A Corrected Expected Improvement Acquisition Function Under Noisy Observations

Han Zhou   HAN.ZHOU@ESAT.KULEUVEN.BE
Dept. ESAT, Center for Processing Speech and Images, KU Leuven, Belgium

Xingchen Ma   XGCHENMA@AMAZON.DE[*]
Amazon Web Services

Matthew B. Blaschko   MATTHEW.BLASCHKO@ESAT.KULEUVEN.BE
Dept. ESAT, Center for Processing Speech and Images, KU Leuven, Belgium

Abstract

Sequential maximization of expected improvement (EI) is one of the most widely used policies in Bayesian optimization because of its simplicity and ability to handle noisy observations. In particular, the improvement function often uses the best posterior mean as the best incumbent in noisy settings. However, the uncertainty associated with the incumbent solution is often neglected in many analytic EI-type methods: a closed-form acquisition function is derived in the noise-free setting, but then applied to the setting with noisy observations. To address this limitation, we propose a modification of EI that corrects its closed-form expression by incorporating the covariance information provided by the Gaussian Process (GP) model. This acquisition function specializes to the classical noise-free result (Jones et al., 1998; Mockus et al., 1978), and we argue should replace that formula in Bayesian optimization software packages, tutorials, and textbooks. This enhanced acquisition provides good generality for noisy and noiseless settings. We show that our method achieves a sublinear convergence rate on the cumulative regret bound under heteroscedastic observation noise. Our empirical results demonstrate that our proposed acquisition function can outperform EI in the presence of noisy observations on benchmark functions for black-box optimization, as well as on parameter search for neural network model compression[1].

Keywords: Sequential maximization, Bayesian optimization, Expected improvement

1. Introduction

Bayesian optimization (BO) is considered as an effective way to search for a global optimum sequentially, especially when optimizing complex black-box objective function $f(\cdot)$ under limited budgets, that is

$$x^* = \arg\max_{x \in \mathcal{X}} f(x) \tag{1}$$

where $\mathcal{X} \in \mathbb{R}^d$ represents the bounded input space. Its optimization procedure relies on a Gaussian Process (GP) model that allows us to relax the assumption of the objective functions, leading to its popularity in many important applications including experimental

---

[*] Work was done at KU Leuven prior to joining Amazon.

1. The source code is available at https://github.com/han678/correctedNoisyEI.

particle physics, material design, and hyper-parameter tuning for machine learning algorithms. Once we have the GP surrogate model, the sequential selection of BO can be made through a decision function, known as the acquisition function that relies on this model. Among existing acquisition functions, Expected Improvement (EI) is one of the most widely used as it preserves a good balance between exploration and exploitation. It is generally defined as the expectation of the improvement function at iteration t: $\mathbb{E}[I_t(x)]$ where the improvement function $I_t(x) = \max\{0, f(x) - \xi\}$ relies on the incumbent solution ξ. Let us denote the set of sampled points up to iteration $t-1$ as $\mathcal{D}_{t-1}$, then in the noiseless setting where $f(x)$ can be easily observed, the incumbent best at iteration t is thus given by $\xi = f(x_t^+)$ with the best point $x_t^+ = \arg\max_{i \leq t-1} f(x_i)$. However, in the framework of noisy observations i.e. $y_t = f(x_t) + \varepsilon_t$, the true value of the objective function is not exactly known due to the noise term ε_t on the observations. Several existing works consider a plug-in estimate as the incumbent best to tailor the improvement function for the noisy environment. Popular substitutes include the best *noisy* observation $\max_{i \leq t-1} y_i$ (Nguyen et al., 2017) and the best value of the GP predictive mean $\arg\max \mu_{t-1}(x_i)$ over the input space $\mathcal{X}$ (Wang and de Freitas, 2014) or the observation set $\mathcal{D}_{t-1}$ (Vazquez et al., 2008; Scott et al., 2011). Then the expectation of those improvement functions can be calculated as that in the noiseless case.

However, existing analytic EI-type acquisitions (Gupta et al., 2022) often treat that plug-in estimate as deterministic throughout each iteration and do not consider its uncertainty when formulating their closed-form expressions, which potentially leads to local search behavior in some circumstances (see Figure 1). Apart from that, noticing the incumbent solution comes with uncertainty, then depending on the type of selected kernel, the covariance information between x_t^+ and other points may need to be considered as well when specifying its analytic expression. For example, if we specify our GP model with a white noise kernel,[2] then this issue can be ignored since this kernel simply assumes all covariances between samples to be zero. However, this assumption does not hold for widely used kernels such as the Matérn and Squared Exponential kernels, especially when the length scale parameter is large. To address these challenges, we introduce a novel acquisition function that effectively incorporates the uncertainty of the incumbent solution. We consider an improvement function with an unknown objective value over the best point that maximizes the GP predictive mean as the best incumbent, akin to the noiseless scenario. Although lacking a deterministic incumbent in our improvement function, we derive an analytical expression for this acquisition function in the presence of noisy observations, which also generalizes that of the noiseless case. Our acquisition is constructed directly from the correct variance, allowing us to take full advantage of the covariance information from the GP model and generate its closed-form representation under noisy observations. Furthermore, we provide a regret bound for this acquisition function under heteroscedastic observation noise. The effectiveness of our method is further demonstrated through our empirical experiments.

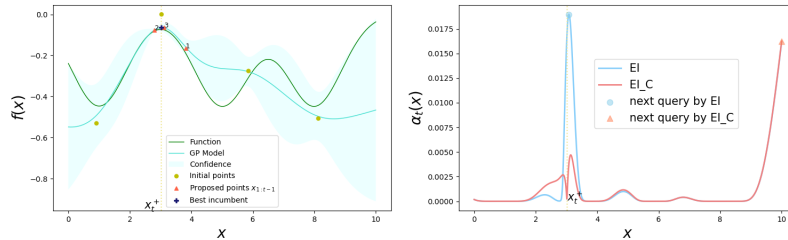2. In this case, BO degenerates to random search.

Figure 1: A scenario example illustrates that the utilization of covariance information has the potential to circumvent local search in a noisy environment. The goal is to locate the maximum of this objective function. Notably, the corrected EI suggests a point that is closer to the global optimizer $x^* = 10$.

2. Related work

For a recent review of Bayesian optimization, we refer interested readers to Garnett (2023). The expected improvement under noisy data is more complicated than in the noiseless case (Jones et al., 1998), due to the uncertainty in the noisy observations. Santner et al. (2003) stressed the importance of measuring the noise variance values and taking them into account when fitting the GP model. In the noisy setting, the best value of the objective function is unknown, thus substituting a plug-in estimate is popular in practice, such as the maximum noisy observation (Nguyen et al., 2017). However, this estimate lacks robustness as it may have high variance (Picheny et al., 2013). An alternative choice is the best GP predictive mean $\max_{x \in \mathcal{S}} \mu_{t-1}(x)$ where $\mathcal{S}$ can be the search space $\mathcal{X}$ (Wang and de Freitas, 2014) or the observation set D_{t-1} (Vazquez et al., 2008; Gupta et al., 2022). The additional computation for the predictive mean adds more cost to the optimization, especially for a large search space $\mathcal{X}$, but may be acceptable for optimizing functions that are very expensive to evaluate. With these plug-in estimators, the noiseless EI acquisition function can be applied also when there are noisy observations. There have been several works that investigate the convergence properties of the regret bound for EI and improved EI-type algorithms. Bull (2011) established an upper bound of the simple regret for EI in the noiseless scenario. Wang and de Freitas (2014) derived a regret bound for expected improvement with the best predictive mean of the GP model as the incumbent best under the noisy setting. Nguyen et al. (2017) proved a regret bound for EI with the best-observed value as the incumbent best. Their regret bound relies on a pre-defined termination threshold κ, which is also assumed in this paper.

Apart from plug-in estimates, Forrester et al. (2006) provided an approximation for EI through re-interpolation, which relies on a noise-free GP using the predicted data made by the noisy GP model. More rigorous approaches estimate EI through Monte Carlo (MC) integration (Williams, 2000; Letham et al., 2019; Balandat et al., 2020), which results in more expensive computation. In particular, Letham et al. (2019) proposed an MC-based EI-type acquisition that handles the uncertainty by averaging over the EI values of a number of noisy-free GP models. On the other hand, Balandat et al. (2020) introduced another MC-based acquisition that considers the uncertainty stemming from the unknown incumbent best. They address this uncertainty by averaging the improvement functions on a set of q test points and previously evaluated points, which allows for parallel (batch-sequential) op-

timization to accelerate the search process. These methods nevertheless remain significantly more computationally expensive than closed-form acquisition functions.

There are several other popular acquisition functions including probability of improvement (PI) (Kushner, 1963), upper confidence bound (UCB), knowledge gradient (KG), entropy search (ES), and predictive entropy search (PES) (Hernández-Lobato et al., 2014). The last three methods are most useful in exotic problems where the assumptions made by EI does hold anymore. Similar to EI, the PI acquisition prefers selecting points near the incumbent best which can potentially result in over-exploitation (Brochu et al., 2010). Recent studies by Ma et al. (2019) combine PI with covariance information from the GP model under a noisy environment. On the other hand, the UCB acquisition function emphasizes exploring areas with higher uncertainty, promoting more exploration. Srinivas et al. (2010) established a regret bound for UCB in the noisy setting using the information capacity γ_i and a parameter β_i that depends on the reproducing kernel Hilbert space (RKHS).

The main contribution of this paper is a modified expected improvement that leverages the covariance information from the heteroscedastic GP model (Le et al., 2005), which is more applicable to noisy environments. In this model, the input measurements are treated as deterministic, leading to the noise variance varying across the input space and allowing us to utilize more precise prior covariance information for noisy observations. To our knowledge, we provide the first corrected expected improvement that directly incorporates the uncertainty of the incumbent solution under noisy observations, along with an upper bound on the regret associated with this acquisition.

3. Bayesian optimization

As mentioned before, Bayesian optimization aims to find the global optimum of a black-box function $f(\cdot)$ on a bounded input space $\mathcal{X}$. Drawing noisy samples from the objective function is typically expensive, making it essential to enhance sampling efficiency. BO tackles this challenge by utilizing a GP surrogate model that could also help balance the exploration-exploitation trade-off during the search process.

3.1. Modelling with Gaussian processes

A Gaussian process inherits the elegant mathematical properties of the multivariate normal distribution and provides a flexible framework for modeling the objective function $f(\cdot)$. Typically, the model specifies $f(x)$ as a Gaussian process $\mathcal{GP}\left(m(x), k(x_i, x_j)\right)$ with a mean function $m(x)$ and a positive semi-definite covariance matrix (or kernel) $k(x_i, x_j)$. In the presence of observation noise ε_i, we observe $y_i = f(x_i) + \varepsilon_i$ instead of the objective value. The noise term is assumed to be $\varepsilon_i \sim \mathcal{N}(0, v_i^2)$ for the purposes of GP regression. Let $\mathcal{D}_{t-1} = \{(x_i, y_i, v_i)\}_{i=1}^{t-1}$ be the set of noisy observations with uncertainty estimates up to iteration $(t-1)$ of Bayesian optimization. Assuming a prior distribution $\mathcal{GP}\left(0, k(x_i, x_j)\right)$ over f, similar to the noiseless case, the posterior distribution also follows a Gaussian distribution $P(f|\mathcal{D}_{t-1}) = \mathcal{GP}(\mu_t(x), \sigma_t^2(x))$ where

$$
\begin{aligned}
\mu_t(x) &= k_{t-1}(x)^T \left(K_{t-1} + \Sigma_{t-1}\right)^{-1} \boldsymbol{y}_{t-1} \\
\sigma_t^2(x) &= k_{t-1}(x, x) - k_{t-1}(x) \left(K_{t-1} + \Sigma_{t-1}\right)^{-1} k_{t-1}(x)^T
\end{aligned}
\tag{2}
$$

with $\boldsymbol{y}_{t-1} = [y_1, \cdots, y_{t-1}]$, covariance matrix $K_{t-1} = [k_{t-1}(x_i, x_j)]_{1 \leq i,j \leq t-1}$ and $k_{t-1}(x) = [k_{t-1}(x_1, x), \cdots, k_{t-1}(x_{t-1}, x)]$. Here $\Sigma_{t-1} = diag(v_1^2, \cdots, v_{t-1}^2)$ is a diagonal matrix formed by the variances of the noise terms. In addition, the covariance between x_i and x_j is

$$\sigma_{t-1}(x_i x_j) = k_{t-1}(x_i, x_j) - k_{t-1}(x_i) \left(K_{t-1} + \Sigma_{t-1}\right)^{-1} k_{t-1}(x_j)^T \tag{3}$$

which indicates the relationship between these two points. As such, the shape (smoothness, amplitude of the predictive variance) of the GP model is significantly influenced by the choice of the kernel as well as the kernel parameters. A large number of covariance kernels are available in the literature (Santner et al., 2003). The Matérn kernel, as one of the most widely used kernels, is defined as

$$k_{\text{Matérn}}(x_i, x_j) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\|x_i - x_j\|_2}{\ell}\right)^{\nu} \mathcal{B}_{\nu}\left(\frac{\|x_i - x_j\|_2}{\ell}\right) \tag{4}$$

where $\mathcal{B}_{\nu}$ is the modified Bessel function, $\Gamma(\cdot)$ is the Gamma function, ν is the smoothness parameter, and ℓ is the scale parameter. A popular Matérn kernel is Matérn-$\frac{5}{2}$ with $\nu = \frac{5}{2}$, which satisfies the twice differentiable property. Another popular kernel is the square exponential kernel, which is given by $k_{se}(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2l^2})$ with length scale ℓ.

3.2. Expected Improvement

Once we have the GP model built on the observation set, we can employ BO alongside an appropriate acquisition function $\alpha_t(x)$ to identify the next point x_t for evaluation. This candidate is obtained by maximizing the acquisition function, that is $x_t = \arg\max_{x \in \mathcal{X}} \alpha_t(x)$. The expected improvement balances exploration and exploitation by maximizing the expectation over the improvement function $I_t(x)$. Some authors (Močkus, 1975; Lizotte, 2008) have introduced an additional parameter ζ to augment this criterion as $I_t(x) = \max\{0, f(x) - \xi - \zeta\}$, but this specific augmentation is out of the scope of this paper and will not be further discussed. In this paper, we consider the improvement function $I_t(x) = \max\{0, f(x) - \mu_{t-1}(x_t^+)\}$ with $x_t^+ = \arg\max_{i \leq t-1} \mu_{t-1}(x_i)$. The incumbent best is defined as the best GP predictive mean over the observation set. This acquisition function can be evaluated in closed form as (Wang and de Freitas, 2014):

$$\alpha_t(x) = \mathbb{E}\left[\max\{0, f(x) - \mu_{t-1}(x_t^+)\}\right] = \sigma_{t-1}(x)\phi(z) + \left(\mu_{t-1}(x) - \mu_{t-1}(x_t^+)\right)\Phi(z) \tag{5}$$

where $z = \left(\mu_{t-1}(x) - \mu_{t-1}(x_t^+)\right)/\sigma_{t-1}(x)$, ϕ is the standard normal PDF, and Φ is the standard normal CDF. The computation cost of this acquisition function is far cheaper than the black box function.

4. Corrected Expected improvement

Our modified expectation improvement aims to maximize the expectation of the improvement function, denoted as $I_t^C(x) = \max\{0, f(x) - f(x_t^+)\}$, s.t.

$$\alpha_t^C(x) = \mathbb{E}(\max\{0, f(x) - f(x_t^+)\}) \tag{6}$$

where $x_t^+ = \arg\max_{i \leq t-1} \mu_{t-1}(x_i)$. We note that the incumbent best $f(x_t^+)$ is defined as the unknown true objective value of the best point among the observation set that maximizes the

predictive mean of the GP model. Let us define $\tilde{\sigma}_{t-1}^2(x) = \sigma_{t-1}^2(x) + \sigma_{t-1}^2(x_t^+) - 2\sigma_{t-1}(xx_t^+)$ where $\sigma_{t-1}^2(\cdot)$ is its corresponding variance, and $\sigma_{t-1}(xx_t^+)$ is the covariance between points x and x_t^+. Despite not knowing the incumbent best in our improvement function, we can still derive a closed-form expression for our acquisition in noisy settings (Section 4.1):

$$\alpha_t^C(x) = \tilde{\sigma}_{t-1}(x)\phi\left(\frac{\mu_{t-1}(x) - \mu_{t-1}(x_t^+)}{\tilde{\sigma}_{t-1}(x)}\right) + \left(\mu_{t-1}(x) - \mu_{t-1}(x_t^+)\right)\Phi\left(\frac{\mu_{t-1}(x) - \mu_{t-1}(x_t^+)}{\tilde{\sigma}_{t-1}(x)}\right) \quad (7)$$

where ϕ and Φ are the density and cumulative distribution functions of the standard normal distribution, respectively. When $\tilde{\sigma}_{t-1}(x) = 0$, we set $\alpha_t^C(x) = 0$. In the noiseless case where $\forall x \neq x_t^+$, $\sigma_{t-1}(x_t^+) = \sigma_{t-1}(xx_t^+) = 0$, we recover the expression for EI. In addition, let us define the function $z_{t-1}(x) = \frac{\mu_{t-1}(x) - \mu_{t-1}(x_t^+)}{\tilde{\sigma}_{t-1}(x)}$ and the function $\tau(z) = z\Phi(z) + \phi(z)$, then we derive another expression of this acquisition function

$$\alpha_t^C(x) = \tilde{\sigma}_{t-1}(x)\tau(z_{t-1}(x)) \quad (8)$$

where $\tilde{\sigma}_{t-1}(x)$ is non-negative and it reaches zero at x_t^+ even in the noisy setting. From this expression, the corrected variance term $\tilde{\sigma}_{t-1}(x)$ tends to zero when $x \to x_t^+$. As a result, the corrected expected improvement of points next to the current best point would be relatively small, avoiding over-exploration around this region. It makes intuitive sense that this acquisition function would search more globally than EI. An example in Figure 1 illustrates this fact.

4.1. Derivation of the Modified Expected Improvement

When $f(x) - f(x_t^+)$ is non-negative, the variable $I_t^C(x)$ is Gaussian distributed with mean $u_{t-1}(x) = \mu_{t-1}(x) - \mu_{t-1}(x_t^+)$ and variance $\tilde{\sigma}_{t-1}^2(x) = \sigma_{t-1}^2(x) + \sigma_{t-1}^2(x_t^+) - 2\sigma_{t-1}(xx_t^+)$ where $\mu_{t-1}(x)$ is the mean of the GP evaluated at x, $\sigma_{t-1}^2(x)$ is its corresponding variance at x, and $\sigma_{t-1}(xx_t^+)$ is the covariance between points x and x_t^+. Thus using the likelihood of $I_t^C(x)$ (for simplicity, we write I), we obtain the expectation of our improvement function:

$$\alpha_t^C(x) = \int_0^\infty \frac{I}{\sqrt{2\pi}\tilde{\sigma}_t(x)}\exp\left(-\frac{1}{2}\left(\frac{I - u_{t-1}(x)}{\tilde{\sigma}_{t-1}(x)}\right)^2\right)dI \quad (9)$$

Let $s = \frac{I - u_{t-1}(x)}{\tilde{\sigma}_{t-1}(x)}$, then $I = s\tilde{\sigma}_{t-1}(x) + u_{t-1}(x)$ and $ds = \frac{1}{\tilde{\sigma}_{t-1}(x)}dI$. Using the above likelihood (9), we obtain our modified expected improvement

$$\begin{aligned}
\alpha_t^C(x) &= \int_{-\frac{u_{t-1}(x)}{\tilde{\sigma}_{t-1}(x)}}^\infty \frac{s\tilde{\sigma}_{t-1}(x) + u_{t-1}(x)}{\sqrt{2\pi}\tilde{\sigma}_{t-1}(x)}\exp\left(-\frac{s^2}{2}\right)\tilde{\sigma}_{t-1}(x)ds \\
&= \frac{\tilde{\sigma}_{t-1}(x)}{\sqrt{2\pi}}\int_{-\frac{u_{t-1}(x)}{\tilde{\sigma}_{t-1}(x)}}^\infty se^{-\frac{s^2}{2}}ds + u_{t-1}(x)\int_{-\frac{u_{t-1}(x)}{\tilde{\sigma}_{t-1}(x)}}^\infty \frac{e^{-\frac{s^2}{2}}}{\sqrt{2\pi}}ds \\
&= \frac{\tilde{\sigma}_{t-1}(x)}{\sqrt{2\pi}}(-e^{-\frac{s^2}{2}})\big|_{-\frac{u_{t-1}(x)}{\tilde{\sigma}_{t-1}(x)}}^\infty + u_{t-1}(x)\Phi\left(\frac{u_{t-1}(x)}{\tilde{\sigma}_{t-1}(x)}\right) \\
&= \tilde{\sigma}_{t-1}(x)\phi\left(\frac{u_{t-1}(x)}{\tilde{\sigma}_{t-1}(x)}\right) + u_{t-1}(x)\Phi\left(\frac{u_{t-1}(x)}{\tilde{\sigma}_{t-1}(x)}\right).
\end{aligned} \quad (10)$$

4.2. Termination Criterion

Similar to Nguyen et al. (2017, Lemma 2), the value of our acquisition function is set to be greater than a positive value κ that is $\max_t \alpha^C(x_t) \geq \kappa$. This termination criterion guarantees the convergence of the regret bound for our acquisition function. We will utilize this property to show our regret bound in Section 4.3. We additionally show here a connection to classical results in economics, thereby giving an interpretation of the value of κ and its scale relative to $f(\cdot)$.

4.2.1. Profit-Cost curve

Let us denote t_κ as the minimum number of iterations to reach the termination criterion κ. Since the improvement function is defined over the best feasible objective, we define the profit as

$$\text{profit}(\kappa) = f(x_{t_\kappa}^+) - \kappa \cdot t_\kappa$$
$$t_\kappa = \min t \text{ s.t. } \alpha_t(x_t) < \kappa \tag{11}$$

where $x_{t_\kappa}^+ = \arg\max_{i \leq t_\kappa - 1} \mu_{t_\kappa - 1}(x_i)$ and κ can be viewed as a cost of evaluating the function set for the optimization routine. A small κ implies the computation cost for the objective function is trivial compared with the improvement in the target function. Meanwhile, with a small κ, the termination will occur after a large number of iterations but also bring an accurate estimation of the target value. Figure 2 presents an empirical profit curve for varying values of κ, showing a consistently higher profit for our proposed acquisition when using a κ threshold termination criterion.
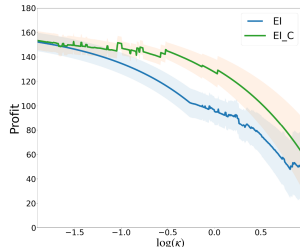

Figure 2: Comparison of EI (blue line) and corrected EI (green line) over the profit for same termination criterion κ on the *Sphere 3d* function. In our simulations, the standard deviation of the Gaussian noise term is set to be 20. The average profit is measured over 15 experiments for each κ.

4.3. Theoretical Properties

In this section, we present the theoretical properties of corrected EI under heteroskedastic noisy outputs, i.e. $y_t = f(x_t) + \varepsilon_t$, which are more natural for real-world applications. The noise term ε_t is assumed to be Gaussian distributed with a known variance proxy v_t^2. The objective function f is assumed to be smooth according to the reproducing kernel Hilbert space (RKHS) associated with a GP kernel. Similar to Srinivas et al. (2010); Wang and de Freitas (2014), the kernel is assumed to be bounded as $k(x, x) \leq 1$. We draw inspiration from Nguyen et al. (2017) and derived the regret bound in a similar way for our proposed acquisition function. We begin this section with a brief introduction to some important lemmas from existing works. Then we show that, under some mild assumptions,

our acquisition function reaches a sublinear convergence rate for the squared exponential kernel similar to the standard EI.

4.3.1. Auxiliary Definitions and Lemmas

Definition 1 *The maximum information gain after T rounds, namely γ_T, is defined as:*

$$\gamma_T := \max_{A \subset D : |A| = T} I(\boldsymbol{y}_A; \boldsymbol{f}_A) = \max_{A \subset D : |A| = T} H(\boldsymbol{y}_A) - H(\boldsymbol{y}_A \mid \boldsymbol{f}_A) \tag{12}$$

where $H(\boldsymbol{y}_A)$ is the marginal entropy of the observations $\boldsymbol{y}_A = (y_1, \cdots, y_T)$ and $H(\boldsymbol{y}_A \mid \boldsymbol{f}_A)$ is the conditional entropy of the observations $\boldsymbol{y}_A$ given the corresponding function values $\boldsymbol{f}_A$.

Lemma 1 *[Theorem 6 of Srinivas et al. (2010)]. Let $\delta \in (0, 1)$ and assume that the noise variables ε_t are uniformly bounded by the maximum standard deviation of the observation noise $\upsilon_{\max} = \max(\upsilon_1, \cdots, \upsilon_T)$. Define $\beta_t = 2\|f\|_k^2 + 300\gamma_t \ln^3 \left(\frac{t}{\delta}\right)$, then*

$$p\left(\forall t, \forall x \in \mathcal{X}, |\mu_{t-1}(x) - f(x)| \leq \sqrt{\beta_t} \sigma_{t-1}(x)\right) \geq 1 - \delta \tag{13}$$

Lemma 7 of Nguyen et al. (2017) has provided the regret bound for the variance of any arbitrary set of points (not just for selected points x_t) with γ_T under the homoskedastic noise setting. In the following lemma, we generalize their findings to accommodate scenarios involving bounded heteroskedastic additive Gaussian observation noise (detailed proof in Appendix A).

Lemma 2 *The sum of the predictive variances is bounded by the maximum information gain γ_T. That is for $\forall x \in \mathcal{X}$, it holds that*

$$\sum_{t=1}^{T} \sigma_{t-1}^2(x) \leq \frac{2}{\log(1 + \upsilon_{\max}^{-2})} \gamma_T \tag{14}$$

where $\upsilon_{\max} = \max(\upsilon_1, \cdots, \upsilon_T)$ is the maximum standard deviation of the additive Gaussian observation noise.

4.3.2. Upper bound for simple regret r_t

Let x_t be the point selected by our acquisition function, then the cumulative regret R_t is the sum of the instantaneous regrets r_t: $R_t = \sum_{i=1}^{t} r_t$ where $r_t = f(x^*) - f(x_t)$. We start the proof sketch by considering breaking down r_t into:

$$r_t = \underbrace{f(x^*) - f(x_t^+)}_{\textbf{term 1}} - \underbrace{(f(x_t) - f(x_t^+))}_{\textbf{term 2}} \tag{15}$$

Lemma 3 *Let $\kappa > 0$ be a pre-defined stopping threshold on the acquisition function $\alpha_t^C(x)$, if $\mu_{t-1}(x_t) \leq \mu_{t-1}(x_t^+)$, then we have that*

$$\mu_{t-1}(x_t^+) - \mu_{t-1}(x_t) \leq \sqrt{C} \tilde{\sigma}_{t-1}(x_t). \tag{16}$$

where $C = \log\left[\frac{2}{\pi\kappa^2}\right]$.

Proof Set $u_{t-1}(x_t) = \mu_{t-1}(x_t) - \mu_{t-1}(x_t^+)$, using our assumption we have $u_{t-1}(x_t) \leq 0$ by the definition of x_t^+. Using Equation (8), we obtain

$$\alpha_t^C(x_t) = \tilde{\sigma}_{t-1}(x_t)\tau\left(\frac{u_{t-1}(x_t)}{\tilde{\sigma}_{t-1}(x_t)}\right) \leq \tilde{\sigma}_{t-1}(x_t)\phi\left(\frac{u_{t-1}(x_t)}{\tilde{\sigma}_{t-1}(x_t)}\right) \tag{17}$$

where the inequality is led by the fact that $\tau(z) \leq \phi(z), \forall z < 0$. Thus we obtain that

$$\frac{u_{t-1}^2(x_t)}{\tilde{\sigma}_{t-1}^2(x_t)} \leq \log\left[\frac{\tilde{\sigma}_{t-1}^2(x_t)}{2\pi\kappa^2}\right] \leq \log\left[\frac{2}{\pi\kappa^2}\right] \tag{18}$$

where the second inequality is led by the fact that $\tilde{\sigma}_{t-1}^2(x) \leq \sigma_{t-1}^2(x) + \sigma_{t-1}^2(x_t^+) - 2\sigma_{t-1}(xx_t^+) \leq 4$ since the kernel satisfies $k(x,x) \leq 1$. Define $C = \log\left[\frac{2}{\pi\kappa^2}\right] \geq 0$, thus we conclude the proof that

$$0 \leq \mu_{t-1}(x_t^+) - \mu_{t-1}(x_t) \leq \sqrt{C}\tilde{\sigma}_{t-1}(x_t). \tag{19}$$

∎

Lemma 4 *Let $\kappa > 0$ be a pre-defined stopping threshold on the acquisition function $\alpha_t^C(x)$, $z_{t-1}(x) = \frac{\mu_{t-1}(x) - \mu_{t-1}(x_t^+)}{\tilde{\sigma}_{t-1}(x)}$ and $\tau(z) = z\Phi(z) + \phi(z)$, we have $\tau(-z_{t-1}(x_t)) \leq 1 + \sqrt{C}$ where $C = \log\left[\frac{1}{\pi\kappa^2}\right]$.*

Proof Notice that function $\tau(z)$ has nice properties depending on the sign of z: $\tau(z) \leq 1 + z, \forall z \geq 0$; $\tau(z) \leq \phi(z), \forall z \leq 0$. Thus, we consider two possible cases for $z_{t-1}(x_t)$:
Case 1: Assume $\mu_{t-1}(x_t) \geq \mu_{t-1}(x_t^+)$ which implies $z_{t-1}(x_t) \geq 0$, thus

$$\tau(-z_{t-1}(x_t)) \leq \phi(-z_{t-1}(x_t)) \leq 1 \tag{20}$$

Case 2: Assume $\mu_{t-1}(x_t) \leq \mu_{t-1}(x_t^+)$ which implies $z_{t-1}(x_t) \leq 0$, thus using Lemma 3, we have that

$$\tau(-z_{t-1}(x_t)) \leq 1 + \frac{\mu_{t-1}(x_t^+) - \mu_{t-1}(x_t)}{\tilde{\sigma}_{t-1}(x_t)} \leq 1 + \sqrt{C} \tag{21}$$

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

The lemma 5 below considers the lower bound for the acquisition function α_t^C under the noisy setting (detailed proof in Appendix A).

Lemma 5 *Let $\delta \in (0,1)$. For $x \in \mathcal{X}, t \in \mathcal{N}$, set $I_t^C(x) = \max\{0, f(x) - f(x_t^+)\}$, then with probability at least $1 - 2\delta$ we have*

$$\alpha_t^C(x) \geq \max\{I_t^C(x) - \sqrt{\beta_t}\left(\sigma_{t-1}(x) + \sigma_{t-1}(x_t^+)\right), 0\}. \tag{22}$$

Then we consider finding an upper bound for term 1 and term 2 in simple regret given by Equation (15). Lemma 6 (see Appendix A) provides an upper bound for term 1.

Lemma 6 *Let $\delta \in (0,1)$. Then with a probability of at least $1 - 2\delta$, we have*

$$f(x^*) - f(x_t^+) \leq \sqrt{\beta_t}\left(\sigma_{t-1}(x^*) + \sigma_{t-1}(x_t^+)\right) + \tilde{\sigma}_{t-1}(x_t)\tau(z_{t-1}(x_t)). \tag{23}$$

For term 2, we have that

$$
\begin{aligned}
f(x_t^+) - f(x_t) &= \tilde{\sigma}_{t-1}(x_t)\left[-z_{t-1}(x_t)\right] \text{ by } z = \tau(z) - \tau(-z) \\
&= \tilde{\sigma}_{t-1}(x_t)\left(\tau\left(-z_{t-1}(x_t)\right) - \tau\left(z_{t-1}(x_t)\right)\right).
\end{aligned}
\tag{24}
$$

Finally, by Lemma 4 and 6, we obtain

$$
\begin{aligned}
r_t &\le f(x^*) - f(x_t^+) + f(x_t^+) - f(x_t) \le \tilde{\sigma}_{t-1}(x_t)\tau\left(-z_{t-1}(x_t)\right) + \sqrt{\beta_t}\left(\sigma_{t-1}(x^*) + \sigma_{t-1}(x_t^+)\right) \\
&\le (1 + \sqrt{C})\tilde{\sigma}_{t-1}(x_t) + \sqrt{\beta_t}\left(\sigma_{t-1}(x^*) + \sigma_{t-1}(x_t^+)\right).
\end{aligned}
\tag{25}
$$

Furthermore, we note that $\sigma_{t-1}^2(x_t)\sigma_{t-1}^2(x_t^+) - \sigma_{t-1}^2(x_tx_t^+) \ge 0$ because of the positive semi-definiteness of the kernel matrix. Thus we have that

$$
\begin{aligned}
\tilde{\sigma}_{t-1}(x_t) = \sqrt{\sigma_{t-1}^2(x_t) + \sigma_{t-1}^2(x_t^+) - 2\sigma_{t-1}(x_tx_t^+)} &\le \sqrt{\sigma_{t-1}^2(x_t) + \sigma_{t-1}^2(x_t^+) + 2\sigma_{t-1}(x_t)\sigma_{t-1}(x_t^+)} \\
&\le \sigma_{t-1}(x_t) + \sigma_{t-1}(x_t^+).
\end{aligned}
\tag{26}
$$

Then we can write an upper bound of the simple regret as

$$
r_t \le \underbrace{(\sqrt{1} + \sqrt{C})\sigma_{t-1}(x_t)}_{A_t} + \underbrace{\sqrt{\beta_t}\sigma_{t-1}(x^*)}_{B_t} + \underbrace{(1 + \sqrt{C} + \sqrt{\beta_t})\sigma_{t-1}(x_t^+)}_{C_t}.
\tag{27}
$$

4.3.3. Upper bounding the cumulative regret R_T

We now look at the cumulative regret $R_T = \sum_{t=1}^{T} r_t$

$$
R_T \le \sum_{t=1}^{T} A_t + \sum_{t=1}^{T} B_t + \sum_{t=1}^{T} C_t.
\tag{28}
$$

Using the Cauchy-Schwartz inequality that $\sum_{i=1}^{n} a_i^2 \le n\left(\sum_{i=1}^{n} a_i^2\right)$, $\beta_T \ge \beta_t, \forall T \ge t$, and Lemma 2, we can bound $\sum_{t=1}^{T} B_t$ with the following

$$
\sum_{t=1}^{T} B_t \le \sqrt{T\sum_{t=1}^{T} B_t^2} \le \sqrt{\frac{2\beta_T T \gamma_T}{\log(1 + v_{\max}^{-2})}}.
\tag{29}
$$

Similarly, we can bound $\sum_{t=1}^{T} C_t$ and $\sum_{t=1}^{T} A_t$ with the following

$$
\begin{aligned}
\sum_{t=1}^{T} A_t^2 &\le 2(1 + C)\sum_{t=1}^{T} \sigma_{t-1}^2(x_t) \le \frac{4(1 + C)\gamma_T}{\log(1 + v_{\max}^{-2})}, \\
\sum_{t=1}^{T} C_t^2 &\le 3(1 + C + \beta_T)\sum_{t=1}^{T} \sigma_{t-1}^2(x_t) \le \frac{6(1 + C + \beta_T)\gamma_T}{\log(1 + v_{\max}^{-2})}.
\end{aligned}
\tag{30}
$$

Using the Cauchy-Schwartz inequality again, we get

$$
\sum_{t=1}^{T} A_t \le \sqrt{T\sum_{t=1}^{T} A_t^2} \le \sqrt{\frac{4(1 + C)T\gamma_T}{\log(1 + v_{\max}^{-2})}} \text{ and } \sum_{t=1}^{T} C_t \le \sqrt{T\sum_{t=1}^{T} C_t^2} \le \sqrt{\frac{6(1 + C + \beta_T)T\gamma_T}{\log(1 + v_{\max}^{-2})}}.
\tag{31}
$$

Combining the above equations, we obtain our regret bound

$$
R_T \le \sqrt{\frac{2T\gamma_T}{\log(1 + v_{\max}^{-2})}}\left(\sqrt{\beta_T} + \sqrt{2(1 + C)} + \sqrt{3(1 + C + \beta_T)}\right),
\tag{32}
$$

where $C = \log \left[\frac{2}{\pi \kappa^2} \right]$, κ is a pre-defined constant to terminate the optimization, β_T is in the form of $\mathcal{O}\left((\log T)^3 \right)$, and $v_{\max} = \max(v_1, \cdots, v_T)$ is the maximum standard deviation of the noise. We can see that the regret bound for our proposed acquisition is equivalent to that of standard EI. The bound for γ_T relies on the chosen kernel i.e. $\gamma_T \sim \mathcal{O}\left((\log T)^{d+1} \right)$ for squared exponential kernel. Therefore, when choosing a squared exponential kernel for the GP model, we achieve a sublinear rate $R_T \sim \mathcal{O}\left(\sqrt{T \left(\log T \right)^{d+4}} \right)$.

5. Experiments

In this section, we present our empirical results using the experiments on the benchmark, synthetic functions, and compression tasks.

5.1. Benchmark objective functions

| (a) *Hartmann3d* | (b) *Griewank(d = 6)* | (c) *Levy(d = 4)* | (d) *Powell(d = 5)* |

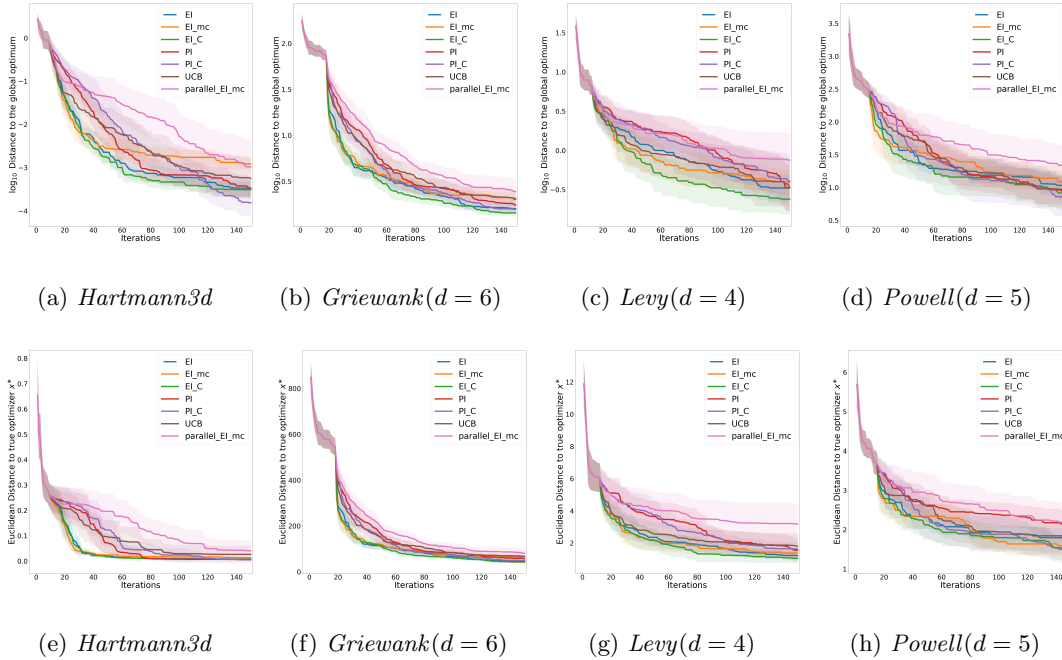| (e) *Hartmann3d* | (f) *Griewank(d = 6)* | (g) *Levy(d = 4)* | (h) *Powell(d = 5)* |

Figure 3: Comparison of methods for Benchmark objective functions. Figures (a)-(d) show how the mean and 95% confidence bound (shaded region) of the distance between the best feasible objective and the global optimum changes with each iteration of optimization. Figures (e)-(f) visualize the variation of the L_2 distance between the best point and the global optimizer x^*.

In this section, we compare our acquisition method with other methods including Corrected-PI (Ma et al., 2019), UCB, EI, PI, MC-based noisy EI (Letham et al., 2019), and parallel MC-based noisy EI (Balandat et al., 2020) on several benchmark objective functions including *Hartmann3d*, *Griewank(d = 6)*, *Levy(d = 4)*, and *Powell(d = 5)*. Our objective is to identify the optimizer that minimizes the values of these functions. We use the framework

of *BoTorch*[3] for implementing those acquisition functions. Two evaluation metrics are considered: the log distance to the global optimum $\log_{10}\left(f\left(x^*\right) - f\left(x_t^+\right)\right)$ and the L_2 distance to the global optimizer x^*. We employed the Matérn kernel for our GP model and set the total number of iterations (T) to 150. The length scale parameter of the kernel is optimized by maximum likelihood. To introduce noise in our experiments, each observation noise ε_t was sampled from a Gaussian distribution with mean 0 and standard deviation υ_t less than or equal to 10% of the range of the objective function $(\max f\left(x\right) - \min f\left(x\right))$. Before fitting the model, the outputs along with observation noises are standardized, and the inputs are normalized to $[0,1]^d$ based on the minimum and maximum values. All experiments were repeated 15 times for each benchmark function with a quasirandom sequence of size $3d$ as the initialization for our GP model.

Results in Figure 3 indicate that the corrected EI outperforms EI for those benchmark objective functions under our noisy settings. For *Powell* function, the corrected PI slightly outperforms our proposed method. Overall, our proposed method demonstrates good performance on those benchmark functions. Furthermore, we assess the computational cost associated with these acquisition functions. Typically, this cost involves inferring the hyperparameters of the GP model, and maximizing the acquisition function in order to propose a candidate point. With analytic acquisition functions, the objective function can be approximated using just one GP model, enabling direct computation of the acquisition value from this model. Consequently, working with this kind of acquisition function is considerably inexpensive. In contrast, MC-type methods are more computationally expensive. For example, MC-based noisy EI relies on multiple noiseless GP models and performs integration by averaging the expected improvement across these models, resulting in higher computational costs. Moreover, parallel MC-based noisy EI assumes that the incumbent best is unknown and uses samples from the joint posterior over the q test points and previously observed points. The integration is computed by averaging the improvements on those samples. In our experiment, we set the number of noiseless models to be 20 (by default) for MC-based noisy EI, and a quasirandom sequence of size $q = 256$ for parallel MC-based noisy EI, resulting in significantly higher computational costs for these two methods compared to analytic acquisition functions.

5.2. Model Compression

Various compression techniques have been proposed for DNN models, leading to a smaller model that can be deployed on edge devices with limited memory and computational resources. Low-rank factorization techniques like Singular Value Decomposition (SVD) and Tensor Decomposition can be utilized to construct the compressed network with a low-rank approximation of the original weight matrices. Their rank parameters denoted as θ can be selected via the BO procedure in order to find a balance between the size and performance of the compressed networks. A scaling scheme proposed by Ma et al. (2019) is applied to transform the rank parameter from discrete space to continuous domain $[0,1]^d$ where d represents the number of dimensions in the parameterization. Let us define f^* as the original model and $\hat{f}_\theta$ as the compressed model, then the objective function with respect to θ can

---

3. *BoTorch* (Balandat et al., 2020) is a state-of-the-art open-source Bayesian optimization software package with support for various acquisition functions.

be specified as

$$\gamma \mathcal{L}(\hat{f}_\theta) + \mathcal{R}(\hat{f}_\theta, f^*) \tag{33}$$

where $\mathcal{R}(\hat{f}_\theta, f^*)$ is the compression ratio which is calculated by dividing the size of the compressed network by the size of the original network, and $\mathcal{L}(\hat{f}_\theta)$ is the error rate of the compressed network. A smaller model with good generalization performance is thus preferred in order to minimize this objective function and γ is the trade-off parameter.

(a) FC3: all iterations　(b) ResNet50: all iterations　(c) VGG-16: all iterations

(d) FC3: best of all iterations　(e) ResNet50: best of all iterations　(f) VGG-16: best of all iterations
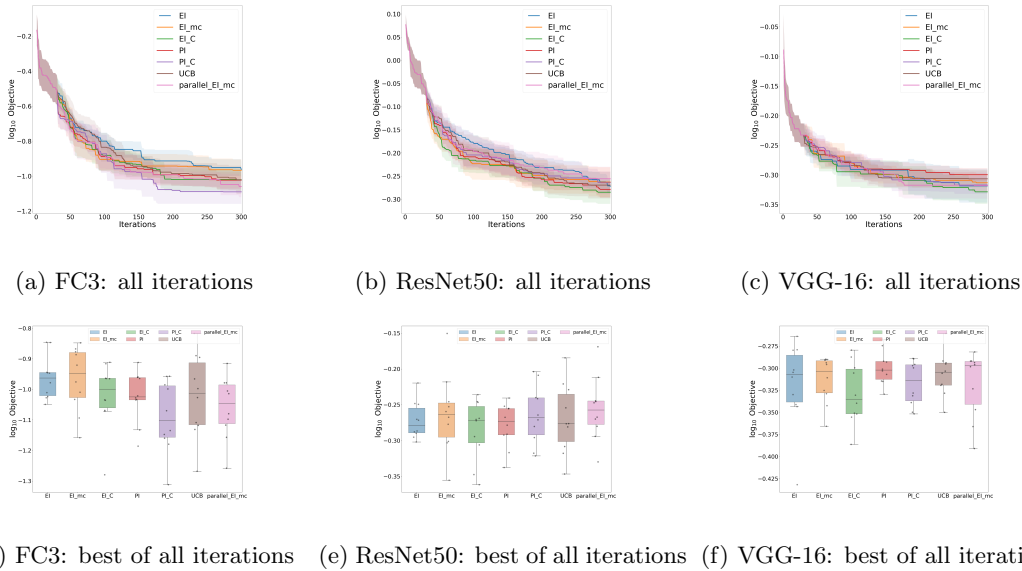
Figure 4: Comparison results on the compression tasks of the pre-trained FC3, ResNet50, and VGG-16 model using BO with different acquisition functions. The results are obtained from 10 independent runs. The plots in the first rows show how the mean of the best observation changes over the iterations of BO together with its 95% confidence interval. The second row shows the best results among all iterations.

In our experiments, we compare our approach with other acquisition functions on the compression tasks of several representative neural networks including a 3-layer fully connected network (FC3, Ma et al. (2019)), ResNet50 (He et al., 2016), and VGG-16 (Simonyan and Zisserman, 2014). We compress all the layers of FC3 using SVD as proposed in Denton et al. (2014). While for the latter two models, we only compress the weights of their convolution layers using tensor decomposition (Tai et al., 2016). Thus there are 3 compression parameters in the compression task of FC3, 13 parameters in that of VGG-16, and 16 parameters for ResNet50. The FC3 model is pre-trained on the MNIST dataset (Deng, 2012) while the latter two models are pre-trained on the ImageNet dataset (Russakovsky et al., 2015). The trade-off parameter γ is set to be 1. During each iteration of optimization, we consistently measure the top-1 error rate on n_t randomly selected samples ($20 \leq n_t \leq 50$), resulting in a noisy evaluation of the objective. For each observation, its noise variance v_t^2 is inversely proportional to the sample size n_t. For FC3, those samples are randomly drawn from the 10,000 testing images of the MNIST dataset. For the latter two models, the samples are selected from 50,000 validation images of the ILSVRC2012 dataset. We continue using the previous settings for the GP model and the acquisition functions. We run all algorithms for 300 iterations with the first 30 iterations being random initialization.

The compression results in Figure 4 show our proposed method gives better performance than standard EI under the noisy setting.

6. Discussion and Conclusions

In this paper, we propose a novel acquisition function that addresses the limitation of the analytic EI-type methods in the presence of noise. We correct the closed-form expression of EI to account for the uncertainty introduced by the incumbent best —a result that has not previously been published despite EI being nearly half a century old (Mockus et al., 1978) and one of the most popular acquisition functions even in the presence of observation noise. Additionally, we show that this modified EI retains a convergence rate similar to that of the standard EI, and results in a profit-maximizing convergence criterion under a linear cost model. Our empirical experiments provide evidence that this approach is effective and competitive compared with some popular acquisition functions when dealing with noisy observations. Although there may be cases where our method underperforms corrected-PI or MC-based methods, we believe that it offers valuable insights into the behavior of EI-type methods under noisy observations. One notable contribution of our work is that it fills a gap in the analytic EI-type approaches by directly incorporating the uncertainty of the incumbent best without relying on MC integration when formulating the closed-form expression. This enhancement improves both the efficiency and performance of BO by leveraging the available covariance information from the GP model.

We also notice that the differences in performance among those acquisitions might be negligible for functions in high dimensions (Ma et al., 2019) or extremely noisy observations (Garnett, 2023). In more complex settings of real-world problems, some details about the data information such as the scale of the observation noise are still unknown to us thus it can be challenging to intuit suitable parameters i.e. prior distribution for the heteroscedastic GP model without a great deal of knowledge about the data. In the worst case with increasing levels of noise, the GP inference reflects more uncertainty in the objective function regardless of the choice of the acquisition functions. Our work also reflects the importance of combining the covariance information provided by this model in Bayesian optimization. Therefore, we expect more advanced techniques for characterizing and capturing the nature of observation noise, which can be instrumental in constructing a more precise GP model.

Acknowledgments

References

Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G Wilson, and Eytan Bakshy. Botorch: A framework for efficient Monte-Carlo Bayesian optimization. *Advances in neural information processing systems*, 33:21524–21538, 2020.

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. *Journal of Machine Learning Research*, 12(10), 2011.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Emily L Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. *Advances in neural information processing systems*, 27, 2014.

Alexander IJ Forrester, Andy J Keane, and Neil W Bressloff. Design and analysis of "noisy" computer experiments. *AIAA journal*, 44(10):2331–2339, 2006.

Roman Garnett. *Bayesian Optimization*. Cambridge University Press, 2023.

Sunil Gupta, Santu Rana, Svetha Venkatesh, et al. Regret bounds for expected improvement algorithms in Gaussian process bandit optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 8715–8737. PMLR, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. *Advances in neural information processing systems*, 27, 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.

Harold J Kushner. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. In *Joint Automatic Control Conference*, number 1, pages 69–79, 1963.

Quoc V Le, Alex J Smola, and Stéphane Canu. Heteroscedastic Gaussian process regression. In *Proceedings of the 22nd international conference on Machine learning*, pages 489–496, 2005.

Benjamin Letham, Brian Karrer, Guilherme Ottoni, and Eytan Bakshy. Constrained Bayesian optimization with noisy experiments. *Bayesian Analysis*, 14(2):495–519, 2019.

Daniel James Lizotte. *Practical Bayesian optimization*. PhD thesis, University of Alberta, 2008.

Xingchen Ma, Amal Rannen Triki, Maxim Berman, Christos Sagonas, Jacques Cali, and Matthew B Blaschko. A Bayesian optimization framework for neural network compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10274–10283, 2019.

Jonas Močkus. On Bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference: Novosibirsk, July 1–7, 1974*, pages 400–404. Springer, 1975.

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. Regret for expected improvement over the best-observed value and stopping condition. In *Asian conference on machine learning*, pages 279–294. PMLR, 2017.

Victor Picheny, Tobias Wagner, and David Ginsbourger. A benchmark of kriging-based infill criteria for noisy optimization. *Structural and multidisciplinary optimization*, 48(3): 607–626, 2013.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Thomas J Santner, Brian J Williams, William I Notz, and Brain J Williams. *The design and analysis of computer experiments*, volume 1. Springer, 2003.

Warren Scott, Peter Frazier, and Warren Powell. The correlated knowledge gradient for simulation optimization of continuous parameters using Gaussian process regression. *SIAM Journal on Optimization*, 21(3):996–1026, 2011.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Proceedings of the International Conference on Learning Representations*, page abs/1409.1556, 2014.

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*, 2010.

Cheng Tai, Tong Xiao, Yi Zhang, Xiaogang Wang, and E Weinan. Convolutional neural networks with low-rank regularization. In *4th International Conference on Learning Representations*, 2016.

Emmanuel Vazquez, Julien Villemonteix, Maryan Sidorkiewicz, and Eric Walter. Global optimization based on noisy evaluations: an empirical study of two statistical approaches. In *Journal of Physics: Conference Series*, volume 135, page 012100, 2008.

Ziyu Wang and Nando de Freitas. Theoretical analysis of Bayesian optimisation with unknown Gaussian process hyper-parameters. *stat*, 1050:30, 2014.

Brian Jonathan Williams. *Sequential design of computer experiments to minimize integrated response functions*. PhD thesis, The Ohio State University, 2000.