On Local Optimizers of Acquisition Functions in Bayesian Optimization

Jungtaek Kim (jtkim@postech.ac.kr)

Department of Computer Science and Engineering, POSTECH,
77 Cheongam-ro, Nam-gu, Pohang 37673,
Gyeongsangbuk-do, Republic of Korea

Presented at ECML-PKDD 2020

Joint work with Seungjin Choi

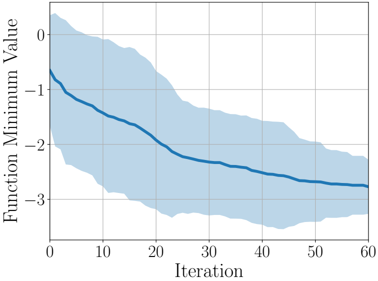
Table of Contents

Overview

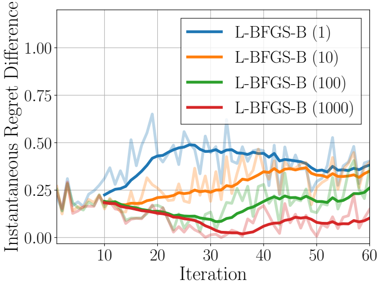
Overview

- **Bayesian optimization**: a sample-efficient method for finding a global optimum of an expensive black-box function.
- A **global optimizer** of acquisition function should be found at each round and selected as the next query point.
- In practice, however, **local optimizers** of acquisition function are also used, since searching for the global optimizer is often a non-trivial or time-consuming task.
- We present a performance analysis on the behavior of local optimizers of those acquisition functions, in terms of **instantaneous regrets** over global optimizers.
- Then, we introduce an analysis, allowing a local optimization method to start from **multiple different initial conditions**.

Intuition



(a) Optimization w/ global optimizers



(b) Instantaneous regret difference over global optimizers

Figure 1: Results on Hartmann6D function.

In-Depth Explanation

Motivation

- Bayesian optimization sequentially finds a global optimum of a **black-box objective function** $f(\mathbf{x}) : \mathcal{X} \to \mathbb{R}$ defined over a compact set $\mathcal{X} \subset \mathbb{R}^d$:

$$\mathbf{x}^\dagger = \underset{\mathbf{x} \in \mathcal{X}}{\arg\min} f(\mathbf{x}). \tag{1}$$

- Instead of optimizing f directly, it optimizes an **acquisition function**, which is defined with a surrogate model (e.g., Gaussian process regression).
- A **global optimizer** of acquisition function should be found at each round. However, in practice, **local optimizers** of acquisition function are used.
- To the best of our knowledge, this work is the first study which analyzes the **difference between global and local optimizers** in terms of instantaneous regrets.

Definitions

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:

$$\mathbf{x}_{t,g} = \overset{global}{\arg\max}_{\mathbf{x}\in\mathcal{X}}\, a(\mathbf{x}|\mathcal{D}_{t-1}). \tag{2}$$

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

Definitions

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*:*

$$\mathbf{x}_{t,l} = \overset{local}{\arg\max}_{\mathbf{x} \in \mathcal{X}} \, a(\mathbf{x}|\mathcal{D}_{t-1}). \tag{3}$$

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

Definitions

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* (3), *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:*

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

Definitions

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*

$$r_t = f(\mathbf{x}_t) - f(\mathbf{x}^\dagger). \tag{5}$$

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* $r_{t,m} = f(\mathbf{x}_{t,m}) - f(\mathbf{x}^\dagger)$.

Definitions

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 $\mathbf{x}_{t,m}$:

$$|r_{t,g} - r_{t,l}| = |f(\mathbf{x}_{t,g}) - f(\mathbf{x}_{t,l})|\,, \quad (6)$$

$$|r_{t,g} - r_{t,m}| = |f(\mathbf{x}_{t,g}) - f(\mathbf{x}_{t,m})|\,, \quad (7)$$

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

Main Theorems

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 $1 - \delta_l$:

$$\mathbb{P}\big(\,|r_{t,g} - r_{t,l}| < \epsilon_l\big) \geq 1 - \delta_l, \tag{8}$$

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.

Main Theorems

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 $1-\delta_m$:

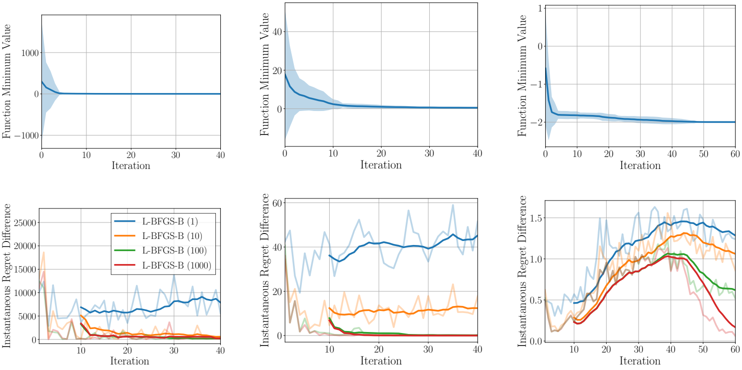
$$\mathbb{P}\big(\,|r_{t,g} - r_{t,m}| < \epsilon_m\big) \geq 1 - \delta_m, \tag{9}$$

where $\delta_m = \frac{\gamma}{\epsilon_3}\left(1-\beta_g\right)^N + \frac{M}{\epsilon_2}$, $\epsilon_m = \epsilon_2\epsilon_3$, $\gamma = \max_{\mathbf{x}_i,\mathbf{x}_j \in \mathcal{X}} \|\mathbf{x}_i - \mathbf{x}_j\|_2$ 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.

Take-Home Message

- As shown in the main theorems, the probability $1 - \delta_l$ is controlled by three statements related to γ, β_g, and M.
- For example, $1 - \delta_l$ is decreased (i) as γ is increased, (ii) as β_g is decreased, and (iii) as M is increased.
- Theorem 2 suggests similar implications with Theorem 1, but their main difference is that δ_m is additionally related to the number of initial points N.
- 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.
- Furthermore, an appropriate multi-started local optimizer can produce a similar convergence quality with the global optimizer, without the expensive computational complexity.

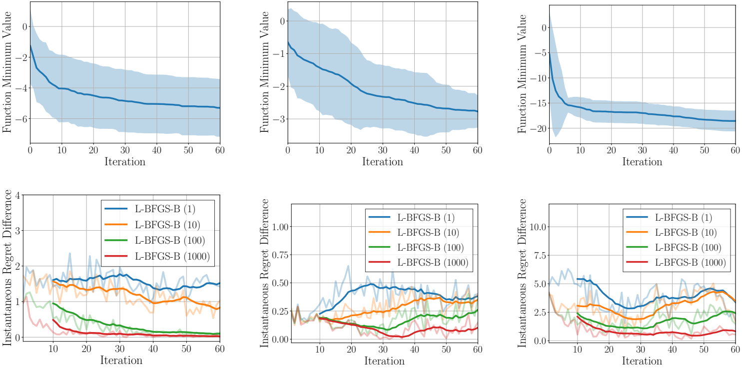
Empirical Analysis



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

Figure 2: Empirical results on Theorem 1 and Theorem 2. For the lower panels, transparent lines are observed instantaneous regret differences and solid lines are moving average (10 steps) of the transparent lines.

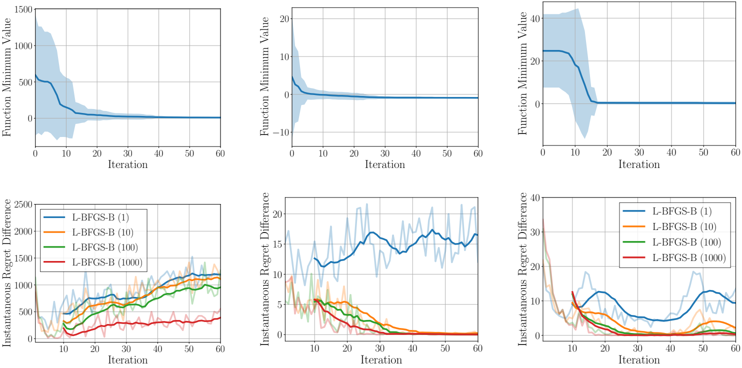
Empirical Analysis



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

Figure 3: Empirical results on Theorem 1 and Theorem 2.

Empirical Analysis



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

Figure 4: Empirical results on Theorem 1 and Theorem 2.

Empirical Analysis

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

	Beale	**Branin**	**Cosines (2 dim.)**
DIRECT	3.434	2.987	2.306
L-BFGS-B (1)	0.010	0.004	0.052
L-BFGS-B (10)	0.096	0.036	0.515
L-BFGS-B (100)	0.977	0.363	5.173
L-BFGS-B (1000)	9.720	3.633	51.818

Empirical Analysis

Table 2: Time (sec.) consumed in optimizing acquisition functions.

	Cosines (8 dim.)	**Hartmann6D**	**Holdertable**
DIRECT	2.508	0.728	2.935
L-BFGS-B (1)	0.023	0.026	0.017
L-BFGS-B (10)	0.224	0.253	0.177
L-BFGS-B (100)	2.224	2.533	1.760
L-BFGS-B (1000)	22.306	25.305	17.629

Empirical Analysis

Table 3: Time (sec.) consumed in optimizing acquisition functions.

	Rosenbrock	Six Hump Camel	Sphere
DIRECT	13.928	4.639	10.707
L-BFGS-B (1)	0.005	0.010	0.030
L-BFGS-B (10)	0.050	0.100	0.311
L-BFGS-B (100)	0.504	0.969	3.048
L-BFGS-B (1000)	5.049	9.682	30.764

Conclusion

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

Thank you for listening!