On Local Optimizers of Acquisition Functions in Bayesian Optimization

Jungtaek Kim[1(✉)] and Seungjin Choi[2]

[1] Pohang University of Science and Technology, Pohang, Republic of Korea
jtkim@postech.ac.kr
[2] Inference Lab, BARO AI, Seoul, Republic of Korea
seungjin@baroai.com

Abstract. Bayesian optimization is a sample-efficient method for finding a global optimum of an expensive-to-evaluate black-box function. A global solution is found by accumulating a pair of query point and its function value, repeating these two procedures: (i) modeling a surrogate function; (ii) maximizing an acquisition function to determine where next to query. Convergence guarantees are only valid when the global optimizer of the acquisition function is found at each round and selected as the next query point. In practice, however, local optimizers of an acquisition function are also used, since searching for the global optimizer is often a non-trivial or time-consuming task. In this paper we consider three popular acquisition functions, PI, EI, and GP-UCB induced by Gaussian process regression. Then we present a performance analysis on the behavior of local optimizers of those acquisition functions, in terms of *instantaneous regrets* over global optimizers. We also introduce an analysis, allowing a local optimization method to start from multiple different initial conditions. Numerical experiments confirm the validity of our theoretical analysis.

Keywords: Global optimization · Bayesian optimization · Acquisition function optimization · Instantaneous regret analysis.

1 Introduction

Bayesian optimization provides an efficient method for finding a global optimum of an objective function $f(\mathbf{x}) : \mathcal{X} \to \mathbb{R}$, defined over a compact set $\mathcal{X} \subset \mathbb{R}^d$:

$$\mathbf{x}^\dagger = \arg\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}), \tag{1}$$

where, in general, $f(\mathbf{x})$ is a black-box function, i.e., its closed-form expression is not available and its gradient is not available either. The value of the function can be computed at a query point $\mathbf{x}$ but the evaluation requires a high cost. In this paper we assume that the objective function $f(\mathbf{x})$ of interest is Lipschitz-continuous. Bayesian optimization searches for a minimum of $f(\mathbf{x})$ to solve the problem (1), gradually accumulating $(\mathbf{x}_t, f(\mathbf{x}_t))$ where input points $\mathbf{x}_t$ are carefully chosen and corresponding function values $f(\mathbf{x}_t)$ are calculated at $\mathbf{x}_t$. It

provides an efficient approach in terms of the number of function evaluations required.

In Bayesian optimization, a global solution to the problem (1) is determined by repeating the following two procedures. At each round, we first train a probabilistic model[1] using the data observed so far to construct a surrogate function for $f(\mathbf{x})$. Then we define an acquisition function [10, 12, 18] over the domain $\mathcal{X}$, which accounts for the utility provided by possible outcomes drawn from the distribution determined by the surrogate model. The maximization of an acquisition function, referred to as an *inner optimization,* yields the selection of the next query point at which to evaluate the objective function. Convergence guarantees are only valid when the global optimizer of the acquisition function is found and selected as the next query point. In practice, however, local optimizers of acquisition functions are also used, since searching for the exact optimizer of the acquisition function is often a non-trivial or time-consuming task.

A recent work [22] has addressed the acquisition function optimization, elucidating gradient-based optimization of Monte Carlo estimates of acquisition functions, as well as on sub-modularity for a family of maximal myopic acquisition functions. However, so far, there is no study on what the performance loss is when a local optimizer of an acquisition function is selected as the next query point. In this paper we attempt to provide an answer to this question on the performance loss brought by local optimizers of acquisition functions over global optimizers, in terms of instantaneous regrets. To this end, we consider three different solutions to the maximization of an acquisition function: (i) a global optimizer; (ii) a local optimizer; (iii) a multi-started local optimizer. For performance analysis of local optimizers, with respect to the global optimizer, we define an *instantaneous regret difference* for a local optimizer as well as for a multi-started local optimizer and present its bound for each case. As expected, the multi-started local optimizer yields a tighter bound on the instantaneous regret difference, compared to the one for a local optimizer.

In this paper we consider three popular acquisition functions, probability of improvement (PI) [10], expected improvement (EI) [12], and Gaussian process upper confidence bound (GP-UCB) [18], each of which is calculated by posterior mean and variance determined by Gaussian process regression. The main contribution of this paper is summarized as:

- We provide an upper bound on the instantaneous regret difference between global and local optimizers, which is given in Theorem 1;
- We provide an upper bound on the instantaneous regret difference when a multi-started local optimization method is employed to search for a local maximum of the acquisition function, which is given in Theorem 2;
- Numerical experiments are provided to justify our theoretical analyses.

[1] Gaussian process regression is used in this paper.

2 Background

In this section, we briefly review Bayesian optimization, the detailed overview of which is referred to [2, 17, 5], and define instantaneous regret difference that is used as a performance measure for local optimizers of acquisition functions. We also explain global and local optimization methods that are popularly used to search for maxima of acquisition functions.

2.1 Bayesian Optimization

The Bayesian optimization strategy solves the problem (1), by gradually selecting queries $\mathbf{x}_1, \ldots, \mathbf{x}_T$ and their corresponding noisy evaluations $y_1, \ldots, y_T$ where $y_t = f(\mathbf{x}_t) + \epsilon_t$ with $\epsilon_t \sim \mathcal{N}(0, \sigma_n^2)$, such that a minimizer of $f(\mathbf{x})$ is determined from $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$. Given the data $\mathcal{D}_{t-1} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{t-1}, y_{t-1})\}$ observed up to round $t-1$, the next point $\mathbf{x}_t$ is chosen as a maximizer of acquisition function $a(\mathbf{x}|\mathcal{D}_{t-1})$, i.e.,

$$\mathbf{x}_t = \arg\max\, a(\mathbf{x}|\mathcal{D}_{t-1}). \tag{2}$$

The acquisition function is the expected utility u of a query $\mathbf{x}$:

$$a(\mathbf{x}|\mathcal{D}_{t-1}) = \int u(\mathbf{x}, y)\, p(y|\mathbf{x}, \mathcal{D}_{t-1})\, \mathrm{d}y, \tag{3}$$

where the posterior distribution $p(y|\mathbf{x}, \mathcal{D}_{t-1})$ is calculated by Gaussian process regression using $\mathcal{D}_{t-1}$ here.

Solving (2) is another optimization problem appearing in the Bayesian optimization task given in (1). We consider three different solutions to the maximization of an acquisition function, defined in detail below.

Definition 1 (Global optimizer). *We denote by $\mathbf{x}_{t,g}$ the optimizer of the acquisition function $a(\mathbf{x}|\mathcal{D}_{t-1})$ at round t, determined by a global optimization method, given a time budget τ:*

$$\mathbf{x}_{t,g} = \overset{global}{\arg\max}_{\mathbf{x}\in\mathcal{X}}\, a(\mathbf{x}|\mathcal{D}_{t-1}). \tag{4}$$

$\mathbf{x}_{t,g}$ is referred to as a global optimizer.

Definition 2 (Local optimizer). *We denote by $\mathbf{x}_{t,l}$ the optimizer of the acquisition function $a(\mathbf{x}|\mathcal{D}_{t-1})$ at round t, determined by an iterative (local) optimization method where the convergence meets $\|\mathbf{x}_{t,l}^{(\tau)} - \mathbf{x}_{t,l}^{(\tau-1)}\|_2 \leq \epsilon_{opt}$ for iteration τ:*

$$\mathbf{x}_{t,l} = \overset{local}{\arg\max}_{\mathbf{x}\in\mathcal{X}}\, a(\mathbf{x}|\mathcal{D}_{t-1}). \tag{5}$$

$\mathbf{x}_{t,l}$ is referred to as a local optimizer.

Definition 3 (Multi-started local optimizer). *Suppose that* $\{\mathbf{x}_{t,l_1}, \ldots, \mathbf{x}_{t,l_N}\}$ *is a set of* N *local optimizers, each of which is determined by a local optimization method* (5)*, starting from a different initial condition. The multi-started local optimizer, denoted by* $\mathbf{x}_{t,m}$*, is the one at which* $a(\mathbf{x}|\mathcal{D}_{t-1})$ *achieves the maximum:*

$$\mathbf{x}_{t,m} = \overset{m\text{-}local}{\arg\max}_{\mathbf{x}\in\mathcal{X}}\ a(\mathbf{x}|\mathcal{D}_{t-1}). \tag{6}$$

With solutions to (2), defined in (4), (5), and (6), we define *instantaneous regret* for each of these solutions and *instantaneous regret difference* for each of local solutions below.

Definition 4 (Instantaneous regret). *Suppose that* $\mathbf{x}^\dagger$ *is the true global minimum of the objective function in* (1)*. Denote by* $\mathbf{x}_t$ *a maximum of acquisition function* $a(\mathbf{x}|\mathcal{D}_{t-1})$ *at round* t*, determined by either a global or local optimization method. The instantaneous regret* r_t *at round* t *is defined as*

$$r_t = f(\mathbf{x}_t) - f(\mathbf{x}^\dagger). \tag{7}$$

Depending on an optimization method (i.e., one of global, local, and multi-started local optimization methods) used to search for a maximum of the acquisition function, we define the following instantaneous regrets: $r_{t,g} = f(\mathbf{x}_{t,g}) - f(\mathbf{x}^\dagger)$*,* $r_{t,l} = f(\mathbf{x}_{t,l}) - f(\mathbf{x}^\dagger)$*, and* $r_{t,m} = f(\mathbf{x}_{t,m}) - f(\mathbf{x}^\dagger)$*.*

Definition 5 (Instantaneous regret difference). *With Definition 4, we define instantaneous regret differences for an local optimizer* $\mathbf{x}_{t,l}$ *and for a multi-started local optimizer* $\mathbf{x}_{t,m}$*:*

$$|r_{t,g} - r_{t,l}| = |f(\mathbf{x}_{t,g}) - f(\mathbf{x}_{t,l})|\,, \tag{8}$$

$$|r_{t,g} - r_{t,m}| = |f(\mathbf{x}_{t,g}) - f(\mathbf{x}_{t,m})|\,, \tag{9}$$

which measures a performance gap with respect to the one induced by $\mathbf{x}_{t,g}$*, at round* t*.*

Henceforth, instantaneous regret and instantaneous regret difference are simply called to *regret* and *regret difference*, respectively.[2]

2.2 Maximization of Acquisition Functions

As described earlier, we may consider either global or local solutions to (2). Famous global optimization methods include DIRECT [8] and CMA-ES [6]. DIRECT is a deterministic Lipschitzian-based derivative-free partitioning method where it observes function values at the centers of rectangles and divides the

[2] In Bayesian optimization, a cumulative regret is usually used to analyze the performance of convergence quality. By Lemma 5.4 of [18] and Theorem 3 of [3], our analysis on instantaneous regret differences can be expanded into the analysis on cumulative regrets. However, to concentrate the scope of this work on the behavior of local optimizers, these analyses are not included in this paper.

rectangles without the Lipschitz constant iteratively. CMA-ES is a stochastic derivative-free method based on evolutionary computing. In this paper we use DIRECT to determine $\mathbf{x}_{t,g}$ in (4).

A local optimization method we used to determine $\mathbf{x}_{t,l}$ or $\mathbf{x}_{t,m}$, given in (5) or (6) is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, which is a quasi-Newton optimization technique. A limited memory version, referred to as L-BFGS [11] and a constrained version known as L-BFGS-B are widely used in the Bayesian optimization literature [13, 20, 5]. Multi-started optimization methods are also widely used [2, 7], where a local optimization method starts from N distinct initializations and such N local solutions started from N distinct initializations are combined to determine the best local solution.

Compared to our work, [22] introduces a reparameterization form to allow differentiability of Monte Carlo acquisition functions to integrate them and query in parallel, which is not related to the topics covered in this paper.

3 Performance Analysis

In this section we present our main contribution on the performance analysis for the local optimizer and the multi-started local optimizer, given in (5) and (6).

3.1 Main Theorems

Before introducing the lemmas used to prove the main theorems, we explain the main theorems and their intuition first. Our theorems are described as follows.

Theorem 1. *Given $\delta_l \in [0, 1)$ and $\epsilon_l, \epsilon_1, \epsilon_2 > 0$, the regret difference for a local optimizer $\mathbf{x}_{t,l}$ at round t, $|r_{t,g} - r_{t,l}|$ is less than ϵ_l with a probability at least $1 - \delta_l$:*

$$\mathbb{P}\big(|r_{t,g} - r_{t,l}| < \epsilon_l \big) \geq 1 - \delta_l, \tag{10}$$

where $\delta_l = \frac{\gamma}{\epsilon_1}(1-\beta_g) + \frac{M}{\epsilon_2}$, $\epsilon_l = \epsilon_1\epsilon_2$, $\gamma = \max_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}} \|\mathbf{x}_i - \mathbf{x}_j\|_2$ is the size of $\mathcal{X}$, β_g is the probability that a local optimizer of the acquisition function collapses with its global optimizer, and M is the Lipschitz constant explained in Lemma 8.

Theorem 1 is extended for a multi-started local optimizer.

Theorem 2. *Given $\delta_m \in [0, 1)$ and $\epsilon_m, \epsilon_2, \epsilon_3 > 0$, a regret difference for a multi-started local optimizer $\mathbf{x}_{t,m}$, determined by starting from N initial points at round t, is less than ϵ_m with a probability at least $1 - \delta_m$:*

$$\mathbb{P}\big(|r_{t,g} - r_{t,m}| < \epsilon_m \big) \geq 1 - \delta_m, \tag{11}$$

where $\delta_m = \frac{\gamma}{\epsilon_3}(1 - \beta_g)^N + \frac{M}{\epsilon_2}$, $\epsilon_m = \epsilon_2\epsilon_3$, $\gamma = \max_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}} \|\mathbf{x}_i - \mathbf{x}_j\|_2$ is the size of $\mathcal{X}$, β_g is the probability that a local optimizer of the acquisition function collapses with its global optimizer, and M is the Lipschitz constant explained in Lemma 8.

As shown in Theorem 1, $|r_{t,g} - r_{t,l}|$ is smaller than ϵ_l with a probability $1-\delta_l$. It implies the probability $1-\delta_l$ is controlled by three statements related to γ, β_g, and M: the probability is decreased (i) as γ is increased, (ii) as β_g is decreased, and (iii) as M is increased. If $\mathcal{X}$ is a relatively small space, γ is naturally small. Moreover, β_g is close to one if converging to global optimum by Definition 3 is relatively easy for some reasons: (i) a small number of local optima exist, or (ii) a global optimum is easily reachable.

Theorem 2 suggests the implications that are similar with Theorem 1 in terms of the control factors of $1-\delta_m$. The main difference of two theorems is that δ_m is related to the number of initial points in Definition 3, N. Because $0 \leq 1-\beta_g < 1$ is given, N can control the bound of (11). Additionally, by this difference, we theoretically reveal how many runs for a multi-started local optimizer are needed to obtain the sufficiently small regret difference over a global optimizer.

3.2 Lemmas

Next, we prove two statements (i) how different the global and local optimizers are (see Lemma 1 to Lemma 7), and (ii) how steep the slope between the global and local optimizers is (see Lemma 8). First, the Lipschitz continuities of acquisition functions are proved in the subsequent lemmas.

Lemma 1 (Lipschitz continuity of PI). *The PI criterion $a(\mathbf{x}|\mathcal{D}_{t-1})$, formed by the posterior distribution calculated by Gaussian process regression on $\mathcal{D}_{t-1}$ is Lipschitz-continuous.*

Proof. $\mathcal{X}$ is a compact subset of d-dimensional space $\mathbb{R}^d$. In this paper, we analyze our theorem with Gaussian process regression as a surrogate function. If we are given $t-1$ covariates $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_{t-1}]^\top$ obtained from $\mathcal{X}$ and their corresponding responses $\mathbf{y} = [y_1 \cdots y_{t-1}] \in \mathbb{R}^{t-1}$, posterior mean and variance functions, $\mu(\mathbf{x})$ and $\sigma^2(\mathbf{x})$ over $\mathbf{x} \in \mathcal{X}$ can be computed, using Gaussian process regression [14]:

$$\mu(\mathbf{x}) = \boldsymbol{k}(\mathbf{x}, \mathbf{X})\tilde{\boldsymbol{K}}^{-1}\mathbf{y}, \tag{12}$$

$$\sigma^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \boldsymbol{k}(\mathbf{x}, \mathbf{X})\tilde{\boldsymbol{K}}^{-1}\boldsymbol{k}(\mathbf{X}, \mathbf{x}), \tag{13}$$

where $k(\cdot, \cdot)$ is a covariance function, $\tilde{\boldsymbol{K}} = \boldsymbol{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \boldsymbol{I}$, and σ_n is an observation noise. $\boldsymbol{k}(\cdot, \cdot)$ accepts a vector and a matrix as two arguments (e.g., $\boldsymbol{k}(\mathbf{x}, \mathbf{X}) = [k(\mathbf{x}, \mathbf{x}_1) \cdots k(\mathbf{x}, \mathbf{x}_{t-1})]$). Similarly, $\boldsymbol{K}(\cdot, \cdot)$ can take two matrices (e.g., $\boldsymbol{K}(\mathbf{X}, \mathbf{X}) = [\boldsymbol{k}(\mathbf{X}, \mathbf{x}_1) \cdots \boldsymbol{k}(\mathbf{X}, \mathbf{x}_{t-1})]$). Before showing the Lipschitz continuity of the acquisition function, we first show the derivatives of (12) and (13). It depends on the differentiability of covariance functions, but the famous covariance functions, which are used in Bayesian optimization are usually at least once differentiable (e.g., squared exponential kernel[3] and Matérn kernel[4]). Thus, the

[3] Squared exponential kernel is infinite times differentiable.
[4] Matérn kernel is $\lceil \nu \rceil - 1$ times differentiable.

derivatives of (12) and (13) are

$$\frac{\partial \mu(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \boldsymbol{k}(\mathbf{x}, \mathbf{X})}{\partial \mathbf{x}} \tilde{\boldsymbol{K}}^{-1} \mathbf{y}, \tag{14}$$

$$\frac{\partial \sigma^2(\mathbf{x})}{\partial \mathbf{x}} = -2 \frac{\partial \boldsymbol{k}(\mathbf{x}, \mathbf{X})}{\partial \mathbf{x}} \tilde{\boldsymbol{K}}^{-1} \boldsymbol{k}(\mathbf{X}, \mathbf{x}), \tag{15}$$

using vector calculus identities. To show (14) and (15) are bounded, each term in both equations should be bounded. For all $i \in \{1, \ldots, t-1\}$, $\mathbf{y}$ and $\boldsymbol{k}(\mathbf{X}, \mathbf{x})$ are obviously bounded:

$$|y_i| < \infty \quad \text{and} \quad |k(\mathbf{x}_i, \mathbf{x})| < \infty, \tag{16}$$

but $\partial \boldsymbol{k}(\mathbf{x}, \mathbf{X}) / \partial \mathbf{x}$ and $\tilde{\boldsymbol{K}}^{-1}$ should be revealed.

First of all, the bound of $\partial \boldsymbol{k}(\mathbf{x}, \mathbf{X}) / \partial \mathbf{x}$ would be proved in Lemma 4. For the latter one, all entries of $\tilde{\boldsymbol{K}}^{-1}$ are bounded by the Kantorovich and Wielandt inequalities [15]. As a result, by (16), Lemma 4, and the Kantorovich and Wielandt inequalities, the following inequalities are satisfied:

$$\left| \frac{\partial \mu(\mathbf{x})}{\partial x_i} \right| < \infty \quad \text{and} \quad \left| \frac{\partial \sigma^2(\mathbf{x})}{\partial x_i} \right| < \infty, \tag{17}$$

for all $i \in \{1, \ldots, d\}$. Thus, we can say

$$\left| \frac{\partial \mu(\mathbf{x})}{\partial x_i} \right| < M_\mu \quad \text{and} \quad \left| \frac{\partial \sigma^2(\mathbf{x})}{\partial x_i} \right| < M_{\sigma^2}, \tag{18}$$

for some $M_\mu, M_{\sigma^2} < \infty$. It implies that $\mu(\mathbf{x})$ and $\sigma^2(\mathbf{x})$ are Lipschitz-continuous with the Lipschitz constants M_μ and M_{σ^2} to each axis direction, respectively. Thus, (12) and (13) are Lipschitz-continuous with the Lipschitz constants dM_μ and dM_{σ^2}, where d is a dimensionality of $\mathbf{x}$, by the triangle inequality.

PI is written with $z(\mathbf{x}) = \left(f(\mathbf{x}^\ddagger) - \mu(\mathbf{x})\right) / \sigma(\mathbf{x})$ if $\sigma(\mathbf{x}) > \sigma_n$, and 0 otherwise, where $\mathbf{x}^\ddagger$ is the current best observation which has a minimum of $\mathbf{y}$. Given the PI criterion $a_{\mathrm{PI}}(\mathbf{x}) = \Phi(z(\mathbf{x}))$, where $\Phi(\cdot)$ is a cumulative distribution function of standard normal distribution, the derivative of PI criterion is

$$\frac{\partial a_{\mathrm{PI}}(\mathbf{x})}{\partial \mathbf{x}} = \phi(z(\mathbf{x})) \frac{\partial z(\mathbf{x})}{\partial \mathbf{x}} = \phi(z(\mathbf{x})) \left(\frac{\mu(\mathbf{x}) - f(\mathbf{x}^\ddagger)}{\sigma^2(\mathbf{x})} \frac{\partial \sigma(\mathbf{x})}{\partial \mathbf{x}} - \frac{1}{\sigma(\mathbf{x})} \frac{\partial \mu(\mathbf{x})}{\partial \mathbf{x}} \right), \tag{19}$$

where $\phi(\cdot)$ is a probability density function of standard normal distribution. By (12), (13), and (18), we can show (19) is bounded, $\left\| \partial a_{\mathrm{PI}}(\mathbf{x}) / \partial \mathbf{x} \right\|_2 < \infty$. □

Lemma 2 (Lipschitz continuity of EI). *The EI criterion $a(\mathbf{x}|\mathcal{D}_{t-1})$, formed by the posterior distribution calculated by Gaussian process regression on $\mathcal{D}_{t-1}$ is Lipschitz-continuous.*

Proof. EI expresses with $z(\mathbf{x}) = \left(f(\mathbf{x}^\ddagger) - \mu(\mathbf{x})\right) / \sigma(\mathbf{x})$ if $\sigma(\mathbf{x}) > \sigma_n$, and 0 otherwise, where $\mathbf{x}^\ddagger$ is the current best observation which has a minimum of $\mathbf{y}$. For the EI criterion:

$$a_{\mathrm{EI}}(\mathbf{x}) = \left(f(\mathbf{x}^\ddagger) - \mu(\mathbf{x})\right) \Phi(z(\mathbf{x})) + \sigma(\mathbf{x}) \phi(z(\mathbf{x})), \tag{20}$$

the derivative of (20) is

$$\frac{\partial a_{\mathrm{EI}}(\mathbf{x})}{\partial \mathbf{x}} = \left(f(\mathbf{x}^{\ddagger}) - \mu(\mathbf{x})\right)\phi(z(\mathbf{x}))\frac{\partial z(\mathbf{x})}{\partial \mathbf{x}} - \frac{\partial \mu(\mathbf{x})}{\partial \mathbf{x}}\Phi(z(\mathbf{x})) + \sigma(\mathbf{x})\phi'(z(\mathbf{x}))\frac{\partial z(\mathbf{x})}{\partial \mathbf{x}} + \frac{\partial \sigma(\mathbf{x})}{\partial \mathbf{x}}\phi(z(\mathbf{x})). \tag{21}$$

Similar to (19), the following inequality,

$$\left\| \frac{\partial a_{\mathrm{EI}}(\mathbf{x})}{\partial \mathbf{x}} \right\|_2 < \infty, \tag{22}$$

is satisfied. □

Lemma 3 (Lipschitz continuity of GP-UCB). *GP-UCB $a(\mathbf{x}|\mathcal{D}_{t-1})$, formed by the posterior distribution calculated by Gaussian process regression on $\mathcal{D}_{t-1}$ is Lipschitz-continuous.*

Proof. GP-UCB [18] and its derivative are

$$a_{\mathrm{UCB}}(\mathbf{x}) = -\mu(\mathbf{x}) + \alpha\sigma(\mathbf{x}), \tag{23}$$

$$\frac{\partial a_{\mathrm{UCB}}(\mathbf{x})}{\partial \mathbf{x}} = -\frac{\partial \mu(\mathbf{x})}{\partial \mathbf{x}} + \alpha\frac{\partial \sigma(\mathbf{x})}{\partial \mathbf{x}}, \tag{24}$$

where α is a coefficient for balancing exploration and exploitation. By (18), the following inequality,

$$\left|\frac{\partial a_{\mathrm{UCB}}(\mathbf{x})}{\partial x_i}\right| = \left|-\frac{\partial \mu(\mathbf{x})}{\partial x_i} + \alpha\frac{\partial \sigma(\mathbf{x})}{\partial x_i}\right| \leq \left|-\frac{\partial \mu(\mathbf{x})}{\partial x_i}\right| + \alpha\left|\frac{\partial \sigma(\mathbf{x})}{\partial x_i}\right| = \left|\frac{\partial \mu(\mathbf{x})}{\partial x_i}\right| + \alpha\left|\frac{\partial \sigma(\mathbf{x})}{\partial x_i}\right| \leq M_{\mu} + \alpha\sqrt{M_{\sigma^2}}, \tag{25}$$

is bounded for $i \in \{1, \ldots, d\}$. Therefore, (24) is bounded. □

Lemma 4. *Given a stationary covariance function $k(\cdot,\cdot)$ that is widely used in Gaussian process regression [21, 4], $\partial \boldsymbol{k}(\mathbf{x}, \mathbf{X})/\partial \mathbf{x}$ is bounded where $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{k}(\mathbf{x}, \mathbf{X}) = [k(\mathbf{x}, \mathbf{x}_1) \cdots k(\mathbf{x}, \mathbf{x}_n)]$.*

Proof. The well-known stationary covariance functions such as squared exponential (SE) and Matèrn kernels are utilized in Gaussian process regression [21, 4]. Since such kernels are additive or multiplicative [4], this lemma can be generalized to most of kernels applied in Gaussian process regression. In this paper, we analyze the cases of SE and Matérn 5/2 kernels. Because the cases of Matérn 3/2 and periodic kernels can be simply extended from the cases analyzed, it is omitted. The SE and Matèrn 5/2 kernels are at least one time differentiable, thus $\partial \boldsymbol{k}(\mathbf{x}, \mathbf{X})/\partial \mathbf{x}$ can be computed. Before explaining in detail, $\partial \boldsymbol{k}(\mathbf{x}, \mathbf{X})/\partial \mathbf{x}$ is written as

$$\frac{\partial \boldsymbol{k}(\mathbf{x}, \mathbf{X})}{\partial \mathbf{x}} = \left[\frac{\partial k(\mathbf{x}, \mathbf{x}_1)}{\partial \mathbf{x}} \cdots \frac{\partial k(\mathbf{x}, \mathbf{x}_n)}{\partial \mathbf{x}}\right], \tag{26}$$

for $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]$. Furthermore, we can define

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^\top \boldsymbol{L}^{-1}(\mathbf{x}_1 - \mathbf{x}_2)}, \tag{27}$$

where $\boldsymbol{L}$ is a diagonal matrix of which entries are lengthscales for each each dimension. For simplicity, $\mathbf{x}_1 - \mathbf{x}_2$ is denoted as $\mathbf{s}_{12}$. The derivative of (27) is

$$\frac{\partial d(\mathbf{x}_1, \mathbf{x}_2)}{\partial \mathbf{x}_1} = \left(\boldsymbol{L}^{-1}\mathbf{s}_{12}\right)\left(\mathbf{s}_{12}^\top \boldsymbol{L}^{-1}\mathbf{s}_{12}\right)^{-\frac{1}{2}}. \tag{28}$$

The derivative of $d^2(\mathbf{x}_1, \mathbf{x}_2)$ is $\frac{\partial d^2(\mathbf{x}_1, \mathbf{x}_2)}{\partial \mathbf{x}_1} = 2\boldsymbol{L}^{-1}\mathbf{s}_{12}$.

First, the SE kernel is $k(\mathbf{x}_1, \mathbf{x}_2) = \sigma_s^2 \exp(-\frac{1}{2}d^2(\mathbf{x}_1, \mathbf{x}_2))$ where σ_s is a signal scale. The derivative of each $\partial k(\mathbf{x}, \mathbf{x}_i)/\partial \mathbf{x}$ is

$$\begin{aligned}\frac{\partial k(\mathbf{x}, \mathbf{x}_i)}{\partial \mathbf{x}} &= \frac{\partial}{\partial \mathbf{x}}\left(\sigma_s^2 \exp\left(-\frac{1}{2}d^2(\mathbf{x}, \mathbf{x}_i)\right)\right) = k(\mathbf{x}, \mathbf{x}_i)\frac{\partial}{\partial \mathbf{x}}\left(-\frac{1}{2}d^2(\mathbf{x}_1, \mathbf{x}_2)\right) \\ &= -\frac{k(\mathbf{x}, \mathbf{x}_i)}{2}\left(2\boldsymbol{L}^{-1}(\mathbf{x} - \mathbf{x}_i)\right) = -k(\mathbf{x}, \mathbf{x}_i)\left(\boldsymbol{L}^{-1}(\mathbf{x} - \mathbf{x}_i)\right). \end{aligned} \tag{29}$$

Because all the terms of (29) are bounded in $\mathcal{X}$, $\|\partial k(\mathbf{x}, \mathbf{x}_i)/\partial \mathbf{x}\|_2 < \infty$ is satisfied for all $i = \{1, \ldots, n\}$. Note that (27) and (28) are bounded, because $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$.

The Matérn 5/2 kernel is

$$k(\mathbf{x}_1, \mathbf{x}_2) = \sigma_s^2\left(1 + \sqrt{5}d(\mathbf{x}_1, \mathbf{x}_2) + \frac{5}{3}d^2(\mathbf{x}_1, \mathbf{x}_2))(-\sqrt{5}d(\mathbf{x}_1, \mathbf{x}_2)\right), \tag{30}$$

where σ_s is a signal scale. Its derivative is

$$\frac{\partial k(\mathbf{x}, \mathbf{x}_i)}{\partial \mathbf{x}} = -\frac{5\sigma_s^2}{3}\left(1 + \sqrt{5}d(\mathbf{x}, \mathbf{x}_i)\right)\exp(-\sqrt{5}d(\mathbf{x}, \mathbf{x}_i))\boldsymbol{L}^{-1}(\mathbf{x} - \mathbf{x}_i). \tag{31}$$

Since the derivative is bounded, $\|\partial \boldsymbol{k}(\mathbf{x}, \mathbf{X})/\partial \mathbf{x}\|_2 < \infty$ is satisfied. Similarly, other kernels can be straightforwardly proved. Thus, this lemma is concluded. □

From now, we show the number of local optima is upper-bounded, using the condition involved in a frequency domain.

Lemma 5. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact set. Given some sufficiently large $|\hat{\boldsymbol{\xi}}| > 0$, a spectral density of stationary covariance function for Gaussian process regression is zero for all $|\boldsymbol{\xi}| > |\hat{\boldsymbol{\xi}}|$ with very high probability. Then, the number of local maxima at iteration t, ρ_t is upper-bounded.*

Proof. By Sard's theorem [16] for a Lipschitz-continuous function [1], critical points (i.e., the points whose gradients are zero) do not exist almost everywhere. Since the number of local maxima ρ_t is upper-bounded by the number of critical points, it can be a starting point to bound ρ_t. Especially, by Lemma 1, Lemma 2,

and Lemma 3, the number of local maxima ρ_t can be restrained in the compact set $\mathcal{X}$. Since it cannot express the upper-bound of ρ_t, we transform a stationary covariance function using a Fourier transform and obtain the spectral density of each covariance function [14, Chapter 4]. Because a spectral density of stationary covariance function is naturally a light-tail function by Bochner's theorem [19], a spectral density of covariance function for Gaussian process regression is zero for all $|\boldsymbol{\xi}| > |\hat{\boldsymbol{\xi}}|$ with very high probability (i.e., exponentially saturated probability over $|\boldsymbol{\xi}|$ due to the form of stationary kernels [14, Chapter 4]), given some sufficiently large $|\hat{\boldsymbol{\xi}}|$. Then, a function has finite local maxima, which implies that the number of local maxima at iteration t is upper-bounded. □

Based on Lemma 5, we can prove the ergodicity of local maxima that are able to be discovered by the local optimization method and coincided with the local optimizers started from different initial points.

Lemma 6. *Let the number of local maxima of acquisition function at iteration t be ρ_t. Since local optimizers which are started from some initial conditions $\in \mathcal{X}$ are ergodic to all the local maxima, the probability of reaching each solution is $\beta_1, \ldots, \beta_{\rho_t} > 0$ such that $\Sigma_{i=1}^{\rho_t} \beta_i = 1$.*

Proof. If we start from some different initial conditions $\in \mathcal{X}$, it is obvious that all the local solutions are reachable. Therefore, all the local optimizers are ergodic, and the probability of reaching each solution is larger than zero and they sum to one: $\sum_{i=1}^{\rho_t} \beta_i = 1$, where $\beta_1, \ldots, \beta_{\rho_t} > 0$. □

We now prove the distance between two points acquired by Definition 1 and Definition 2 is bounded with a probability.

Lemma 7. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact space where $\gamma = \max_{\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}} \|\mathbf{x}_1 - \mathbf{x}_2\|_2$. Then, for $\gamma > \epsilon_1 > 0$, we have*

$$\mathbb{P}\big(\|\mathbf{x}_{t,g} - \mathbf{x}_{t,l}\|_2 \geq \epsilon_1\big) \leq \frac{\gamma}{\epsilon_1}(1 - \beta_g), \tag{32}$$

where β_g is the probability that some local optimizer is coincided with the global optimizer of the acquisition function.

Proof. By Markov inequality for $\epsilon_1 > 0$, we have

$$\mathbb{P}\big(\|\mathbf{x}_{t,g} - \mathbf{x}_{t,l}\|_2 \geq \epsilon_1\big) \leq \frac{1}{\epsilon_1}\mathbb{E}\big[\|\mathbf{x}_{t,g} - \mathbf{x}_{t,l}\|_2\big]. \tag{33}$$

Following from Lemma 6 and (33), the expectation in the right-hand side of (33) is calculated as

$$\begin{aligned}
&\frac{1}{\epsilon_1}\mathbb{E}\big[\|\mathbf{x}_{t,g} - \mathbf{x}_{t,l}\|_2\big] \\
&= \frac{1}{\epsilon_1}\beta_g \|\mathbf{x}_{t,g} - \mathbf{x}_{t,l}\|_2\Big|_{\mathbf{x}_{t,g} = \mathbf{x}_{t,l}} + \frac{1}{\epsilon_1}(1 - \beta_g)\|\mathbf{x}_{t,g} - \mathbf{x}_{t,l}\|_2\Big|_{\mathbf{x}_{t,g} \neq \mathbf{x}_{t,l}} \\
&\leq \frac{1}{\epsilon_1}(1 - \beta_g)\gamma,
\end{aligned} \tag{34}$$

which completes the proof. □

The lower-bound of $\frac{|f(\mathbf{x}_{t,1})-f(\mathbf{x}_{t,2})|}{\|\mathbf{x}_{t,1}-\mathbf{x}_{t,2}\|_2}$ can be expressed with a probability as follows.

Lemma 8. *Given any $\epsilon_2 > 0$, the probability that $\frac{|f(\mathbf{x}_{t,g})-f(\mathbf{x}_{t,l})|}{\|\mathbf{x}_{t,g}-\mathbf{x}_{t,l}\|_2} \geq \epsilon_2$ is less than $\frac{M}{\epsilon_2}$:*

$$\mathbb{P}\left(\frac{|f(\mathbf{x}_{t,g}) - f(\mathbf{x}_{t,l})|}{\|\mathbf{x}_{t,g} - \mathbf{x}_{t,l}\|_2} \geq \epsilon_2\right) \leq \frac{M}{\epsilon_2}. \tag{35}$$

Proof. By Markov's inequality, we can express

$$\mathbb{P}\left(\frac{|f(\mathbf{x}_{t,g}) - f(\mathbf{x}_{t,l})|}{\|\mathbf{x}_{t,g} - \mathbf{x}_{t,l}\|_2} \geq \epsilon_2\right) \leq \frac{1}{\epsilon_2}\mathbb{E}\left[\frac{|f(\mathbf{x}_{t,g}) - f(\mathbf{x}_{t,l})|}{\|\mathbf{x}_{t,g} - \mathbf{x}_{t,l}\|_2}\right] \leq \frac{M}{\epsilon_2}, \tag{36}$$

where M is the Lipschitz constant of function f, because f is M-Lipschitz continuous and $\mathbf{x}_{t,g}, \mathbf{x}_{t,l} \in \mathcal{X}$. □

3.3 Proof of Theorem 1

Now we present the proof of Theorem 1 here.

Proof. The probability of $|r_{t,g} - r_{t,l}| < \epsilon_l$ can be written as

$$\begin{aligned}\mathbb{P}\big(|r_{t,g} - r_{t,l}| < \epsilon_l\big) &= \mathbb{P}\big(|(f(\mathbf{x}_{t,g}) - f(\mathbf{x}^\dagger)) - (f(\mathbf{x}_{t,l}) - f(\mathbf{x}^\dagger))| < \epsilon_l\big) \\ &= \mathbb{P}\left(\|\mathbf{x}_{t,g} - \mathbf{x}_{t,l}\|_2 \cdot \frac{|f(\mathbf{x}_{t,g}) - f(\mathbf{x}_{t,l})|}{\|\mathbf{x}_{t,g} - \mathbf{x}_{t,l}\|_2} < \epsilon_l\right).\end{aligned} \tag{37}$$

We define two events:

$$E_1 = \big(\|\mathbf{x}_{t,g} - \mathbf{x}_{t,l}\|_2 < \epsilon_1\big) \quad \text{and} \quad E_2 = \left(\frac{|f(\mathbf{x}_{t,g}) - f(\mathbf{x}_{t,l})|}{\|\mathbf{x}_{t,g} - \mathbf{x}_{t,l}\|_2} < \epsilon_2\right). \tag{38}$$

Then, (37) can be expressed as

$$\mathbb{P}\big(|r_{t,g} - r_{t,l}| < \epsilon_l\big) = \mathbb{P}\big(E_1 \cap E_2\big), \tag{39}$$

where $\epsilon_l = \epsilon_1\epsilon_2$. Thus, (37) can be written as

$$\mathbb{P}\big(E_1 \cap E_2\big) = 1 - \mathbb{P}\big(E_1^c \cup E_2^c\big) \geq 1 - \mathbb{P}\big(E_1^c\big) - \mathbb{P}\big(E_2^c\big), \tag{40}$$

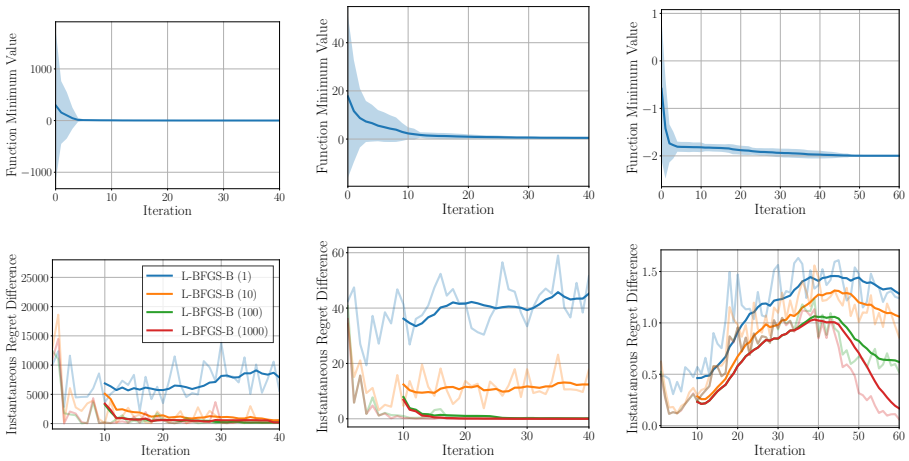
since $\mathbb{P}\big(E_1^c \cup E_2^c\big) \leq \mathbb{P}\big(E_1^c\big) + \mathbb{P}\big(E_2^c\big)$ by Boole's inequality. Then, we have

$$\begin{aligned}\mathbb{P}\big(E_1 \cap E_2\big) &\geq 1 - \mathbb{P}\big(\|\mathbf{x}_{t,g} - \mathbf{x}_{t,l}\|_2 \geq \epsilon_1\big) - \mathbb{P}\left(\frac{|f(\mathbf{x}_{t,g}) - f(\mathbf{x}_{t,l})|}{\|\mathbf{x}_{t,g} - \mathbf{x}_{t,l}\|_2} \geq \epsilon_2\right) \\ &\geq 1 - \frac{\gamma}{\epsilon_1}(1 - \beta_g) - \frac{M}{\epsilon_2},\end{aligned} \tag{41}$$

where Lemma 7 and Lemma 8 are used to arrive at the last inequality. Therefore, the proof is completed:

$$\mathbb{P}\big(|r_{t,g} - r_{t,l}| < \epsilon_l\big) \geq 1 - \delta_l, \tag{42}$$

where $\delta_l = \frac{\gamma}{\epsilon_1}(1 - \beta_g) + \frac{M}{\epsilon_2}$. □



(a) Beale (b) Branin (c) Cosines (2 dim.)

Fig. 1. Empirical results on Theorem 1 and Theorem 2. The caption of each figure indicates a target function. The upper panel of each figure is an optimization result for a global optimizer, and the lower panel is the regret differences between a global optimizer and four types of local optimizer (i.e., multi-started local optimizers found by starting from {1, 10, 100, 1000} initial points). Some legends of the lower panels are missed not to interfere with the graphs, but all the legends are same. For the lower panels, transparent lines are observed instantaneous regret differences and solid lines are moving average (10 steps) of the transparent lines. All experiments are repeated 50 times.

As described above, Theorem 1 implies that the regret difference is basically controlled by γ, β_g, and M. For example, if ρ_t is close to one, the regret difference is tight with high probability. On the other hand, if ρ_t goes to infinity, the difference is tight with low probability.

3.4 On Theorem 2

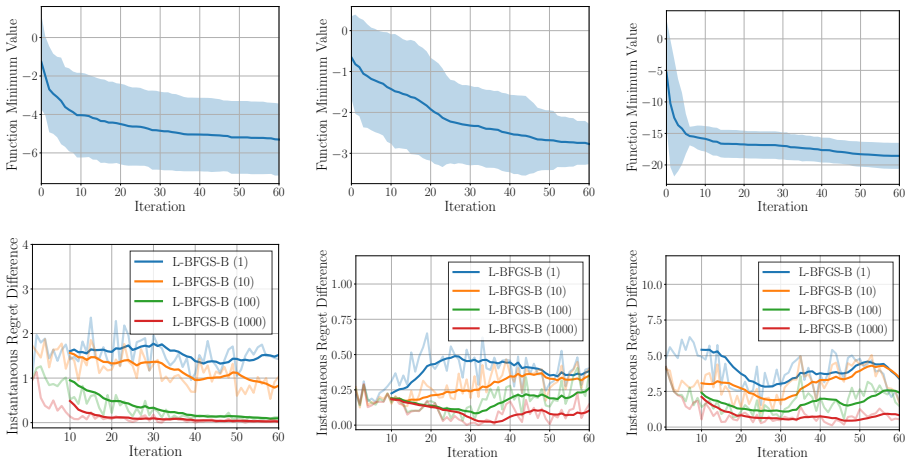
We extend Theorem 1 into the version for a multi-started local optimizer defined in Definition 3. To prove the next theorem, we need to prove Lemma 9.

Lemma 9. *Let the number of initial points for a multi-started local optimizer be N. A global optimizer and a multi-started local optimizer are different with a probability:*

$$\mathbb{P}\big(\mathbf{x}_{t,g} \neq \mathbf{x}_{t,m}\big) = (1-\beta_g)^N\,, \tag{43}$$

where $\mathbf{x}_{t,m}$ is determined by (6).

Proof. Since each initial condition of local optimizer is independently sampled, N local optimization methods started from different initial points are independently run. Therefore, the proof is obvious. □



(a) Cosines (8 dim.) (b) Hartmann6D (c) Holdertable

Fig. 2. Empirical results on Theorem 1 and Theorem 2. All settings follow the settings described in Fig. 1.

Because $N \geq 1$ and (43) is less than one, (43) is decreased as N is increased. For instance, these are satisfied:

$$(1-\beta_g)^N \leq 1-\beta_g \quad \text{and} \quad \lim_{N\to\infty}(1-\beta_g)^N = 0. \tag{44}$$

By Lemma 9, we can prove the theorem for the local optimization method started from multiple initial points. Before introducing Theorem 2, we simply derive Corollary 1.

Corollary 1. *l_2 distance between the acquired points $\mathbf{x}_{t,g}$ and $\mathbf{x}_{t,m}$ from (4) and (6) at iteration t is larger than any $\gamma > \epsilon_3 > 0$ with a probability:*

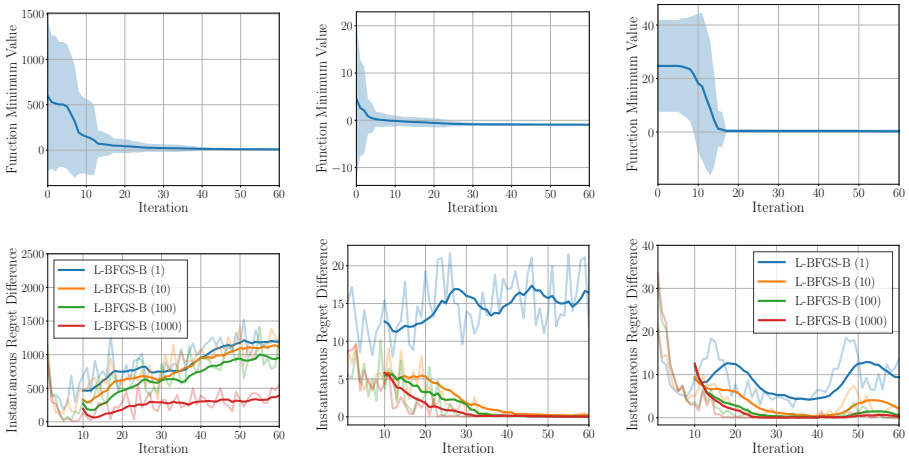
$$\mathbb{P}\big(\|\mathbf{x}_{t,g} - \mathbf{x}_{t,m}\|_2 \geq \epsilon_3\big) \leq \frac{\gamma}{\epsilon_3}(1-\beta_g)^N. \tag{45}$$

Proof. Because it can be proved in the same manner of Lemma 7, it is trivial. □

We provide the proof of Theorem 2 with the above lemmas.

Proof. It is an extension of Theorem 1. By Lemma 8 and Corollary 1, it is proved in the same way. □

As we mentioned before, because the equations in (44) are satisfied, we can emphasize a lower-bound on the probability of the case using a multi-started local optimizer is tighter than the case using a local optimizer. It implies an appropriate multi-started local optimizer can produce a similar convergence quality with the global optimizer without expensive computational complexity.



(a) Rosenbrock (b) Six Hump Camel (c) Sphere

Fig. 3. Empirical results on Theorem 1 and Theorem 2. All settings follow the settings described in Fig. 1.

4 Empirical Analysis

We present empirical analyses for Theorem 1 and Theorem 2, demonstrating the acquisition function optimization with global, local, and multi-started local optimizers on various examples: Beale, Branin, Cosines (2 dim. and 8 dim.), Hartmann6D, Holdertable, Rosenbrock, Six Hump Camel, and Sphere functions, which are widely used as benchmark functions in the Bayesian optimization literature. We use Gaussian process regression with Mátern 5/2 kernel as a surrogate function and EI as an acquisition function. In addition, the hyperparameters (e.g., signal scale and lengthscales) of Gaussian process regression are found by maximizing the marginal likelihood. All the experiments are implemented with `bayeso` [9].

To show regret differences, we need to sync up the historical observations for all the methods. The Bayesian optimization results via DIRECT are compared to the results via local optimization methods, considering each of the results by DIRECT as a true global solution for the acquisition function given. At each iteration, four L-BFGS-B algorithms started from {1, 10, 100, 1000} different initial points find the next point to measure a regret difference. The points found by L-BFGS-B are only used to compute the regret differences. The transparent lines described in the lower panels of Fig. 1 and Fig. 2 are the observed regret differences, and the solid lines are the moving averages of the transparent lines, each of which is computed as the unweighted mean of the previous 10 steps.

As shown in Fig. 1 and Fig. 2, the regret difference at each iteration is decreased as N is increased, which supports the main theorems. For some cases, the regret differences are slightly increased as the optimization step is repeated. It

Table 1. Time (sec.) consumed in optimizing acquisition functions.

	1(a)	1(b)	1(c)	2(a)	2(b)	2(c)	3(a)	3(b)	3(c)
DIRECT	3.434	2.987	2.306	2.508	0.728	2.935	13.928	4.639	10.707
L-BFGS-B (1)	0.010	0.004	0.052	0.023	0.026	0.017	0.005	0.010	0.030
L-BFGS-B (10)	0.096	0.036	0.515	0.224	0.253	0.177	0.050	0.100	0.311
L-BFGS-B (100)	0.977	0.363	5.173	2.224	2.533	1.760	0.504	0.969	3.048
L-BFGS-B (1000)	9.720	3.633	51.818	22.306	25.305	17.629	5.049	9.682	30.764

means that the acquisition function at the latter iteration has relatively many local optima, which is usually observed in the Bayesian optimization procedures. Furthermore, Table 1 shows Bayesian optimization with a multi-started optimizer is a fair and efficient choice for most of the cases. However, the cases using 1000-started optimizer tends to be slower than the ones using DIRECT, which implies choosing the adequate number of initial conditions for a multi-started local optimizer is significant and it should be carefully selected.

5 Conclusion

In this paper, we theoretically and empirically analyze the upper-bound of instantaneous regret difference between two regrets occurred by global and local optimizers for an acquisition function. The probability on this bound becomes tighter, using a multi-started local optimizer instead of the local optimizer. Our experiments show our theoretical analyses can be supported.

References

1. Barbet, L., Dambrine, M., Daniilidis, A., Rifford, L.: Sard theorems for Lipschitz functions and applications in optimization. Israel Journal of Mathematics **212**(2), 757–790 (2016)
2. Brochu, E., Cora, V.M., de Freitas, N.: A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv preprint arXiv:1012.2599 (2010)
3. Chowdhury, S.R., Gopalan, A.: On kernelized multi-armed bandits. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 844–853. Sydney, Australia (2017)
4. Duvenaud, D.: Automatic model construction with Gaussian processes. Ph.D. thesis, University of Cambridge (2014)
5. Frazier, P.I.: A tutorial on Bayesian optimization. arXiv preprint arXiv:1807.02811 (2018)
6. Hansen, N.: The CMA evolution strategy: A tutorial. arXiv preprint arXiv:1604.00772 (2016)

7. Hutter, F., Hoos, H.H., Leyton-Brown, K.: Sequential model-based optimization for general algorithm configuration. In: Proceedings of the International Conference on Learning and Intelligent Optimization (LION). pp. 507–523. Rome, Italy (2011)
8. Jones, D.R., Perttunen, C.D., Stuckman, B.E.: Lipschitzian optimization without the Lipschitz constant. Journal of Optimization Theory and Applications **79**(1), 157–181 (1993)
9. Kim, J., Choi, S.: bayeso: A Bayesian optimization framework in Python. `http://bayeso.org` (2017)
10. Kushner, H.J.: 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 (1964)
11. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Mathematical Programming **45**(3), 503–528 (1989)
12. Moćkus, J., Tiesis, V., Źilinskas, A.: The application of Bayesian methods for seeking the extremum. Towards Global Optimization **2**, 117–129 (1978)
13. Picheny, V., Gramacy, R.B., Wild, S., Le Digabel, S.: Bayesian optimization under mixed constraints with a slack-variable augmented Lagrangian. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 29, pp. 1435–1443. Barcelona, Spain (2016)
14. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006)
15. Robinson, P.D., Wathen, A.J.: Variational bounds on the entries of the inverse of a matrix. IMA Journal of Numerical Analysis **12**(4), 463–486 (1992)
16. Sard, A.: The measure of the critical values of differentiable maps. Bulletin of the American Mathematical Society **48**(12), 883–890 (1942)
17. Shahriari, B., Swersky, K., Wang, Z., Adams, R.P., de Freitas, N.: Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE **104**(1), 148–175 (2016)
18. Srinivas, N., Krause, A., Kakade, S., Seeger, M.: Gaussian process optimization in the bandit setting: No regret and experimental design. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 1015–1022. Haifa, Israel (2010)
19. Stein, M.L.: Interpolation of Spatial Data: Some Theory for Kriging. Springer New York (1999)
20. Wang, Z., Gehring, C., Kohli, P., Jegelka, S.: Batched large-scale Bayesian optimization in high-dimensional spaces. In: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS). pp. 745–754. Lanzarote, Spain (2018)
21. Wilson, A.G., Adams, R.P.: Gaussian process kernels for pattern discovery and extrapolation. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 1067–1075. Atlanta, Georgia, USA (2013)
22. Wilson, J.T., Hutter, F., Deisenroth, M.P.: Maximizing acquisition functions for Bayesian optimization. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 31, pp. 9906–9917. Montreal, Quebec, Canada (2018)