of conditioning with respect to $Z(X_{1:n})$ in (D.1) has no effect on Z'. The only thing that matters is the effect of the conditioning on the design points $X_{1:n}$. Hence, (D.1) is reduced to

$$\mathbb{P}\left(\sup_{x\in\Omega}\frac{Z'(x)-\mathcal{I}_{\Phi,X_{1:n}}Z'(x)}{\sigma_n(x)\log^{1/2}(e\sigma/\sigma_n(x))} - C\sqrt{p(1\vee\log(A_0D_\Omega))} > t|X_{1:n}\right). \tag{D.2}$$

Clearly, we can regard the points $X_{1:n}$ in the formula above as a fixed design. Then the probability (D.2) is bounded above by $e^{-t^2/2}$ as asserted by Corollary 1.

E PROOF OF THEOREM A.1

This proof is similar to Theorem 1 of Wang et al. (2020) but with a few technical improvements.

Because $\mu(x)$ is a linear combination of $Z(x_i)$'s, $\mu(x)$ is also a Gaussian process. The main idea of the proof is to invoke a maximum inequality for Gaussian processes, which states that the supremum of a Gaussian process is no more than a multiple of the integral of the covering number with respect to its natural distance $\mathfrak{d}$. See Adler and Taylor (2009); van der Vaart and Wellner (1996) for related discussions.

Let $g(x)=Z(x)-\mu(x)$. For any $x,x'\in A$, because A is deterministic, we have

$$\begin{aligned}
\mathfrak{d}(x,x')^2 = & \mathbb{E}(g(x)-g(x'))^2 \\
= & \mathbb{E}(Z(x)-\mu(x)-(Z(x')-\mu(x')))^2 \\
= & \sigma^2(\Psi(x-x)-r^T(x)K^{-1}r(x)+\Psi(x'-x')-r^T(x')K^{-1}r(x') \\
& -2[\Psi(x-x')-r^T(x')K^{-1}r(x)]),
\end{aligned}$$

where $r(\cdot)=(\Psi(\cdot-x_1),\ldots,\Psi(\cdot-x_n))^T$, $K=(\Psi(x_j-x_k))_{jk}$.

The rest of our proof consists of the following steps. In step 1, we bound the covering number $N(\epsilon,A,\mathfrak{d})$. Next we bound the diameter D. In step 3, we obtain a bound for the entropy integral. In the last step, we invoke Lemmas A.1 and A.2 to obtain the desired results.

Step 1: Bounding the covering number

Let $h(\cdot) = \Psi(x - \cdot) - \Psi(x' - \cdot)$. It can verified that

$$\mathfrak{d}(x, x')^2 = -\sigma^2[h(x') - \mathcal{I}_{\Psi,X} h(x')] + \sigma^2[h(x) - \mathcal{I}_{\Psi,X} h(x)].$$

By Theorem 11.4 of Wendland (2004),

$$\mathfrak{d}(x, x')^2 \leq 2\sigma^2(\sigma_A/\sigma \|h\|_{\mathcal{N}_\Psi(\mathbf{R}^d)}) = 2\sigma\sigma_A \|h\|_{\mathcal{N}_\Psi(\mathbf{R}^d)}, \tag{E.1}$$

where

$$\sigma_A^2 = \sup_{x \in A} \sigma(x)^2 = \sigma^2 \sup_{x \in A}(\Psi(x - x) - r^T(x)K^{-1}r(x)).$$

Denote the Euclidean norm by $\|\cdot\|$. Then, by the definition of the spectral density and the mean value theorem, we have

$$\begin{aligned}
\|h\|^2_{\mathcal{N}_\Psi(\mathbf{R}^d)} &= \Psi(x - x) - 2\Psi(x' - x) + \Psi(x' - x') \\
&= 2\int_{\mathbf{R}^d} (1 - \cos((x - x')^T \omega))\tilde{\Psi}(\boldsymbol{\omega}) d\boldsymbol{\omega} \\
&\leq \left(2\int_{\mathbf{R}^d} \|\omega\| \tilde{\Psi}(\boldsymbol{\omega}) d\boldsymbol{\omega}\right) \|x - x'\| \\
&\leq 2A_0 \|x - x'\|,
\end{aligned} \tag{E.2}$$

where the last inequality follows from the fact that $\|\omega\| \leq \|\omega\|_1$. Combining (E.1) and (E.2) yields

$$\mathfrak{d}(x, x')^2 \leq 2A_0^{1/2} \sigma\sigma_A \|x - x'\|^{1/2}. \tag{E.3}$$

Therefore, the covering number is bounded above by

$$\log N(\epsilon, A, \mathfrak{d}) \leq \log N\left(\frac{\epsilon^4}{4A_0\sigma^2\sigma_A^2}, A, \|\cdot\|\right). \tag{E.4}$$

The right side of (E.4) involves the covering number of a Euclidean ball, which is well understood in the literature. See Lemma 4.1 of Pollard (1990). This result leads to the bound

$$\log N(\epsilon, A, \mathfrak{d}) \leq p \log\left(\frac{48A_0 D_A \sigma^2\sigma_A^2}{\epsilon^4} + 1\right) \leq p \log\left(\frac{48A_0 D_\Omega \sigma^2\sigma_A^2}{\epsilon^4} + 1\right), \tag{E.5}$$

where $D_A = \operatorname{diam}(A)$ and $D_\Omega = \operatorname{diam}(\Omega)$ are the Euclidean diameter of A and Ω, respectively.

Step 2: Bounding the diameter D

Define the diameter under metric $\mathfrak{d}$ by $D = \sup_{x,x' \in A} \mathfrak{d}(x, x')$. For any $x, x' \in A$,

$$\begin{aligned}
\mathfrak{d}(x, x')^2 =& \mathbb{E}(g(x) - g(x'))^2 \leq 4 \sup_{x \in A} \mathbb{E}(g(x))^2 \\
=& 4 \sup_{x \in A} \mathbb{E}(Z(x) - \mathcal{I}_{\Psi,\mathbf{X}} Z(x))^2 \\
=& 4\sigma^2 \sup_{x \in A}(\Psi(x - x) - r^T(x)K^{-1}r(x)) = 4\sigma_A^2.
\end{aligned} \tag{E.6}$$

Thus we conclude that

$$D \leq 2\sigma_A. \tag{E.7}$$

Step 3: Bounding the entropy integral

By (E.5) and (E.7),

$$\begin{aligned}\int_0^{D/2}\sqrt{\log N(\epsilon,A,\mathfrak{d})}d\epsilon &\leq \int_0^{\sigma_A}\sqrt{p\log\left(\frac{48A_0D_\Omega\sigma^2\sigma_A^2}{\epsilon^4}+1\right)}d\epsilon\\ &\leq\left(\int_0^{\sigma_A}d\epsilon\right)^{1/2}\left(\int_0^{\sigma_A}p\log\left(\frac{48A_0D_\Omega\sigma^2\sigma_A^2}{\epsilon^4}+1\right)d\epsilon\right)^{1/2}\\ &=\left(\int_0^{\sigma_A}d\epsilon\right)^{1/2}\left(\sigma\int_0^{\sigma_A/\sigma}p\log\left(\frac{48A_0D_\Omega\sigma_A^2}{u^4\sigma^2}+1\right)du\right)^{1/2}\\ &\leq\sigma_A^{1/2}\left(\sigma\int_0^{\sigma_A/\sigma}p\log\left(\frac{48A_0D_\Omega\sigma_A^2}{u^4\sigma^2}+\frac{\sigma_A^2}{u^4\sigma^2}\right)du\right)^{1/2}\\ &\leq\sqrt{2p}\sigma_A\sqrt{\log(e^2\sqrt{1+48A_0D_\Omega}\sigma/\sigma_A))}\\ &\leq\sqrt{4p}\sigma_A\sqrt{\log(e\sqrt{1+48A_0D_\Omega})}\sqrt{\log(e\sigma/\sigma_A))}\\ &\leq c\sqrt{p(1\vee\log(A_0D_\Omega))}\sigma_A\sqrt{\log(e\sigma/\sigma_A))},\end{aligned}\tag{E.8}$$

where $c=\sqrt{6\log(7e)}$.

Step 4: Bounding $\mathbb{P}(\sup_{x\in A}Z(x)-\mu(x)>\eta\int_0^{D/2}\sqrt{\log N(\epsilon,A,\mathfrak{d})}d\epsilon+u)$

By Lemmas A.1 and A.2, we have

$$P\left(\sup_{x\in A}Z(x)-\mu(x)>\eta\int_0^{D/2}\sqrt{\log N(\epsilon,A,\mathfrak{d})}d\epsilon+t\right)\leq e^{-t^2/(2\sigma_A^2)}.\tag{E.9}$$

By plugging (E.8) into (E.9), we obtain the desired inequality with $\eta_1=c\eta$, which completes the proof.

F DETAILS OF CALIBRATING C VIA SIMULATION

An upper bound of the constant C in Theorem 1 can be obtained by examine the proof of Lemma A.1 and Theorem A.1. However, this theoretical upper bound can be too large for practical use. In this section, we consider estimating C via numerical simulation.

According to Part 1 of Theorem 1,

$$C_0\geq\mathbb{E}M/\sqrt{p(1\vee\log(A_0D_\Omega))},$$

where $M=\sup_{x\in\Omega}\frac{Z(x)-\mu(x)}{\sigma(x)\log^{1/2}(e\sigma/\sigma(x))}$, A_0 is as in (1), and D_Ω is the Euclidean diameter of Ω. For a specific Gaussian process, $\mathbb{E}M/\sqrt{p(1\vee\log(A_0D_\Omega))}$ is a constant and can be obtained by Monte Carlo. Let $\mathcal{M}$ be the collection of Gaussian processes satisfying the conditions of Theorem 1. Then

$$C_0=\sup_{M\in\mathcal{M}}\mathbb{E}M/\sqrt{p(1\vee\log(A_0D_\Omega))}=:\sup_{M\in\mathcal{M}}H(M).$$

The idea is to consider various Gaussian processes and find the maximum value of $\mathbb{E}M/\sqrt{p(1\vee\log(A_0D_\Omega))}$. This value can be close to C when we cover a broad range of Gaussian processes.

In the numerical studies, we consider $\Omega=[0,1]^p$ for $p=1,2,3$. We consider different A_0 values to get different A_0D_Ω's. In each Monte Carlo sampling, we approximate M using

$$M_1=\sup_{x\in\Omega_1}\frac{Z(x)-\mu(x)}{\sigma(x)\log^{1/2}(e\sigma/\sigma(x))},$$

where Ω_1 is the first $100, 1000, 2000$ points of the Halton sequence (Niederreiter, 1992) for $p=1,2,3$, respectively. We calculate the average of $M_1/\sqrt{p(1\vee\log(A_0D_\Omega))}$ over all the simulated realizations of each Gaussian process.

Specifically, We simulate 1000 realizations of the Gaussian processes for $p = 1$, 100 realizations for $p = 2, 3$ and consider the following four cases. In Cases 1-3, we use maximin Latin hypercube designs (Santner et al., 2003), and use independent samples from the uniform distribution in Case 4.

Case 1: We consider $p = 1$ with 20 and 50 design points. We consider the Gaussian correlation functions and Matérn correlation functions with $\nu = 1.5, 2.5, 3.5$. The results are presented in Table F.1.

Case 2: We consider $p = 2$ with 20, 50, and 100 design points. We consider the Gaussian correlation functions and product Matérn correlation functions with $\nu = 1.5, 2.5, 3.5$. The results are presented in Table F.2.

Case 3: We consider $p = 3$ with 20, 50, 100 and 500 design points. We consider the product Matérn correlation functions with $\nu = 1.5, 2.5, 3.5$. The results are shown in Table F.3.

Case 4: We consider $p = 2$ with 20, 50, and 100 design points. We consider the product Matérn correlation functions with $\nu = 1.5, 2.5, 3.5$. The results are shown in Table F.4.

Table F.1: Simulation Results of Case 1

	design points	$A_0D_\Omega = 1$	$A_0D_\Omega = 3$	$A_0D_\Omega = 5$	$A_0D_\Omega = 10$	$A_0D_\Omega = 25$
Gaussian	20	0.11640290	0.1978563	0.2450737	0.4542654	0.859318
	50	0.08102775	0.0916648	0.1206034	0.1683377	0.422786
$\nu = 1.5$	20	0.9640650	1.065597	0.9537634	0.9429957	1.0197966
	50	0.9442937	1.009187	0.8981430	0.8331926	0.8372607
$\nu = 2.5$	20	0.7432965	0.8554707	0.7804686	0.8371662	1.0074204
	50	0.7304104	0.8218710	0.7346077	0.6987832	0.7563067
$\nu = 3.5$	20	0.6054239	0.7248086	0.6833789	0.7711124	0.9608837
	50	0.3367513	0.6941391	0.6244660	0.6278185	0.6928741

Table F.2: Simulation Results of Case 2

	design points	$A_0D_\Omega = 1$	$A_0D_\Omega = 3$	$A_0D_\Omega = 5$	$A_0D_\Omega = 10$	$A_0D_\Omega = 25$
Gaussian	20	0.2801128	0.4767259	0.5644628	0.7408401	1.0554507
	50	0.1465512	0.2927036	0.3789438	0.5683807	0.9309326
	100	0.1156139	0.1961319	0.2436626	0.4189444	0.7641615
$\nu = 1.5$	20	0.8106718	0.9528429	0.8748865	0.9365989	1.0894451
	50	0.8114071	0.9299506	0.8568070	0.8576984	0.9964256
	100	0.8137517	0.9108342	0.8224467	0.7951887	0.9168643
$\nu = 2.5$	20	0.6072854	0.7709362	0.7411921	0.8540687	1.0933120
	50	0.6316136	0.7218077	0.7218077	0.7690956	0.9703693
	100	0.5651732	0.6677120	0.6677120	0.7090934	0.8791792
$\nu = 3.5$	20	0.5243251	0.6881401	0.6915576	0.8290974	1.0876019
	50	0.3947094	0.6420423	0.6434791	0.7030224	0.9494486
	100	0.2898865	0.6279639	0.6036111	0.6420049	0.8373886

Table F.3: Simulation Results of Case 3

Cases	$H(M)$
20 design points, $\nu = 1.5$, $A_0D_\Omega = 1$	0.6977030
500 design points, $\nu = 3.5$, $A_0D_\Omega = 5$	0.4961581
100 design points, $\nu = 2.5$, $A_0D_\Omega = 3$	0.6628567
50 design points, $\nu = 1.5$, $A_0D_\Omega = 10$	0.7632713

From Tables F.1-F.4, we find the following patterns:

Table F.4: Simulation Results of Case 4

Cases	$H(M)$
100 design points, $\nu = 3$, $A_0 D_\Omega = 3$	0.6778535
50 design points, $\nu = 1.5$, $A_0 D_\Omega = 1$	0.8144700
20 design points, $\nu = 2.5$, $A_0 D_\Omega = 5$	0.7735112
100 design points, $\nu = 1.5$, $A_0 D_\Omega = 10$	0.8164859

- All numerical values ($H(M)$) in Tables F.1-F.4 are less than 1.10. Only eight entries are greater than one.
- In most scenarios, the obtained values are decreasing in ν. This implies that $H(M)$ is smaller when M is smoother.
- $H(M)$ is not monotonic in $A_0 D_\Omega$, which implies a more complicated function relationship between $H(M)$ and $A_0 D_\Omega$.
- In most scenarios, $H(M)$ decreases as the dimension p increases.
- The obtained values are decreasing in the number of design points.

In summary, the largest $H(M)$ values are observed when the sample size is small, the smoothness is low and the dimension is low. Therefore, we believe that our simulation study covers the largest possible $H(M)$ values and our suggestion of choosing $C_0 = 1$ can be used in most practical situations.

G MORE FIGURES OF NUMERICAL EXPERIMENTS FOR MISSPECIFIED GAUSSIAN PROCESSES

Here we present more figures of numerical experiments when Gaussian process is misspecified, as shown in Figures G.1 and G.2.

H DETAILS OF NUMERICAL EXPERIMENTS FOR DETERMINISTIC FUNCTIONS

Deterministic function 1 The first deterministic function we consider is

$$f_1(x) = 1.5\sin(2\pi x/2) - 0.2\sin(2\pi x/2.5) - (x-1)^2/120,$$

where $x \in [0, 8]$, which is a modification of the function used in Higdon (2002). The modification is made because the original function is quite easy to be optimized. The maximum of f_1 is taken on the point $x^* = 0.520$, and the maximum is $f_1(x^*) = 1.5953$. The initial points are selected as 30 equally spaced points on the interval $[0, 8]$. The results are collected in Table H.1 in Appendix H.

Deterministic function 2 The third deterministic function we consider is the test function in Keane et al. (2008):

$$f_3(x) = -(6x-2)^2\sin(12x-4), x \in [0, 1].$$

The maximum of f_3 is taken on the point $x^* = 0.7575$, with $f_3(x^*) = 6.0207$. The initial points are selected as 30 equally spaced points on the interval $[0, 8]$. We use Ω_1 to approximate Ω, where Ω_1 is a set of grid points with grid length 1/2499 (thus, there are 2500 test points in total).

Deterministic function 3 The fourth deterministic function we consider is the rescaled form of the Branin-Hoo function (Picheny et al., 2013b):

$$f_4(x) = -\frac{1}{51.95}\left(\left(\bar{x}_2 - \frac{5.1\bar{x}_1^2}{4\pi^2} + \frac{5\bar{x}_1}{\pi} - 6\right)^2 + \left(10 - \frac{10}{8\pi}\right)\cos(\bar{x}_1) - 44.82\right),$$

where $\bar{x}_1 = 15x_1 - 5, \bar{x}_2 = 15x_2$, and $x = (x_1, x_2)^T \in [0, 1]^2$. The settings are the same of that in Deterministic function 2.

Table H.1: Simulation Results of Deterministic Function 1. The following abbreviations are used: IN = iteration number, CI = confidence interval. The notation ✓stands for "cover" and ✗stands for "not cover".

	CI	IN = 5	IN = 10	IN = 15	IN = 20	IN = 25	IN = 30
$\nu = 1.5$	$CI_t^{\mathbf{seq}}$	✓	✓	✓	✓	✓	✓
	CI_G	✓	✓	✓	✓	✓	✓
$\nu = 2.5$	$CI_t^{\mathbf{seq}}$	✓	✓	✓	✓	✓	✓
	CI_G	✓	✗	✗	✗	✗	✗
$\nu = 3.5$	$CI_t^{\mathbf{seq}}$	✓	✓	✓	✓	✓	✓
	CI_G	✓	✗	✗	✗	✗	✗
$\nu = 5.5$	$CI_t^{\mathbf{seq}}$	✗	✗	✗	✗	✗	✗
	CI_G	✗	✗	✗	✗	✗	✗

Table H.2: Simulation Results of Deterministic Function 2. The following abbreviations are used: IN = iteration number, CI = confidence interval. The notation ✓stands for "cover" and ✗stands for "not cover".

	CI	IN = 5	IN = 10	IN = 15	IN = 20	IN = 25	IN = 30
$\nu = 1.5$	$CI_t^{\mathbf{seq}}$	✓	✓	✓	✓	✓	✓
	CI_G	✓	✓	✓	✓	✓	✓
$\nu = 2.5$	$CI_t^{\mathbf{seq}}$	✓	✓	✓	✓	✓	✓
	CI_G	✓	✓	✓	✓	✓	✓
$\nu = 4$	$CI_t^{\mathbf{seq}}$	✓	✓	✓	✓	✓	✓
	CI_G	✓	✓	✗	✗	✗	✗
$\nu = 5.5$	$CI_t^{\mathbf{seq}}$	✓	✓	✓	✓	✓	✓
	CI_G	✓	✓	✗	✗	✗	✗

Table H.3: Simulation Results of Deterministic Function 3. The following abbreviations are used: IN = iteration number, CI = confidence interval. The notation ✓stands for "cover" and ✗stands for "not cover".

	CI	IN = 5	IN = 10	IN = 15	IN = 20	IN = 25	IN = 30
$\nu = 1.1$	$CI_t^{\mathbf{seq}}$	✓	✓	✓	✓	✓	✓
	CI_G	✓	✓	✓	✓	✓	✓
$\nu = 2.3$	$CI_t^{\mathbf{seq}}$	✓	✓	✓	✓	✓	✓
	CI_G	✓	✓	✓	✓	✓	✓
$\nu = 2.8$	$CI_t^{\mathbf{seq}}$	✓	✓	✓	✓	✓	✓
	CI_G	✓	✓	✓	✓	✓	✓
$\nu = 4$	$CI_t^{\mathbf{seq}}$	✓	✓	✓	✓	✓	✓
	CI_G	✓	✓	✓	✓	✓	✓

I ILLUSTRATION OF CONFIDENCE REGIONS

We plot confidence regions for one realizations of Gaussian process with smoothness $\nu = 1.5$. The iteration number is 30. The results are shown in Figure I.1. It can be seen from Figure I.1 that our confidence region is more conservative than the naive one.

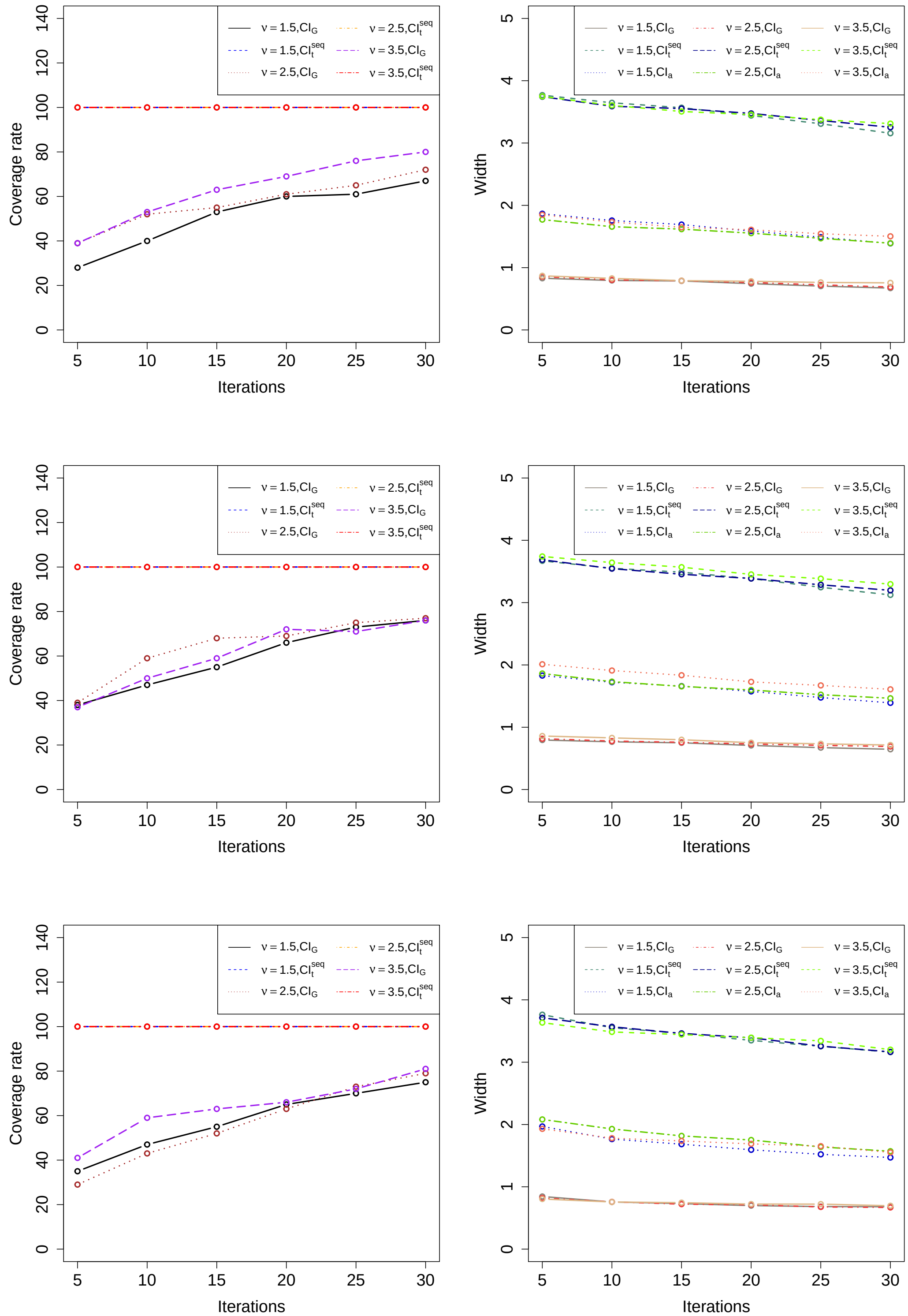

Figure G.1: Coverage rates of $CI_t^{\mathbf{seq}}$ and CI_G (Panels 1, 3, 5, 7) and widths of $CI_t^{\mathbf{seq}}$, CI_G, and CI_a (Panels 2, 4, 6, 8) under different scenarios. The nominal confidence level is 95%. **Panels 1 and 2:** The Gaussian process is well specified. **Panels 1 and 2:** The Gaussian process is misspecified with $\nu_0 = 1.5$. **Panels 3 and 4:** The Gaussian process is misspecified with $\nu_0 = 2.5$. **Panels 5 and 6:** The Gaussian process is misspecified with $\nu_0 = 3.5$.

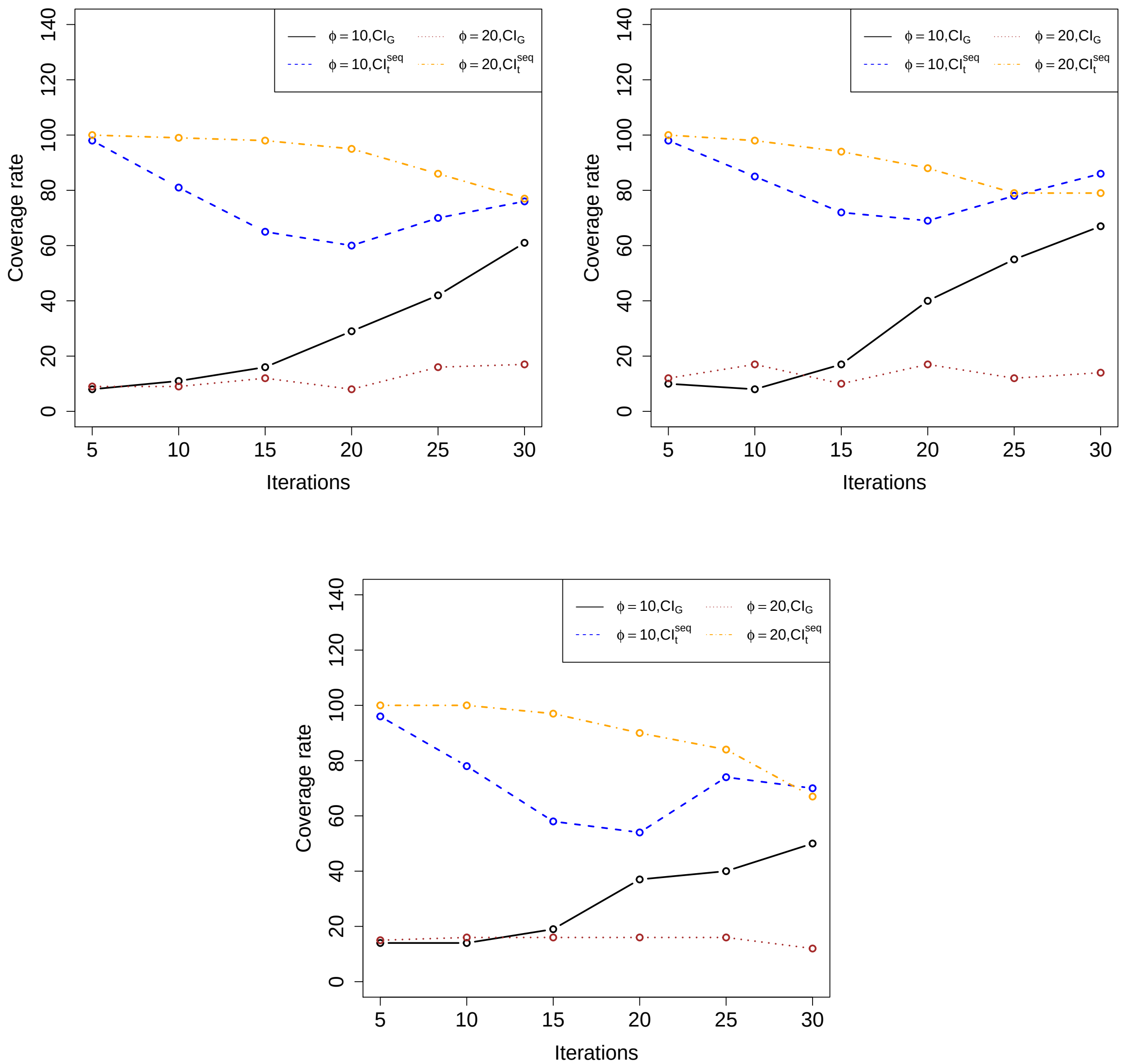

Figure G.2: Coverage rates of $CI_t^{\mathbf{seq}}$ and CI_G under different scenarios, where a rational quadratic correlation function is used for prediction. The nominal confidence level is 95%. The underlying true correlation function is Matérn with smoothness parameter ν_0. **Panel 1:** $\nu_0 = 1.5$. **Panel 2:** $\nu_0 = 2.5$. **Panel 2:** $\nu_0 = 3.5$.

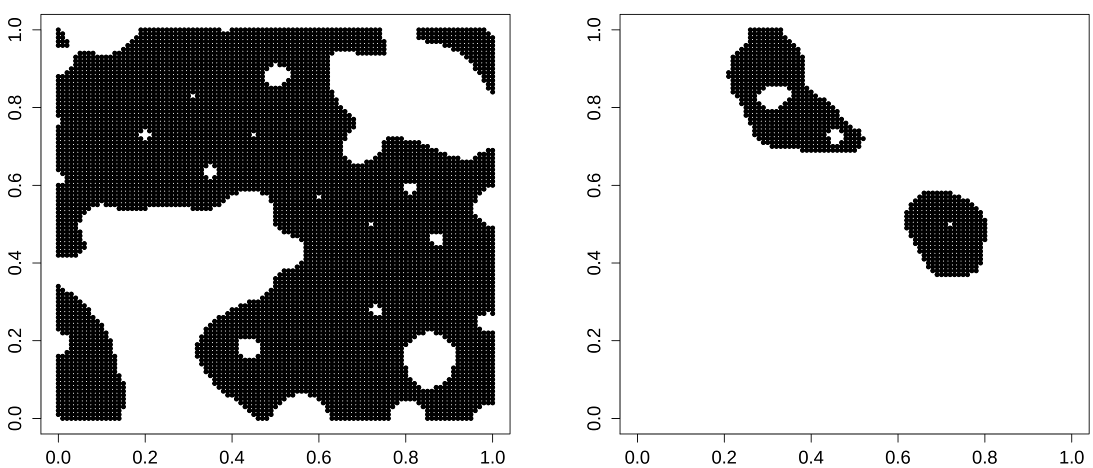

Figure I.1: Confidence region of $CI_t^{\mathbf{seq}}$ (Panel 1) and CI_G (Panel 2). The nominal confidence level is 95%.