A stopping criterion for Bayesian optimization by the gap of expected minimum simple regrets

Hideaki Ishibashi
Kyushu Institute of Technology

Masayuki Karasuyama
Nagoya Institute of Technology

Ichiro Takeuchi
Nagoya University/RIKEN AIP

Hideitsu Hino
The Institute of Statistical Mathematics/RIKEN AIP

Abstract

Bayesian optimization (BO) improves the efficiency of black-box optimization; however, the associated computational cost and power consumption remain dominant in the application of machine learning methods. This paper proposes a method of determining the stopping time in BO. The proposed criterion is based on the difference between the expectation of the minimum of a variant of the simple regrets before and after evaluating the objective function with a new parameter setting. Unlike existing stopping criteria, the proposed criterion is guaranteed to converge to the theoretically optimal stopping criterion for any choices of arbitrary acquisition functions and threshold values. Moreover, the threshold for the stopping criterion can be determined automatically and adaptively. We experimentally demonstrate that the proposed stopping criterion finds reasonable timing to stop a BO with a small number of evaluations of the objective function.

1 INTRODUCTION

This paper focuses on deriving a criterion to determine a reasonable stopping time for Bayesian optimization (BO). The aim of BO is to find the global optimal parameters of an unknown and costly-to-evaluate objective function efficiently while balancing the search for a surrogate model to approximate the objective function and using the current surrogate model to minimize it[1]. It has a wide range of applications, from hyperparameter searching in deep learning (Snoek et al., 2012) to the development of materials with desirable properties (Frazier and Wang, 2015).

In a BO procedures, the black box objective function is evaluated at the point (or parameter setting) selected based on the acquisition function. Evaluation of the black box function sometimes requires an iterative procedure; hence we consider BO to be composed of two loops in principle. For example, suppose the aim of BO is to determine the hyperparameter of a predictive model that minimizes the generalization error. In this case, BO selects a hyperparameter of the predictive model (e.g., network architecture, regularization parameter) in the *outer-loop*, and the training procedure of a model with the fixed hyperparameter that corresponds to the *inner-loop*. Many studies have focused on accelerating BO via sophisticated inner-loop optimization, most of which are considered as an early stopping methods (Prechelt, 2012).

A complementary approach to the early stopping of the inner-loop is to accelerate the BO by stopping the outer-loop (i.e., the BO itself) at an appropriate timing. BO is a method of efficiently searching for optimal points (parameter settings) while avoiding an exhaustive search; it is desirable to stop the search after finding the parameter that sufficiently minimizes the objective function. If we stop the optimization before establishing a sufficiently good surrogate model, we will lose the opportunity to find a better parameter. Terminating BO at an appropriate time is important not only to reduce the computational time, but, possibly more importantly, also to reduce the electric power consumption.

When BO is used for the hyperparameter optimization (HPO) of a predictive model, Makarova et al. (2022) proposes a stopping criterion based on the simple regret. They estimated a stochastic upper bound of the simple regret and

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

[1]Without loss of generality, we consider minimization problem in this paper.

compared it with the estimated standard deviation (SD) of the generalization error by cross-validation (CV). Although their criterion gives us a theoretical guarantee that the simple regret of BO is lower than the estimated SD of the generalization error with high probability, there is no guarantee that the upper bound will converge to zero, and there remains a possibility that the criterion will fail to terminate the BO. In addition, it remains difficult to adaptively determine the threshold in general BO settings other than HPO.

We propose a stopping criterion for BO. This method is based on the gap between the expectation of the minimum of a variant of simple regret and our main contributions are summarized as follows:
1: We develop a stopping criterion for BO based on the difference between the expected minimum simple regret before and after the evaluation of the objective function at a new point, which quantifies the amount of decrease in the expectation of the minimizer of the sampled surrogate function (similar to the Thompson sampling) from a posterior distribution at each iteration. We guarantee that the value of the proposed criterion will converge to zero with high probability under certain assumptions. Therefore, the proposed method terminates the BO when the search is sufficiently progressed. The proposed criterion is applicable to BO with any acquisition function as long as the surrogate function is a Gaussian process.

2: We propose two threshold determination methods that can be used universally: One determines the threshold adaptively and the other determines the threshold heuristically by using a parameter that is robust to changes of task and dataset.

2 BAYESIAN OPTIMIZATION

Let Θ be the domain of the point θ, and let the bounded unknown and costly to evaluate objective function to be minimized with respect to $\theta \in \Theta$ be $f : \Theta \to \mathbb{R}$. We do not assume f has a unique minimum, but we consider one of its minima for the sake of simplicity. The aim of BO is to find the point θ^* that minimizes the objective function:

$$\theta^* = \operatorname*{argmin}_{\theta \in \Theta} f(\theta).$$

We denote the minimum value of the objective function as $f^* = f(\theta^*)$. To find the optimal point θ^*, BO uses a surrogate function—denoted by $\hat{f}$—and iterates the following two procedures: (i) Estimate the distribution $p(\hat{f})$ of the surrogate function using the points and their values measured in the past, and (ii) select the point that maximizes the acquisition function $\alpha : \Theta \to \mathbb{R}$ defined based on the the estimated distribution $p(\hat{f})$, and evaluate the objective function at that selected point.

In theoretical analysis, the goodness of the explored points is generally evaluated using the simple regret. Let Θ_t be the set of points explored in time t. The best point in Θ_t is $\theta_t^* = \operatorname{argmin}_{\theta \in \Theta_t} f(\theta)$, and the corresponding true best value is $f(\theta_t^*)$. Then, the simple regret for θ_t^* is defined as follows:

$$r(\theta_t^*) = f(\theta_t^*) - f^*.$$

Herein, we assume that a distribution of the surrogate function $\hat{f}$ is represented by a Gaussian process (GP) (Rasmussen and Williams, 2006). Note that we distinguish the function value $\hat{f}(\theta)$ from its observed value y. Suppose that a set of N distinct points and corresponding observed values $S = \{(\theta_n, y_n)\}_{n=1}^N$ are given. Let $p(\hat{f}(\theta)) = \mathcal{N}(\hat{f}(\theta) | m(\theta), k(\theta, \theta))$ be a prior of $\hat{f}$, where m and k are the mean function and the kernel function, and $p(y|\theta, \hat{f}) = \mathcal{N}(y|\hat{f}(\theta), \lambda^{-1})$ be the likelihood, where λ is a precision parameter of noise. Given a new point θ, the joint distribution of $\hat{f}(\theta)$ and $\mathbf{y} \coloneqq (y_1, y_2, \ldots, y_N)$ is written as:

$$\begin{bmatrix} \mathbf{y} \\ \hat{f}(\theta) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m} \\ m(\theta) \end{bmatrix}, \begin{bmatrix} \tilde{\mathbf{K}} & \mathbf{k}(\theta) \\ \mathbf{k}^{\mathrm{T}}(\theta) & k(\theta, \theta) \end{bmatrix} \right),$$

where $\tilde{\mathbf{K}} = \mathbf{K} + \lambda^{-1}\mathbf{I}$, $[\mathbf{K}]_{i,j} = k(\theta_i, \theta_j)$, $\mathbf{k}(\theta) = (k(\theta_n, \theta))_{n=1}^N \in \mathbb{R}^N$, and $\mathbf{m} = (m(\theta_n))_{n=1}^N \in \mathbb{R}^N$.

Then, the posterior distribution $p(\hat{f}(\theta) \mid \mathbf{y}) = \mathcal{N}(\hat{f}(\theta) \mid \mu(\theta), \sigma^2(\theta, \theta))$ is written in closed form with

$$\mu(\theta) = m(\theta) + \mathbf{k}^{\mathrm{T}}(\theta) \left(\mathbf{K} + \lambda^{-1}\mathbf{I}\right)^{-1} (\mathbf{y} - \mathbf{m}),$$
$$\sigma^2(\theta, \theta) = k(\theta, \theta) - \mathbf{k}^{\mathrm{T}}(\theta) \left(\mathbf{K} + \lambda^{-1}\mathbf{I}\right)^{-1} \mathbf{k}(\theta).$$

In the following, $\sigma^2(\theta, \theta)$ is abbreviated as $\sigma^2(\theta)$. In BO based on GPs, at each iteration, the point that maximizes the acquisition function $\alpha : \Theta \to \mathbb{R}$ is chosen where the acquisition function is defined using the posterior distribution $p(\hat{f} \mid \mathbf{y})$ as

$$\theta_t = \operatorname*{argmax}_{\theta \in \Theta} \alpha(\theta; p(\hat{f} \mid \mathbf{y})).$$

Commonly used acquisition functions include the probability of improvement (PI) (Kushner, 1964), expected improvement (EI) (Mockus et al., 1978; Jones et al., 1998) and Gaussian process upper confidence bound (GP-UCB) (Srinivas et al., 2010). There are also the information-based approach called the entropy search (Hennig and Schuler, 2012; Hernández-Lobato et al., 2014), the randomized strategy known as Thompson sampling (Thompson, 1933), and the "one-step" analysis-based method known as the knowledge gradient (Frazier et al., 2009). Novel acquisition functions are continually being proposed, e.g., Siemenn et al. (2022).

3 EXISTING STOPPING CRITERIA FOR BAYESIAN OPTIMIZATION

To realize an efficient parameter search, the timing of terminating the BO is important. The most common method is

to search a predetermined number of times, but the appropriate number of searches varies depending on the objective function, its surrogate function, the given dataset, and the acquisition function. Another possible way is to stop searching when the value of the objective function becomes lower than a predetermined threshold. However, applying this method is difficult when the desired or acceptable value of the objective function is unknown. Even when the desired value is known; when the efficiency of the searching is low, the desired value becomes unobtainable and so the search may be continued endlessly.

Lorenz et al. (2015) proposed a stopping criterion that stops the search when the probability of improvement (PI) of the input point selected by the EI acquisition function falls below a threshold. They claim that this can be regarded as a hypothesis test where the null hypothesis is $H_0 : \mu(\theta) < f(\theta_t)$, and the threshold is regarded as the acceptable type-I error for the test. It is advantageous that the threshold is interpreted as the acceptable type-I error. However, since this method assumes that the posterior distribution is correct, there is a possibility of erroneous stopping, when the mean function of the posterior distribution does not sufficiently represent the true function.

Nguyen et al. (2017) assumed that the acquisition function is EI and proposed stopping BO when the EI of the newly selected point falls below a threshold. No objective method for setting the threshold has been provided.

Makarova et al. (2022) focused on the problem of HPO of predictive models, and they proposed a stopping criterion for BO based on the simple regret, $r(\theta_t^*)$, of the chosen point. Specifically, for any $\delta \in (0, 1)$, they showed that the following inequality holds with probability $1 - \delta$:

$$r(\theta_t^*) \leq \min_{\theta \in \Theta_t} \mathrm{UCB}_\delta(\theta) - \min_{\theta \in \Theta} \mathrm{LCB}_\delta(\theta), \qquad (1)$$

where $\mathrm{UCB}_\delta(\theta) = \mu(\theta) + \beta_t^{1/2}\sigma(\theta)$, $\mathrm{LCB}_\delta(\theta) = \mu(\theta) - \beta_t^{1/2}\sigma(\theta)$ and $\beta_t^{1/2}$ is a trade-off parameter between exploration and exploitation depending on δ. They propose to stop the search when this upper bound falls below a threshold, s_t. The notable advantage of this method is that the threshold s_t can be automatically and adaptively determined when the BO is used in the HPO as is explained in Section 4.3. However, there is no guarantee that the upper bound will converge to zero with certainty, and it remains difficult to determine the threshold adaptively in a general BO setting other than HPO. The stopping criteria for other optimization methods are explained in Section E in the appendix.

BOs have been extensively used for HPO. Therefore, several BO stopping criteria particularly designed for HPO tasks were proposed. Swersky et al. (2014) explored a diverse collection of hyperparameter settings at the initial stage by training their predictive models with a small number of epochs, and then gradually focusing on a small number of promising settings by managing learning processes with different hyperparameters. Dai et al. (2019) combined BO with Bayesian optimal stopping (Ferguson, 2006). In the context of multi-fidelity BO, several methods have been proposed to reduce the resource consumption of BO by utilizing low-fidelity functions that can be obtained by using a subset of the training data or by training the machine learning model for just a few epochs (Kandasamy et al., 2016, 2017; Li et al., 2021). It is also proposed to dynamically allocate a smaller budget to the less-promising hyperparameter setting while giving a larger budget to the promising hyperparameter (Li et al., 2017; Falkner et al., 2018). Other approaches for the early stopping of the inner-loop include the extrapolation of the generalization error curve (Domhan et al., 2015; Klein et al., 2017).

4 PROPOSED STOPPING CRITERION

To address problems with existing stopping criteria, we propose a criterion based on the *difference between the expectation of the minimum of a variant of the simple regret*. The proposed criterion terminates the BO in a finite number of iterations with high probability since the proposed criterion converges to zero under certain assumptions. In addition, we propose two threshold determination methods; one determines the threshold automatically, and the other determines the threshold heuristically by using a parameter that is robust to the changes in the task and dataset.

Instead of considering the standard simple regret, we consider the following quantity:

$$R_t = \mathbb{E}_{p(\hat{f}|\mathbf{y}_t)}[\min_{\theta \in \Theta}\{\hat{f}(\theta)\}] - f^*.$$

Herein, R_t is called the *expected minimum simple regret* since R_t indicates the expectation of a variant of the simple regret when the optimal point is selected by the "oracle" for each $\hat{f}$ generated from $p(\hat{f}|\mathbf{y}_t)$. The quantity R_t can also be regarded as the expectation of the minimum obtained by sampling from the posterior (i.e., Thompson sampling) minus f^*, as shown in Figure 1. When BO has not yet explored around the optimal solutions yet, R_t does not converge to zero because there is a low probability of generating functions that attains the optimal function value f^* from $p(\hat{f}|\mathbf{y}_t)$. Therefore, when this quantity converges to zero sufficiently, it indicates that BO has found the optimal solution. In addition, this quantity only evaluates the minimum value of functions sampled from $p(\hat{f}|\mathbf{y}_t)$; hence it can deal with an objective function with multiple minima.

4.1 Upper Bound of the Gap between the Expected Minimum Simple Regrets

We propose a stopping criterion for BO by evaluating the convergence of R_t. To this end, we consider the difference between the expected minimum simple regrets $\Delta R_t :=$

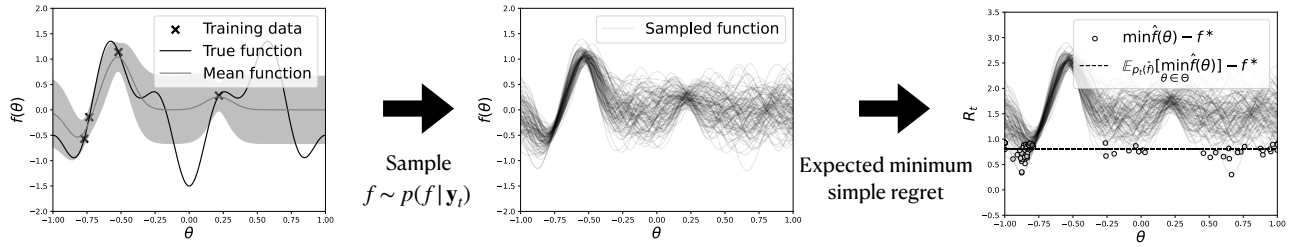


Figure 1: The concept of the proposed stopping criterion. The expected minimum simple regret is independent of the acquisition function, as it does not depend on how the next point is selected.

$|R_{t-1} - R_t|$ before and after adding (θ_t, y_t) to the training data. By evaluating the difference between the expected minimum simple regrets, we can stop BO without knowing f^* because it indicates that the search efficiency is low and there is almost no improvement in the objective value. However, it is generally difficult to calculate ΔR_t analytically. Therefore, this study derives an inequality shown in the following theorem to evaluate the upper bound of ΔR_t.

Let $D_{\mathrm{KL}}[p||q]$ be the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) between the probability distributions p and q. Then, the following theorem holds.

Theorem 1. *Let $p_t(\hat{f}) \coloneqq p(\hat{f}|\mathbf{y}_t)$ be the posterior distribution for $\hat{f}$ when $S_t = \{(\Theta_t, \mathbf{y}_t)\}$ is observed, and let μ_t and σ_t be the mean and covariance functions of $p_t(\hat{f})$. Pick $\delta \in (0,1)$, and let $\kappa_{t-1} \coloneqq \min_{\theta\in\Theta_{t-1}} \mathrm{UCB}_\delta(\theta) - \min_{\theta\in\Theta} \mathrm{LCB}_\delta(\theta)$. We denote the probability distribution function and cumulative distribution function of standard normal distribution by $\phi(x)$ and $\Phi(x)$, respectively. Then, the following inequality holds with probability $\geq 1-\delta$:*

$$\begin{aligned}\Delta R_t \leq & v(\phi(g) + g\Phi(g)) \\ & + |\Delta\mu_t^*| + \kappa_{t-1}\sqrt{\frac{1}{2}D_{\mathrm{KL}}[p_t(\hat{f})||p_{t-1}(\hat{f})]} \\ =: & \Delta\tilde{R}_t, \end{aligned} \tag{2}$$

where $\Delta\mu_t^* \coloneqq \mu_{t-1}(\theta_{t-1}^*) - \mu_t(\theta_t^*)$, $v \coloneqq \sqrt{\sigma_t^2(\theta_t^*) - 2\sigma_t^2(\theta_t^*, \theta_{t-1}^*) + \sigma_t^2(\theta_{t-1}^*)}$ and $g \coloneqq (\mu_t(\theta_t^*) - \mu_{t-1}(\theta_{t-1}^*))/v$.

Proof. We transform $\mathbb{E}_{p_t(\hat{f})}[\min_{\theta\in\Theta}\{\hat{f}(\theta)\}]$ to $\mathbb{E}_{p_t(\hat{f})}[\min_{\theta\in\Theta}\{\min\{\hat{f}(\theta), \hat{f}(\theta_t^*)\}\}]$ and use the relations $\max\{a,0\} = -\min\{-a,0\}$ and $\max\{a+b,0\} \leq \max\{a,0\} + \max\{b,0\}$. Then, we have

$$\begin{aligned}&\Delta R_t \\ \leq & |\Delta\mu_t^*| + \mathbb{E}_{p_t(\hat{f})}[\max\{\hat{f}(\theta_t^*) - \hat{f}(\theta_{t-1}^*), 0\}] \\ & + |\mathbb{E}_{p_t(\hat{f})}[\max_{\theta\in\Theta}\{\max\{\hat{f}(\theta_{t-1}^*) - \hat{f}(\theta), 0\}\}] \\ & - \mathbb{E}_{p_{t-1}(\hat{f})}[\max_{\theta\in\Theta}\{\max\{\hat{f}(\theta_{t-1}^*) - \hat{f}(\theta), 0\}\}]|. \end{aligned} \tag{3}$$

For any two-variate Gaussian distribution, the following equation holds.

$$\begin{aligned}&\mathbb{E}_{p(x,y)}[\max\{x-y, 0\}] \\ =&\sqrt{\sigma_{xx}^2 - 2\sigma_{xy}^2 + \sigma_{yy}^2}\,\phi\left(\frac{(\mu_x - \mu_y)}{\sqrt{\sigma_{xx}^2 - 2\sigma_{xy}^2 + \sigma_{yy}^2}}\right) \\ &+ (\mu_x - \mu_y)\Phi\left(\frac{(\mu_x - \mu_y)}{\sqrt{\sigma_{xx}^2 - 2\sigma_{xy}^2 + \sigma_{yy}^2}}\right), \end{aligned}$$

where μ_x and μ_y are mean of x and y, and σ_{xx}^2, σ_{yy}^2, σ_{xy}^2 are the variance of x, variance of y and covariance of x and y, respectively. Therefore, the second term of R.H.S. of the Eq. (3) is $v(\phi(g) + g\Phi(g))$.

For any measurable function $\mathcal{L}(\hat{f}) \in [a,b]$, $|\mathbb{E}_{p(\hat{f})}[\mathcal{L}(\hat{f})] - \mathbb{E}_{q(\hat{f})}[\mathcal{L}(\hat{f})]| \leq (b-a)\sqrt{1/2 D_{\mathrm{KL}}[p||q]}$ holds. Let $\mathcal{L}(\hat{f}) = \max_{\theta\in\Theta}\{\max\{\hat{f}(\theta_{t-1}^*) - \hat{f}(\theta), 0\}\} = \hat{f}(\theta_{t-1}^*) - \min_{\theta\in\Theta}\hat{f}(\theta)$. By using $|\hat{f}(\theta_t) - \mu_{t-1}(\theta_t)| \leq \beta_t^{1/2}\sigma_{t-1}(\theta_t)$, $\mathcal{L}(\hat{f})$ can be bound by κ_{t-1} with high probability. Meanwhile, the minimum value of $\mathcal{L}(\hat{f})$ is zero. From this, the sum of the third term and fourth terms of R.H.S. in Eq. (3) is bound by $\kappa_{t-1}\sqrt{1/2 D_{\mathrm{KL}}[p_t(\hat{f})||p_{t-1}(\hat{f})]}$, which proofs the claim. Detailed proof is shown in Section A in the appendix. □

Algorithm 1 is the concrete algorithm for computing the proposed stopping criterion. We propose stopping BO when $\Delta\tilde{R}_t$ becomes less than or equal to a threshold, s_t. To calculate the upper bound, we have to calculate θ_t^*, which cannot be strictly calculated when there is noise in the objective function. In this study, θ_t^* is approximated by using the explored point whose objective function is the minimum overall observed points.

We can guarantee that the derived upper bound becomes tighter and converges to zero as the BO process becomes sufficiently advanced by the following theorem.

Theorem 2. *Assume $\theta_t^* = \theta_{t-1}^*$ when $t \to \infty$, the acquisition function of the BO is GP-LCB, a hyperparameter of the GP is the same before and after adding data, the kernel*

function for GP is any of finite dimensional linear, squared exponential, and Matérn kernels satisfying $k(\theta,\theta) \leq 1$ for any $\theta \in \Theta$, and $\kappa_{t-1} < C$ holds for some constant value, C. For a sample $\hat{f}$ of GP with the kernel function $k(\theta,\theta')$, assume $\hat{f}$: $\exists a, b > 0, Pr\{\sup_{\theta\in\Theta} |\partial f/\partial\theta_j| > L\} \leq a\exp^{-(L/b)^2}$ holds. Then, the Eq. (2) *converges to zero with probability with high probability.*

Proof. Since we assume $\theta_t^* = \theta_{t-1}^*$, $\nu(\phi(g) + g\Phi(g)) = 0$ holds. Regarding $p(\hat{f}_t|\mathbf{y}_t)$ as the posterior whose prior is $p(\hat{f}_t|\mathbf{y}_{t-1})$ observing (θ_t, y_t), the mean function corresponding to θ are derived as

$$\mu_t(\theta) = \mu_{t-1}(\theta) + \frac{\sigma_{t-1}^2(\theta,\theta_t)}{\sigma_{t-1}^2(\theta_t,\theta_t)+\lambda^{-1}}(y_t - \mu_{t-1}(\theta_t)).$$

We decompose an observed value of an objective function for θ_t into a value of a true function corresponding to θ_t and noise term, that is $y_t = f(\theta_t) + \epsilon_t$. Then, $|f(\theta_t) - \mu_{t-1}(\theta_t)| \leq \beta_t^{1/2}\sigma_{t-1}(\theta_t)$ and $\epsilon_t \leq c/\lambda^{1/2}$ hold with probability $\geq 1-\delta$, respectively, where $c = \sqrt{-2\log\delta}$. From the assumption of $\theta_t^* = \theta_{t-1}^*$, we have $|\mu_{t-1}(\theta_t^*) - \mu_{t-1}(\theta_{t-1}^*)| = 0$. From these results, the following inequality holds.

$$|\Delta\mu_t^*| \leq \frac{\beta_t^{1/2}\sigma_{t-1}(\theta_t^*)\sigma_{t-1}^2(\theta_t)}{\sigma_{t-1}^2(\theta_t)+\lambda^{-1}} + \frac{\sigma_{t-1}(\theta_t^*)\sigma_{t-1}(\theta_t)c}{\lambda^{1/2}(\sigma_{t-1}^2(\theta_t)+\lambda^{-1})}.$$

Under assumption that a prior of $p_t(\hat{f})$ is the same as that of $p_t(\hat{f})$, the KL divergence between the GP posterior distribution $p_t(\hat{f})$ given S_t and the GP posterior distribution $p_{t-1}(\hat{f})$ given S_{t-1} can be bound as follows:

$$\begin{aligned}
&D_{\mathrm{KL}}[p_t(\hat{f})||p_{t-1}(\hat{f})] \\
=&\frac{1}{2}\log\left(1+\lambda\sigma_{t-1}^2(\theta_t)\right) - \frac{1}{2}\frac{\sigma_{t-1}^2(\theta_t)}{\sigma_{t-1}^2(\theta_t)+\lambda^{-1}} \\
&+\frac{1}{2}\frac{\sigma_{t-1}^2(\theta_t)(y_t-\mu_{t-1}(\theta_t))^2}{(\sigma_{t-1}^2(\theta_t)+\lambda^{-1})^2}. \\
\leq&\frac{1}{2}\log\left(1+\lambda\sigma_{t-1}^2(\theta_t)\right) - \frac{1}{2}\frac{\sigma_{t-1}^2(\theta_t)}{\sigma_{t-1}^2(\theta_t)+\lambda^{-1}} \\
&+\frac{1}{2}\frac{\sigma_{t-1}^2(\theta_t)\left(\lambda^{1/2}\beta_t^{1/2}\sigma_{t-1}(\theta_t)+c\right)^2}{\lambda(\sigma_{t-1}^2(\theta_t)+\lambda^{-1})^2},
\end{aligned}$$

where we used $|f(\theta_t) - \mu_t(\theta_t)| \leq \beta_t^{1/2}\sigma_{t-1}(\theta_t)$ and $\epsilon_t \leq c/\lambda^{1/2}$ in the equation transformation between line 2 and 3. If we use GP-LCB as an acquisition function, $\beta_t^{1/2}\sigma_t(\theta_t)$ converges to zero when $t \to \infty$ under some assumptions. Hence, $\Delta\tilde{R}_t$ converges to zero with high probability. Section C in the appendix provides a more detailed proof. □

We discuss the intuitive interpretation of the proposed upper bound. The first term means the expected improvement when θ_t^* changes, which is zero when $\theta_t^* = \theta_{t-1}^*$. The second term on the right-hand side of Eq. (2) can be interpreted as a change in the simple regret when we regard the mean function as a true function, which tends to become a small value as shown in Section F in the appendix. In the last term, $D_{\mathrm{KL}}[p_t(\hat{f})||p_{t-1}(\hat{f})]$ becomes small when there is no change in the posterior distribution, even when new points are added to the search. Meanwhile, κ_{t-1} is an upper bound of the simple regret, which also becomes small when the BO process advances. Therefore, there are two possible situations in which the proposed stopping criterion stops the BO procedure. 1) When a suboptimal parameter is found and the simple regret becomes sufficiently small to stop the search, even if the posterior distribution has not converged, because $r(\theta_t^*)$ becomes small; 2) When the posterior distribution has converged. This means that the parameter search stops even when the search efficiency is low and the simple regret remains large. For example, when the hyperparameter for exploration is too small in UCB, only limited regions with high objective values are explored before sufficiently exploring the entire search space; therefore, the posterior distribution does not change and $D_{\mathrm{KL}}[p_t(\hat{f})||p_{t-1}(\hat{f})]$ becomes small and the search is terminated. On the other hand, when the simple regret does not decrease while searching a region that is not well-explored, neither κ_{t-1} nor $D_{\mathrm{KL}}[p_t(\hat{f})||p_{t-1}(\hat{f})]$ decrease and the problem of terminating the search too early does not occur. Note that although the proposed method can stop a search procedure effectively, there would be situations in which we want to obtain a better solution even if it takes more time. In such cases, the proposed method can be used as a trigger to restart the search from a different initial point. We experimentally show behaviors of the proposed upper bound in Section F in the appendix.

Roughly speaking, the proposed criterion evaluates the BO by using a product of the upper bound of the simple regret and KL divergence between GPs. Since the KL divergence becomes large when the gap between the true function and the mean function of the posterior is large, the proposed criterion does not stop the BO. Therefore, the risk that the proposed criterion stops the BO by mistake is lower than the conventional methods. However, the possibility still remains that the proposed criterion could stop the BO by mistake because the upper bound of the simple regret may not hold when the gap between the true function and mean function of the posterior is large.

4.2 Formula for Computing The Upper Bound

Calculating the upper bound (2) requires calculating the KL divergence between GPs. Generally, the KL divergence does not necessarily have a finite value, because GPs have infinite-dimensional parameters. However, when the prior is common between two posteriors of GP, it can be shown that the KL divergence is equivalent to that between multi-

Algorithm 1 stopping criterion for Bayesian optimization

Input: thresholds $s_t > 0$, initial dataset $S_0 = \{\Theta_0, \mathbf{y}_0\}$
$\mu_0(\theta) \leftarrow m(\theta) + \mathbf{k}^{\mathrm{T}}(\theta)\left(\mathbf{K} + \lambda^{-1}\mathbf{I}\right)^{-1}(\mathbf{y}_0 - \mathbf{m})$
$\sigma_0^2(\theta,\theta) \leftarrow k(\theta,\theta) - \mathbf{k}^{\mathrm{T}}(\theta)\left(\mathbf{K} + \lambda^{-1}\mathbf{I}\right)^{-1}\mathbf{k}(\theta)$
for $t = 1, 2, \ldots$ **do**
  $\theta_t \leftarrow \mathrm{argmax}_{\theta\in\Theta}\, \alpha(\theta; p(\hat{f}|\mathbf{y}_t))$
  $S_t \leftarrow S_{t-1} \cup \{(\theta_t, y_t)\}$
  $\mu_t(\theta) \leftarrow m(\theta) + \mathbf{k}^{\mathrm{T}}(\theta)\left(\mathbf{K} + \lambda^{-1}\mathbf{I}\right)^{-1}(\mathbf{y}_t - \mathbf{m})$
  $\sigma_t^2(\theta,\theta) \leftarrow k(\theta,\theta) - \mathbf{k}^{\mathrm{T}}(\theta)\left(\mathbf{K} + \lambda^{-1}\mathbf{I}\right)^{-1}\mathbf{k}(\theta)$
  $\Delta\tilde{R}_t \leftarrow \Delta\mu_t^* + v\phi(g) + vg\Phi(g) + \kappa_{t-1}\sqrt{\frac{1}{2}D_{\mathrm{KL}}[p_t(\hat{f})||p_{t-1}(\hat{f})]}$
  if $\Delta\tilde{R}_t \le s_t$ **then**
    Terminate BO loop
  end if
end for

variate Gaussian distributions of finite dimension (Ishibashi and Hino, 2020). Specifically, by denoting the data set at time t as $S_t = \{(\Theta_t, \mathbf{y}_t)\}$, the KL divergence between the GP posterior distribution $p_t(\hat{f})$ given S_t and the GP posterior distribution $p_{t-1}(\hat{f})$ given S_{t-1} is as follows:[2]

$$\begin{aligned} &D_{\mathrm{KL}}[p_t(\hat{f})||p_{t-1}(\hat{f})] \\ =&\frac{1}{2}\log\left(1+\lambda\sigma_{t-1}^2(\theta_t)\right) - \frac{1}{2}\frac{\sigma_{t-1}^2(\theta_t)}{\sigma_{t-1}^2(\theta_t)+\lambda^{-1}} \\ &+\frac{1}{2}\frac{\sigma_{t-1}^2(\theta_t)(y_t-\mu_{t-1}(\theta_t))^2}{(\sigma_{t-1}^2(\theta_t)+\lambda^{-1})^2}. \end{aligned} \tag{4}$$

We note that μ_{t-1} and σ_{t-1}^2 denote the posterior mean and variance functions of GP, respectively, which are already computed when calculating the GP posterior.

4.3 Threshold Determination

To the best of our knowledge, all BO stopping criteria are based on the comparisons of certain quantities with a threshold to determine the appropriate stop time. Regarding the method based on the probability of improvement (Lorenz et al., 2015), although it has an interpretation for the type-I error of the statistical test, still the user must still determine the threshold. Nguyen et al. (2017) treated the threshold purely as a hyperparameter and left determining the threshold.

By restricting the problem to HPO of predictive models, Makarova et al. (2022) used an estimate of the SD of the generalization error of the model obtained by K-fold CV as the threshold (Nadeau and Bengio, 2000):

$$s_t = \sqrt{\mathrm{Var}[\hat{f}(\theta)]} = \sqrt{\frac{1}{K} + \frac{|D_{\mathrm{val}}|}{|D_{\mathrm{tr}}|}}\, s_{\mathrm{cv}}(\theta), \tag{5}$$

[2] See Section B in the appendix for derivation.

where D_{val} and D_{tr} are the validation and training datasets, respectively, and $|D|$ denotes the size of the set D. s_{cv}^2 is the empirical variance of the generalization error for the predictive model with hyperparameter θ estimated by K-fold CV. This enables determining the threshold to be determined adaptively at each iteration of BO. It is also advantageous that the threshold has the interpretation that once the maximum plausible improvement becomes less than the SD of the generalization error, further evaluations will not reliably improve the generalization error.

This work proposes another adaptive way of determining the threshold utilizing the particular form of our problem formulation. For $\Delta\tilde{R}_t$, we have the following upper bound under the assumptions of Theorem 2:

$$\begin{aligned} \Delta\tilde{R}_t \le& \frac{\beta_t^{1/2}\sigma_{t-1}(\theta_t^*)\sigma_{t-1}^2(\theta_t)}{\sigma_{t-1}^2(\theta_t)+\lambda^{-1}} + \frac{\sigma_{t-1}(\theta_t^*)\sigma_{t-1}(\theta_t)c}{\lambda^{1/2}(\sigma_{t-1}^2(\theta_t)+\lambda^{-1})} \\ &+\frac{\kappa_{t-1}}{2\lambda^{1/2}(\sigma_{t-1}^2(\theta_t)+\lambda^{-1})}\left(\lambda^2\sigma_{t-1}^6(\theta_t)+\lambda\sigma_{t-1}^4(\theta_t)\right. \\ &\left.+\lambda\beta_t\sigma_{t-1}^4(\theta_t)+2\lambda^{1/2}c\beta_t^{1/2}\sigma_{t-1}^3(\theta_t)+c^2\sigma_{t-1}^2(\theta_t)\right)^{1/2}, \\ =:& \Delta\hat{R}_t \end{aligned} \tag{6}$$

where $c = \sqrt{-2\log\delta}$. Section D in the appendix shows a detailed derivation. Then, we have

$$s_t = \frac{(\sigma_{t-1}(\theta_t^*) + \kappa_{t-1}/2)\sigma_{t-1}(\theta_t)c}{\sqrt{\lambda}(\sigma_{t-1}^2(\theta_t)+\lambda^{-1})} \tag{7}$$

by evaluating each term's convergence speed. Although each term of Eq. (6) converges to zero, Eq. (7) converges to zero slower than the other terms of the inequality. It is notable that this value s_t is only determined only by the fluctuation of observations and estimated by using the observed data points. We adopt Eq. (7) as the threshold and terminate the search when $\Delta\tilde{R}_t$ is less than this value. Since determining the threshold does not require additional information, we can use this in the general BO setting. Note that both $\Delta\tilde{R}_t$ and s_t are less than $\Delta\hat{R}_t$.

It is of a great practical benefit to be able to determine the threshold automatically and adaptively. However, we may sometimes want to find a point that is as close to the optimal point as possible, even if it increases the cost slightly. In such cases, it is desirable to be able to set the threshold such that it is as independent of the task as possible. For such a purpose, we also propose a heuristic to determine the threshold. It is unlikely that the parameter search by BO will end after the initial few iterations; therefore, we record the value of $\Delta\tilde{R}_t, t = 1, \ldots, T_{\mathrm{ini}}$ for the initial T_{ini} searches and terminate the search when we expect only an improvement of about several percent over the initial search. Concretely, we consider the median $\mathrm{Med}(\{\Delta\tilde{R}_t\}_{t=1}^{T_{\mathrm{ini}}})$ of the recorded upper bounds $\Delta\tilde{R}_t$ in the initial T_{ini} trials, and set the threshold as $s = \eta \times \mathrm{Med}(\{\Delta\tilde{R}_t\}_{t=1}^{T_{\mathrm{ini}}})$, where $\eta \in (0, 1)$ is a threshold determined by a user depending on their purpose. The T_{ini}

should be large enough to see a trend of the initial upper bound values, and we found $T_{ini} \simeq 20$ is a good rule-of-thumb. The coefficient of η only means "about $100 \times \eta\%$ compared with the gain in the initial search" and is only a guideline for how much patience is needed to continue the search.

4.4 Possible Limitation and Extension

As explained in §4.2, $D_{\mathrm{KL}}[p_t(\hat{f})||p_{t-1}(\hat{f})]$ is the KL divergence between GPs and there is no guarantee that the values are bounded, so it is generally difficult to calculate $D_{\mathrm{KL}}[p_t(\hat{f})||p_{t-1}(\hat{f})]$. Hence, it was necessary to guarantee that $D_{\mathrm{KL}}[p_t(\hat{f})||p_{t-1}(\hat{f})]$ is computable by assuming that the prior distribution of GP is always the same in the BO process. However, the hyperparameters of GPs are generally updated in BO. A practical approximation that relaxes the assumption that the prior distribution of GP remains unchanged in the BO process is calculating $D_{\mathrm{KL}}[p_t(\hat{f}|\nu_t)||p_{t-1}(\hat{f}|\nu_t)]$ instead of $D_{\mathrm{KL}}[p_t(\hat{f}|\nu_t)||p_{t-1}(\hat{f}|\nu_{t-1})]$, where the hyperparameter of GP at time t is denoted as ν_t. We adopt this approximation when implementing our proposed method. It is natural to assume that the change in the posterior of GP decreases along with the progress of BO in general, which leads to a decrease in the change of ν_t. Therefore, the error of the approximation is expected to decrease as the BO progresses. We demonstrate that the assumption tends to hold in Section G in the appendix.

Our proposed method is currently restricted to the case where a GP is used as the surrogate of the objective function while usable with any acquisition function. GP is widely adopted in many BO implementations, but it is not suitable for discrete variables. Future work will explore the use of other surrogate models (Bergstra et al., 2011; Jenatton et al., 2017; Garrido-Merchán and Hernández-Lobato, 2020; Tiao et al., 2021) suitable for discrete variables (or combinations of continuous and discrete variables).

The proposed criterion approximates θ_t^* using the minimum value for all observed values of the objective function. It could result in an inaccurate evaluation for the upper bound and stop the BO by mistakenly where the variance of the noise remains large.

5 EXPERIMENTAL RESULTS

We demonstrate the effectiveness of the proposed criterion through two experiments: test function minimization and HPO of machine learning methods. In these experiments, the BO is implemented by using GPyOpt (The GPyOpt authors, 2016).

5.1 Test Function Minimization

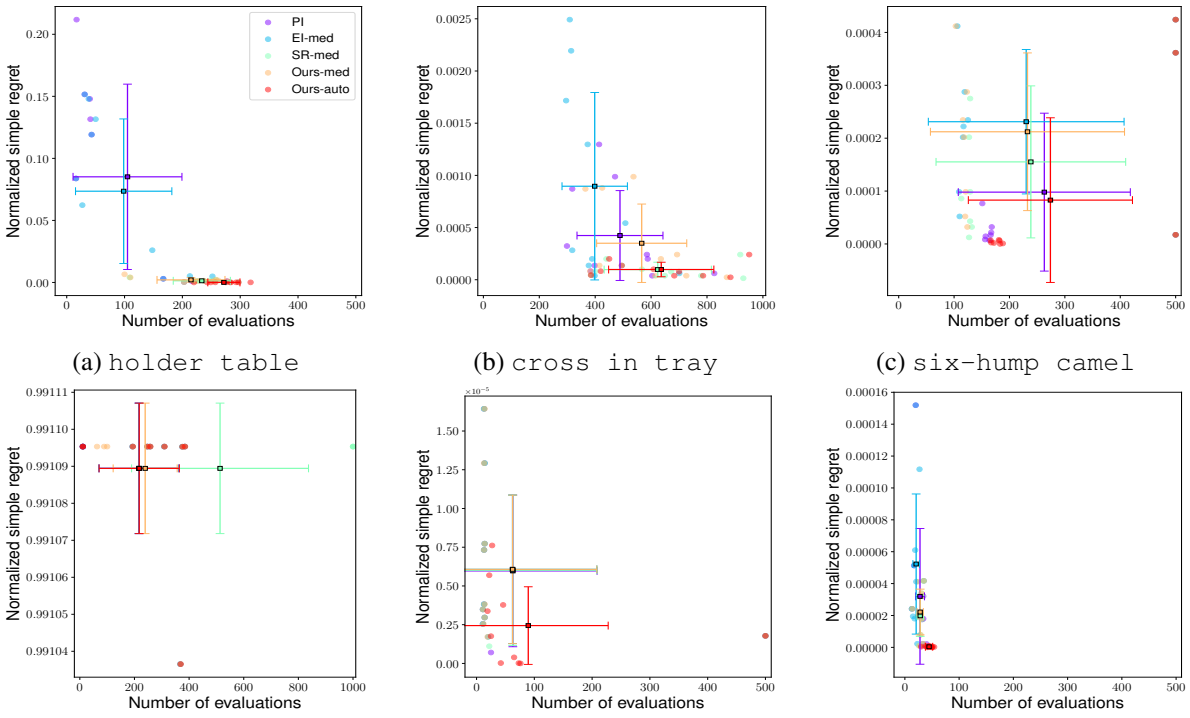
In this experiment, we evaluate stopping criteria using a set of two-dimensional test functions[3] as benchmark data when it is minimized by BO. The test functions used are `holder table`, `cross in tray`, `six-hump camel`, `easom`, `rosenbrock`, and `booth`. The experimental results are evaluated by plotting the mean and SD of the stop timing and the simple regret calculated by varying the random number seeds of the initial sampling 10 times, where the simple regret is normalized such that its range is $[0, 1]$. Thus, we evaluate whether the proposed criterion could stop BO sufficiently early while also evaluating whether the optimal parameters found by the stopped time are global minima.

We compare the stopping time determined by our proposed method with those determined by the stopping criteria based on probability of improvement (PI) (Lorenz et al., 2015), expected improvement (EI) (Nguyen et al., 2017), and simple regret (SR) (Makarova et al., 2022). The thresholds of EI and SR are set to $\eta = 0.01$ times the median of the values of initial $T_{\mathrm{ini}} = 20$ searches in each criterion, and the threshold of PI is set to 0.01. In this experiment, these criteria are denoted by EI-med, SR-med, and PI, respectively. For the proposed method, we consider two threshold determinations: i) the threshold is automatically determined by using Eq. (7) and ii) it is set to $\eta = 0.01$ times the median of the values of initial $T_{\mathrm{ini}} = 20$. These are denoted by Ours-auto and Ours-med, respectively.

The simple regret (SR)-based method and the proposed method requires setting β_t, a trade-off parameter between exploration and exploitation, appropriately to calculate the UCB and LCB. According to the previous work (Makarova et al., 2022), we can use the theoretical value $\beta_t = 2\log(|\Theta|t^2\pi^2/6\delta)$, where $1-\delta$ is the probability that the upper bound in Eq. (1) holds, and $\delta = 0.1$ is used. The mean function of GP is zero, and the covariance function is the Matérn kernel, and the hyperparameters are updated each time the search is performed by maximizing the marginal likelihood. We use GP-LCB as the acquisition function.

Figure 2 shows the normalized simple regret of BO and the stopped timing for each criterion in each test function minimization. From these results, PI and EI-med tend to stop the BO procedure earlier than the other criteria; they may stop before finding the optimal point, as shown in Fig. 2(a). SR-med, Ours-auto, and Ours-med tend to terminate the BO after discovering the optimal point. In the results of the `rosenbrock` and `booth` functions, the stopped timing of Ours-auto tends to be slower than the other stopping criteria. It is difficult to explore the optimal solution efficiently for GP-based BO because the shape of the `easom`

[3] `https://www.sfu.ca/~ssurjano/optimization.html`



(a) `holder table` (b) `cross in tray` (c) `six-hump camel`

(d) `easom` (e) `rosen brock` (f) `booth`

Figure 2: Stopped timing for each test function. "med" of the label name in the figure means that the threshold is determined based on the median of initial 20 searches, and "auto" means that the threshold is determined automatically by Eq. (7).

function is flat over most of the region and is a sharp function near the optimal solution as shown in Fig. 8(d). Therefore, the exploring by BO becomes less efficient, which leads to the proposed criterion stopping BO early. However, the SR-based criterion does not stop BO because the simple regret is insufficiently small. Section H.1 in the appendix shows the results in more detail.

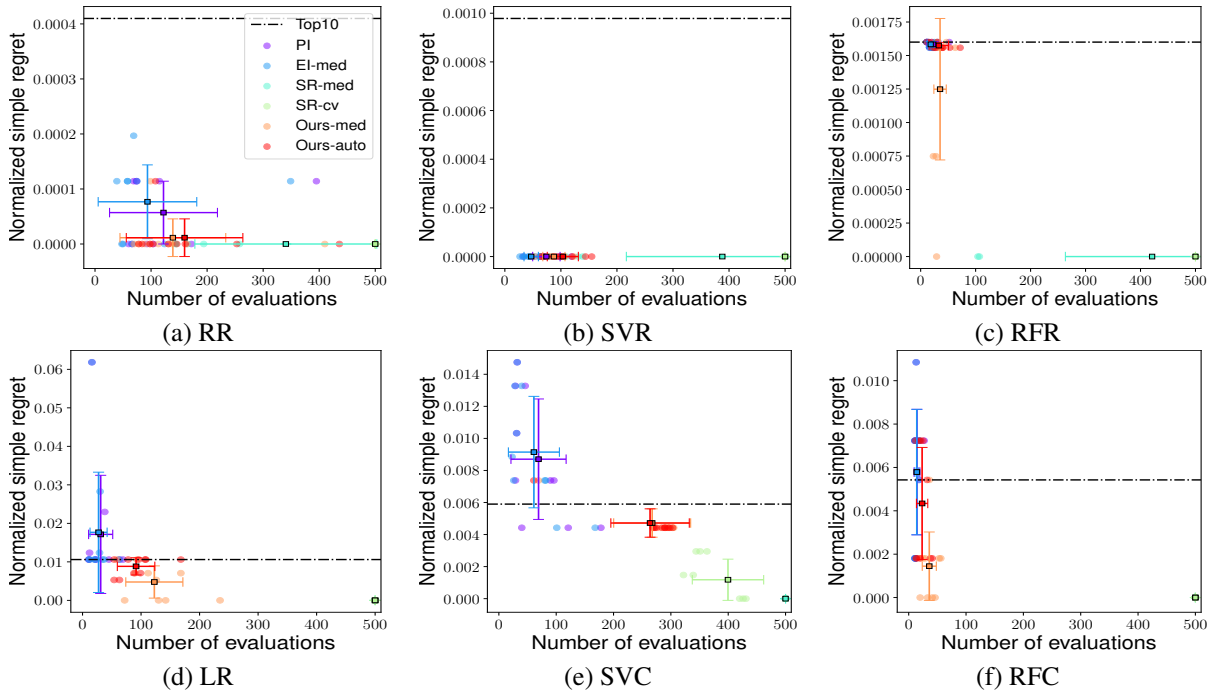
5.2 Hyperparameter Optimization

We consider the hyperparameter tuning problem of predictive models. The heuristic threshold is applied in the EI- and SR-based methods, which are denoted by EI-med and SR-med, respectively. We set s_t to $\eta = 0.01$ times the median of the values of initial $T_{\mathrm{ini}} = 10$ trials in each criterion. For the PI-based method, the threshold is set to 0.01. In addition, we consider the stopping criterion based on SR whose threshold is determined by the SD of the CV error, which is denoted by SR-cv. To see the quality of the obtained stopping timing, the 10th smallest simple regret value among the candidate hyperparameters is plotted as a reference.

We consider logistic regression (LR) with an additive RBF basis, support vector classification (SVC) (Cortes and Vapnik, 1995) with sigmoid kernel and random forest classification (RFC) (Breiman, 2001) for classification problems, and ridge regression (RR) (Hoerl and Kennard, 1970) with additive RBF basis, support vector regression (SVR) (Drucker et al., 1996) with RBF kernel and random forest regression (RFR) for regression problems. These methods are implemented by using scikit learn (Pedregosa et al., 2011). Table 1 shows the hyperparameters and their value ranges for each predictive model. We used the `gas`

Table 1: Description of predictors' hyperparameters.

	Hyperparameter	Variable type	Range
LR	# of basis functions	discrete	[2,30]
	scale of RBF	continuous	[0.01,100]
	Regularization coeff.	continuous	[0.01,100]
SVC	Kernel coefficient	continuous	[1e-3, 1e+3]
	coef0	continuous	[1e-3,1e+3]
	Regularization coeff.	continuous	[1e-3, 1e+3]
RFC	# of trees	discrete	[1, 20]
	Max depth	discrete	[1, 20]
	Min samples split	continuous	[0.01,0.5]
RR	# of basis functions	discrete	[2, 30]
	scale of RBF	continuous	[0.01, 100]
	Regularization coeff.	continuous	[1e-3,1e+3]
SVR	Kernel coefficient	continuous	[1e-3, 1e+3]
	Regularization coeff.	continuous	[1e-3, 1e+3]
	epsilon	continuous	[1e-3, 1e+3]
RFR	# of trees	discrete	[1,20]
	Max depth	discrete	[1,20]
	Min samples split	continuous	[0.01,0.5]



(a) RR (b) SVR (c) RFR

(d) LR (e) SVC (f) RFC

Figure 3: The number of evaluations and regret at the time of termination. Average and SD are calculated with 20 trials with different random initializations. "Top10" is the 10th smallest normalized simple regret. (a)–(c): `skin` data for classification, (d)–(f): `gas turbine` data for regression. The mean and SD for each stopping criterion are shown as rectangles and error bars, respectively, with the corresponding colors. "med" in the label name in the legend indicates that the threshold was determined based on median of initial 10 searches, while "auto" indicates that the threshold was determined automatically, and "cv" means that the threshold was determined by the SD of the CV error.

`turbine` dataset for the regression problem and `skin` dataset for the classification problem from the UCI repository (Dua and Graff, 2017) to train each predictive model. Section H.2 in the appendix shows experimental results on other datasets.

The hyperparameters contain discrete values, which are simply treated as continuous values. The hyperparameters of GP were updated after each search using the marginal likelihood maximization. GP-LCB was used as the acquisition function. To calculate the simple regret, we discretized the space of hyperparameters and optimized the hyperparameter on the space by using BO.

Figure 3 shows the normalized simple regret and the number of evaluated points at the stopped timings. The stopping criteria based on EI and PI terminate the BO earlier than the other criteria, and they sometimes terminate BO before finding a hyperparameter with the 10th smallest simple regret, as shown in Fig. 3(d) and (e). Meanwhile, SR-cv terminates the BO when the optimal hyperparameter is found in SVC, as shown in Fig. 3(e); however, the criterion cannot terminate the BO with other models, as shown in Fig. 3(a)–(d) and (f). SR-med also terminates the BO when the optimal hyperparameter is found for RR, SVR, and RFR—as shown in Fig. 3(a)–(c)—but it cannot terminate the BO with the other models. Ours-auto and Ours-med tend to terminate the BO after finding a hyperparameter with the 10th smallest simple regret, and their stopping timings tend to be earlier than the SR-based criterion.

6 CONCLUSION

This work considered the important question of when to stop a parameter search by BO, which is very important to ensure the efficiency of parameter search in black-box functions and of great significance for Green-AI research (Strubell et al., 2020; Schwartz et al., 2019). Experimentally, it was confirmed that the search could stop at a sufficiently early stage with reasonable performance for the resultant model.

One downside of the application of stopping criteria including our proposed one, is the possibility of missing better solutions or important findings. There is always a trade-off between cost and solution quality, and care must be taken when using the stopping criteria of BO.

Acknowledgements

The authors express special thanks to the anonymous reviewers, area chair, and senior area chair whose comments valuably improved this paper. Part of this work is supported by JSPS KAKENHI Grant Number 22K17951, JST CREST JPMJCR1761, JPMJCR2015, JST Mirai program JPMJMI21G2 and NEDO JPNP18002.

References

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf.

Peter I. Frazier and Jialei Wang. Bayesian optimization for materials design. *Springer Series in Materials Science*, page 45–75, Dec 2015. ISSN 2196-2812. doi: 10.1007/978-3-319-23871-5_3. URL http://dx.doi.org/10.1007/978-3-319-23871-5_3.

Lutz Prechelt. *Early Stopping – But When?*, pages 53–67. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_5. URL https://doi.org/10.1007/978-3-642-35289-8_5.

Anastasia Makarova, Huibin Shen, Valerio Perrone, Aaron Klein, Jean Baptiste Faddoul, Andreas Krause, Matthias Seeger, and Cedric Archambeau. Automatic termination for hyperparameter optimization. In *First Conference on Automated Machine Learning (Main Track)*, 2022. URL https://openreview.net/forum?id=BNeNQWaBIgq.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006. ISBN 026218253X.

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

Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. The application of Bayesian methods for seeking the extremum. *Towards Global Optimization*, 2(117-129):2, 1978.

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998. doi: 10.1023/A:1008306431147. URL https://scholar.google.com/scholar?cluster=13328031565180796595.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 1015–1022, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Philipp Hennig and Christian J. Schuler. Entropy search for information-efficient global optimization. *J. Mach. Learn. Res.*, 13:1809–1837, 2012. URL http://dl.acm.org/citation.cfm?id=2343701.

José Miguel Hernández-Lobato, Matthew W. Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 918–926, 2014.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. ISSN 00063444. URL http://www.jstor.org/stable/2332286.

Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing*, 21(4):599–613, 2009. URL http://joc.journal.informs.org/cgi/reprint/21/4/599.

Alexander Siemenn, Zekun Ren, Qianxiao Li, and Tonio Buonassisi. Fast Bayesian optimization of needle-in-a-haystack problems using zooming memory-based initialization, 2022. URL https://doi.org/10.21203/rs.3.rs-2003069/v1.

Romy Lorenz, Ricardo Pio Monti, Inês Violante, Aldo Faisal, Christoforos Anagnostopoulos, Robert Leech, and Giovanni Montana. Stopping criteria for boosting automatic experimental design using real-time fmri with Bayesian optimization. 11 2015.

Vu Nguyen, Sunil Gupta, Santu Rana, Cheng Li, and Svetha Venkatesh. Regret for expected improvement over the best-observed value and stopping condition. In Min-Ling Zhang and Yung-Kyun Noh, editors, *Proceedings of the Ninth Asian Conference on Machine Learning*, volume 77 of *Proceedings of Machine Learning Research*, pages 279–294, Yonsei University, Seoul, Republic of Korea, 15–17 Nov 2017. PMLR. URL https://proceedings.mlr.press/v77/nguyen17a.html.

Kevin Swersky, Jasper Snoek, and Ryan P Adams. Freeze-Thaw Bayesian Optimization, 2014. URL http://arxiv.org/abs/1406.3896.

Zhongxiang Dai, Haibin Yu, Bryan Kian Hsiang Low, and Patrick Jaillet. Bayesian optimization meets Bayesian

optimal stopping. In *36th International Conference on Machine Learning, ICML 2019*, volume 2019-June, pages 2701–2725, 2019. ISBN 9781510886988.

Thomas S. Ferguson. Optimal stopping and applications, 2006. URL https://www.math.ucla.edu/~tom/Stopping/Contents.htm.

Kirthevasan Kandasamy, Gautam Dasarathy, Junier B Oliva, Jeff Schneider, and Barnabas Poczos. Gaussian process bandit optimisation with multi-fidelity evaluations. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Kirthevasan Kandasamy, Gautam Dasarathy, Jeff Schneider, and Barnabás Póczos. Multi-fidelity Bayesian optimisation with continuous approximations. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1799–1808. JMLR.org, 2017.

Shibo Li, Robert Kirby, and Shandian Zhe. Batch multi-fidelity Bayesian optimization with deep auto-regressive networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=wF-llA3k32.

Lisha Li, Kevin G. Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: Bandit-based configuration evaluation for hyperparameter optimization. In *ICLR*, 2017. URL https://openreview.net/forum?id=ry18Ww5ee.

Stefan Falkner, Aaron Klein, and Frank Hutter. BOHB: Robust and Efficient Hyperparameter Optimization at Scale. In *35th International Conference on Machine Learning, ICML 2018*, volume 4, pages 2323–2341, 2018. ISBN 9781510867963.

Tobias Domhan, Jost Tobias Springenberg, and Frank Hutter. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 3460–3468. AAAI Press, 2015. ISBN 9781577357384.

Aaron Klein, Stefan Falkner, Simon Bartels, Philipp Hennig, and Frank Hutter. Fast Bayesian Optimization of Machine Learning Hyperparameters on Large Datasets. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 528–536. PMLR, 20–22 Apr 2017.

S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 1951.

Hideaki Ishibashi and Hideitsu Hino. Stopping criterion for active learning based on deterministic generalization bounds. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 386–397. PMLR, 26–28 Aug 2020.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000. URL https://proceedings.neurips.cc/paper/1999/file/7d12b66d3df6af8d429c1a357d8b9e1a-Paper.pdf.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

Rodolphe Jenatton, Cedric Archambeau, Javier González, and Matthias Seeger. Bayesian optimization with tree-structured dependencies. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1655–1664. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/jenatton17a.html.

Eduardo C. Garrido-Merchán and Daniel Hernández-Lobato. Dealing with categorical and integer-valued variables in Bayesian optimization with Gaussian processes. *Neurocomputing*, 380:20–35, 2020. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2019.11.004. URL https://www.sciencedirect.com/science/article/pii/S0925231219315619.

Louis C Tiao, Aaron Klein, Matthias W Seeger, Edwin V. Bonilla, Cedric Archambeau, and Fabio Ramos. Bore: Bayesian optimization by density-ratio estimation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10289–10300. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/tiao21a.html.

The GPyOpt authors. GPyOpt: A Bayesian optimization framework in python. http://github.com/SheffieldML/GPyOpt, 2016.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. doi: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

Leo Breiman. Random forests. *Machine Learning*, 45 (1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A%3A1010933404324.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. In M.C. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. URL https://proceedings.neurips.cc/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13693–13696, Apr. 2020. doi: 10.1609/aaai.v34i09.7123. URL https://ojs.aaai.org/index.php/AAAI/article/view/7123.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI, 2019. URL http://arxiv.org/abs/1907.10597.

D. Russo and B. V. Roy. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016. URL http://jmlr.org/papers/v17/14-087.html.

D. B. Owen. A table of normal integrals. *Communications in Statistics - Simulation and Computation*, 9(4):389–419, 1980. doi: 10.1080/03610918008812164. URL https://doi.org/10.1080/03610918008812164.

Huong Ha, Santu Rana, Sunil Gupta, Thanh Nguyen, Hung Tran-The, and Svetha Venkatesh. Bayesian optimization with unknown search space. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/ccf0304d099baecfbe7ff6844e1f6d91-Paper.pdf.

Newsha Ghoreishi, Anders Clausen, and Bo Nørregaard Jørgensen. Termination criteria in evolutionary algorithms: A survey. In *Proceedings of 9th International Joint Conference on Computational Intelligence*, volume 1, pages 373–384. SCITEPRESS Digital Library, 2017. doi: 10.5220/0006577903730384. 9th International Joint Conference on Computational Intelligence ; Conference date: 01-11-2017 Through 03-11-2017.

Brijnesh J. Jain, Hartmut Pohlheim, and Joachim Wegener. On termination criteria of evolutionary algorithms. In *Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation*, GECCO'01, page 768, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607749.

Carlos A. Coello Coello, Gary B. Lamont, and David A. Van Veldhuizen. *Evolutionary Algorithms for Solving Multi-Objective Problems (Genetic and Evolutionary Computation)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387332545.

E. Zitzler, L. Thiele, M. Laumanns, C.M. Fonseca, and V.G. da Fonseca. Performance assessment of multiobjective optimizers: an analysis and review. *IEEE Transactions on Evolutionary Computation*, 7(2):117–132, 2003. doi: 10.1109/TEVC.2003.810758.

Heike Trautmann, Uwe Ligges, Jörn Mehnen, and Mike Preuss. A convergence criterion for multiobjective evolutionary algorithms based on systematic statistical testing. In Günter Rudolph, Thomas Jansen, Nicola Beume, Simon Lucas, and Carlo Poloni, editors, *Parallel Problem Solving from Nature – PPSN X*, pages 825–836, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-87700-4.

José L. Guerrero, Jesus Garcia, Luis Marti, José Manuel Molina, and Antonio Berlanga. A stopping criterion based on kalman estimation techniques with several progress indicators. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation*, GECCO '09, page 587–594, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605583259. doi: 10.1145/1569901.1569983. URL https://doi.org/10.1145/1569901.1569983.

Anastasia Makarova, Huibin Shen, Valerio Perrone, Aaron Klein, Jean Baptiste Faddoul, Andreas Krause, Matthias Seeger, and Cedric Archambeau. Overfitting in Bayesian optimization: an empirical study and early-stopping solution. ICLR Workshop on Neural Architecture Search, May 2021. URL https://iclr.cc/virtual/2021/workshop/2145#wse-detail-3968.

A Proof of the Theorem 1

To prove the Theorem 1, we use the following lemmas.

Lemma 1. *Let P and Q be any probability distributions such that P is absolutely continuous with respect to Q. Then, for any random variable X and any measurable function $\mathcal{L}(X) \in [a, b]$, we have*

$$|\mathbb{E}_P[\mathcal{L}(X)] - \mathbb{E}_Q[\mathcal{L}(X)]| \leq (b-a)\sqrt{\frac{1}{2}D_{\mathrm{KL}}[P||Q]}.$$

Proof. Let Ω be a countable set. Let $h(\omega) = \mathcal{L}(X(\omega)) - (b+a)/2$ so that $h : \Omega \to [-(b-a)/2, (b-a)/2]$. Choose a base measure μ so that P and Q are absolutely continuous with respect to μ. Then, the following Pinsker's inequality holds.

$$\sqrt{\frac{1}{2}D_{\mathrm{KL}}[P||Q]} \geq \int \frac{1}{2}\left|\frac{dP}{d\mu} - \frac{dQ}{d\mu}\right| d\mu.$$

By using this inequality, we have

$$\begin{aligned}
(b-a)\sqrt{\frac{1}{2}D_{\mathrm{KL}}[P||Q]} \geq & \frac{b-a}{2}\int \left|\frac{dP}{d\mu} - \frac{dQ}{d\mu}\right| d\mu \\
\geq & \frac{b-a}{2}\int \left|\frac{2}{b-a}\left(\frac{dP}{d\mu} - \frac{dQ}{d\mu}\right) h\right| d\mu \\
\geq & \left|\int h dP - \int h dQ\right| \\
= & \left|\mathbb{E}_P[\mathcal{L}(X)] - \mathbb{E}_Q[\mathcal{L}(X)]\right|.
\end{aligned}$$

This proof is the same as the proof of fact 9 in Russo and Roy (2016) except for the assumption that $\sup_X \mathcal{L} - \inf_X \mathcal{L} = (b-a)$ and the transformation from line 2 to line 3, which is used $|\int h d\mu| \leq \int |h| d\mu$. □

Lemma 2. *(Srinivas et al., 2010) For any function f generated by a Gaussian process with mean function μ and covariance function σ^2,*

$$|f(\theta) - \mu(\theta)| \leq \beta_t^{1/2}\sigma(\theta)$$

holds with probability $\geq 1 - \delta$.

Lemma 3. *Let $p(x, y)$ be a two-dimensional Gaussian distribution, that is,*

$$p(x,y) = \mathcal{N}\left(\begin{bmatrix} x \\ y \end{bmatrix} \mid \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_{xx}^2 & \sigma_{yx}^2 \\ \sigma_{xy}^2 & \sigma_{yy}^2 \end{bmatrix}\right).$$

We denote the probability distribution function and cumulative distribution function of standard normal distribution by $\phi(x)$ and $\Phi(x)$, respectively. Then, for any measurable function $\mathcal{L}(f) \in [a, b]$, we have

$$\mathbb{E}_{p(x,y)}[\max\{x - y, 0\}] = v\phi(g) + vg\Phi(g),$$

where $v = \sqrt{\sigma_{xx}^2 - 2\sigma_{xy}^2 + \sigma_{yy}^2}$, and $g = \frac{\mu_x - \mu_y}{v}$.

Proof. From

$$p(x,y) = \mathcal{N}\left(\begin{bmatrix} x \\ y \end{bmatrix} \mid \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_{xx}^2 & \sigma_{yx}^2 \\ \sigma_{xy}^2 & \sigma_{yy}^2 \end{bmatrix}\right),$$

letting $\mu_{y|x} = \mu_y + \frac{\sigma_{yx}^2}{\sigma_{xx}^2}(x-\mu_x)$ and $\sigma_{y|x}^2 = \sigma_{yy}^2 - \frac{\sigma_{yx}^4}{\sigma_{xx}^2}$, we can write $p(y|x) = \mathcal{N}(y \mid \mu_{y|x}, \sigma_{y|x}^2)$. Let $\tilde{y} = (y-\mu_{y|x})/\sigma_{y|x}$ and $u = (x-\mu_{y|x})/\sigma_{y|x}$, then the following equation holds.

$$\begin{aligned}
\mathbb{E}_{p(x,y)}[\max\{x-y,0\}] =& \mathbb{E}_{p(x)}[\mathbb{E}_{p(y|x)}[\max\{x-y,0\}]] \\
=& \mathbb{E}_{p(x)}\left[\int_{-\infty}^{u}(x-\sigma_{y|x}\tilde{y}-\mu_{y|x})\phi(\tilde{y})d\tilde{y}\right] \\
=& \mathbb{E}_{p(x)}\left[(x-\mu_{y|x})\int_{-\infty}^{u}\phi(\tilde{y})d\tilde{y} - \sigma_{y|x}\int_{-\infty}^{u}\tilde{y}\phi(\tilde{y})d\tilde{y}\right] \\
=& \mathbb{E}_{p(x)}\left[(x-\mu_{y|x})\Phi(u) - \sigma_{y|x}\left[-\phi(\tilde{y})\right]_{-\infty}^{u}\right] \\
=& \mathbb{E}_{p(x)}\left[\sigma_{y|x}u\Phi(u) + \sigma_{y|x}\phi(u)\right].
\end{aligned}$$

Let $\tilde{x} = \frac{x-\mu_x}{\sigma_{xx}}$. By applying $x = \sigma_{xx}\tilde{x} + \mu_x$, we have

$$\begin{aligned}
u =& \frac{x-\mu_{y|x}}{\sigma_{y|x}} \\
=& \frac{x-\mu_y}{\sigma_{y|x}} - \frac{\sigma_{yx}^2}{\sigma_{y|x}\sigma_{xx}^2}(x-\mu_x) \\
=& \frac{\sigma_{xx}^2-\sigma_{yx}^2}{\sigma_{y|x}\sigma_{xx}^2}x - \frac{\mu_y}{\sigma_{y|x}} + \frac{\sigma_{yx}^2}{\sigma_{y|x}\sigma_{xx}^2}\mu_x \\
=& \frac{\sigma_{xx}^2-\sigma_{yx}^2}{\sigma_{y|x}\sigma_{xx}}\tilde{x} + \frac{\sigma_{xx}^2-\sigma_{yx}^2}{\sigma_{y|x}\sigma_{xx}^2}\mu_x - \frac{\mu_y}{\sigma_{y|x}} + \frac{\sigma_{yx}^2}{\sigma_{y|x}\sigma_{xx}^2}\mu_x \\
=& \frac{\sigma_{xx}^2-\sigma_{yx}^2}{\sigma_{y|x}\sigma_{xx}}\tilde{x} + \frac{\mu_x-\mu_y}{\sigma_{y|x}} =: b\tilde{x} + a.
\end{aligned}$$

Therefore, the following equation holds.

$$\begin{aligned}
\mathbb{E}_{p(x,y)}[\max\{x-y,0\}] =& \mathbb{E}_{p(\tilde{x})}\left[\sigma_{y|x}(b\tilde{x}+a)\Phi(b\tilde{x}+a) + \sigma_{y|x}\phi(b\tilde{x}+a)\right] \\
=& \sigma_{y|x}b\int_{-\infty}^{\infty}\tilde{x}\Phi(b\tilde{x}+a)\phi(\tilde{x})d\tilde{x} + \sigma_{y|x}a\int_{-\infty}^{\infty}\Phi(b\tilde{x}+a)\phi(\tilde{x})d\tilde{x} \\
&+ \sigma_{y|x}\int_{-\infty}^{\infty}\phi(b\tilde{x}+a)\phi(\tilde{x})d\tilde{x}. \qquad (8)
\end{aligned}$$

The following equalities hold for Gaussian distributions (Owen, 1980).

$$\begin{aligned}
\int_{-\infty}^{\infty}\tilde{x}\Phi(b\tilde{x}+a)\phi(\tilde{x})d\tilde{x} =& \frac{b}{\sqrt{b^2+1}}\phi(\frac{a}{\sqrt{b^2+1}}), \\
\int_{-\infty}^{\infty}\Phi(b\tilde{x}+a)\phi(\tilde{x})d\tilde{x} =& \Phi(\frac{a}{\sqrt{b^2+1}}), \\
\int_{-\infty}^{\infty}\phi(b\tilde{x}+a)\phi(\tilde{x})d\tilde{x} =& \frac{1}{\sqrt{b^2+1}}\phi(\frac{a}{\sqrt{b^2+1}}).
\end{aligned}$$

Using the above equalities, we have

$$\begin{aligned}
\mathbb{E}_{p(x,y)}[\max\{x-y,0\}] =& \frac{\sigma_{y|x}b^2}{\sqrt{b^2+1}}\phi\left(\frac{a}{\sqrt{b^2+1}}\right) + \sigma_{y|x}a\Phi\left(\frac{a}{\sqrt{b^2+1}}\right) + \frac{\sigma_{y|x}}{\sqrt{b^2+1}}\phi\left(\frac{a}{\sqrt{b^2+1}}\right) \\
=& \sigma_{y|x}\sqrt{b^2+1}\phi\left(\frac{a}{\sqrt{b^2+1}}\right) + \sigma_{y|x}a\Phi\left(\frac{a}{\sqrt{b^2+1}}\right).
\end{aligned}$$

Then, the following equation holds.

$$\begin{aligned}
\sigma_{y|x}\sqrt{b^2+1} =& \sigma_{y|x}\sqrt{\frac{(\sigma_{xx}^2-\sigma_{yx}^2)^2+\sigma_{y|x}^2\sigma_{xx}^2}{\sigma_{y|x}^2\sigma_{xx}^2}} \\
=& \sqrt{\frac{(\sigma_{xx}^2-\sigma_{yx}^2)^2+\sigma_{y|x}^2\sigma_{xx}^2}{\sigma_{xx}^2}}, \\
=& \sqrt{\frac{\sigma_{xx}^4-2\sigma_{xx}^2\sigma_{yx}^2+\sigma_{yx}^4+\sigma_{yy}^2\sigma_{xx}^2-\sigma_{yx}^4}{\sigma_{xx}^2}} \\
=& \sqrt{\sigma_{xx}^2-2\sigma_{yx}^2+\sigma_{yy}^2} \\
\frac{a}{\sqrt{b^2+1}} =& \frac{(\mu_x-\mu_y)}{\sigma_{y|x}}\sqrt{\frac{\sigma_{y|x}^2\sigma_{xx}^2}{(\sigma_{xx}^2-\sigma_{yx}^2)^2+\sigma_{yy}^2\sigma_{xx}^2-\sigma_{yx}^4}} \\
=& \frac{(\mu_x-\mu_y)\sigma_{yx}}{\sqrt{\sigma_{xx}^4-2\sigma_{xx}^2\sigma_{yx}^2+\sigma_{yx}^4+\sigma_{yy}^2\sigma_{xx}^2-\sigma_{yx}^4}} \\
=& \frac{\mu_x-\mu_y}{\sqrt{\sigma_{xx}^2-2\sigma_{yx}^2+\sigma_{yy}^2}}.
\end{aligned}$$

Substituting Eq. (8) into the above equation, we have

$$\mathbb{E}_{p(x,y)}[\max\{x-y,0\}] = \sqrt{\sigma_{xx}^2-2\sigma_{yx}^2+\sigma_{yy}^2}\,\phi\left(\frac{\mu_x-\mu_y}{\sqrt{\sigma_{xx}^2-2\sigma_{yx}^2+\sigma_{yy}^2}}\right) + (\mu_x-\mu_y)\Phi\left(\frac{\mu_x-\mu_y}{\sqrt{\sigma_{xx}^2-2\sigma_{yx}^2+\sigma_{yy}^2}}\right),$$

which proves Lemma 3. □

The difference in the simple regret ΔR_t can be expanded as follows:

$$\begin{aligned}
&|\mathbb{E}_{p_{t-1}(\hat{f})}[\min_{\theta\in\Theta}\{\hat{f}(\theta)\}] - f^* - \mathbb{E}_{p_t(\hat{f})}[\min_{\theta\in\Theta}\{\hat{f}(\theta)\}] + f^*| \\
=&|\mathbb{E}_{p_{t-1}(\hat{f})}[\min_{\theta\in\Theta}\{\min\{\hat{f}(\theta),\hat{f}(\theta_{t-1}^*)\}\}] - \mathbb{E}_{p_t(\hat{f})}[\min_{\theta\in\Theta}\{\min\{\hat{f}(\theta),\hat{f}(\theta_t^*)\}\}]| \\
\leq&|\Delta\mu_t^*| + |\mathbb{E}_{p_{t-1}(\hat{f})}[\min_{\theta\in\Theta}\{\min\{\hat{f}(\theta)-\hat{f}(\theta_{t-1}^*),0\}\}] - \mathbb{E}_{p_t(\hat{f})}[\min_{\theta\in\Theta}\{\min\{\hat{f}(\theta)-\hat{f}(\theta_t^*),0\}\}]| \\
=&|\Delta\mu_t^*| + |\mathbb{E}_{p_t(\hat{f})}[\max_{\theta\in\Theta}\{\max\{\hat{f}(\theta_t^*)-\hat{f}(\theta),0\}\}] - \mathbb{E}_{p_{t-1}(\hat{f})}[\max_{\theta\in\Theta}\{\max\{\hat{f}(\theta_{t-1}^*)-\hat{f}(\theta),0\}\}]| \\
\leq&|\Delta\mu_t^*| + \mathbb{E}_{p_t(\hat{f})}[\max\{\hat{f}(\theta_t^*)-\hat{f}(\theta_{t-1}^*),0\}] + |\mathbb{E}_{p_t(\hat{f})}[\max_{\theta\in\Theta}\{\max\{\hat{f}(\theta_{t-1}^*)-\hat{f}(\theta),0\}\}] \\
&- \mathbb{E}_{p_{t-1}(\hat{f})}[\max_{\theta\in\Theta}\{\max\{\hat{f}(\theta_{t-1}^*)-\hat{f}(\theta),0\}\}]|.
\end{aligned}$$

Let $\mathcal{L}(\hat{f}) \coloneqq \max_{\theta\in\Theta}\{\max\{\hat{f}(\theta_{t-1}^*)-\hat{f}(\theta),0\}\} = \hat{f}(\theta_{t-1}^*) - \min_{\theta\in\Theta}\hat{f}(\theta)$. By using Lemma 2, we can bound $\mathcal{L}(\hat{f})$ as follows with probability $1-\delta$:

$$\mathcal{L}(\hat{f}) \leq \min_{\theta\in\Theta_{t-1}} \mathrm{UCB}_\delta(\theta) - \min_{\theta\in\Theta}\mathrm{LCB}_\delta(\theta) =: \kappa_{t-1}.$$

From the above inequality and Lemmas 1, and 3, we have

$$\Delta R_t \leq |\Delta\mu_t^*| + v\phi\left(g\right) + vg\Phi\left(g\right) + \kappa_{t-1}\sqrt{\frac{1}{2}D_{\mathrm{KL}}[p_t(\hat{f})||p_{t-1}(\hat{f})]}$$

hence the Theorem holds.

B KL divergence between GP posteriors

We derive formula (4) in the main text used for computing the KL divergence between GP posteriors. For any GP posterior, the following property holds.

Lemma 4 (Ishibashi and Hino (2020))**.** *Let $p(\hat{f}|\mathbf{y})$ and $p(\hat{f}|\mathbf{y}')$ be the posteriors with respect to $\hat{f}$ given $S = \{\Theta, \mathbf{y}\}$ and $S' = \{\Theta', \mathbf{y}'\}$, respectively. Assume that the prior of $p(\hat{f}|\mathbf{y})$ and that of $p(\hat{f}|\mathbf{y}')$ are the same. Then, the following equality holds:*

$$D_{\mathrm{KL}}[p(\hat{f}|\mathbf{y})||p(\hat{f}|\mathbf{y}')] = D_{\mathrm{KL}}[p(\hat{\mathbf{f}}_+|\mathbf{y}))||p(\hat{\mathbf{f}}_+|\mathbf{y}'))],$$

where $\Theta_+ \coloneqq \Theta \cup \Theta'$ and $\hat{\mathbf{f}}_+ \coloneqq f(\Theta_+)$.

Let $p(\hat{f}|\mathbf{y}_t)$ and $p(\hat{f}|\mathbf{y}_{t-1})$ be GP posteriors given $S_{t-1} = \{\Theta_{t-1}, \mathbf{y}_{t-1}\}$ and $S_t = S_{t-1} \cup \{(\theta_t, y_t)\}$, respectively. From the above Lemma 4, $D_{\mathrm{KL}}[p(\hat{f}|\mathbf{y}_t)||p(\hat{f}|\mathbf{y}_{t-1})]$ is derived as

$$\begin{aligned}
D_{\mathrm{KL}}[p(\hat{f}|\mathbf{y}_t)||p(\hat{f}|\mathbf{y}_{t-1})] =& D_{\mathrm{KL}}[p(\hat{\mathbf{f}}_t|\mathbf{y}_t)||p(\hat{\mathbf{f}}_t|\mathbf{y}_{t-1})] \\
=& \mathop{\mathbb{E}}_{p(\hat{\mathbf{f}}_t|\mathbf{y}_t)}\left[\log \frac{p(y_t|\hat{\mathbf{f}}_t)p(\hat{\mathbf{f}}_t|\mathbf{y}_{t-1})}{p(\hat{\mathbf{f}}_t|\mathbf{y}_{t-1})p(y_t|\mathbf{y}_{t-1})}\right] \\
=& \int p(\hat{f}_t|\mathbf{y}_t)\log p(y_t|\hat{f}_t)d\hat{f}_t - \log \int p(\hat{f}_t|\mathbf{y}_{t-1})p(y_t|\hat{f}_t)d\hat{f}_t, \qquad (9)
\end{aligned}$$

where $\hat{\mathbf{f}}_t \coloneqq \hat{f}(\Theta_t)$, $\hat{f}_t \coloneqq \hat{f}(\theta_t)$ and $p(y_t|\mathbf{y}_{t-1}) = \int p(y_t|\hat{f}_t)p(\hat{f}_t|\mathbf{y}_{t-1})d\hat{f}_t$. The first term of Eq. (9) can be rewritten as

$$\begin{aligned}
\int p(\hat{f}_t|\mathbf{y}_t)\log p(y_t|\hat{f}_t)d\hat{f}_t =& -\mathop{\mathbb{E}}_{p(\hat{f}_t|\mathbf{y}_t)}\left[\frac{\lambda}{2}(y_t - \hat{f}_t)^2\right] - \frac{1}{2}\log 2\pi\lambda^{-1} \\
=& -\frac{\lambda}{2}\left(y_t^2 - 2y_t\mathbb{E}[\hat{f}_t] + \mathbb{E}[\hat{f}_t^2]\right) - \frac{1}{2}\log 2\pi\lambda^{-1} \\
=& -\frac{\lambda}{2}(y_t - \mu_t(\theta_t))^2 - \frac{\lambda}{2}\sigma_t^2(\theta_t, \theta_t) - \frac{1}{2}\log 2\pi\lambda^{-1}.
\end{aligned}$$

The second term of Eq. (9) becomes the logarithm of a normal distribution since $p(y_t|\hat{f}_t)$ and $p(\hat{f}_t|\mathbf{y}_t)$ are normal distributions. Therefore, the following equation holds:

$$\begin{aligned}
\log \int p(y_t|\hat{f}_t)p(\hat{f}_t|\mathbf{y}_{t-1})d\hat{f}_t =& \log \mathcal{N}(y_t|\mu_{t-1}(\theta_t), \sigma_{t-1}^2(\theta_t, \theta_t) + \lambda^{-1}) \\
=& -\frac{(y_t - \mu_{t-1}(\theta_t))^2}{2(\sigma_{t-1}^2(\theta_t, \theta_t) + \lambda^{-1})} - \frac{1}{2}\log 2\pi(\sigma_{t-1}^2(\theta_t, \theta_t) + \lambda^{-1}).
\end{aligned}$$

Regarding $p(\hat{f}_t|\mathbf{y}_t)$ as the posterior whose prior is $p(\hat{f}_t|\mathbf{y}_{t-1})$ observing (θ_t, y_t), the mean and covariance functions corresponding to θ_t are derived as

$$\begin{aligned}
\mu_t(\theta_t) =& \mu_{t-1}(\theta_t) + \frac{\sigma_{t-1}^2(\theta_t, \theta_t)}{\sigma_{t-1}^2(\theta_t, \theta_t) + \lambda^{-1}}(y_t - \mu_{t-1}(\theta_t)) \\
=& \frac{\lambda^{-1}\mu_{t-1}(\theta_t) + \sigma_{t-1}^2(\theta_t, \theta_t)y_t}{\sigma_{t-1}^2(\theta_t, \theta_t) + \lambda^{-1}}, \\
\sigma_t^2(\theta_t, \theta_t) =& \sigma_{t-1}^2(\theta_t, \theta_t) - \frac{\sigma_{t-1}^4(\theta_t, \theta_t)}{\sigma_{t-1}^2(\theta_t, \theta_t) + \lambda^{-1}} \\
=& \frac{\lambda^{-1}\sigma_{t-1}^2(\theta_t, \theta_t)}{\sigma_{t-1}^2(\theta_t, \theta_t) + \lambda^{-1}}.
\end{aligned}$$

From the result, the second term of Eq. (9) is described as

$$\int p(\hat{f}_t|\mathbf{y}_t)\log p(y_t|\hat{f}_t)d\hat{f}_t = -\frac{1}{2}\frac{\lambda^{-1}(y_t - \mu_{t-1}(\theta_t))^2}{(\sigma_{t-1}^2(\theta_t, \theta_t) + \lambda^{-1})^2} - \frac{1}{2}\frac{\sigma_{t-1}^2(\theta_t, \theta_t)}{\sigma_{t-1}^2(\theta_t, \theta_t) + \lambda^{-1}} - \frac{1}{2}\log 2\pi\lambda^{-1}.$$

Therefore, the following equation holds as claimed:

$$\begin{aligned}
D_{\mathrm{KL}}[p(\hat{f}|\mathbf{y}_t)||p(\hat{f}|\mathbf{y}_{t-1})] =& \int p(\hat{f}_t|\mathbf{y}_t)\log p(y_t|\hat{f}_t,\theta_t)d\hat{f}_t - \log\int p(\hat{f}_t|\mathbf{y}_{t-1})p(y_t|\hat{f}_t,\theta_t)d\hat{f}_t \\
=& -\frac{1}{2}\frac{\lambda^{-1}(y_t-\mu_{t-1}(\theta_t))^2}{(\sigma^2_{t-1}(\theta_t,\theta_t)+\lambda^{-1})^2} - \frac{1}{2}\frac{\sigma^2_{t-1}(\theta_t,\theta_t)}{\sigma^2_{t-1}(\theta_t,\theta_t)+\lambda^{-1}} - \frac{1}{2}\log 2\pi\lambda^{-1} \\
&+ \frac{1}{2}\frac{(y_t-\mu_{t-1}(\theta_t))^2}{\sigma^2_{t-1}(\theta_t,\theta_t)+\lambda^{-1}} + \frac{1}{2}\log 2\pi(\sigma^2_{t-1}(\theta_t,\theta_t)+\lambda^{-1}) \\
=& \frac{1}{2}\log\left(1+\lambda\sigma^2_{t-1}(\theta_t,\theta_t)\right) - \frac{1}{2}\frac{\sigma^2_{t-1}(\theta_t,\theta_t)}{\sigma^2_{t-1}(\theta_t,\theta_t)+\lambda^{-1}} + \frac{1}{2}\frac{\sigma^2_{t-1}(\theta_t,\theta_t)(y_t-\mu_{t-1}(\theta_t))^2}{(\sigma^2_{t-1}(\theta_t,\theta_t)+\lambda^{-1})^2}.
\end{aligned}$$

C Proof of Theorem 2

We prove Theorem 2 using the following lemmas.

Lemma 5. *(Ha et al., 2019) Suppose the Bayesian optimization is performed with the GP-LCB as the acquisition function. Let the parameter space Θ be the d dimensional cube with side l, and the class of kernel functions for GP as a surrogate function be those composed of finite dimensional linear, squared exponential, and Matérn kernels. For a sample f of GP with the kernel function $k(\theta,\theta')$, assume $f : \exists a, b > 0, Pr\{\sup_{\theta\in\Theta}|\partial f/\partial\theta_j| > L\} \leq a\exp^{-(L/b)^2}$ holds. Let $\delta \in (0,1)$ and, $\beta_t = 2\log(t^2 2\pi^2/3\delta) + 2d\log(t^2 dbl\sqrt{\log(4da/\delta)})$. Then, for any $\epsilon > 0$, with probability at least $1-\delta$, for any $t > T$, there exists T satisfying*

$$2\beta_t^{1/2}\sigma_{t-1}(\theta_t,\theta_t) \leq \epsilon.$$

Lemma 6. *Let $\epsilon \sim \mathcal{N}(0,\lambda)$ and $\delta \in (0,1)$. Then,*

$$Pr\left(|\epsilon| \leq \frac{c}{\lambda^{1/2}}\right) \geq 1-\delta.$$

holds with probability $\geq 1-\delta$, where $c = \sqrt{-2\log\delta}$.

Proof. Setting $r \sim \mathcal{N}(0,1)$, the following equation holds.

$$\begin{aligned}
Pr(r \geq c) =& (2\pi)^{-\frac{1}{2}}\int_c^\infty e^{-r^2/2} \\
=& e^{-\frac{c^2}{2}}(2\pi)^{-\frac{1}{2}}\int_c^\infty e^{-(r-c)^2/2-c(r-c)}dr.
\end{aligned}$$

Since $e^{-c(r-c)} \leq 1$ holds for any $c > 0$ and $r > c$, the following equation is derived:

$$\begin{aligned}
e^{-\frac{c^2}{2}}(2\pi)^{-\frac{1}{2}}\int_c^\infty e^{-(r-c)^2/2-c(r-c)}dr \leq& e^{-\frac{c^2}{2}}Pr(r>0) \\
=& \frac{1}{2}e^{-\frac{c^2}{2}}.
\end{aligned}$$

Let $r = \epsilon_t/\sigma$. Since $Pr(|r| \leq c) \geq e^{-\frac{c^2}{2}}$,

$$Pr\left(|\epsilon_t| \leq \frac{c}{\lambda^{1/2}}\right) \geq 1-\delta$$

holds. Therefore, $\epsilon_t \leq c/\lambda^{1/2}$ holds with probability $\geq 1-\delta$. The proof is similar to Lemma 5.1 in Srinivas et al. (2010). □

We assume that a hyperparameter of the GP at time t is the same as that of the GP at time $t-1$. Then, for any posterior of Gaussian process, the following equation holds:

$$\mu_t(\theta) = \mu_{t-1}(\theta) + \frac{\sigma^2_{t-1}(\theta, \theta_t)(y_t - \mu_{t-1}(\theta_t))}{\sigma^2_{t-1}(\theta_t, \theta_t) + \lambda^{-1}}.$$

We decompose an observed value of an objective function for θ_t into a value of a true function corresponding to θ_t and noise term, that is $y_t = f(\theta_t) + \epsilon_t$. Then, by using Lemma 2 and 6, for any $\theta, \theta' \in \Theta$,

$$\begin{aligned}
|\mu_t(\theta) - \mu_{t-1}(\theta')| \leq& \left| \mu_{t-1}(\theta) - \mu_{t-1}(\theta') + \frac{\sigma^2_{t-1}(\theta, \theta_t)(y_t - \mu_{t-1}(\theta_{t-1}))}{\sigma^2_{t-1}(\theta_t) + \lambda^{-1}} \right| \\
\leq& |\mu_{t-1}(\theta) - \mu_{t-1}(\theta')| + \frac{|\sigma^2_{t-1}(\theta, \theta_t)||f(\theta_t) - \mu_{t-1}(\theta_t)| + |\sigma^2_{t-1}(\theta, \theta_t)||\epsilon_t|}{\sigma^2_{t-1}(\theta_t) + \lambda^{-1}} \\
\leq& |\mu_{t-1}(\theta) - \mu_{t-1}(\theta')| + \frac{\beta_t^{1/2}|\sigma^2_{t-1}(\theta, \theta_t)|\sigma_{t-1}(\theta_t)}{\sigma^2_{t-1}(\theta_t) + \lambda^{-1}} + \frac{|\sigma^2_{t-1}(\theta, \theta_t)|c}{\lambda^{1/2}(\sigma^2_{t-1}(\theta_t) + \lambda^{-1})} \\
\leq& |\mu_{t-1}(\theta) - \mu_{t-1}(\theta')| + \frac{\beta_t^{1/2}\sigma_{t-1}(\theta)\sigma^2_{t-1}(\theta_t)}{\sigma^2_{t-1}(\theta_t) + \lambda^{-1}} + \frac{\sigma_{t-1}(\theta)\sigma_{t-1}(\theta_t)c}{\lambda^{1/2}(\sigma^2_{t-1}(\theta_t) + \lambda^{-1})}
\end{aligned}$$

holds with probability $\geq 1 - \delta$. The transformation between the fourth and fifth rows uses the relationship between the covariance $\sigma^2_{t-1}(\theta, \theta')$ and correlation coefficient ρ, that is, $\sigma^2_{t-1}(\theta, \theta') = \rho\sigma_{t-1}(\theta)\sigma_{t-1}(\theta') \leq \sigma_{t-1}(\theta)\sigma_{t-1}(\theta')$.

Recall that

$$D_{\mathrm{KL}}[p(f|\mathbf{y}_t)||p(f|\mathbf{y}_{t-1})] = \frac{1}{2}\log\left(1 + \lambda\sigma^2_{t-1}(\theta_t)\right) - \frac{1}{2}\frac{\sigma^2_{t-1}(\theta_t)}{\sigma^2_{t-1}(\theta_t) + \lambda^{-1}} + \frac{1}{2}\frac{\sigma^2_{t-1}(\theta_t)(y_t - \mu_{t-1}(\theta_t))^2}{(\sigma^2_{t-1}(\theta_t) + \lambda^{-1})^2}.$$

Then, the third term can be decomposed as follows:

$$\begin{aligned}
(y_t - \mu_{t-1}(\theta_t))^2 =& (f(\theta_t) + \epsilon_t - \mu_{t-1}(\theta_t))^2 \\
\leq& (f(\theta_t) - \mu_{t-1}(\theta_t))^2 + 2|f(\theta_t) - \mu_{t-1}(\theta_t)||\epsilon_t| + \epsilon_t^2
\end{aligned}$$

Using Lemmas 2 and 6,

$$\begin{aligned}
(f(\theta_t) - \mu_{t-1}(\theta_t))^2 + 2|f(\theta_t) - \mu_{t-1}(\theta_t)||\epsilon_t| + \epsilon_t^2 \leq& \beta_t\sigma^2_{t-1}(\theta_t) + 2\beta_t^{1/2}\sigma_{t-1}(\theta_t)\frac{c}{\lambda^{1/2}} + \frac{c^2}{\lambda} \\
=& \frac{1}{\lambda}(\lambda^{1/2}\beta_t^{1/2}\sigma_{t-1}(\theta_t) + c)^2
\end{aligned}$$

holds with probability $\geq 1 - 2\delta$.

Since $\theta_t^* = \theta_{t-1}^*$ when $t \to \infty$, $|\mu_{t-1}(\theta_t^*) - \mu_{t-1}(\theta_{t-1}^*)| = 0$ and $v\phi(g) = vg\Phi(g) = 0$ holds. Summarizing the above argument, the following inequality holds with probability $\geq 1 - 2\delta$:

$$\begin{aligned}
\Delta\tilde{R}_t \leq& \frac{\beta_t^{1/2}\sigma_{t-1}(\theta_t^*)\sigma^2_{t-1}(\theta_t)}{\sigma^2_{t-1}(\theta_t) + \lambda^{-1}} + \frac{\sigma_{t-1}(\theta_t^*)\sigma_{t-1}(\theta_t)c}{\lambda^{1/2}(\sigma^2_{t-1}(\theta_t) + \lambda^{-1})} \\
&+ \kappa_t\sqrt{\frac{1}{4}\log\left(1 + \lambda\sigma^2_{t-1}(\theta_t)\right) - \frac{1}{4}\frac{\sigma^2_{t-1}(\theta_t)}{\sigma^2_{t-1}(\theta_t) + \lambda^{-1}} + \frac{1}{4}\frac{\sigma^2_{t-1}(\theta_t)\left(\lambda^{1/2}\beta_t^{1/2}\sigma_{t-1}(\theta_t) + c\right)^2}{\lambda(\sigma^2_{t-1}(\theta_t) + \lambda^{-1})^2}}. \qquad (10)
\end{aligned}$$

Using Lemma 5, $2\beta_t^{1/2}\sigma_{t-1}(\theta_t, \theta_t)$ goes to zero. In addition, since $\beta_t^{1/2}$ is an increasing function, $\sigma_{t-1}(\theta_t, \theta_t)$ goes to zero. Therefore, by using the assumption that $\kappa_t \leq C$, we can confirm that the proposed upper bound converges to zero.

D Bottleneck of the convergence of Eq. (10)

In this section, we compare the convergence rate for each term of Eq. (10) and show that Eq. (7) in the main text is a bottleneck for the convergence rate of Eq. (10).

By using $\log(1+x) \leq x$, we have

$$\begin{aligned}
\Delta\tilde{R}_t \leq & \frac{\beta_t^{1/2}\sigma_{t-1}(\theta_t^*)\sigma_{t-1}^2(\theta_t)}{\sigma_{t-1}^2(\theta_t)+\lambda^{-1}} + \frac{\sigma_{t-1}(\theta_t^*)\sigma_{t-1}(\theta_t)c}{\lambda^{1/2}(\sigma_{t-1}^2(\theta_t)+\lambda^{-1})} \\
& + \kappa_{t-1}\sqrt{\frac{1}{4}\frac{\lambda^2\sigma_{t-1}^2(\theta_t)(\sigma_{t-1}^2(\theta_t)+\lambda^{-1})^2}{\lambda(\sigma_{t-1}^2(\theta_t)+\lambda^{-1})^2} - \frac{1}{4}\frac{\lambda\sigma_{t-1}^2(\theta_t)(\sigma_{t-1}^2(\theta_t)+\lambda^{-1})}{\lambda(\sigma_{t-1}^2(\theta_t)+\lambda^{-1})^2} + \frac{1}{4}\frac{\sigma_{t-1}^2(\theta_t)\left(\lambda^{1/2}\beta_t^{1/2}\sigma_{t-1}(\theta_t)+c\right)^2}{\lambda(\sigma_{t-1}^2(\theta_t)+\lambda^{-1})^2}} \\
= & \frac{\beta_t^{1/2}\sigma_{t-1}(\theta_t^*)\sigma_{t-1}^2(\theta_t)}{\sigma_{t-1}^2(\theta_t)+\lambda^{-1}} + \frac{\sigma_{t-1}(\theta_t^*)\sigma_{t-1}(\theta_t)c}{\lambda^{1/2}(\sigma_{t-1}^2(\theta_t)+\lambda^{-1})} \\
& + \frac{\kappa_{t-1}}{2}\frac{\sqrt{\sigma_{t-1}^2(\theta_t)(\lambda^2\sigma_{t-1}^4(\theta_t)+\lambda\sigma_{t-1}^2(\theta_t)+\lambda\beta_t\sigma_{t-1}^2(\sigma_t)+2\lambda^{1/2}c\beta_t^{1/2}\sigma_{t-1}(\theta_t)+c^2)}}{\lambda^{1/2}(\sigma_{t-1}^2(\theta_t)+\lambda^{-1})}.
\end{aligned}$$

Since $\beta_t^{1/2}\sigma_t(\theta_t)$ and $\sigma_t(\theta_t)$ converge to zero and c and λ are constant, the convergence rate of the third term is $\kappa_{t-1}c\sigma_{t-1}(\theta_t)/2\lambda^{1/2}(\sigma_{t-1}^2(\theta_t)+\lambda^{-1})$, and the second term converges to zero slower than the first term. Therefore, the bottleneck of the convergence of Eq. (10) is

$$s_t \coloneqq \frac{(\sigma_{t-1}(\theta_t^*)+\kappa_{t-1}/2)\sigma_{t-1}(\theta_t)c}{\lambda^{1/2}(\sigma_{t-1}^2(\theta_t)+\lambda^{-1})}.$$

E An overview of the stopping criteria for other optimization methods

There are other approaches other than BO for finding the point that minimizes the objective function besides BO. This section outlines how to determine when to stop in evolutionary algorithms (EA). In evolutionary algorithms, the various methods that have been proposed for the stopping criterion, which can be classified into seven approaches (Ghoreishi et al., 2017): Direct termination criteria, derived termination criteria, cluster-based termination criteria, operator-based termination criteria, performance indicator termination criteria, progress indicator termination criteria, and termination criteria in hybrid multi-objective evolutionary algorithms.

The simplest approach is to decide whether to stop searching depending on predetermined termination criteria without using statistics or models obtained in the search process, which is called direct termination criteria. For example, the maximal time budget stops EA after running a predetermined CPU time, the maximum number of objective function evaluations stops EA after evaluating a predetermined number of evaluations, and K-iterations stops EA when the optimal value does not update a predetermined number of times consecutively (Jain et al., 2001). These stopping criteria can be used in other optimization methods, but there is no theoretical guarantee of the quality of the search results.

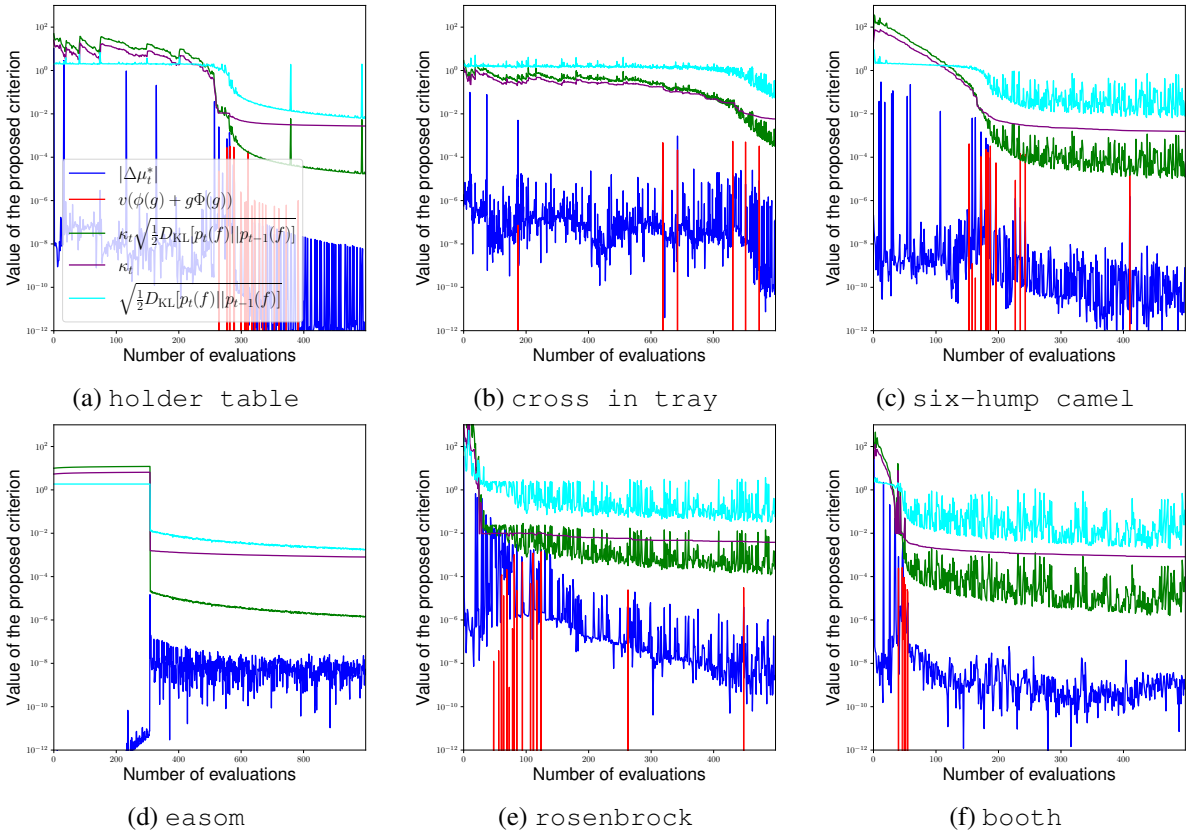
Another approach is to terminate EA when some kind of performance indicators exceed a predefined desirable value, which is called performance indicator termination criteria. For example, the hyper volume metric and the additive epsilon indicator have been proposed (Coello et al., 2006; Zitzler et al., 2003). However, there are many cases in which the desired performance indicators are not predefined. In such cases, the approach that stops EA based on the change in the performance indicators before and after the search is called the progress indicator termination criteria (Trautmann et al., 2008; Guerrero et al., 2009). In these classifications, the proposed criteria is one of the progress indicator termination criteria, because it stops BO based on the upper bound of the changes in the expected minimum simple regret before and after the search.

F Behavior of the proposed bound

In this section, we discuss the behavior of the following upper bound for test function minimization:

$$\Delta R_t \leq |\Delta\mu_t^*| + v(\phi(g)+g\Phi(g)) + \kappa_{t-1}\sqrt{\frac{1}{2}D_{\mathrm{KL}}[p_t(\hat{f})||p_{t-1}(\hat{f})]},$$

where $\kappa_{t-1} \coloneqq \min_{\hat{\theta}\in\Theta_t}\mathrm{UCB}_\delta(\hat{\theta}) - \min_{\hat{\theta}\in\Theta}\mathrm{LCB}_\delta(\theta)$, $\Delta\mu_t^* \coloneqq \mu_{t-1}(\theta_{t-1}^*) - \mu_t(\theta_t^*)$, $v = \sqrt{\sigma_t^2(\theta_t^*) - 2\sigma_t^2(\theta_t^*,\theta_{t-1}^*) + \sigma_t^2(\theta_{t-1}^*)}$ and $g = (\mu_t(\theta_t^*) - \mu_t(\theta_{t-1}^*))/v$.



(a) `holder table` (b) `cross in tray` (c) `six-hump camel`

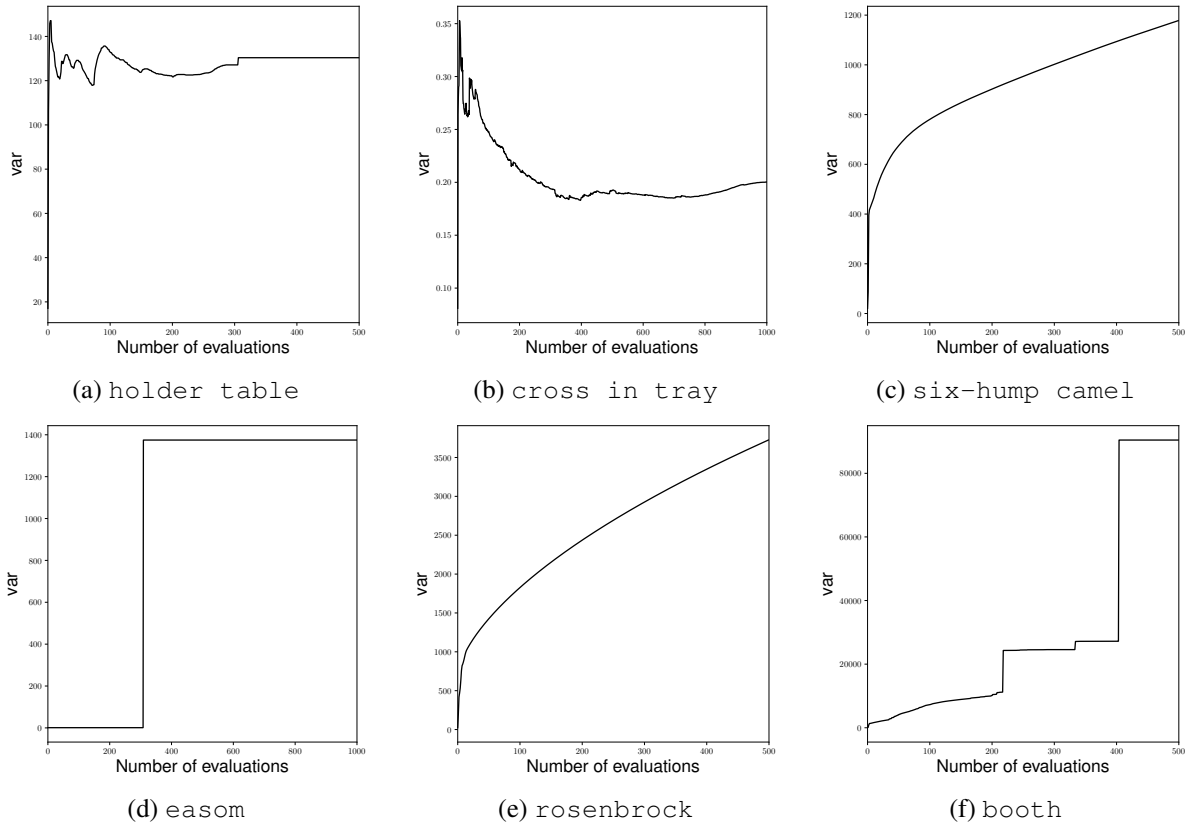
(d) `easom` (e) `rosenbrock` (f) `booth`

Figure 4: Sequences of each term of the proposed bound in LCB acquisition function.

The transitions of each R.H.S. term in the above inequality during the process of BO applied to minimizing test functions are shown in Fig. 4. Because the third term is more significant than the first and second terms for each test function, we can see that the third term strongly affects the determination of the stopping timing. It is important to note that the third term is an upper bound product of the upper bound of the simple regret and the KL divergence between GPs. Furthermore, the KL divergence between GPs tends to be small because the change in the posterior of GP becomes smaller with BO progress. On the other hand, the upper bound of the simple regret is difficult to decrease, even when the BO procedure is progressed. Therefore, the proposed upper bound tends to decrease faster than that of the simple regret as BO is progressed, and it is an intuitive reasoning for the consequence of Theorem 2 (i.e., the proposed method can terminate BO with high probability).

G Convergence of the sequence of estimated hyperparameters

The proposed criterion has a limitation in that we cannot calculate a KL divergence between GPs when its hyperparameter ν_t changes in each iteration. To avoid the problem, we assume that the change of the GP posterior is decreasing simultaneously with the progress of BO in general, which leads to a decrease of the change of ν_t, and can be approximately calculated by $D_{\mathrm{KL}}[p_t(\hat{f}|\nu_t)||p_{t-1}(\hat{f}|\nu_t)]$ instead of $D_{\mathrm{KL}}[p_t(\hat{f}|\nu_t)||p_{t-1}(\hat{f}|\nu_{t-1})]$. In this section, we demonstrate that the assumption tends to hold. Figures 5 and 6 show the sequence of the variance and length scale of the Matérn kernel in the test function optimization, respectively. From these results, in the `holder table`, `cross in tray`, `easom` and `booth` functions, the variance and length scale converge to a value, respectively. In the `rosenbrock` function, the length scale converges to a value, while the variance does not. In the `six-hump camel` function, the sequences of variance and length scale do not yet converge, but they tend to decrease as the number of searches increases.



(a) `holder table` (b) `cross in tray` (c) `six-hump camel`

(d) `easom` (e) `rosenbrock` (f) `booth`

Figure 5: Sequences of the variance of each stopping criterion's kernel function in LCB acquisition function.

H Additional experiments

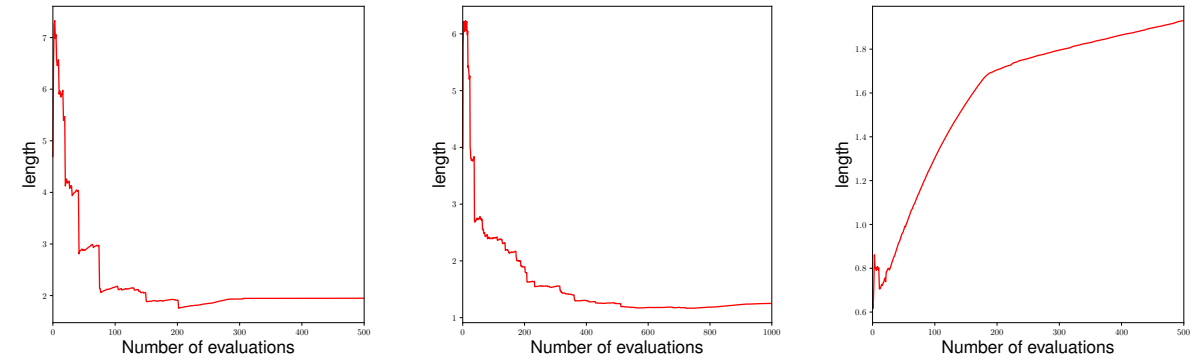
H.1 Test function optimization

Figures 8 and 9 show the global optimal solutions and the explored points discovered by each stopping criterion in each test function, respectively. Every stopping criterion can find the global optima for the `cross in tray` and `six-hump camel` functions, while EI- and PI-based criteria may not find the optimal points for the `holder in table` function. In Fig. 7, the values of EI and PI are shown to rapidly decrease at the timings before the optimal points have been found. In the `rosenbrock` and `booth` functions, the discovered points by Ours-auto tend to be closer than those of the other criteria. As shown in Fig. 10, the BO was not found to be the optimal solution in any of the trials, even with the full budget in the `easom` function. In addition, the search by the GP-based BO is clearly inefficient, as it repeatedly searches the edges of the search area. In such a situation, it is desirable to stop BO early and then seek another search method.

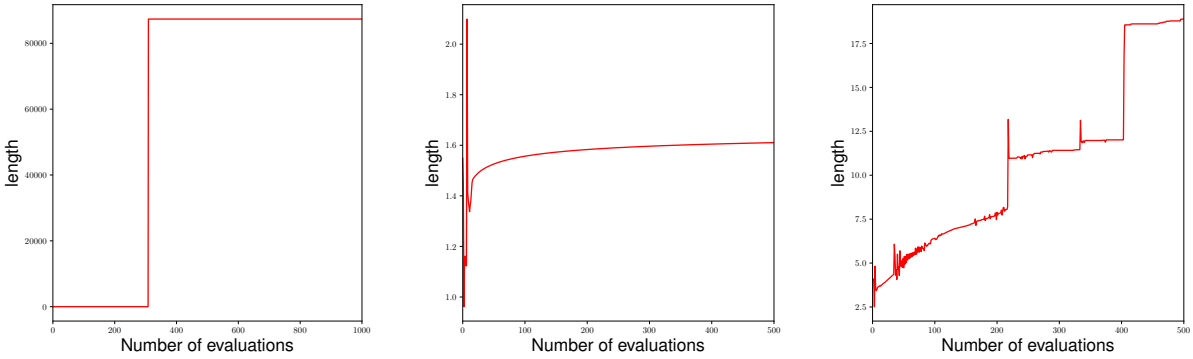
Next, we demonstrate the results of the comparison of the stopping criteria when we using the other acquisition functions, EI and PI. Figures 11, 12, 13, and 14 show the result from each stopping criterion's timing and the explored points discovered by them. These results are almost identical to the results from when the LCB is used. Therefore, the proposed criterion can be applied to any acquisition function.

H.2 Hyperparameter optimization

We evaluate the stopped timing of BO for the hyperparameter search. The stopping criteria and the settings are the same as those of the experiment conducted in the main text. The details of the data used in the experiments are shown in Table 2. In this experiment, to enable the SR-based stopping criterion to automatically determine thresholds, a predictive model is conducted by the selected hyperparameter, and the SD of the generalization error is estimated by using a 10-fold CV and the threshold in Eq. (5) of the main text is computed. Errors in test data are evaluated to assess the actual prediction accuracy of the model. In the regression task, 4,000 points are randomly sampled from the entire dataset: 2,000 samples are used for 10-fold CV, and the remaining samples are used as the test dataset. Here, the input data were normalized to a



(a) `holder table` (b) `cross in tray` (c) `six-hump camel`



(a) `easom` (b) `rosenbrock` (c) `booth`

Figure 6: Sequences of the length scale of each stopping criterion's kernel function in LCB acquisition function.

mean of 0 and a variance of 1. In the classification task, 6,000 points are randomly sampled from the entire dataset: 3,000 samples are used for 10-fold CV, and the remaining samples are used as the test dataset. Here, the input and output data were normalized to a mean of 0 and a variance of 1. In this experiment, the BO is executed after sampling the 10 initial points randomly.

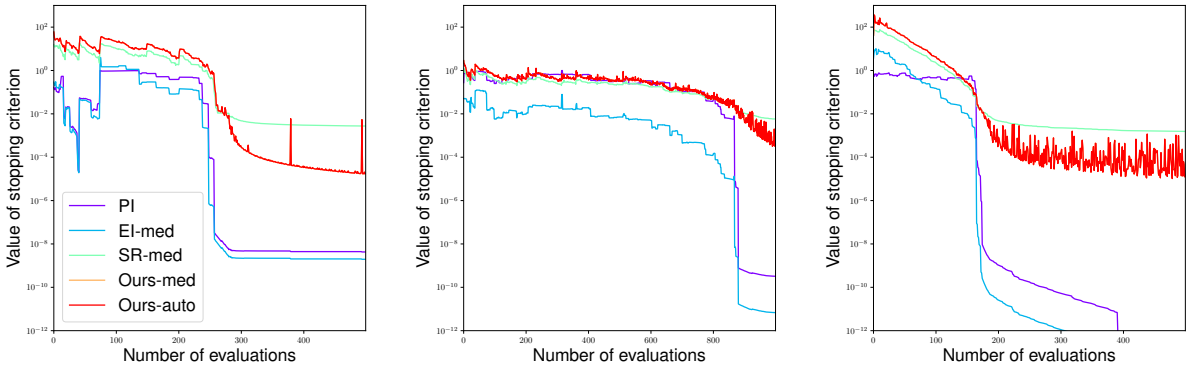
The experimental results for each data are shown in Fig. 15 and Fig. 16, which consistently demonstrate that the EI and PI criteria tend to terminate quickly, and the SR criterion tends to terminate BO at the slowest timing among the criteria. In addition, Ours-auto and Ours-med tend to terminate BO after discovering hyperparameters that are within the 10th smallest simple regret.

Table 2: Description of datasets.

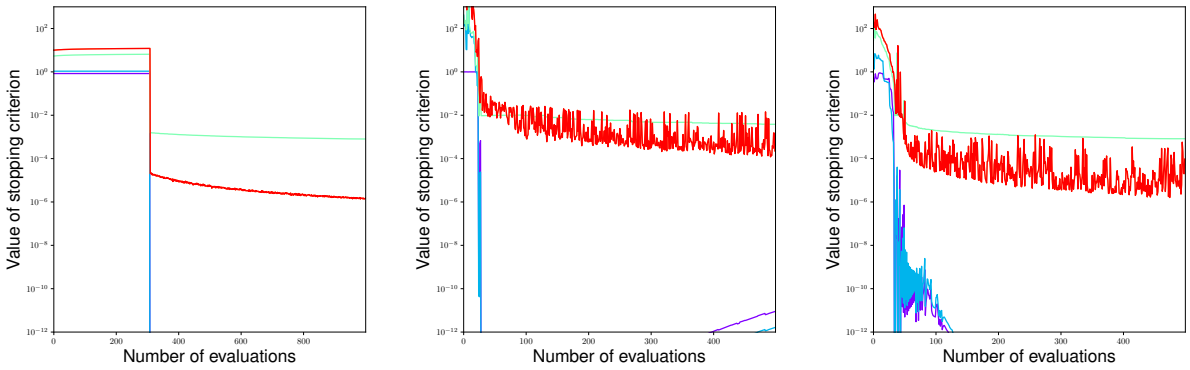
	dataset name	# of whole samples	# of features
classification	HTRU2	17898	9
	electrical grid stability	10000	14
regression	protein	45730	9
	power plant	9568	4

In the above experiment, we considered minimizing the CV errors when we search an optimal hyperparameter for predictive models. However, even though hyperparameters minimize the CV error, BO can overfit the validation dataset and test errors may be significant. Therefore, it is not always desirable to terminate BO at the timing when simple regret is minimized. To address this problem, Makarova et al. (2021) proposed using the stopping criterion to terminate BO early. In this experiment, we evaluate how much overfitting on the dataset for CV can be avoided by stopping BO early using this method. For this purpose, relative test error change (RYC) and relative time change (RTC) were used to evaluate the stopping criteria (Makarova et al., 2021). This study also employs RYC and RTC to verify whether the proposed method can stop BO at a time that reduces the test error.

The test error of the optimal hyperparameter at the stopped timing determined by each stopping criterion (early stopped) is



(a) `holder table` (b) `cross in tray` (c) `six-hump camel`



(a) `easom` (b) `rosenbrock` (c) `booth`

Figure 7: The representative result on sequences of each stopping criterion's value in LCB acquisition function.

y_{es}, and y_T is the test error of the optimal hyperparameter when all budgets T are used. Then, RYC is defined as

$$\mathrm{RYC} = \frac{y_T - y_{\mathrm{es}}}{\max\left(y_T, y_{\mathrm{es}}\right)}.$$

The range of RYC is $[-1, 1]$, where $\mathrm{RYC} > 0$ indicates that the test error due to early stopping is smaller than that achieved when all budgets are consumed, and $\mathrm{RYC} < 0$ vice versa. $\mathrm{RYC} \simeq 0$ indicates that the stopping criterion has successfully terminated BO. Usually, the greater the number of budgets used, the smaller the regret becomes, so $\mathrm{RYC} > 0$ does not occur. However, in the hyperparameter tuning of predictive models, the predictive model found by BO may be over-fitted to the validation data, and the actual test error may be large even when the error in the validation data is small (Makarova et al., 2021). In this case, there is a possibility that $\mathrm{RYC} > 0$.

t_{es} denotes the total cost of computation at the time of stopping determined by a stopping criterion, and t_T is the total cost when all budgets of T are consumed[4]. Then, RTC is defined as

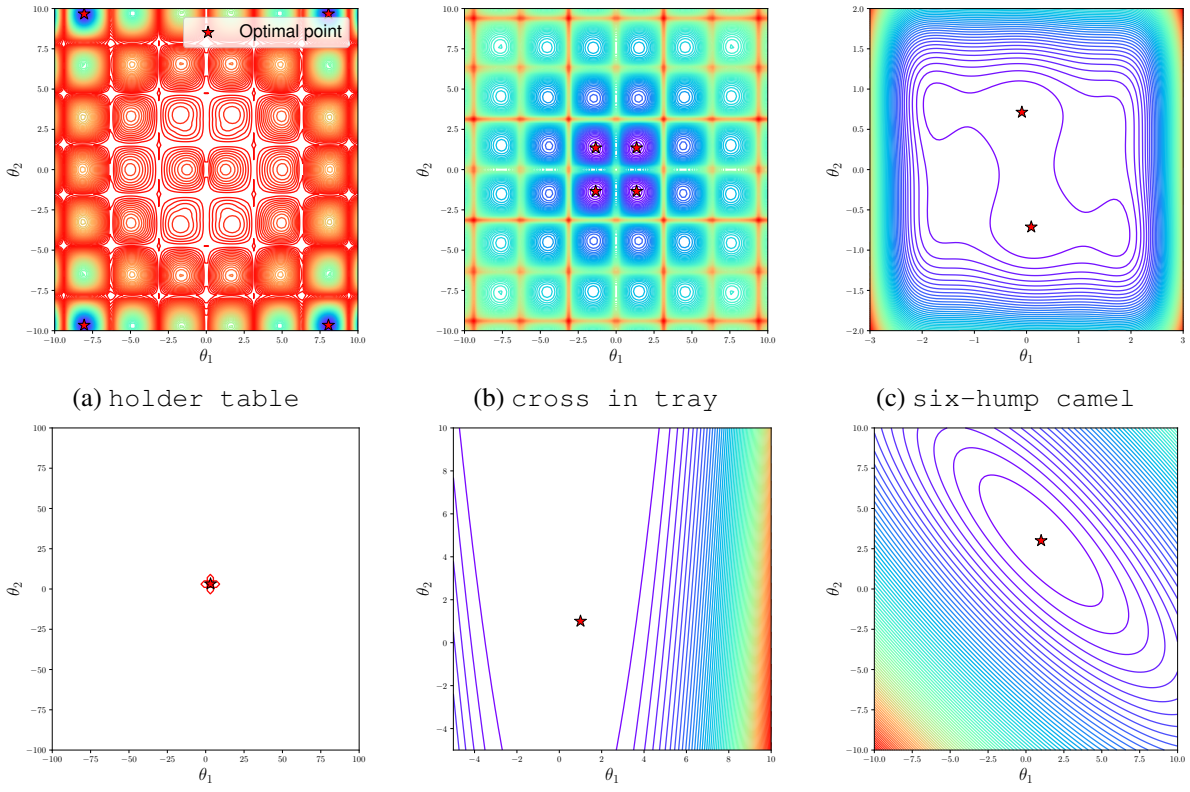
$$\mathrm{RTC} = \frac{t_T - t_{\mathrm{es}}}{t_T}.$$

The range of the RTC is $[0, 1]$, where the closer the RTC is to 1, the lower the cost. For simplicity, we assume that the cost of evaluating a hyperparameter at any time is always constant; hence, $\mathrm{RTC} = (T - T^*)/T$, where T^* is the number of outer-loop iterations when the BO is terminated by a stopping criterion.

Since RYC and RTC are trade-offs, we evaluate both RYC and RTC for each stopping criterion when BO is performed; $I = 10$ times with different random seeds.

As shown in Figs. 18 and 19, we could not obtain consistent results (i.e., the RYC of Ours-med and Ours-auto is below or above zero in some cases). Therefore, the early stopping in the proposed method does not always lead to avoiding

[4]Namely, t_{es} is the cumulative sum of time spent in evaluating the objective function and acquisition function when the outer-loop of BO is terminated by early stopping, and t_T is that when the outer-loop of BO is iterated T times and then terminated.

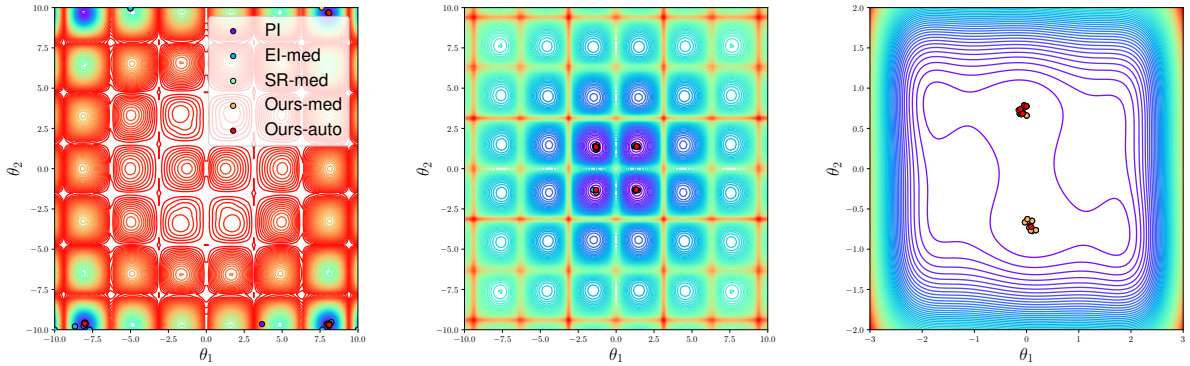


(a) `holder table` (b) `cross in tray` (c) `six-hump camel`

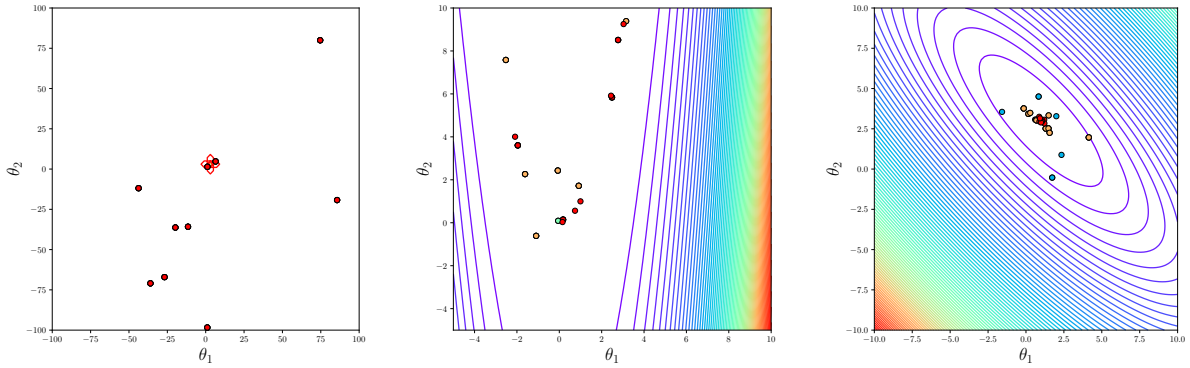
(d) `easom` (e) `rosenbrock` (f) `booth`

Figure 8: Optimal points in each test function.

overfitting. This is because the proposed stopping criterion aims to stop when it finds a parameter that minimizes the error for CV, which is a different objective.

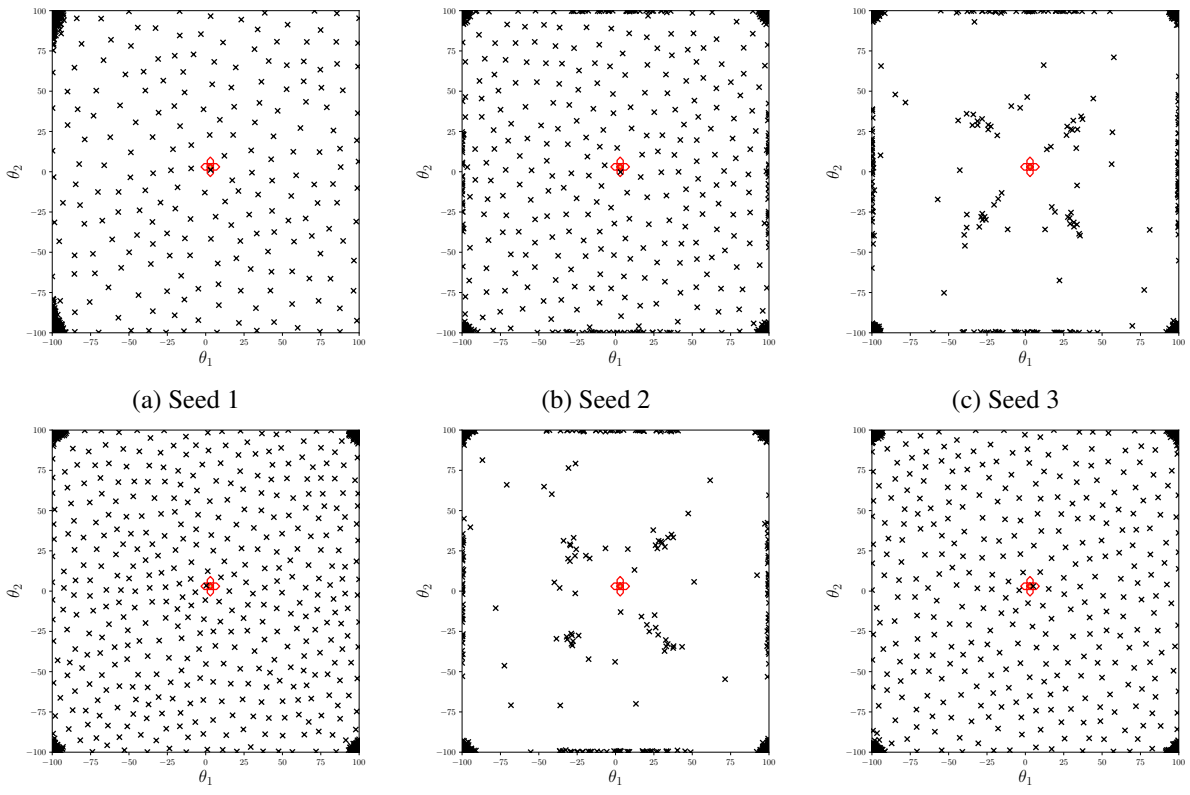


(a) `holder table` (b) `cross in tray` (c) `six-hump camel`



(d) `easom` (e) `rosenbrock` (f) `booth`

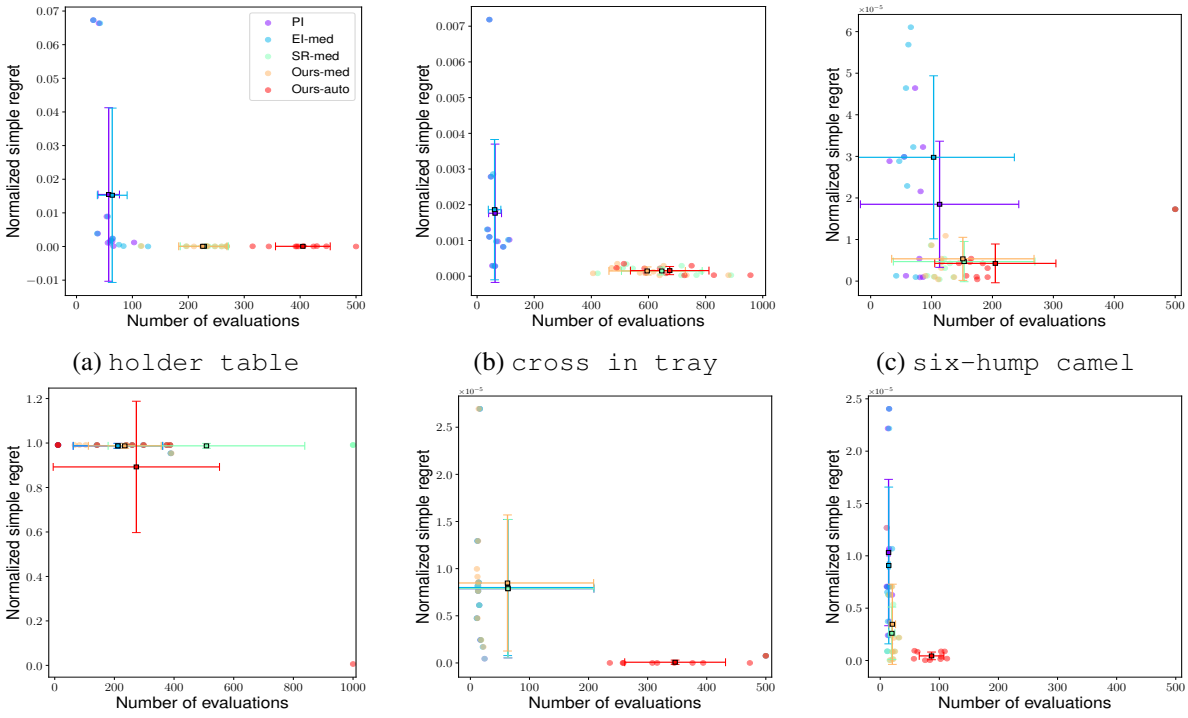
Figure 9: Discovered points by each stopping criterion in each test function.



(a) Seed 1 (b) Seed 2 (c) Seed 3

(c) Seed 4 (b) Seed 5 (c) Seed 6

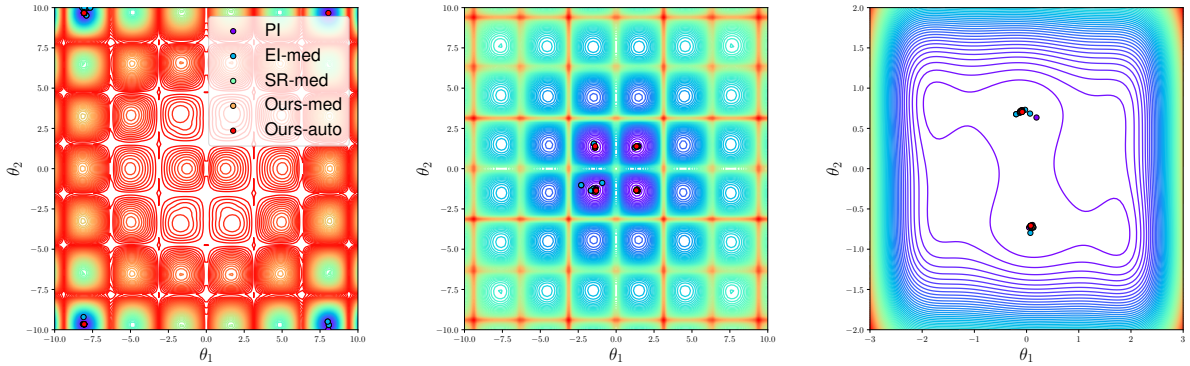
Figure 10: Sample points in the `easom` function. "Seed" represents the random seed provided when randomly generating the initial samples.



(a) `holder table` (b) `cross in tray` (c) `six-hump camel`

(d) `easom` (e) `rosenbrock` (f) `booth`

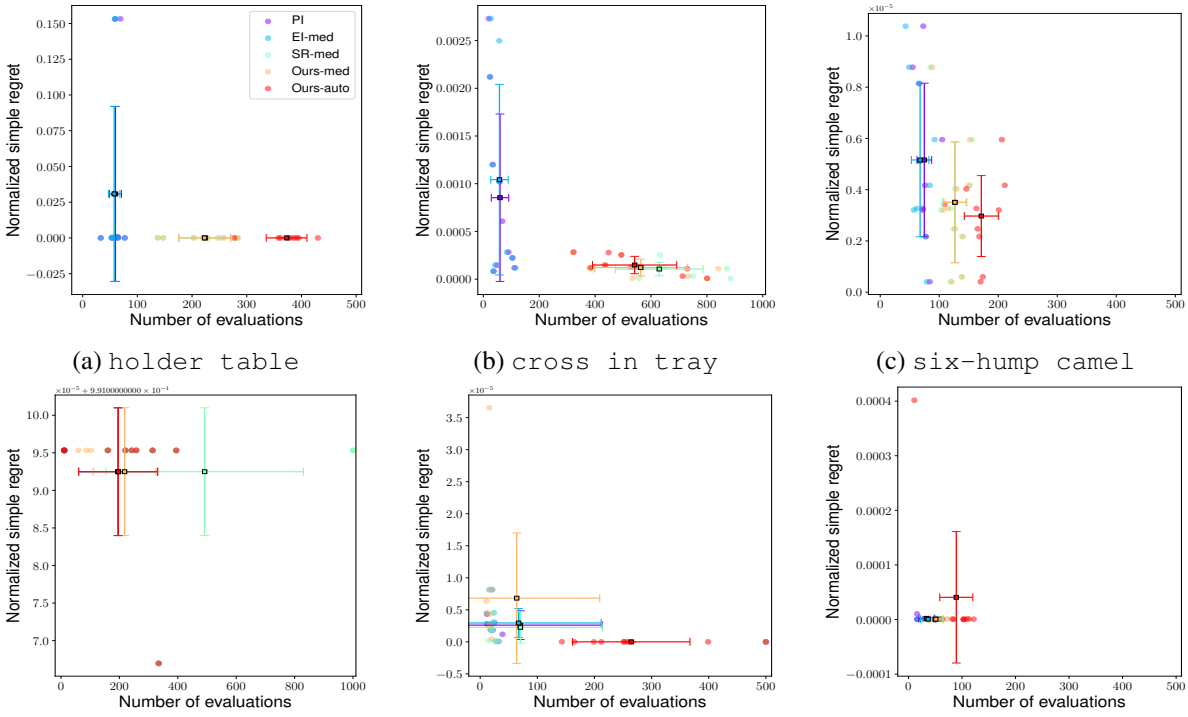
Figure 11: Stopped timing for each test function in the EI acquisition function; "med" denotes that the threshold is determined based on the median of the initial 20 searches, and "auto" indicates that the threshold was determined automatically by Eq. (7).



(a) `holder table` (b) `cross in tray` (c) `six-hump camel`

(d) `easom` (e) `rosenbrock` (f) `booth`

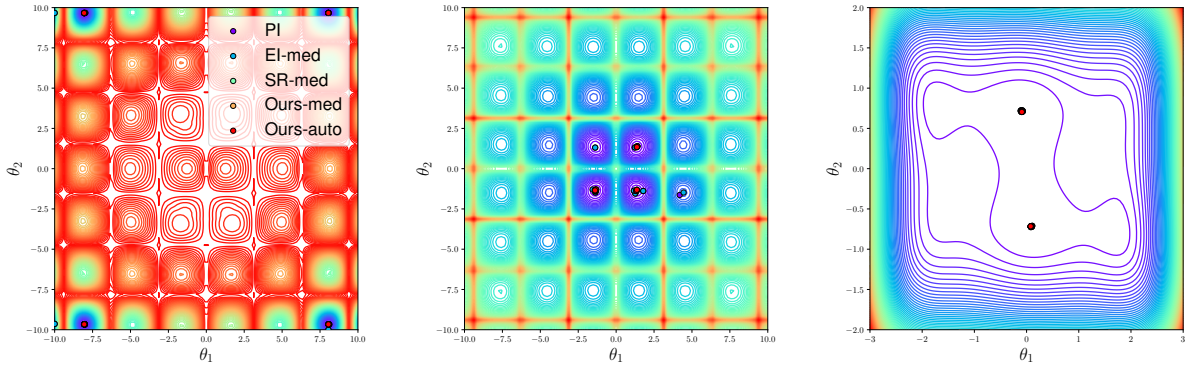
Figure 12: Discovered points by each stopping criterion in the EI acquisition function.



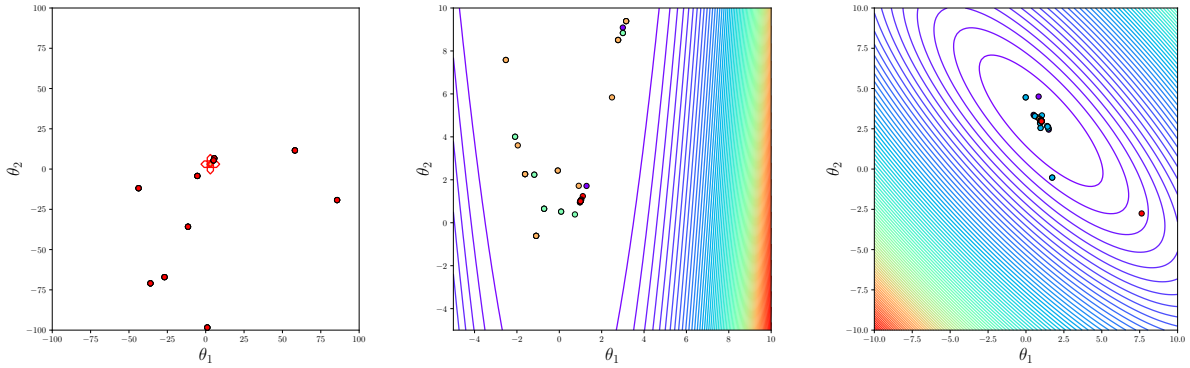
(a) `holder table` (b) `cross in tray` (c) `six-hump camel`

(d) `easom` (e) `rosenbrock` (f) `booth`

Figure 13: Stopped timing for each test function in the PI acquisition function. "med" denotes that the threshold is determined based on the median of the initial 20 searches, and "auto" indicates that the threshold was determined automatically by Eq. (7).

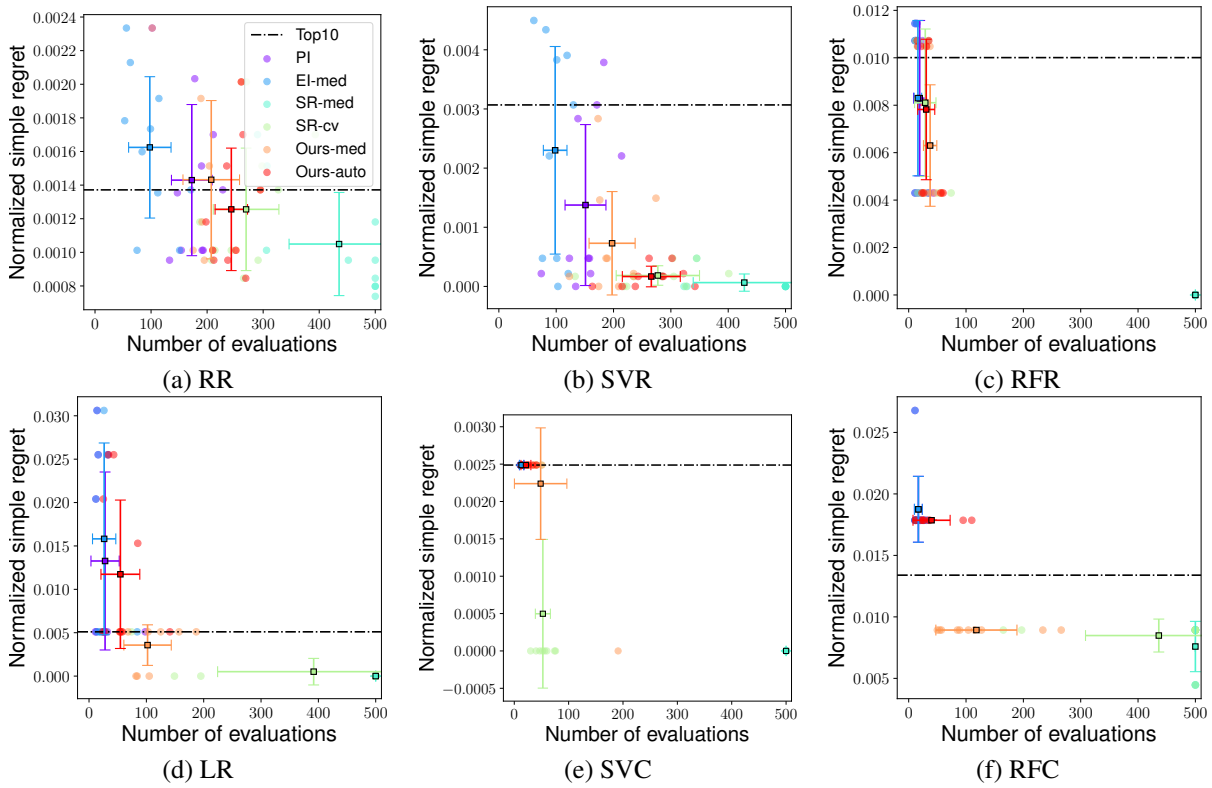


(a) `holder table` (b) `cross in tray` (c) `six-hump camel`



(d) `easom` (e) `rosenbrock` (f) `booth`

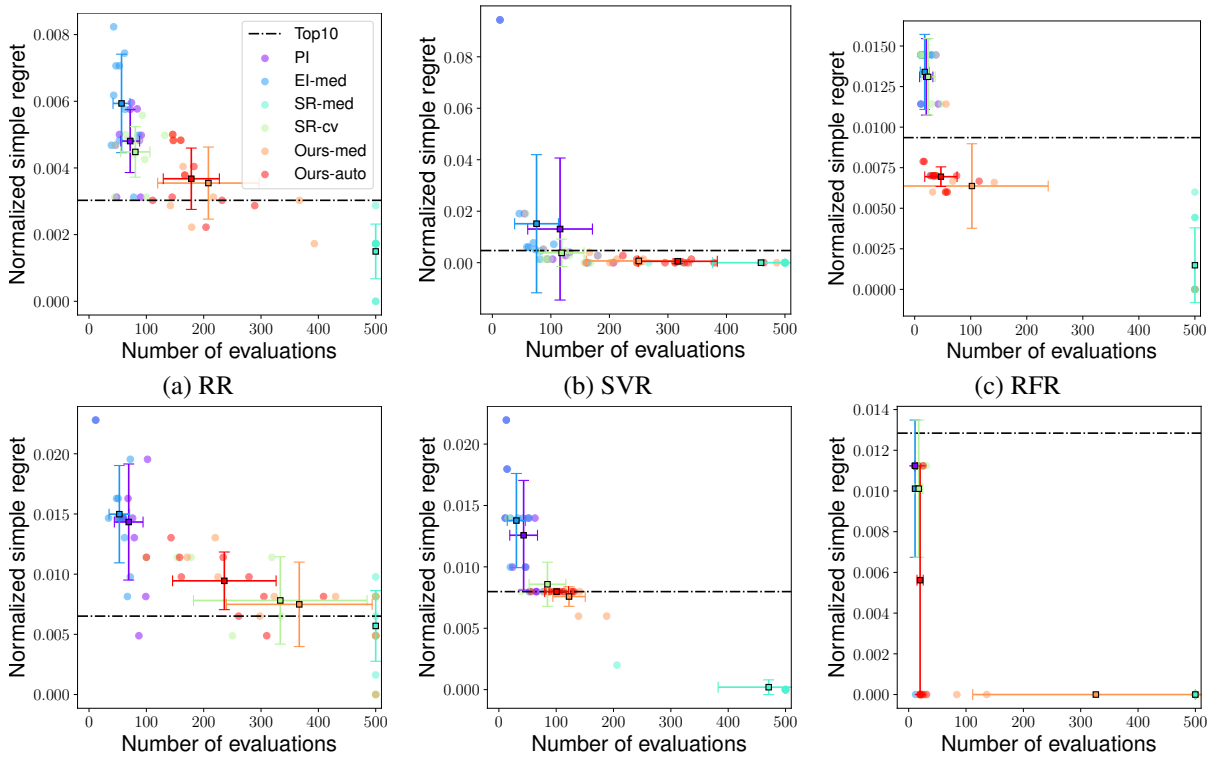
Figure 14: Discovered points by each stopping criterion in the PI acquisition function.



(a) RR (b) SVR (c) RFR

(d) LR (e) SVC (f) RFC

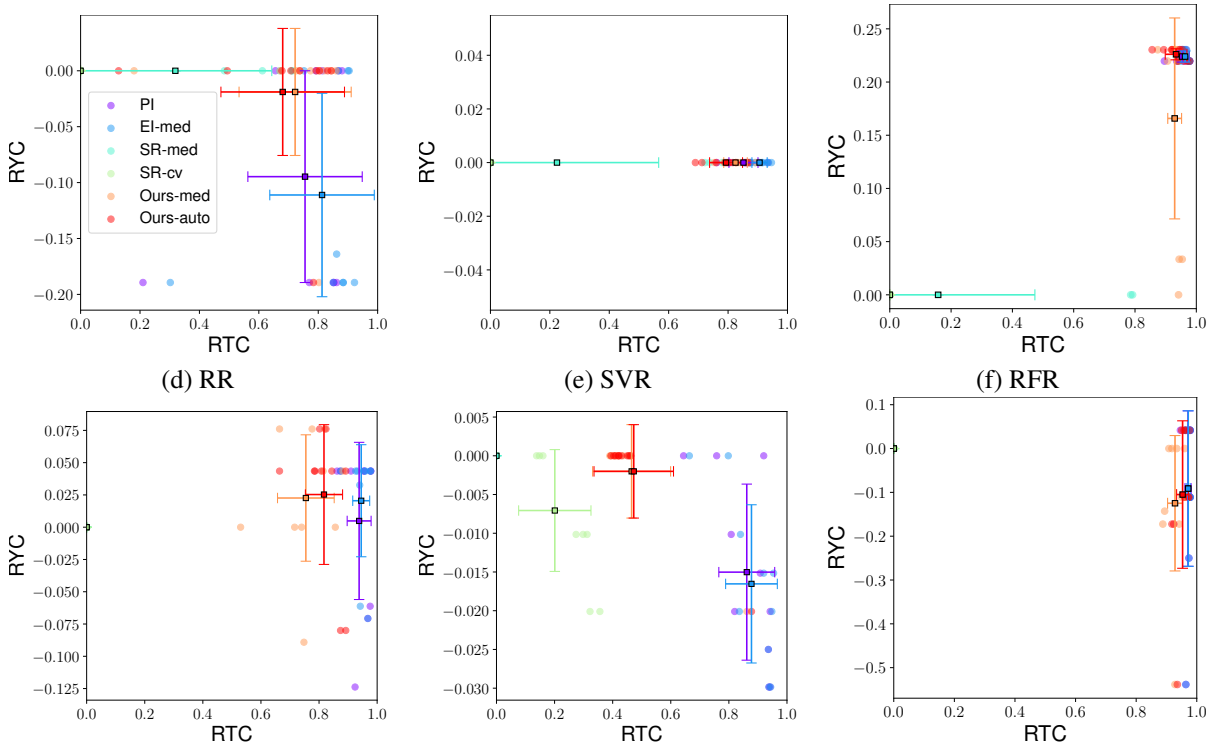
Figure 15: The number of evaluations and regret at each criterion's stopped timing in the predictive model search during 20 trials with different random initialization. (a)–(c): `HTRU2` data for classification, (d)–(f): `power plant` data for regression. The mean and SD for each stopping criterion are respectively shown as rectangles and error bars, respectively, in the corresponding colors.



(a) RR (b) SVR (c) RFR

(d) LR (e) SVC (f) RFC

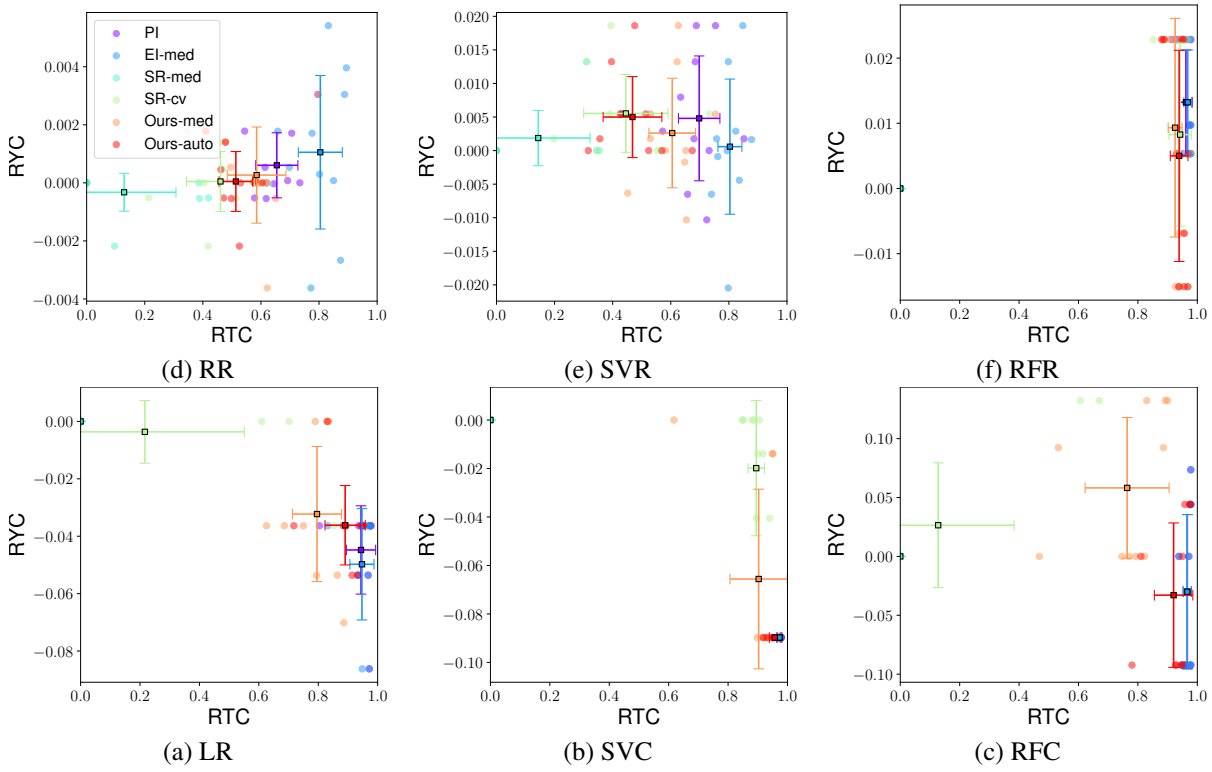
Figure 16: The number of evaluations and regret at each stopping criterion's stopped timing in the predictive model search during 20 trials with different random initialization. (a)–(c): `electrical grid stability` data for classification, (d)–(f): `protein` data for regression. The mean and SD for each stopping criterion are respectively shown as rectangles and error bars, respectively, in the corresponding colors.



(d) RR (e) SVR (f) RFR

(a) LR (b) SVC (c) RFC

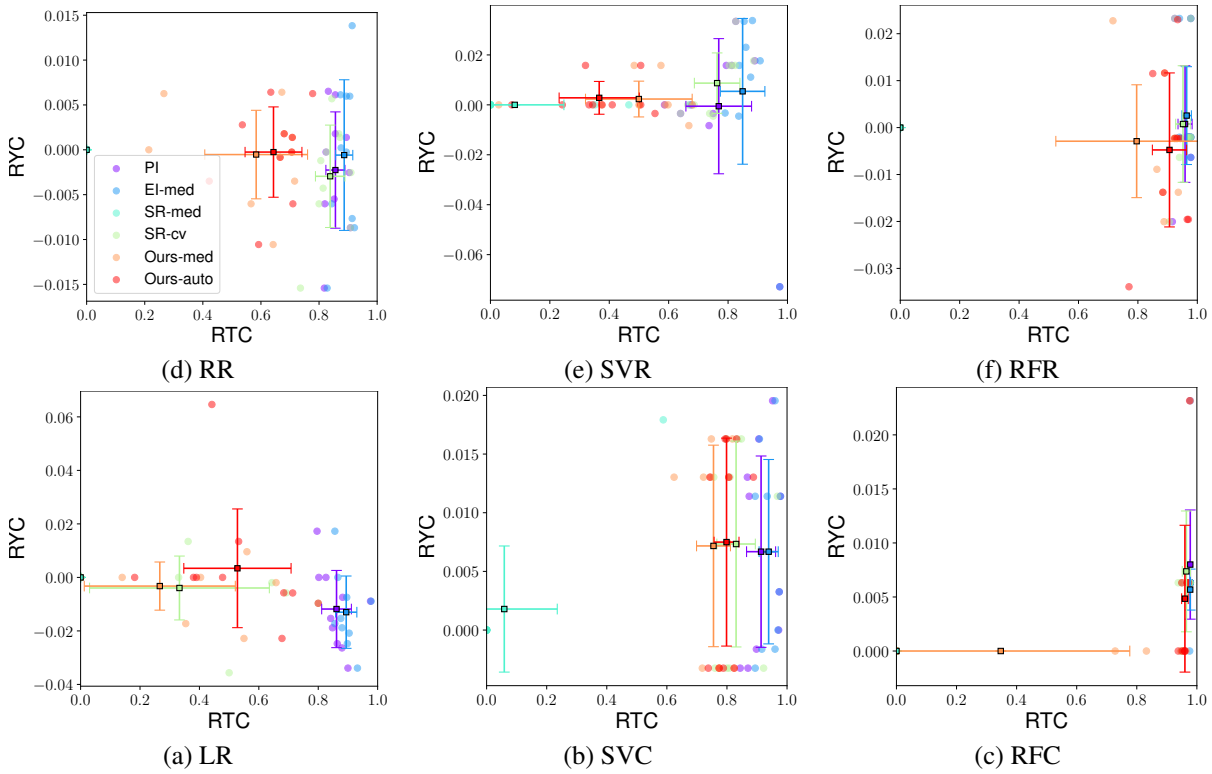
Figure 17: The RTC and RYC of the stopped timing determined by each stopping criterion in the predictive model search during 20 trials with different random initializations. (a)–(c): `skin` data for classification, (d)–(f): `gas turbine` data for regression. The mean and SD for each stopping criterion are respectively shown as rectangles and error bars, respectively, in the corresponding colors.



(d) RR (e) SVR (f) RFR

(a) LR (b) SVC (c) RFC

Figure 18: The RTC and RYC of the stopped timing determined by each stopping criterion in the predictive model search during 20 trials with different random initializations. (a)–(d): `HTRU2` data for classification, (e)–(h): `power plant` data for regression. The mean and SD for each stopping criterion are respectively shown as rectangles and error bars, respectively, in the corresponding colors.



(d) RR (e) SVR (f) RFR

(a) LR (b) SVC (c) RFC

Figure 19: The RTC and RYC of the stopped timing determined by each stopping criterion in the predictive model search during 20 trials with different random initializations. (a)–(d): `electrical grid stability` data for classification, (e)–(h): `protein` data for regression. The mean and SD for each stopping criterion are respectively shown as rectangles and error bars, respectively, in the corresponding colors.