UNCERTAINTY QUANTIFICATION FOR BAYESIAN OPTIMIZATION

Anonymous authors
Paper under double-blind review

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 criteria.

1.1 PROBLEM OF INTEREST AND OUR RESULTS

Let \( f \) be an underlying deterministic continuous function over \( \Omega \), 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 \( x_{\text{max}} \) which satisfies \( f(x_{\text{max}}) = \max_{x \in \Omega} f(x) \). 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; Mockus et al., 1978), materials and drug design (Frazier & Wang, 2016; Negoescu et al., 2011; Solomou et al., 2018), chemistry (Häse et al., 2018), deep neural networks (Díaz 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., 2013), Gaussian process upper confidence bound (Azimi et al., 2010; Contal et al., 2013; Desautels et al., 2014; Srinivas et al., 2010), predictive entropy search (Hernández-Lobato et al., 2014), entropy search portfolio (Shahriari et al., 2014), knowledge gradient (Scott et al., 2011; Wu & 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 \( Z \), \( x_{\text{max}} \) and \( f(x_{\text{max}}) \) should also be random. However, the highly nontrivial distributions of \( x_{\text{max}} \) and \( f(x_{\text{max}}) \) make uncertainty quantification rather challenging. Monte Carlo approaches can be employed to compute the posterior distributions of \( x_{\text{max}} \) and \( f(x_{\text{max}}) \), but they are usually computationally expensive, especially when a large number of observations are available.

**Our results.** We develop efficient methods to construct confidence regions of \( x_{\text{max}} \) and \( f(x_{\text{max}}) \) 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 2 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. 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 \( \sigma \) 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 (2005; 1997), Ryzhov (2016), Vazquez & Bect (2010), Yarotsky (2013). These analyses 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, e.g., Bull (2011), Astudillo & Frazier (2019), Yarotsky (2013), where the underlying function \( f \) is assumed to be a deterministic function satisfying pre-specified smoothness conditions.

## 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 \( \sigma^2 \) and correlation function \( \Psi \), i.e., \( \text{Cov}(Z(x), Z(x')) = \sigma^2 \Psi(x - x') \) 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 \( \Psi \), 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 \( \Psi \) the spectral density of \( \Psi \).

Some popular choices of correlation functions and their spectral densities are discussed in Section 3.1. We further assume \( \Psi \) satisfies the following condition. For a vector \( \omega = (\omega_1, \ldots, \omega_p)^T \), define its \( 1 \)-norm as \( \|\omega\|_1 = |\omega_1| + \ldots + |\omega_p| \).

**Condition 1** The correlation function \( \Psi \) has a spectral density, denoted by \( \tilde{\Psi} \), and

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

**Remark 1** The \( 1 \)-norm in Equation [1] can be replaced by the usual Euclidean norm. However, we use the former here because they usually have explicit expressions. See Section 3.1 for details.

**Remark 2** Condition [1] imposes a smoothness condition on the correlation function \( \Psi \), 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 \( \Psi \).
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 the design points. The following results are well-known and can be found in Rasmussen & 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

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

(2)

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

(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 called 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** $n = 1, 2, \ldots, \text{do}$
3: Find $x_{n+1} = \arg \max_{x \in \Omega} a(x; X_n, Y_n)$, 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 $a_{EI}(x; X_n, Y_n) := \mathbb{E}((Z(x) - y_n^*)1(Z(x) - y_n^*)|X_n, Y_n)$, where $1(\cdot)$ is the indicator function, and $y_n^* = \max_{1 \leq i \leq n} f(x_i)$.

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

3. Predictive entropy search (Hernández-Lobato et al. 2014), with the acquisition function $a_{PES}(x; X_n, Y_n) := f^{(n)}(x)$, where $f^{(n)}$ is an approximate simulation via spectral sampling (Lázaro-Gredilla et al. 2010, Rahimi & Recht 2008) from GP($0, \Psi|X_n, Y_n$).

Among the above acquisition functions, $a_{EI}$ and $a_{UCB}$ are deterministic functions of $(x, X_n, Y_n)$, whereas $a_{PES}$ is random because it 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 & Ji 2017). Although stopping criteria plays no role in the analysis of the algorithms’ asymptotic behaviors, they can greatly affect the output uncertainty.
3 Uncertainty Quantification for Bayesian Optimization

In this section, we present our uncertainty quantification methodology for Bayesian optimization in Section 3.1. In Section 3.2, we provide theoretical guarantees for the proposed uncertainty quantification method.

3.1 Methodology

Although the conditional distribution of \( Z(x) \) is simple as shown in Equation 2 and Equation 3, those for \( \max(x) \) and \( Z(\max(x)) \) 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, as presented in Algorithm 2. In the rest of this work, let \( T \) be the number of iterations when an instance of Algorithm 1 stops and \( D_\Omega \) be the diameter of \( \Omega \). 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. We will use the notion \( a \vee b := \max(a, b) \).

Algorithm 2 Confidence regions for \( \max(x) \) and \( f(\max(x)) \)

1: Input: Significance parameter \( t \), data \( X_{1:T}, Y_{1:T} \) collected from an instance of Bayesian optimization algorithm.
2: For any 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

\[
\begin{align*}
\mu_T(x) &= r(x)^T K^{-1} Y_{1:T}, \quad (5) \\
\sigma_T(x) &= \sqrt{\sigma^2 (1 - r(x)^T K^{-1} r(x))}. \quad (6)
\end{align*}
\]

3: Compute

\[
\text{UPPERCL}(x, t, X_{1:T}, Y_{1:T}) = \mu_T(x) + s_T(x) \sqrt{\log(e \sigma / s_T(x)) \left( C \sqrt{p} (1 \vee \log(A_0 D_\Omega)) + t \right)},
\]

where \( A_0 \) is as in Condition 1 and \( C \) is a universal constant.
4: Calculate

\[
CR_t^{\text{seq}} := \left\{ x \in \Omega : \text{UPPERCL}(x, t, X_{1:T}, Y_{1:T}) \geq \max_{1 \leq i \leq m_T} f(x_i) \right\}, \quad (7)
\]

\[
CI_t^{\text{seq}} := \left[ \max_{1 \leq i \leq m_T} f(x_i), \max_{x \in \Omega} \text{UPPERCL}(x, t, X_{1:T}, Y_{1:T}) \right]. \quad (8)
\]

5: Output: The confidence region \( CR_t^{\text{seq}} \) for \( \max(x) \) and the confidence interval \( CI_t^{\text{seq}} \) for \( f(\max(x)) \).

In Section 3.2, we will show that under the condition that \( f \) is a realization of \( Z \), \( CR_t^{\text{seq}} \) and \( CI_t^{\text{seq}} \) are confidence regions of \( \max(x) \) and \( f(\max(x)) \), respectively, with a simultaneous confidence level at least \( 1 - e^{-t^2/2} \), respectively. In particular, to obtain a 95% confidence region, we use \( t = 2.448 \).

Calculating \( A_0 \). For an arbitrary \( \Psi \), calculation of \( A_0 \) in Equation 1 can be challenging. Fortunately, for two most popular correlation functions in one dimension, namely the Gaussian and the Matérn correlation functions (Rasmussen & Williams 2006; Santner et al. 2003), \( A_0 \) can be calculated in closed form. The results are summarized in Table 1.

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
ties of Fourier transform that
\[ \hat{\Psi}(x) = \prod_{i=1}^{P} \hat{\Psi}_i(x_i). \] Let \( X_i \) be a random variable with probability density function \( \Psi_i \). 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 \( \Psi_i \) is either a Gaussian or Matérn correlation function, then \( \mathbb{E}[|X_i|] \)'s can be read from Table 1.

### Calibrating \( C \) via simulation studies.
To use Equation 7 and Equation 8, we need to specify the constant \( C \). In this work, we identify \( C \) by numerical simulations. 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 \( C_0 D_{11} \), 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 1 corresponding to their specific setting, or run similar numerical studies as in Appendix F to calibrate their own \( C \).

### 3.2 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 \( \Psi \). 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 \( n \)th iteration as \( X_n \) and \( Y_n \), respectively. Let \( X_{1:n} \) and \( Y_{1:n} \) be as in Section 3.1. Then, we can write the data we obtain after the \( n \)th iteration as \( I_n = (X_{1:n}, Y_{1:n}) \). Because \( X_{1:n} \) and \( Y_{1:n} \) are random, the data \( I_n \) is associated with the \( \sigma \)-algebra \( \mathcal{F}_n \), defined as the \( \sigma \)-algebra generated by \( (X_{1:n}, Y_{1:n}) \). When the algorithm just starts, no data is gain and we set \( I_0 = \emptyset \). The empty \( I_0 \) is associated with the trivial \( \sigma \)-algebra \( \mathcal{F}_0 \), which consists of only the empty set and the entire probability space. 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. This strategy can be deterministic or random, and may vary at different stages of the process. For example, one can choose initial designs and subsequent sampling points with different strategies. Clearly, such strategy should not depend on unobserved data. After each sampling-evaluation iteration, a stopping criterion is checked 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 & van der Vaart (2015); Yoo et al. (2016); Yang et al. (2017); Kuriki et al. (2019); Azimonti et al. (2019) Azais et al. (2010), investigate uncertainty quantification methods which are not

Table 1: Gaussian and Matérn correlation families, where \( \Gamma(\cdot) \) is the Gamma function and \( K_0(\cdot) \) is the modified Bessel function of the second kind.

<table>
<thead>
<tr>
<th>Correlation family</th>
<th>Gaussian</th>
<th>Matérn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation function</td>
<td>( \exp{-(x/\theta)^2} )</td>
<td>( \frac{1}{\Gamma(\nu)2^{\nu-1}} \left( \frac{2\sqrt{\pi}</td>
</tr>
<tr>
<td>Spectral density</td>
<td>( \frac{\theta}{2\pi} \exp{-(\omega^2 \theta^2)/4} )</td>
<td>( \frac{\Gamma(\nu+1/2)}{\pi^\nu \Gamma(\nu)} \left( \frac{4\nu}{\pi} \right)^\nu (\omega^2 + 4\nu)^{-\nu-1/2} \frac{\sqrt{\pi} \nu}{\sqrt{2\nu-1}}</td>
</tr>
</tbody>
</table>
within the Bayesian optimization or sequential sampling scheme, and cannot be directly applied to quantify the uncertainties of outputs of Bayesian optimization.

3.2.1 Fixed designs

We start with a simpler case, where we choose all the input points before we evaluate \( f \) at any of them. Although sequential samplings are more popular in Bayesian optimization, the fixed designs situation will serve as an important intermediate step to the general problem in Section 3.2.2. Let \( X = \{ x_1, \ldots, x_n \} \) be fixed design points, and \( Y = ( f(x_1), \ldots, f(x_n) )^T \). The confidence region for \( x_{\text{max}} \) is then defined as

\[
CR_t := \left\{ x \in \Omega : \text{UPPERCL}(x, t, X, Y) \geq \max_{1 \leq i \leq n} f(x_i) \right\}.
\]

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

\[
CI_t := \left[ \max_{1 \leq i \leq n} f(x_i), \max_{x \in \Omega} \text{UPPERCL}(x, t, X, Y) \right].
\]

Also, we shall use the convention 0/0 = 0 in all statements in this article related to error bounds. The following theorem states a uniform error bound for Gaussian process regression, which is the first theoretical result of this kind, to the best of our knowledge. The proof of Theorem 1 can be found in Appendix B.

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

1. \( E M \leq C_0 \sqrt{p(1 \vee \log(A_0 D_\Omega))} \), where \( C_0 \) is a universal constant, \( A_0 \) as in Condition 2 and \( D_\Omega = \text{diam}(\Omega) \) is the Euclidean diameter of \( \Omega \).
2. For any \( t > 0 \), \( P(M - EM > t) \leq e^{-t^2/2} \).

In practice, Part 2 of Theorem 1 is hard to use directly because \( EM \) is difficult to calculate accurately. Instead, we can replace \( EM \) 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 \( EM \leq C \sqrt{p(1 \vee \log(A_0 D_\Omega))} \), we have

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

for any \( t > 0 \), where the constants \( A_0 \) and \( D_\Omega \) are the same as those in Theorem 1.

It is worth noting that the probability in Corollary 1 is not a posterior probability. Therefore, the regions given by Equation 9 and Equation 10 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 Equation 9 and \( CI_t \) in Equation 10 are confidence region of \( x_{\text{max}} \) and \( f(x_{\text{max}}) \) with confidence level \( 1 - e^{-t^2/2} \), respectively.

**Corollary 2** Let \( C(\Omega) \) be the space of continuous functions on \( \Omega \) and \( P_Z \) be the law of \( Z \). Then there exists a set \( B \subset C(\Omega) \) so that \( P_Z(B) \geq 1 - e^{-t^2/2} \) and for any \( f \in B \), its maximum point \( x_{\text{max}} \) is contained in \( CR_t \) defined in Equation 9 and \( f(x_{\text{max}}) \) is contained in \( CI_t \) defined in Equation 10.

In practice, the shape of \( CR_t \) can be highly irregular and representing the region of \( CR_t \) can be challenging. If \( \Omega \) is of one or two dimensions, we can choose a fine mesh over \( \Omega \) and call \( \text{UPPERCL}(x, t, X, Y) \) for each mesh grid point \( x \). In a general situation, we suggest calling \( \text{UPPERCL}(x, t, X, Y) \) with randomly chosen \( x \)'s and using the \( k \)-nearest neighbors algorithm to represent \( CR_t \).
3.2.2 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. Similar to Section 3.2.1, 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

$$\mu_n(x) := r_n^T(x)K_n^{-1}Y_{1:n},$$

$$\sigma_n^2(x) := \sigma^2(1 - r_n^T(x)K_n^{-1}r_n(x)),$$

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

**Theorem 2 (Uncertainty quantification for sequential samplings)** Suppose Condition 7 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 Equation 11 and Equation 12 respectively. Then for any $t > 0$,

$$\mathbb{P}(M_T - C \sqrt{\text{t} \log(A_0D_\Omega)}) > t) \leq e^{-t^2/2},$$

where $C, A_0, D_\Omega$ are the same as in Corollary 3.

The proof of Theorem 2 can be found in Appendix D. The probability bound Equation 13 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 3.1 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.

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 $\Omega$ 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 $CI_t^{seq}$ defined in Equation 7 and $f(x_{max})$ is contained in $CI_t^{seq}$ defined in Equation 8.

4 Numerical Experiments

We compare the performance between the proposed confidence interval $CI_t^{seq}$ as in Equation 8 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.,

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

where $q_{0.05}$ is the 0.95 quantile of the standard normal distribution, $\mu_T(x)$ and $\sigma_T(x)$ are given in Equation 5 and Equation 6 respectively. As suggested in Section 2.2, we use $C_0 = 1$ and $t = 2.448$ in $CI_t^{seq}$. We consider the Matern correlation functions (see Table 1) with $\nu = 1.5, 2.5, 3.5$, and $A_0D_\Omega = 25$.

We simulate Gaussian processes on $\Omega = [0, 1]^2$ for each $\nu$. We use optimal Latin hypercube designs (Stocki, 2005) to generate 5 initial points. We employ $\alpha_{\text{UCB}}$ (defined in Section 2.2) as the acquisition function, in which the parameter $\beta_n$ is chosen as 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 \( f(x_{\text{max}}) \in CI_{t}^{\text{seq}} \) or \( f(x_{\text{max}}) \in CI_{G} \). We also consider the "optimal upper bound" in the sense that we choose a constant \( a_\nu \) and the confidence upper bound

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

such that the relative frequency of the event \( f(x_{\text{max}}) \in CI_a \) is exactly 95\%, where \( a_\nu \) only depends on \( \nu \). Then we plot the coverage rate of \( CI_{t}^{\text{seq}} \) and \( CI_{G} \), and the width of \( CI_{t}^{\text{seq}}, CI_{G} \), and \( CI_a \) under 5, 10, 15, 20, 25, 30 iterations, respectively.

The comparison results are shown in Figure 1. 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 left plot in Figure 1, we find that the coverage rate of \( CI_{t}^{\text{seq}} \) is almost 100\% for all the experiments, while \( CI_{G} \) has a lower coverage rate no more than 82\%. 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. The coverage results support our theory and conclusions made in Section 3.2. As shown by the right plot in Figure 1, the widths of \( CI_{t}^{\text{seq}} \) 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.

Figure 1: **Panel 1**: Coverage rates of \( CI_{t}^{\text{seq}} \) and \( CI_{G} \). The nominal confidence level is 95\%. **Panel 2**: Widths of \( CI_{t}^{\text{seq}}, CI_{G}, \) and \( CI_a \).

5 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.
REFERENCES


