Proceedings of Machine Learning Research 77:279–294, 2017 ACML 2017

Regret for Expected Improvement over the Best-Observed Value and Stopping Condition

Vu Nguyen V.NGUYEN@DEAKIN.EDU.AU
Sunil Gupta SUNIL.GUPTA@DEAKIN.EDU.AU
Santu Rana SANTU.RANA@DEAKIN.EDU.AU
Cheng Li CHENG.L@DEAKIN.EDU.AU
Svetha Venkatesh SVETHA.VENKATESH@DEAKIN.EDU.AU
Deakin University, Geelong, Australia, Center for Pattern Recognition and Data Analytics

Editors: Yung-Kyun Noh and Min-Ling Zhang

Abstract

Bayesian optimization (BO) is a sample-efficient method for global optimization of expensive, noisy, black-box functions using probabilistic methods. The performance of a BO method depends on its selection strategy through the acquisition function. Expected improvement (EI) is one of the most widely used acquisition functions for BO that finds the expectation of the improvement function over the incumbent. The incumbent is usually selected as the best-observed value so far, termed as $y^{\max}$ (for the maximizing problem). Recent work has studied the convergence rate for EI under some mild assumptions or zero noise of observations. Especially, the work of Wang and de Freitas (2014) has derived the sublinear regret for EI under a stochastic noise. However, due to the difficulty in stochastic noise setting and to make the convergent proof feasible, they use an alternative choice for the incumbent as the maximum of the Gaussian process predictive mean, $\mu^{\max}$. This modification makes the algorithm computationally inefficient because it requires an additional global optimization step to estimate $\mu^{\max}$ that is costly and may be inaccurate. To address this issue, we derive a sublinear convergence rate for EI using the commonly used $y^{\max}$. Moreover, our analysis is the first to study a stopping criteria for EI to prevent unnecessary evaluations. Our analysis complements the results of Wang and de Freitas (2014) to theoretically cover two incumbent settings for EI. Finally, we demonstrate empirically that EI using $y^{\max}$ is both more computationally efficiency and more accurate than EI using $\mu^{\max}$.

1. Introduction

Global optimization is fundamental to diverse real-world problems where parameter settings and design choices are pivotal - as an example, in algorithm hyper-parameter tuning (Nguyen et al., 2017) or engineering design (Frazier and Wang, 2016; Rana et al., 2017)). In particular, these algorithms and designs can be viewed as an optimization problem of a black-box objective function. Here, the input of the black-box are the hyper-parameters, and the objective function value is the output performance such as accuracy. This requires us to optimize a non-convex objective function using sequential and noisy observations. Critically, the objective functions are unknown and expensive to evaluate. The challenge is to find the maximum of such expensive objective functions in few sequential queries, thus minimizing time and cost.

Bayesian optimization (BO) has become a popular choice for improving the performance of machine learning algorithms and laboratory experiments (Brochu et al., 2010; Snoek et al., 2012;

Dai Nguyen et al., 2016; Shahriari et al., 2016; Li et al., 2017). BO finds a solution of an expensive black-box function $x^* = \underset{x \in \mathcal{X}}{\text{argmax}}\, f(x)$ by making a series of evaluations $x_1, ..., x_T$ of f . Fundamentally, BO builds and sequentially updates a surrogate model, typically through a Gaussian process (GP) (Rasmussen, 2006). Then, given the GP posterior update, BO defines a decision function - known as the acquisition function - to select a next experimental setting.

Although many acquisition functions have been proposed (e.g. Hennig and Schuler (2012); Hernández-Lobato et al. (2014); Srinivas et al. (2010)), the expected improvement (EI) (Mockus et al., 1978; Jones et al., 1998)is considered as one of the most popular acquisition function for Bayesian optimization and remains the default choice in BO packages, such as Spearmint (Snoek et al., 2012). EI balances the exploration and exploitation by taking the expectation of the improvement function over the *incumbent* ξ, i.e. $\mathbb{E}[\max\{0, f(x) - \xi\}]$. The incumbent ξ is often set to the best-observed value up to an iteration t, i.e. $\xi = y^{\max} = \max_{y_i \in \mathscr{D}_{t-1}} y_i$ (Brochu et al., 2010; Snoek et al., 2012) where $\mathscr{D}_t = \left\{x_i \in \mathbb{R}^d, y_i \in \mathbb{R}\right\}_{i=1}^{t}$ is the observation set including the feature x_i and the outcome $y_i = f(x_i) + \varepsilon_i$ given $f(.)$ which is the black-box function.

Theoretical analyses for the expected improvement (EI) have been studied in recent work under some mild conditions (Vazquez and Bect, 2010; Bull, 2011; Ryzhov, 2016). Since Bayesian optimization needs to accommodate the noise measurement that EI also needs to be able cope with. Theoretical analysis of EI for the noisy setting is notably studied in (Wang and de Freitas, 2014). However, due to the difficulty in stochastic setting and to make the proof feasible, instead of choosing the standard $\xi = y^{\max}$, (Wang and de Freitas, 2014) *uses the alternative choice of the incumbent as the maximum of the GP predictive mean* $\xi = \mu^{\max} = \max_{x \in \mathscr{X}} \mu_{t-1}(x)$ to derive the convergence bound. Due to this modification, the Bayesian optimization algorithm needs an additional step of global optimization to estimate $\mu^{\max}$ in $\mathbb{R}^d$. As a result, the computational cost of the whole BO process increases. In addition, the estimation of global optimization is not always accurate especially for high dimension problems. This computational cost and inaccuracy in the incumbent $\mu^{\max}$ can degrade the efficiency of the EI and Bayesian optimization as a whole.

Moreover, none of the existing work considers the stopping criteria for EI. Stopping criteria is critical in Bayesian optimization to control when to stop the search. This is because BO algorithm is often run and terminated after a finite number of iterations in practice. However, this stopping condition is not theoretically studied in the literature (Lorenz et al., 2015). A possible stopping criteria can be set as the maximum possible value of the black-box function if we know it in advance. For example in optimizing the hyper-parameter for a machine learning algorithm to get the highest F1-score, a simple stopping criteria is when the F1-score reaches to 1. However, this maximum possible value of F1-score is never reachable due to the imperfection of the learning algorithm and the given data. Hence, the Bayesian optimization may waste time and resource for evaluations even after it visits the best location. Thus, BO needs a principle criteria to stop the search.

To address these gaps, we consider the original version of EI that uses the best-observed function value $y^{\max}$. Allowing noisy observations, we prove the sublinear convergence rate of BO using EI over $y^{\max}$. In addition, we are the first to present and connect the stopping criteria for BO using EI into the regret. We show the convergence rate which is sublinear in the number of iteration T. Therefore, our analysis is **complementary** to (Wang and de Freitas, 2014). Finally, we empirically show that our EI (using $y^{\max}$) is more efficient in computational time and accuracy than the $\mu^{\max}$ counterpart in (Wang and de Freitas, 2014) because our EI version gets rid of the additional optimization step of estimating $\mu^{\max}$.

Algorithm 1 Bayesian optimization using expected improvement (EI) with stopping condition.

Input: Max iteration T, a stopping criteria κ (a small positive constant, e.g., 10^{-9})

1: Initialize the data $\mathscr{D}_0$
2: **for** $t = 1$ to T and $\alpha_{t-1}^{\mathrm{EI}}(x_{t-1}) = \max_{\forall x \in \mathscr{X}} \alpha_{t-1}^{\mathrm{EI}}(x) \geq \kappa$ **do**
3: Fit a GP to compute the predictive mean $\mu_{t-1}(.)$ and variance $\sigma_{t-1}(.)$ from the data $\mathscr{D}_{t-1}$.
4: Original setting (Mockus et al., 1978) using the best-observed value: $\xi = \max y_i, \forall y_i \in \mathscr{D}_{t-1}$. This step is obtained from $\mathscr{D}_{t-1}$ without performing optimization.
5: Alternative setting (Wang and de Freitas, 2014) using maximum GP predictive mean: $\xi = \arg\max_{x \in \mathscr{X}} \mu_{t-1}(x)$. This step requires to perform a global optimization in $\mathscr{R}^d$.
6: Define $\alpha_t^{\mathrm{EI}}(x) = \sigma_{t-1}(x)\,\phi(z) + [\mu_{t-1}(x) - \xi]\,\Phi(z)$ where $z(x) = \frac{\mu_{t-1}(x) - \xi}{\sigma_{t-1}(x)}$.
7: Optimize $x_t = \arg\max_{x \in \mathscr{X}} \alpha_t^{\mathrm{EI}}(x)$
8: Evaluate the function $y_t = f(x_t)$ and augment the data $\mathscr{D}_t = \mathscr{D}_{t-1} \cup (x_t, y_t)$
9: **end for**

Output: $x_{\max}, y_{\max}$

2. Bayesian Optimization

Our goal is solving the global optimization problem $x^* = \underset{x \in \mathscr{X}}{\operatorname{argmax}} f(x)$. Bayesian optimization is one approach to solve the above optimization problem by making a series of evaluations $x_1, ..., x_T$ of f such that the maximum of f is found in the fewest iterations (Shahriari et al., 2016). Bayesian optimization reasons about f by building a surrogate model through evaluations, typically a Gaussian process (Rasmussen, 2006). This flexible distribution allows us to associate a normally distributed random variable at every point in the continuous input space. Formally, a GP is given by $f(x) \sim GP(m(x), k(x, x'))$, where $m(x)$ is the mean, and $k(\mathbf{x}, \mathbf{x}')$ contains the covariance of any two observations. A popular choice for the covariance function is the squared exponential function: $k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left[\frac{-(\mathbf{x}-\mathbf{x}')^2}{2l^2}\right]$ where l is the length scale and σ_f^2 is the output variance. The length scale defines the "region of influence" of a point within the parameter space that the influence of an observation decreases as one considers points farther away from this observation. We get the predictive distribution for a new observation $p(f_{N+1} \mid X_{1:N}, \mathbf{y}_{1:N}, x_{N+1})$ that also follows a Gaussian distribution (Rasmussen, 2006) - its mean and variance are given by:

$$\mu(x_{N+1}) = \mathbf{k}_* \mathbf{K}^{-1} \mathbf{y} \qquad \sigma^2(x_{N+1}) = k_{**} - \mathbf{k}_* \mathbf{K}^{-1} \mathbf{k}_*^T$$

where the covariance matrices are defined as the following $\mathbf{k}_* = [k(x_1, x^*), k(x_2, x^*), ..., k(x_N, x^*)]$, $\mathbf{K} = [k(x_i, x_j)]_{\forall x_i, x_j \in \mathscr{D}_t}$ and $k_{**} = k(x_{N+1}, x_{N+1})$.

Acquisition function. As the original function is expensive to evaluate, the acquisition function acts as a surrogate that determines which point should be selected next. Therefore, instead of maximizing the original function f, we maximize the acquisition function to select the next point $x_{t+1} = \arg\max_{x \in \mathscr{X}} \alpha_t(x)$. In particular, the acquisition function takes into account the mean and variance of the GP predictions. The decision represents an automatic trade-off between exploration (where the objective function is very uncertain) and exploitation (where the objective function is expected to be high). This exploration-exploitation trade-off has the nice property that it aims to

minimize the number of objective function evaluations. Moreover, it is likely to do well even in settings where the objective function has multiple local maxima (Brochu et al., 2010). Unlike the original objective function $f(x)$, the acquisition function $\alpha(x)$ can be cheaply sampled. Therefore, we can utilize some standard optimization packages.

3. Expected Improvement for Bayesian Optimization

Although many acquisition functions have been proposed, the expected improvement (EI) is considered as one of the most popular acquisition function for Bayesian optimization and remains the default choice in BO packages, such as Spearmint (Snoek et al., 2012). In particular, EI considers the expectation over the improvement function. The improvement function is defined over the *incumbent* ξ as $I_t(x) = \max\{0, f(x) - \xi\}$ where $\xi = y_{t-1}^{\max} = \max_{y_i \in \mathscr{D}_{t-1}} y_i$ is the maximum in the observation set.

Let us denote by $z = z_{t-1}(x) = \frac{\mu_{t-1}(x) - y_{t-1}^{\max}}{\sigma_{t-1}(x)}$, we obtain the closed-form acquisition function by taking the expectation over the improvement function as $\mathbb{E}[I_t(x)]$ (refer to Appendix A for the derivation). Thus, this strategy is called expected improvement,

$$\alpha_t^{\mathrm{EI}}(x) = \mathbb{E}[I_t(x)] = \sigma_{t-1}(x)\,\phi(z) + \left[\mu_{t-1}(x) - y_{t-1}^{\max}\right]\Phi(z) \tag{1}$$

where ϕ is the standard normal p.d.f. and Φ is the c.d.f. When $\sigma_t(x) = 0$, we set $\alpha^{\mathrm{EI}}(x) = 0$.

We define the function $z_{t-1}(x) = \frac{\mu_{t-1}(x) - y_{t-1}^{\max}}{\sigma_{t-1}(x)}$ and the function $\tau(z) = z\Phi(z) + \phi(z)$ where Φ and ϕ are the c.d.f. and the p.d.f. of the standard normal distribution.

Lemma 1 *The acquisition function of EI can be expressed as* $\alpha_t^{EI}(x) = \sigma_{t-1}(x)\,\tau(z_{t-1}(x))$ *and* $\alpha_t^{EI}(x) \leq \tau(z_{t-1}(x))$ *where* $\tau(z) = z\Phi(z) + \phi(z)$ *with* Φ *and* ϕ *are the c.d.f. and the p.d.f. of the standard normal distribution.*

Proof Given $z_{t-1}(x) = \frac{\mu_{t-1}(x) - y_{t-1}^{\max}}{\sigma_{t-1}(x)}$, from Eq. (1) we have $\alpha_t^{\mathrm{EI}}(x) = \sigma_{t-1}(x)\,[z \times \Phi(z) + \phi(z)] = \sigma_{t-1}(x)\,\tau(z_{t-1}(x))$. In addition, $\sigma_{t-1}(x) \leq 1$, then $\alpha_t^{\mathrm{EI}}(x) \leq \tau(z_{t-1}(x))$. ■

3.1. Existing Convergence Analysis for Expected Improvement

The convergent property of EI has been studied in recent work. (Vazquez and Bect, 2010) provides the convergent property of EI by assuming some mild assumptions on the mean and covariance functions that the EI strategy produces a dense sequence of evaluation points in the search domain. However, these assumptions may not always hold. (Ryzhov, 2016) characterizes the asymptotic rates of EI for a finite decision space (e.g., the case of multi-arm bandit). The work of (Bull, 2011) considers a global optimization problem with a continuous, d-dimensional decision space, and derives a convergence rate of $O\left(T^{-\frac{1}{d}}\right)$. However, this work assumes zero noise for the outcomes and may not be applicable to practical scenario where observations always have measurement noises. Recently, (Wang and de Freitas, 2014) extends the work of (Bull, 2011) to stochastic objective functions with noise. However, due to the difficulty in stochastic setting and to make the proof feasible, instead of choosing $\xi = y^{\max}$, they use the alternative choice of the incumbent as $\xi = \mu^{\max} = \max_{x \in \mathscr{X}} \mu_{t-1}(x)$ to derive the convergence bound. Due to this modification, the Bayesian optimization algorithm needs to use an additional global optimization step to estimate $\mu^{\max}$ in $\mathbb{R}^d$. As a result, the computational cost of the whole BO process (unnecessarily) increases. In addition, the estimation of global

optimization is not always accurate, especially in high dimensions. This computational cost and inaccurate incumbent $\mu^{\max}$ can degrade the efficiency of EI strategy and the Bayesian optimization in general. Moreover, none of the existing work has considered the stopping criteria for EI.

3.2. Stopping Criteria for Expected Improvement

In Bayesian optimization, although the theoretical analysis assumes the number of iteration goes to infinity $T \to \infty$, we always run a finite number of iterations in practice. Thus, there should be a stopping point where to terminate the search, such as when there is no more promising locations to visit. Such stopping criteria is essential because each evaluation in Bayesian optimization comes at a cost. Despite of its importance, this stopping criteria is not well studied in the literature for Bayesian optimization. In this section, we consider the stopping criteria for Expected improvement strategy. This stopping condition is later integrated into our derivation of the cumulative regret.

The acquisition function value at the selected point $\alpha^{\text{EI}}(x_t) = \max_{x \in \mathscr{X}} \alpha(x)$ encodes the maximum improvement that the BO can make. Because EI is a function of the variance and the improvement quantity (encoded by σ and τ in Lem. 1), the value of EI tends to decrease w.r.t. iterations (cf. Fig. 1 for the trend). Therefore, we can stop the algorithm to prevent from unnecessary evaluations when the maximum amount of improvement is smaller than a threshold κ.

Lemma 2 *The value of the EI acquisition function at the selected point should be positive for a valid optimization, i.e.* $\forall x_t \in \mathscr{D}_t, \alpha_t^{EI}(x_t) \geq \kappa > 0$ *where* κ *is a small positive constant. If this condition is violated, the optimization should be stopped.*

Proof Using Lem. 1 and the fact that $\sigma(x) \geq 0$ and $\tau(.) \geq 0$, we have $\forall x \in \mathscr{X}, \alpha^{\text{EI}}(x) \geq 0$. In addition, we select a point by maximizing the acquisition function as $x_t = \underset{x \in \mathscr{X}}{\text{argmax}} \alpha_t^{\text{EI}}(x)$. Thus, the value of the selected point should be positive $\alpha_t^{\text{EI}}(x_t) > 0$. As a result, there exists a small positive constant κ such that $\forall x_t \in \mathscr{D}_t, \alpha_t^{\text{EI}}(x_t) \geq \kappa > 0$. ■

Lem. 2 prevents BO from selecting the sub-optimal locations where the maximum amount of improvement is less than a user-defined threshold $\alpha^{\text{EI}}(x_t) < \kappa$. Based on this condition, we also have the predictive variance at the selected points positive $\forall x_t \in \mathscr{D}_t, \sigma_{t-1}(x_t) \geq \frac{\kappa}{\tau(.)} > 0$ which means that the selected location is informative to explore. We note that this stopping criteria is specific for the EI that may not directly be applicable to other acquisition functions.

4. Convergence Rate for EI over the Best-Observed Value with Stopping Condition

Our general goal in Bayesian optimization is to maximize the rewards. Equivalently, we could also minimize the (cumulative) regret. Similar to (Srinivas et al., 2010; Wang and de Freitas, 2014), we use the regret to measure the convergence. We consider the case with noisy output, i.e., $y_t = f(x_t) + \varepsilon_t$ and assume that the noise process ε_t is sub-Gaussian, and that the black-box function f is smooth according to the reproducing kernel Hilbert space (RKHS) associated with a GP kernel $k()$. Let the global maximum point be $x^* = \underset{x \in \mathscr{X}}{\text{argmax}} f(x)$ and our choice at iteration t be x_t. The cumulative regret R_T after iteration T is the sum of the instantaneous regrets: $R_T = \sum_{t=1}^{T} r_t$ where $r_t = f(x^*) - f(x_t)$. We are going to derive that R_T grows almost at a sublinear rate, i.e. $\lim_{T \to \infty} \frac{R_T}{T} = 0$.

We follow (Srinivas et al., 2010) to define the maximum information gain in the following.

Definition 3 *Given $A = \{x_1, \cdots, x_T\} \subset X$, let $f_A = \{f(x_i)\}$, $y_A = f_A + \varepsilon_A$ and I be the mutual information. The maximum information gain after T iterations is defined as $\gamma_T = \max_{A \in X, |A|=T} I(y_A; f_A)$.*

We now present the main theorem of the paper that provides the sublinear cumulative regret for EI using best-found value $y^{\max}$ as the incumbent.

Theorem 4 *Let $\kappa > 0$ be a pre-defined small constant as a stopping criteria, $\gamma_T \sim \mathcal{O}\left((\log T)^{d+1}\right)$ be the maximum information gain for the squared exponential kernel, σ^2 be the measurement noise variance, $C \triangleq \log\left[\frac{1}{2\pi\kappa^2}\right]$, $\beta_T = 2||f||_k^2 + 300\gamma_T \log^3\left(\frac{T}{\delta}\right)$ and $\delta \in (0,1)$. Then with probability at least $1-\delta$, the cumulative regret of EI, using best-found value $y^{\max} = \max_{y_i \in \mathscr{D}_t} y_i$ as the incumbent, obeys the following sublinear rate $R_T \lesssim \sqrt{T\beta_T\gamma_T} \sim \mathcal{O}(\sqrt{T \times (\log T)^{d+4}}$.*

For clarity in the notation, we denote the noise variance as σ^2 s.t. $\varepsilon_t \sim \mathcal{N}\left(0, \sigma^2\right)$ and the predictive variance as $\sigma_t^2(x)$. Without loss of generality, we assume $k(x,x) = 1$.

Lemma 5 *(Theorem 6 of (Srinivas et al., 2010)) Let $\delta \in (0,1)$ and assume that the noise variables ε_t are uniformly bounded by σ. Define $\beta_t = 2||f||_k^2 + 300\gamma_t \ln^3\left(\frac{t}{\delta}\right)$, then*

$$p\left(\forall t, \forall x \in \mathscr{X}, |\mu_t(x) - f(x)| \leq \sqrt{\beta_t}\sigma_t(x)\right) \geq 1 - \delta$$

Lemma 6 *The improvement function $I_t(x) = \max\left\{0, f(x) - y_{t-1}^{max}\right\}$ and the acquisition function $\alpha_t^{EI}(x) = \mathbb{E}[I_t(x)]$ satisfy the inequality such that $I_t(x) - \sqrt{\beta_t}\sigma_{t-1}(x) \leq \alpha_t^{EI}(x)$.*

Proof If $\sigma_{t-1}(x) = 0$ then $\alpha(x) = I_t(x) = 0$ which makes the result trivial. We now assume that $\sigma_{t-1}(x) > 0$. We set $q = \frac{f(x) - y_{t-1}^{\max}}{\sigma_{t-1}(x)}$ and $z = \frac{\mu_{t-1}(x) - y_{t-1}^{\max}}{\sigma_{t-1}(x)}$. Using Lem. 1 and Lem. 5, we then express the acquisition function as follows

$$\begin{aligned}
\alpha_t^{\mathrm{EI}}(x) &\geq \sigma_{t-1}(x)\,\tau\left(q - \sqrt{\beta_t}\right) \\
&\geq \sigma_{t-1}(x)\left(q - \sqrt{\beta_t}\right) \qquad\qquad \text{by } \tau(z) \geq z
\end{aligned}$$

If $I_t(x) = 0$, the result is trivial. Thus, we can assume $I_t(x) > 0$ and conclude the proof

$$\alpha_t^{\mathrm{EI}}(x) \geq I_t(x) - \sqrt{\beta_t}\sigma_{t-1}(x).$$

■

Using the results of (Srinivas et al., 2010), we have the maximum information gain for common kernels, such as $\gamma_T \in \mathcal{O}\left((\log T)^{d+1}\right)$ for the SE kernel, $\gamma_T \in \mathcal{O}\left((\log T)^{d}\right)$ for linear kernel. Lemma 5.4 of (Srinivas et al., 2010) has provided the bound of the variance of the selected points with γ_T. However, we can generalize for any arbitrary set of points (not just for selected points x_t) because the maximum information gain γ_T quantifies the maximum possible information gain achievable by sampling T points in a GP with kernel function $k(.)$ (Srinivas et al., 2010; Krause and Ong, 2011). Therefore, we have the following lemma.

Lemma 7 *The sum of the predictive variances is bounded by the maximum information gain γ_T. That is $\forall x \in \mathscr{X}, \sum_{t=1}^{T} \sigma_{t-1}^2(x) \leq \frac{2}{\log(1+\sigma^{-2})}\gamma_T$.*

Proof Using the fundamental inequality of logarithm $\frac{x}{\log(1+x)} \geq 1$, we have $\frac{1}{\sigma^2\log(1+\sigma^{-2})} \geq 1$. Since $\frac{s^2}{\log(1+s^2)} \leq \frac{1}{\sigma^2\log(1+\sigma^{-2})}$ for $s^2 \in \left[0, \sigma^{-2}\right]$ and $\sigma^{-2}\sigma_{t-1}^2(x) \leq \sigma^{-2}$ since $\sigma_{t-1}^2(x) \leq k(x,,x) = 1$.

$$\begin{aligned}
\forall x \in \mathscr{X}, \sum_{t=1}^{T} \sigma_{t-1}^2(x) &= \sum_{t=1}^{T} \sigma^2 \underbrace{\sigma^{-2}\sigma_{t-1}^2(x)}_{s^2} \leq \sum_{t=1}^{T} \sigma^2 \left[\frac{\log\left(1+s^2\right)}{\sigma^2\log\left(1+\sigma^{-2}\right)}\right] \\
&= \frac{2}{\log\left(1+\sigma^{-2}\right)} \frac{1}{2} \sum_{t=1}^{T} \log\left(1+\sigma^{-2}\sigma_{t-1}^2(x)\right) \\
&= \frac{2}{\log\left(1+\sigma^{-2}\right)} I(f; y_A) \leq \frac{2}{\log\left(1+\sigma^{-2}\right)} \gamma_T
\end{aligned}$$

where the last equation utilizes the property that $I(f; y_A) = \frac{1}{2}\sum_{t=1}^{T} \log\left[1+\sigma^{-2}\sigma_{t-1}(x)\right]$.

■

Lemma 8 *Let $\kappa > 0$ be a pre-defined stopping threshold on the acquisition function $\alpha_{t-1}^{EI}(x)$, if $y_{t-1}^{max} - \mu_{t-1}(x_t) > 0$, we have $y_{t-1}^{max} - \mu_{t-1}(x_t) \leq \sigma_{t-1}(x_t)\sqrt{C}$ where $C \triangleq \log\left[\frac{1}{2\pi\kappa^2}\right]$.*

Proof Let us define a constant $C \triangleq \log\left[\frac{1}{2\pi\kappa^2}\right] \geq 0$. By using Lem. 2, we write

$$\kappa \leq \sigma_{t-1}(x_t)\,\tau\left(z_{t-1}(x_t)\right). \tag{2}$$

Using the lemma assumption of $\mu_{t-1}(x_t) - y_{t-1}^{\max} \leq 0$, we get $\tau\left(z_{t-1}(x_t)\right) \leq \phi\left(z_{t-1}(x_t)\right)$ by utilizing the property of the τ function that $\tau(z) \leq \phi(z), \forall z < 0$. It means

$$\begin{aligned}
\kappa &\leq \sigma_{t-1}(x_t) \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} z_{t-1}^2(x_t)\right] \\
\frac{\sqrt{2\pi}\kappa}{\sigma_{t-1}(x_t)} &\leq \exp\left[-\frac{1}{2} z_{t-1}^2(x_t)\right].
\end{aligned}$$

Taking the logarithm both sides, we obtain $z_{t-1}^2(x_t) \leq 2\log\frac{\sigma_{t-1}(x_t)}{\sqrt{2\pi}\kappa}$ and thus

$$0 \leq z_{t-1}^2(x_t) = \left[\frac{\mu_{t-1}(x_t) - y_{t-1}^{\max}}{\sigma_{t-1}(x_t)}\right]^2 \leq \log\left[\frac{\sigma_{t-1}^2(x_t)}{2\pi\kappa^2}\right] \leq \log\left[\frac{1}{2\pi\kappa^2}\right].$$

In addition, using the lemma condition that $y_{t-1}^{\max} - \mu_{t-1}(x_t) > 0$, we conclude the proof

$$y_{t-1}^{\max} - \mu_{t-1}(x_t) \leq \sigma_{t-1}(x_t)\sqrt{C}.$$

■

We next bound the function $\tau(-z(x))$ which is later used to prove the main theorem.

Lemma 9 *Let $\kappa > 0$ be a pre-defined stopping criteria, $z_{t-1}(x) = \frac{\mu_{t-1}(x) - y_{t-1}^{max}}{\sigma_{t-1}(x)}$ and $\tau(z) = z\Phi(z) + \phi(z)$, we have $\tau\left(-z_{t-1}(x_t)\right) \leq 1 + \sqrt{C}$ where $C \triangleq \log\left[\frac{1}{2\pi\kappa^2}\right]$.*

Proof The property of $\tau(z)$ depends on the sign of z. Thus, we consider two cases of $y_{t-1}^{\max} - \mu_{t-1}(x_t) > 0$ and $y_{t-1}^{\max} - \mu_{t-1}(x_t) \leq 0$, respectively. We denote a constant $C \triangleq \log\left[\frac{1}{2\pi\kappa^2}\right]$.

Case 1: For the first case we assume $y_{t-1}^{\max} - \mu_{t-1}(x_t) > 0$. Thus, we can utilize the property of $\tau(z) \leq 1 + z, \forall z \geq 0$ and write

$$\tau(-z_{t-1}(x_t)) \leq 1 + \frac{y_{t-1}^{\max} - \mu_{t-1}(x_t)}{\sigma_{t-1}(x_t)} \leq 1 + \sqrt{C} \qquad \text{by Lem. 8}$$

Case 2: For the second case we assume $y_{t-1}^{\max} - \mu_{t-1}(x_t) \leq 0$. We utilize $\forall z \leq 0, \tau(z) \leq \phi(z) \leq 1$,

$$\tau(-z_{t-1}(x_t)) \leq \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} z_{t-1}^2(x_t)\right] \leq 1.$$

Clearly, for both cases, we have $\tau(-z_{t-1}(x_t)) \leq 1 + \sqrt{C}$. ■

We now prove the main Theorem 4 that provides the sublinear cumulative regret for EI using best-found value $y^{\max}$ as the incumbent.

Proof Let $x_t = \underset{x \in X}{\operatorname{argmax}}\ \alpha_t^{\mathrm{EI}}(x)$ be the choice at iteration t, the instantaneous regret is defined as:

$$r_t = f(x^*) - f(x_t) = \underbrace{f(x^*) - y_{t-1}^{\max}}_{A_t} + \underbrace{y_{t-1}^{\max} - f(x_t)}_{B_t}. \tag{3}$$

We bound r_t with the GP posterior variance so that we later connect it to the maximum information gain γ_T. From the definition of x_t and Lem. 1, we have $\alpha_t^{\mathrm{EI}}(x^*) \leq \alpha_t^{\mathrm{EI}}(x_t) = \sigma_{t-1}(x_t)\,\tau(z_{t-1}(x_t))$. Then, by using Lem. 6 we write

$$\begin{aligned} A_t &\leq \alpha^{\mathrm{EI}}(x^*) + \sqrt{\beta_t}\sigma_{t-1}(x^*) \leq \alpha^{\mathrm{EI}}(x_t) + \sqrt{\beta_t}\sigma_{t-1}(x^*) \\ &= \sigma_{t-1}(x_t)\,\tau(z_{t-1}(x_t)) + \sqrt{\beta_t}\sigma_{t-1}(x^*) \qquad \text{by Lem. 5} \end{aligned}$$

Next, we express the second term in Eq. (3) as follows

$$\begin{aligned} B_t =& y_{t-1}^{\max} - \mu_{t-1}(x_t) + \mu_{t-1}(x_t) - f(x_t) \\ \leq& \sigma_{t-1}(x_t)(-z_{t-1}(x_t)) + \sigma_{t-1}(x_t)\sqrt{\beta_t} \qquad \text{by Lem. 5} \\ =& \sigma_{t-1}(x_t)\left[\tau(-z_{t-1}(x_t)) + \sqrt{\beta_t} - \tau(z_{t-1}(x_t))\right] \qquad \text{by } z = \tau(z) - \tau(-z) \end{aligned}$$

Continuing from Eq. (3), we have $r_t = A_t + B_t$ that is

$$r_t \leq \sigma_{t-1}(x_t)\left[\sqrt{\beta_t} + \tau(-z_{t-1}(x_t))\right] + \sqrt{\beta_t}\sigma_{t-1}(x^*) \tag{4}$$

Using the bound of $\tau(-z_{t-1}(x_t))$ in Lem. 9, we obtain

$$r_t \leq \underbrace{\sigma_{t-1}(x_t)\left[\sqrt{\beta_t} + 1 + \sqrt{C}\right]}_{L_t} + \underbrace{\sqrt{\beta_t}\sigma_{t-1}(x^*)}_{U_t}$$

where $C \triangleq \log\left[\frac{1}{2\pi\kappa^2}\right]$. We then simplify L_t and U_t, respectively. Taking the sum of squared regret and utilizing the Cauchy-Schwartz inequality that $(a+b+c)^2 \leq 3\left(a^2+b^2+c^2\right)$, we have

$$\begin{aligned}\sum_{t=1}^{T} L_t^2 &\leq \sum_{t}^{T} \sigma_{t-1}^2(x_t)\, 3\left(\beta_t+1+C\right) \leq 3\left(\beta_T+1+C\right)\sum_{t}^{T}\sigma_{t-1}^2(x_t) && \text{by } \beta_T \geq \beta_t, \forall t \leq T \\ &\leq \frac{6\left(\beta_T+1+C_T\right)\gamma_T}{\log\left(1+\sigma^{-2}\right)} && \text{by Lem. 7}\end{aligned}$$

Again, using the Cauchy-Schwartz inequality, we obtain

$$\sum_{t=1}^{T} L_t \leq \sqrt{T}\sqrt{\sum_{t=1}^{T} L_t^2} \leq \sqrt{\frac{6T\left(\beta_T+1+C\right)\gamma_T}{\log\left(1+\sigma^{-2}\right)}}.$$

We further utilize Lem. 7 and the Cauchy-Schwartz again to simplify U_t that

$$\sum_{t=1}^{T} U_t \leq \beta_T \sum_{t=1}^{T} \sigma_{t-1}\left(x^*\right) \leq \sqrt{\frac{2T\beta_T\gamma_T}{\log\left(1+\sigma^{-2}\right)}}.$$

Finally, we get the cumulative regret $R_T \leq \sum_{t=1}^{T}\left(L_t+U_t\right)$,

$$R_T \leq \sqrt{\frac{2T\gamma_T}{\log\left(1+\sigma^{-2}\right)}}\left[\sqrt{3\left(\beta_T+1+C\right)}+\sqrt{\beta_T}\right]$$

where $C \triangleq \log\left[\frac{1}{2\pi\kappa^2}\right]$, $\kappa > 0$ is a (constant) pre-defined stopping criteria, σ^2 is the measurement noise variance, and β_T is also in the form of $\mathcal{O}(\log T)^3$. The bound of γ_T is kernel specific. To have concrete regret bound, we consider the squared exponential kernel $\gamma_T \sim \mathcal{O}\left((\log T)^{d+1}\right)$. Therefore, we can write a sublinear rate $R_T \sim \mathcal{O}\left(\sqrt{T \times (\log T)^{d+4}}\right)$ which vanishes in the limit as $\lim_{T\to\infty}\frac{R_T}{T} = 0$. ■

As the stopping criteria κ decreases, it allows more evaluations with low improvement values, thus the cumulative regret R_T increases. In contrast, if κ increases, we will not take evaluations with low improvement, then R_T decreases. We note that our convergence rate is analogous to the sublinear regret rate for GP-UCB algorithm (Srinivas et al., 2010) and EI using maximum predictive GP mean derived by (Wang and de Freitas, 2014) which is $\lim_{T\to\infty}\frac{R_T}{T} = 0$. More significantly, we are the first to incorporate the stopping criteria into the regret analysis.

5. Experiments

We compare two variants of EI using the incumbents (1) as the maximum of GP mean function $\mu_t^{\max} = \max_{x\in\mathcal{X}}\mu_{t-1}(x)$ (Wang and de Freitas, 2014) and (2) as the maximum on the observation set $y_t^{\max} = \max_{y_i\in\mathcal{D}_t} y_i$ (Mockus et al., 1978; Jones et al., 1998) as in our analysis.

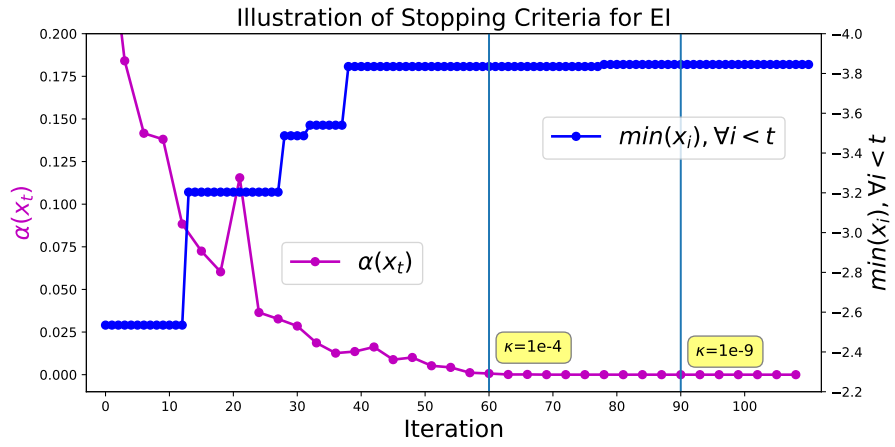

Figure 1: Example of stopping criteria using Hartmann 3D function. The acquisition function value at the selected point $\alpha^{\text{EI}}(x_t) = \max_{x \in \mathcal{X}} \alpha(x)$ encodes the maximum improvement which tends to decrease w.r.t iterations (see the magetan curve). Therefore, we can stop the algorithm to prevent from unnecessary evaluations when the amount of improvement is smaller than a pre-specified threshold κ. We consider $\kappa = 10^{-4}$ (stop at iteration 60) and $\kappa = 10^{-9}$ (stop at iteration 90). We note that the BO will not find any significantly better value after $t = 40$ (see the blue curve). Thus, if BO continues searching, it may waste time and resource for little gain.

Experimental setting. Throughout the experiment, we use GP priors with the squared exponential kernels $k(x,x') = \exp(-\frac{||x-x'||^2}{l})$ where l is set to the dimension size d as the default setting of RBF kernel in LibSVM. In addition, as a common practice in implementation, we scale the feature $x \in [0,1]^d$ so that each dimension of the feature vector is treated equally. Further, the output is standardized $y \sim \mathcal{N}(0,1)$. The results are averaged over 20 independent runs with different initializations. All implementations are in Python. The performance of the algorithms is compared for a fixed number of iterations $T = 10d$ and the initialization point $n_0 = 3d$. The stopping criteria is set small as $\kappa = 10^{-9}$. The UCB parameter is set as $\sqrt{\beta_t} = 2$ as used in (Nguyen et al., 2016b). We optimize the acquisition function using L-BFGS-B algorithm (multi-start).

5.1. Illustration of Stopping Criteria in EI

We illustrate the behavior of Bayesian optimization using stopping criteria to prevent from unnecessary evaluations in Fig. 1. These unnecessary evaluations can happen after the function is well-learned (or densely covered) by Bayesian optimization and there is no promising location to explore. In this situation, Bayesian optimization should be stopped for saving cost and resource. However, without a stopping mechanism, Bayesian optimization continues running and exploiting until the maximum evaluation budget T is reached.

Because EI is a function of the variance and the improvement quantity (encoded by σ and τ in Lem. 1), the value of EI tends to decrease w.r.t iterations (see Fig. 1). We consider two cases of the stopping criteria $\kappa = 10^{-4}$ and $\kappa = 10^{-9}$ when the amount of improvement is smaller than a

Func	Hartmann	Ackley	Hartmann	Alpine2	gSobol	gSobol
Dim	3	5	6	10	10	12
Stopping criteria is active	yes	yes	no	no	no	no
EI $\mu_t^{\max}$	**-3.57±.2**	13.299±1.0	-2.87±.04	-519±353	2k±1k	3k±2k
EI $y_t^{\max}$	-3.46±.3	**9.754±3.2**	**-2.93±.05**	**-922±608**	**367±273**	**2k±1k**

Table 1: Performance comparison using Best-Found-Value ($\max_{\forall t \leq T} y_t$) on the benchmark minimization problems between two variants of EI. Our setting using $y_t^{\max}$ outperforms the counterpart using $\mu_t^{\max}$ in accuracy. We use $T = 10d$ and $\kappa = 10^{-9}$. Stopping criteria row indicates if the BO algorithm is terminated before reaching T due to violating the condition of $\alpha^{\text{EI}}(x_t) \leq \kappa$.

threshold κ. These two settings will force the algorithm to stop at iteration 60 and 90, respectively. This termination is essential for efficiency purpose (saving time and resource) and to ensure that BO will not over-exploit the function. We note that the choice of kernel parameter will affect the value of the acquisition function. However, we will not investigate it in the current paper since our primary focus is on the convergent analysis for EI over the best-observed value.

5.2. Comparison on Benchmark Functions

We assess the performance of the Bayesian optimization using EI in finding the optimum of the chosen benchmark functions in a range of dimensions D from 3 to 12. We empirically show that our setting consistently performs better than the setting of $\mu_t^{\max}$ (Wang and de Freitas, 2014) both in computational complexity and accuracy. The performance is in line with our intuition about taking another global optimization step for $\mu_t^{\max}$. This is because estimating $\mu_t^{\max}$ in $\mathbb{R}^d$ is expensive and sometimes inaccurate while finding $y_t^{\max}$ from the observed set of $\{y_1 ... y_t\} \in \mathscr{D}_t$ is much cheaper and accurate. Therefore, the error in estimating $\mu_t^{\max}$ at some iterations can result in misleading acquisition function α_t^{EI} and be inefficient for the whole process. We present the quantitative results in Table 1 for optimizing the minimization problems.

We also empirically observe that the stopping criteria is only active for low dimensional functions (e.g., $d \leq 5$). This is because the search space goes exponentially large with the dimension. Thus, it requires a very large number of evaluations to cover the high dimensional space. As a result, BO may take more iterations (than the currently used $10d$ iterations) so that the stopping criteria is violated.

5.3. Computational Time

We study the computational time spent per iteration w.r.t. increasing dimensions from 5 to 12. For fair comparison, all simulations are done using the same Windows machine Core i7, Ram 24GB. We learn that the extra step for estimating $\mu^{\max}$ occurs with an additional computation than the original counterpart of EI (with $y^{\max}$). As a result, the CPU time per iteration of EI using $\mu^{\max}$ (Wang and de Freitas, 2014) will be more expensive than our EI using $y^{\max}$. We note that the stopping criteria will not affect the computational time per iteration for a fair comparison. We present the numerical comparison in Fig. 2.

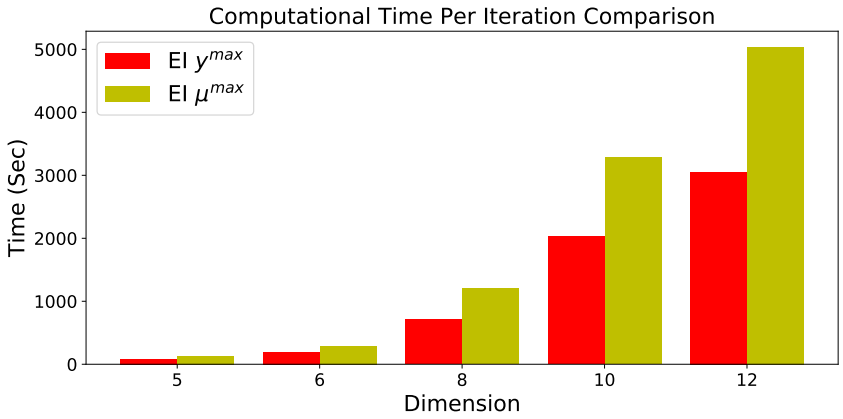

Figure 2: EI using $\mu^{\max}$ takes an additional optimization step of estimating $\mu^{\max}$. On the other hand, EI using $y^{\max}$ does not require the optimization step to find the incumbent. Thus, EI $y^{\max}$ is computationally more efficient with sublinear theoretical guarantee on convergence.

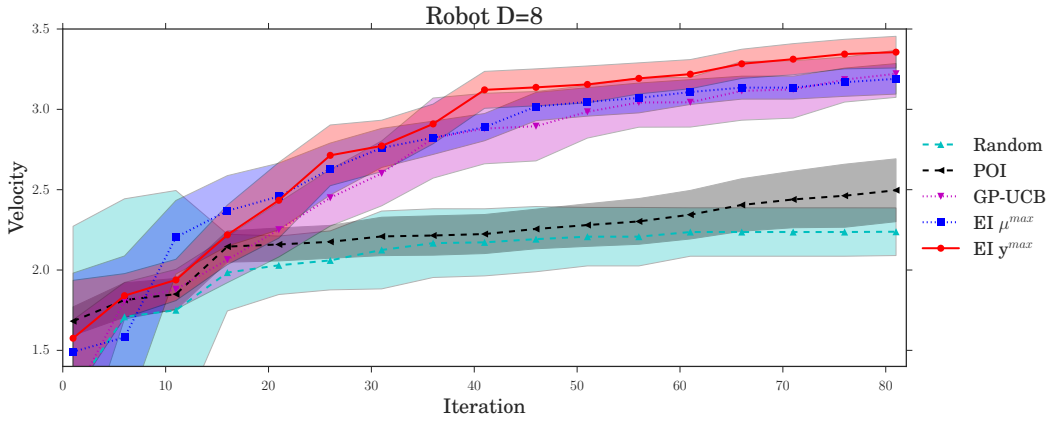

Figure 3: Configuration tuning for robot design. We maximize the horizontal velocity.

5.4. Real-world Applications

We consider tuning configurations for robot control and hyper-parameters for machine learning algorithm. We again aim to highlight that EI strategy using the original $y^{\max}$ is more robust and computational efficient than EI using $\mu^{\max}$ (Wang and de Freitas, 2014). In this experiment, the stopping criteria $\kappa = 10^{-9}$ is not violated due to the relatively high dimensional functions of $d = 6$ and $d = 8$, respectively. Thus, the algorithm will run until the maximum budget $T = 10d$ is reached.

ROBOT CONTROL CONFIGURATION TUNING

One of the key challenges in robotic bipedal locomotion is finding gait parameters that optimize a desired performance metric, such as speed. Typically, gait optimization requires extensive robot experiments and specific expert knowledge. Instead, in this setting, we utilize Bayesian optimization

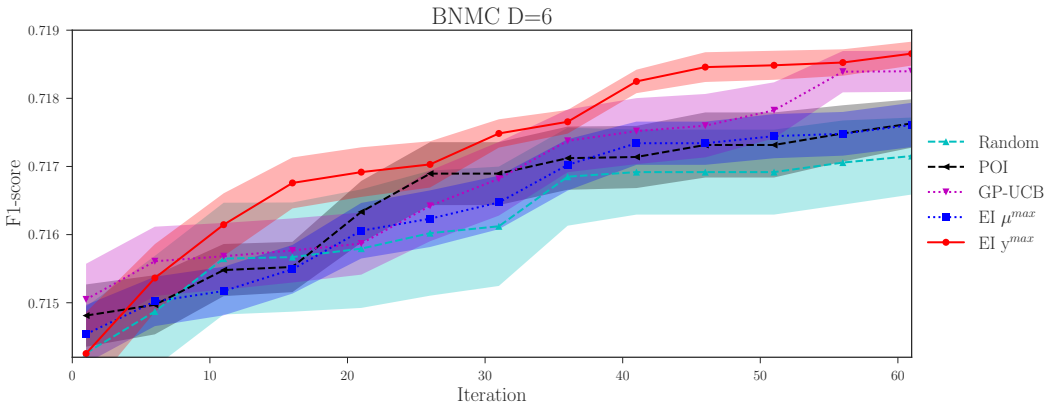

Figure 4: Machine learning hyper-parameter tuning.

to search for the best configuration to speed up the process of gait optimization. In particular, we consider the Walker, an eight-dimensional control problem where the inputs are fed into a simulator which returns the horizontal velocity of a bipedal robot (Westervelt et al., 2007). We use the released Matlab source code available at [1] and convert the algorithm to a black-box function for optimization.

MACHINE LEARNING HYPER-PARAMETER TUNING

We select to tune the hyper-parameters for the multi-label classification machine learning algorithm of BNMC (Nguyen et al., 2016a)[2]. The main advantage of BNMC is that it maintains high accuracy while training in a fraction of the time compared to the previous state-of-the-art. The BNMC algorithm has 6 parameters and the performance depends on these hyper-parameters to a reasonable amount. In particular, our task is to optimize these 6 hyper-parameters: Dirichlet symmetric for feature and for label, learning rate for SVI and for SGD, truncation threshold and stick-breaking parameter. We aim to maximize the F1 score. While the authors of (Nguyen et al., 2016a) run extensive tests on a variety of datasets, we pick a Scene dataset for our experiment.

RESULTS

We compare the performances of two EI variants using $y^{\max}$ and using $\mu^{\max}$ on real-word tasks. We also include the performance of *Random*, POI and GP-UCB in Fig. 4. We see that EI and GP-UCB performs generally better than POI because POI tends to exploit quite aggressively. In addition, the EI using $y^{\max}$ is more stable and performs better than the $\mu^{\max}$ counterpart. Although the performance gap between two EI variants is marginal, the original version of EI using $y^{\max}$ is advantageous since we do not need an additional optimization step for estimating $\mu^{\max}$.

6. Conclusion

We have derived the sublinear convergence rate for the expected improvement using the best-observed value as the incumbent. The previous analyses of expected improvement either assume

1. http://web.eecs.umich.edu/∼grizzle/biped_book_web/
2. https://github.com/ntienvu/ACML2016_BNMC

zero noise of the observations or using the maximum of the Gaussian process predictive mean function as the incumbent. The maximum of the GP predictive mean is more expensive to compute and is rarely used in practice. For the first time, we take the original EI and prove the sublinear regret under noise setting and stopping condition. Our experiments on benchmark functions and real experiments indicate that the original setting of EI (using the best-observed value) performs relatively better than the maximum GP mean counterpart while being more efficient.

7. Acknowledgments

This research was partially funded by the Australian Government through the Australian Research Council (ARC) and the Telstra-Deakin Centre of Excellence in Big Data and Machine Learning. Prof Venkatesh is the recipient of an ARC Australian Laureate Fellowship (FL170100006).

Appendix A. Derivation of the Expected Improvement

Let $\mathscr{D}_t = \left\{x_i \in \mathscr{R}^d, y_i \in \mathscr{R}\right\}_{i=1}^t$ be the observation set including a feature x_i and an outcome y_i. We define the improvement function $I^{\mathrm{EI}}(x) = \max\{0, f(x) - y^+\}$ where $y^+ = \max_{y_i \in \mathscr{D}_t} y_i$. The likelihood of improvement $I^{\mathrm{EI}}(x)$ (for brevity, we write I) on a normal posterior distribution is as follows

$$p(I) = \frac{1}{\sqrt{2\pi}\sigma(x)} \exp\left(-\frac{1}{2}\frac{[\mu(x) - y^+ - I]^2}{\sigma^2(x)}\right). \tag{5}$$

The expected improvement (EI) is then defined as $\alpha^{\mathrm{EI}}(x) = \mathbb{E}\left[I^{\mathrm{EI}}(x)\right]$. Using the likelihood function in Eq. (5), we obtain

$$\alpha^{\mathrm{EI}}(x) = \int_0^\infty I \times \frac{1}{\sqrt{2\pi}\sigma(x)} \exp\left(-\frac{1}{2}\frac{[\mu(x) - y^+ - I]^2}{\sigma^2(x)}\right) dI.$$

Let $t = \frac{\mu(x) - y^+ - I}{\sigma(x)}$, then $I = -t \times \sigma(x) + \mu(x) - y^+$ and $dt = -\frac{1}{\sigma(x)} dI$. We write $\alpha^{\mathrm{EI}}(x)$ as

$$\begin{aligned}
\alpha^{\mathrm{EI}}(x) &= \int_{t=\frac{\mu(x)-y^+}{\sigma(x)}}^{-\infty} \left[-t \times \sigma(x) + \mu(x) - y^+\right] \frac{1}{\sqrt{2\pi}\sigma(x)} \exp(-\frac{1}{2}t^2) \times [-\sigma(x)]\, dt \\
&= \sigma(x) \int_{t=\frac{\mu(x)-y^+}{\sigma(x)}}^{-\infty} \frac{t}{\sqrt{2\pi}} \exp(-\frac{1}{2}t^2) dt + \left[\mu(x) - y^+\right] \int_{-\infty}^{t=\frac{\mu(x)-y^+}{\sigma(x)}} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}t^2) dt. \qquad (6)
\end{aligned}$$

Let denote $u = t^2 = \left[\frac{\mu(x) - y^+ - I}{\sigma(x)}\right]^2, du = 2z dz$, we compute the first term in Eq. (6) as follows

$$\begin{aligned}
\sigma(x) \int_{t=\frac{\mu(x)-y^+}{\sigma(x)}}^{-\infty} \frac{t}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt &= \frac{\sigma(x)}{\sqrt{2\pi}} \int_{u=t^2}^{-\infty} \exp\left(-\frac{1}{2}u\right) \frac{du}{2} \\
&= \frac{\sigma(x)}{\sqrt{2\pi}} \left[-\exp\left(-\frac{1}{2}\left[\frac{\mu(x) - y^+ - I}{\sigma(x)}\right]^2\right)\right]_{I=0}^{I=-\infty} \\
&= \sigma(x) \mathscr{N}\left(\frac{\mu(x) - y^+}{\sigma(x)} \mid 0, 1\right).
\end{aligned}$$

We compute the second term in Eq. (6) as

$$\left[\mu(x)-y^{+}\right]\int_{-\infty}^{t=\frac{\mu(x)-y^{+}}{\sigma(x)}}\frac{\exp\left(-\frac{1}{2}t^{2}\right)}{\sqrt{2\pi}}dt=\left[\mu(x)-y^{+}\right]\int_{\infty}^{0}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left[\frac{\mu(x)-y^{+}-I}{\sigma(x)}\right]^{2}\right)dz$$
$$=\left[\mu(x)-y^{+}\right]\Phi\left(\frac{\mu(x)-y^{+}}{\sigma(x)}\right).$$

Explicitly, denoting $z=\frac{\mu(x)-y^{+}}{\sigma(x)}$, we obtain the acquisition function as follows

$$\alpha^{\mathrm{EI}}(x)=\sigma(x)\,\phi(z)+\left[\mu(x)-y^{+}\right]\Phi(z) \tag{7}$$

where $\phi(z)=\mathcal{N}(z\mid 0,1)$ is the standard normal pdf and $\Phi(z)$ is the cdf.

References

Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

Adam D Bull. Convergence rates of efficient global optimization algorithms. *The Journal of Machine Learning Research*, 12:2879–2904, 2011.

Thanh Dai Nguyen, Sunil Gupta, Santu Rana, Vu Nguyen, Svetha Venkatesh, Kyle J Deane, and Paul G Sanders. Cascade Bayesian optimization. In *Australasian Joint Conference on Artificial Intelligence*, pages 268–280. Springer, 2016.

Peter I Frazier and Jialei Wang. Bayesian optimization for materials design. In *Information Science for Materials Discovery and Design*, pages 45–75. Springer, 2016.

Philipp Hennig and Christian J Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13:1809–1837, 2012.

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*, pages 918–926, 2014.

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.

Andreas Krause and Cheng S Ong. Contextual gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*, pages 2447–2455, 2011.

Cheng Li, Sunil Gupta, Santu Rana, Vu Nguyen, Svetha Venkatesh, and Alistair Shilton. High dimensional bayesian optimization using dropout. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 2096–2102, 2017.

Romy Lorenz, Ricardo P Monti, Ines R Violante, Aldo A Faisal, Christoforos Anagnostopoulos, Robert Leech, and Giovanni Montana. Stopping criteria for boosting automatic experimental design using real-time fmri with bayesian optimization. *arXiv preprint arXiv:1511.07827*, 2015.

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

Vu Nguyen, Sunil Gupta, Santu Rana, Cheng Li, and Svetha Venkatesh. A Bayesian nonparametric approach for multi-label classification. In *Proceedings of The 8th Asian Conference on Machine Learning*, pages 254–269, 2016a.

Vu Nguyen, Santu Rana, Sunil Gupta, Cheng Li, and Svetha Venkatesh. Budgeted batch Bayesian optimization. In *IEEE 16th International Conference on Data Mining (ICDM)*, pages 1107–1112. IEEE, 2016b.

Vu Nguyen, Sunil Gupta, Santu Rana, Cheng Li, and Svetha Venkatesh. Bayesian optimization in weakly specified search space. In *IEEE 17th International Conference on Data Mining (ICDM)*. IEEE, 2017.

Santu Rana, Cheng Li, Sunil Gupta, Vu Nguyen, and Svetha Venkatesh. High dimensional Bayesian optimization with elastic gaussian process. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 2883–2891, 2017.

Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

Ilya O Ryzhov. On the convergence rates of expected improvement methods. *Operations Research*, 2016.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1): 148–175, 2016.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

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 Machine Learning (ICML-10)*, pages 1015–1022, 2010.

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

Ziyu Wang and Nando de Freitas. Theoretical analysis of bayesian optimisation with unknown gaussian process hyper-parameters. *arXiv preprint arXiv:1406.7758*, 2014.

Eric R Westervelt, Jessy W Grizzle, Christine Chevallereau, Jun Ho Choi, and Benjamin Morris. *Feedback control of dynamic bipedal robot locomotion*, volume 28. CRC press, 2007.