Exploration Enhanced Expected Improvement for Bayesian Optimization

Julian Berk, Vu Nguyen, Sunil Gupta, Santu Rana, Svetha Venkatesh

Deakin University, Centre for Pattern Recognition and Data Analytics, Geelong, Australia

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 an acquisition function. This must balance improving our understanding of the function in unknown regions (exploration) with locally improving on known promising samples (exploitation). Expected improvement (EI) is one of the most widely used acquisition functions for BO. Unfortunately, it has a tendency to over-exploit, meaning that it can be slow in finding new peaks. We propose a modification to EI that will allow for increased early exploration while providing similar exploitation once the system has been suitably explored. We also prove that our method has a sub-linear convergence rate and test it on a range of functions to compare its performance against the standard EI and other competing methods.

1 Introduction

There are numerous situations, both in research and industry, where it is necessary to know the optimal input to a black box function but sampling it is either difficult or expensive. *Bayesian optimization* is one of the most evaluation efficient methods for finding the input, x^*, that will produce the optimal value of such systems [11]. It has been successfully applied to many problems in a wide range of fields including materials science, biomedical science, and even other computer science problems. An example of an application in materials science is the development of new polymer fibres [9]. In biomedical science, it has been used for many applications including studying how age effects time perception [19] and synthetic gene design [4]. Another application is the selection of hyperparameters for other machine learning algorithms [17].

Bayesian optimization methods work by fitting a probabilistic model to the available data (evaluation locations for the objective and corresponding function values). This model provides a distribution of all possible functions within a set range of possible inputs, $\mathcal{X} \in \mathbb{R}^d$. The most common model among these is the *Gaussian process* [15].

The predictions of the probabilistic model above are then used to make intelligent decisions about where to evaluate the objective function next, so that its optimum is found by using a reduced number of function evaluations. These intelligent decisions are made through an *acquisition function*. This maps $\forall x \in \mathcal{X}$ to some property that describes how useful sampling at that point will be in determining the black box function's true optima. Optimizing the acquisition function will therefore allow the best possible sample to be made from the black box function given the data and prior knowledge.

Based on the decision of the acquisition function, the new sample is collected and evaluated. This sample can then be used to update the model, allowing the point after that to be determined from the updated acquisition function. This data-driven decision allows the optima of the function to be found in far fewer iterations than if samples had been taken at random [3].

The choice of acquisition function can greatly impact the number of iterations necessary to find the optimal input. As such, poor acquisition functions can lead to sup-optimal results or the need for a larger number of costly iterations. A good acquisition function needs to balance between trying to generalize from known good points (exploitation) and trying to search for new peaks in unexplored regions (exploration). There are currently many choices of acquisition functions that provide various degrees of exploration and exploitation. Two popular choices are *expected improvement* (EI) [7] and *Gaussian process upper confidence bound* (GP-UCB) [18], with EI being more exploitative without the need to choose hyperparameters and GP-UCB having more exploration but requiring the specification of several hyperparameters. These hyperparameters can reduce optimization performance if they are not suited to the problem and determining them is both computationally costly and potentially inaccurate. As such, EI is more popular.

We propose a modification to EI that will improve its exploration in the early stages of the experiment, but converge to its previous level of exploitation at later stages. This method is detailed in Section 3 along with a proof that it has a sub-linear convergence rate. In Section 4 we discuss results from several experiments performed using our method. First, we verify that our method has increased exploration by testing it against competing methods on a synthetic function designed to favour high-exploration methods. We then test its performance in comparison to these methods on several benchmark functions and a machine learning hyperparameter tuning problem. Finally, we discuss results concerning the analytical properties of our method.

2 Bayesian Optimization and Expected Improvement

Below we first provide a background of Bayesian optimization and Gaussian processes. Then we discuss acquisition functions with a focus on expected improvement.

2.1 Bayesian Optimization

Bayesian optimization is an efficient method for optimizing noisy, expensive black box-functions [7]. More formally, the ultimate goal of the method is to find the input,

$$x^* = \underset{x \in \mathcal{X}}{\text{argmax}} f(x), \tag{1}$$

that maximises the black-box function, $f(x)$, in the bounded input space, $\mathcal{X} \subset \mathbb{R}^d$. It is possible to directly draw potentially noisy samples from the function: $y_t = f(x_t) + \epsilon$ where ϵ a random noise term, $\epsilon \sim \mathcal{N}(0, \sigma_n)$ with some unknown σ_n. However, doing so is expensive so we wish to determine x^* in as few samples as possible. To do this, Bayesian optimization uses a statistical model for the black box function to construct a surrogate function that is cheaper to sample. The statistical model is generated from all current information about the system, including all prior knowledge and all t sampled input-output pairs, $D_t = \{x_i, y_i\}_{i=1}^t$.

The statistical model is often chosen to be a Gaussian process due to its flexibility and analytic properties [2]. While the statistical model is generally not accurate enough to directly locate the optima of the function, it gives a probabilistic estimate of the function with epistemic uncertainties. This means that it can be used to select the “best” new point to sample. This is done by finding the optima of a surrogate function called an acquisition function which emphasises characteristics that are desirable for the new point to have. Acquisition functions are discussed further in section 2.3.

Once the new samples are found, they can then be used to improve the model, allowing us to find a new, potentially better point to sample. This process is iterated until a predetermined stopping condition has been met. As samples are costly, it is common to choose a maximum number of iterations as a stopping criteria, but other stopping criteria can be used as well, such as stopping when a satisfactory result has been found or when the possible improvement predicted by the model becomes too small. A more detailed review of Bayesian optimization can be found in [2].

2.2 Gaussian Process

A Gaussian process is a statistical model of the black-box function. It represents the function values, $f(x)$, at each point, $x \in \mathcal{X}$ as infinitely many correlated Gaussian random variables. As such, it is completely characterized by its mean and covariance functions, $m(x)$ and $k(x_i, x_j)$. More formally, we assume that $f(x) \sim \mathcal{GP}(m(x), k(x_i, x_j))$. The covariance function, also called a kernel, has a profound impact on the shape of the resulting Gaussian process. As such, the use of an appropriate kernel is vital.

One of the most popular kernels is the *square exponential kernel*. This is given by $k_{SE}(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2l^2}\right)$. Here, the length scale is completely determined by a single hyperparameter, l. This kernel was chosen because it is simple and translation invariant. This property is important for the generation of Thompson samples as discussed in Section 3. Other popular kernel functions are discussed in Rasmussen *et al.* [15].

Using our data with this kernel gives a posterior distribution over the function, $f(x)|D_t \sim \mathcal{GP}(m(x), \mathbf{K}_t)$. Here $\mathbf{K}_t = [k(x_i, x_j)_{\forall x_i, x_j \in D_t}]$ is the kernel matrix, which acts as the covariance matrix for the distribution. The posterior, in turn, can be used to calculate a predictive distribution, $p(f(x) \mid D_t, x) = \mathcal{N}(\mu_t(x), \sigma_t(x))$. The predictive distribution allows us to estimate the function value at any point x by calculating the predictive mean, $\mu_t(x)$, and variance, $\sigma_t^2(x)$. These are given by $\mu_t(x) = \mathbf{k}_*(\mathbf{K}_t + \sigma_n \mathbf{I})^{-1}\boldsymbol{y}$ and $\sigma_t^2(x) = k_t(x, x) - \mathbf{k}_*(\mathbf{K}_t + \sigma_n \mathbf{I})^{-1}\mathbf{k}_*^T$ with $\mathbf{k}_* = [k(x_1, x), k(x_2, x), \ldots, k(x_t, x)]$. Here $\mathbf{I}$ is the identity matrix with the same dimensions as $\mathbf{K}_t$ and σ_n is the function noise standard deviation.

2.3 Acquisition Functions

Once the Gaussian process has been built, it is used to select the optimal next point to sample from the black box function, x_t. However, exactly what qualifies a point to be x_t is non-trivial. As such, there are many potentially desirable properties that could be used to select x_t. Once a desired property is chosen, an *acquisition function*, $\alpha(x)$, is used to used to calculate it. This is generally far cheaper to evaluate than the black box function to the point where it is efficient to perform a global optimization on it $\forall x \in \mathcal{X}$ to determine a single sample of the black box function. More formally, the optimal next point is given by

$$x_t = \arg\max_{x \in X} \alpha(x) \tag{2}$$

Improvement Based Acquisition Functions One of the most basic families of acquisition functions are the improvement based acquisition functions. These use the potential improvement over what is believed to be the current maxima, called the *incumbent*. The incumbent is often taken as the current best observed value, $y^+ = \max_{i \leq t}(y_i)$. The improvement is therefore given by $I(x) = \max(f(x) - y^+, 0)$.

Probability of Improvement A simple acquisition function is the probability of improvement (PI) [8], which gives the probability that a given point will have an improvement over the incumbent. Despite being an intuitive and simple formulation, PI often favours points near the incumbent [2]. As a result, the algorithm tends to over exploit. This can lead to the algorithm failing to quickly find promising peaks away from the incumbent, reducing the optimization efficiency in cases where there is more that one peak in the black box function. This lack of exploration can be improved by maximizing the expected improvement instead of PI [6].

Expected Improvement As $f(x)$ can be approximated by its Gaussian process predictive distribution, $I(x)$ can likewise be approximated as a function of this random variable. This allows us to take the expectation over this to find the expected amount of improvement at any point $x \in \mathcal{X}$, giving us the expected improvement acquisition function: $\alpha^{EI}(x) = \mathbb{E}[I(x)]$. This can be expressed with the Gaussian process predictive mean and variance in the following closed form [7]:

$$\alpha^{EI}(x) = \begin{cases} (\mu(x) - y^+)\varPhi(z) + \sigma(x)\phi(z), & \text{if } \sigma(x) > 0 \\ 0 & \text{if } \sigma(x) = 0 \end{cases} \tag{3}$$

where $z = \frac{\mu(x) - y^+}{\sigma(x)}$, ϕ is the standard normal PDF, and $\varPhi$ is the standard normal CDF. For a full analytical derivation of EI, we refer interested readers to [14].

EI has better exploration than PI but still tends to over-exploit in many situations, such as when it hits a local optimum. Despite this, EI is currently the most common acquisition functions due to its consistent performance without the need to choose additional hyperparameters.

A Heuristic Approach for boosting the exploration of EI: ζ-EI It is a common belief that artificially increasing the incumbent by some positive ζ will reduce the value of the acquisition function near the currently sampled points, boosting exploration [10]. However, this method does not work well in practice as it is not easy to choose the right value of ζ. If this value is large, the algorithm will significantly over-explore. This often leads to inefficiency in optimization performance.

3 The Proposed E^{3}I Method

In this section we will outline our modification to EI that will improve its exploration without causing it to significantly over-explore. We will then prove that, under some mild assumptions, it has a sub-linear regret bound.

3.1 Thompson Sampling

For our method, we wish to generate full random approximations of the black-box function. Hernández-Lobato *et al.* [5] have developed a method for doing this through Thompson sampling. For a shift invariant kernel such as the square exponential kernel we are using, Bochner's theorem [1] states that it has a Fourier dual, $s(w)$, which is equal to the spectral density of $k(x_i, x_j)$. Normalizing this as $\hat{s}(w) = s(w)/\beta$ allows us to represent the kernel as

$$k(x_i, x_j) = 2\beta \mathbb{E}_{\hat{s}(w)} \left[\cos(w^T x_i + b) \cos(w^T x_j + b) \right] \tag{4}$$

where $b \sim \mathcal{U}[0, 2\pi]$. If we draw V random samples of w and let $\phi(x) = \sqrt{\frac{2\beta}{V}} \{\cos(Wx + b), \sin(Wx + b)\}$, we can approximate the kernel with $k(x_i, x_j) \approx \phi(x_i)^T \phi(x_j)$. By setting $\varPhi = [\phi(x_1), \ldots, \phi(x_V)]$, we can also approximate the kernel matrix with $\mathbf{K} \approx \varPhi\varPhi^T + \sigma^2 \mathbf{I}$, where $\mathbf{I}$ is the $V \times V$ matrix identity. These estimates have been augmented by the random samples of w so they can be viewed as having random probable points added to them. This means that, if V is sufficiently large, the corresponding predictive mean can be viewed as a complete estimate of the black box function given the current data:

$$f(x) \approx g(x) = \phi(x)^T (\varPhi\varPhi^T + \sigma^2 \mathbf{I})^{-1} \varPhi^T \boldsymbol{y} \tag{5}$$

The process for generating and finding the optima of these Thompson samples is outlined in Algorithm 1.

3.2 Exploration Enhanced Expected Improvement (E^{3}I)

The Thompson sample functions have two useful properties. Firstly, without noise they will agree with the currently sampled points exactly (i.e. $g(x_i) = y_i, \forall (x_i, y_i) \in D_t$). This means that $g(x^+) = f(x^+) = y^+$. As such, either the maximum of $g(x)$ will occur at x^+, in which case $g^* = \max_x g(x) = y^+$, or it will occur elsewhere, in which case $g^* > y^+$. Secondly, the Thompson sample functions will also converge to the

Algorithm 1 Thompson Sampling

```
Input:$D_{t-1} = \{x_i, y_i\}_{i=1}^{t-1}$, #random feature dimension, $V$, #Thompson samples, $M$
for $m = 1$ to $M$ do
    Randomly generate $b \sim \mathcal{U}[0, 2\pi]$ and $V$ weights, $w_i \sim \mathcal{N}(0, \mathbf{I}_{d \times d})\, \forall i = 1 \dots V$
    Let $W = [w_1, \dots, w_V] \in \mathbb{R}^{V \times d}$
    Let $\phi(x) = \sqrt{\frac{2\beta}{V}}\,(\cos(Wx + b), \sin(Wx + b))$ and $\varPhi = [\phi(x_1), \dots, \phi(x_V)]$
    Thompson samples are given by $g_m(x) = \phi(x)^T (\varPhi\varPhi^T + \sigma^2\mathbf{I})^{-1}\varPhi^T \boldsymbol{y}$
    Use a global optimizer to find $g_m^* = \max_{x \in \mathcal{X}} g_m(x)$
end for
Output: $g_1^*, \dots, g_M^*$
```

true function as the number of iterations increase. This means that g^* should converge towards y^+.

These two properties allow us to use g^* as the incumbent in EI instead of y^+. As $g^* \geq y^+$, the algorithm will have greater exploration. However, as the Thompson samples are sensible approximations of the underlying function, the method does not have the same risk of over-exploration as artificially increasing the incumbent does. As $g^* \to y^+$, it should also explore less at later stages in the algorithm when exploration is less important.

This approach assumes that any given Thompson sample is a good approximation of the black box function given the current data. Due to variations between Thompson samples, it is possible that any given Thompson sample may be an outlier, voiding this assumption. As such, we instead look at the distribution of possible Thompson samples. This makes the new acquisition function a function of this distribution. To obtain the best point, we take the expected value of this distribution, i.e. $\alpha^{E^3I}(x) = \mathbb{E}_g[\mathbb{E}_x[I(x, g^*)]]$. Unfortunately, determining this directly is difficult. As such, we instead generate M Thompson samples, $g_1^*, g_2^*, \dots, g_M^*$, and find the sample mean instead. Setting $z = \frac{\mu(x) - g_m^*}{\sigma(x)}$ and $\tau(z) = z\varPhi(z) + \phi(z)$ we get

$$\alpha^{E^3I}(x) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_x[I(x, g_m^*)] = \begin{cases} \frac{\sigma(x)}{M}\left[\sum_{m=1}^{M} \tau(z)\right], & \text{if } \sigma(x) > 0 \\ 0 & \text{if } \sigma(x) = 0 \end{cases} \tag{6}$$

We outline the E^{3}I routine in Algorithm 2.

3.3 Convergence

The regret bound is one of the basic criteria for evaluating the performance of an optimization algorithm. EI has been shown to converge at a sub-linear rate under a variety of assumptions. Bull [3], and Ryzhov[16] both derive convergence rates in the absence of noise. Wang and de Freitas [20] were able to derive a convergence rate in the noisy setting, but they needed to use $\mu^+(x) = \max_x \mu(x)$ as the incumbent. As E^{3}I utilizes a different incumbent, it is not compatible with this approach. Nguyen *et al.* [14] have

Algorithm 2 Bayesian optimization with E^{3}I

```
Input: $D_{t-1} = \{x_i, y_i\}_{i=1}^{t-1}$, #Weights, $V$, #Thompson samples, $M$, #Iterations, $T$
for $t = 1$ to $T$ do
    Generate the $M$ Thompson sample optima, $g_1^*, \ldots, g_M^*$, using Algorithm 1
    Use a global optimizer to find $x_t = \arg\max_{x \in \mathcal{X}} \left( \alpha_t^{E^3 I}(x) \right)$
    Query the black box function with $x_t$ to get $y_t = f(x_t)$
    Augment the current data: $D_t = D_{t-1} \cup (x_t, y_t)$
end for
Output: $(x^*, y^*) = \arg\max_y D_T$
```

shown that EI has a sub-linear convergence rate if a minimum improvement stopping condition is used. The proof is valid in the noisy setting and does not require a modified incumbent. As such we extend and employ it to show that this is also true for our method. Many of the lemmas used in their proof can be directly applied to our method.

We start our derivation for the regret bound of E^{3}I as follows.

Lemma 1. *(Srinivas et al. [18]) 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*

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

Lemma 2. *The improvement function,* $I_{t,m}(x) = \max\left(0, f(x) - g_{t,m}^*\right)$*, and the acquisition function,* $\alpha_t^{E^3 I}(x) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}[I_{t,m}(x)]$ *satisfy the inequality* $\frac{1}{M}\sum_{m=1}^{M} I_{t,m}(x) - \sqrt{\beta_t}\sigma_{t-1}(x) \leq \alpha_t^{E^3 I}(x)$*.*

Proof. In the case that $\sigma_{t-1}(x) = 0$, we have $\mathbb{E}[I_{t,m}(x)] = I_{t,m}(x), \forall m \in [1, M]$. This means that $\alpha_t^{E^3 I}(x) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}[I_{t,m}(x)] = \frac{1}{M}\sum_{m=1}^{M} I_{t,m}(x)$ so the result is trivial. For $\sigma_{t-1}(x) > 0$, the proof is as follows. Using Lemma 1, $q_{t-1,m} = \frac{f(x) - g_{t-1,m}^*}{\sigma_{t-1}(x)}$, and $z_{t-1,m} = \frac{\mu_{t-1}(x) - g_{t-1,m}^*}{\sigma_{t-1}(x)}$, we can get the following result:

$$\begin{aligned}
\alpha_t^{E^3 I}(x) &\geq \frac{1}{M}\sum_{m=1}^{M} \sigma_{t-1}(x)\tau\left(q_{t-1,m} - \sqrt{\beta_t}\right) \\
&\geq \frac{1}{M}\sum_{m=1}^{M} \sigma_{t-1}(x)\left(q_{t-1,m} - \sqrt{\beta_t}\right) \qquad \text{by } \tau(z) \geq z
\end{aligned}$$

If $I_t(x) = 0$, the lemma becomes $\alpha_t^{E^3 I}(x) \geq -\sqrt{\beta_t}\sigma_{t-1}(x)$. As $\alpha_t^{E^3 I}(x) \geq 0$, $\sqrt{\beta_t} \geq 0$, and $\sigma_{t-1}(x) \geq 0$, this is always true. For $I_t(x) > 0$ we have $q_{t-1,m} = \frac{I_t(x)}{\sigma_{t-1}(x)}$. This gives us

$$\alpha_t^{E^3 I}(x) \geq \frac{1}{M}\sum_{m=1}^{M} I_{t,m}(x) - \sqrt{\beta_t}\sigma_{t-1}(x)$$

which concludes our proof. □

We now prove the main theorem:

Theorem 3. *Let $\kappa > 0$ be a predefined small constant as a stopping criteria, σ^2 be the measurement noise variance, $C \triangleq \log\left[\frac{1}{2\pi\kappa^2}\right]$, $\beta_t = 2\left\|f\right\|_k^2 + 300\gamma_t \ln^3\left(\frac{t}{\delta}\right)$ and $\delta \in (0,1)$. Then, with probability at least $1-\delta$, after T iterations the cumulative regret of E^{3}I using a collection of maxima samples g_m^* drawn from Thompson sampling as the incumbents obeys the following sublinear rate: $R_T \lesssim \sqrt{T\beta_T\gamma_T} \sim \mathcal{O}\left(\sqrt{T \times (\log T)^{d+4}}\right)$, where $\gamma_T \sim \mathcal{O}\left((\log T)^{d+1}\right)$ is the maximum information gain for the squared exponential kernel.*

Proof. Let $x_t = \underset{x\in\mathcal{X}}{\text{argmax}} \alpha_t^{E^3I}(x)$ be the choice at iteration t, the instantaneous regret is:

$$\begin{aligned}
Mr_t &= Mf(x^*) - Mf(x_t)\\
&= Mf(x^*) - Mf(x_t) + \sum_{m=1}^{M} g_{t-1,m}^* - \sum_{m=1}^{M} g_{t-1,m}^*\\
&= \underbrace{\sum_{m=1}^{M}\left[f(x^*) - g_{t-1,m}^*\right]}_{A_t} + \underbrace{\sum_{m=1}^{M}\left[f(x_t) + g_{t-1,m}^*\right]}_{B_t}
\end{aligned}$$

We need to connect this with the maximum information gain, γ_T. This can be done by bounding r_t with the GP posterior variance. We bound A_t with the using Lemma 2, Lemma 1, and the fact that $\alpha_t^{E^3I}(x^*) \leq \alpha_t^{E^3I}(x_t)$ to get

$$\begin{aligned}
A_t &= \sum_{m=1}^{M}\left[f(x^*) - g_{t-1,m}^*\right] = \sum_{m=1}^{M} I_{t,m}(x)\\
A_t &\leq M\left[\alpha_t^{E^3I}(x^*) + \sqrt{\beta_t}\sigma_{t-1}(x^*)\right]\\
&\leq M\left[\alpha_t^{E^3I}(x_t) + \sqrt{\beta_t}\sigma_{t-1}(x^*)\right] && \text{by Lemma 2}\\
&= M\left[\sigma_{t-1}(x_t)\tau(z_{t-1}(x_t)) + \sqrt{\beta_t}\sigma_{t-1}(x^*)\right] && \text{by Lemma 1}
\end{aligned}$$

Likewise, we bound B_t with the following:

$$\begin{aligned}
B_t &= \sum_{m=1}^{M}\left[g_{t-1,m}^* - \mu_{t-1}(x_t) + \mu_{t-1}(x_t) - f(x_t)\right]\\
&\leq \sum_{m=1}^{M}\left[\sigma_{t-1}(x_t)(-z_{t-1}(x_t)) + \sigma_{t-1}(x)\sqrt{\beta_t}\right] && \text{by Lemma 1}\\
&= M\sigma_{t-1}(x_t)\left[\tau(-z_{t-1}(x_t)) + \sqrt{\beta_t} - \tau(z_{t-1}(x_t)\right] && \text{by } z = \tau(z) - \tau(-z)
\end{aligned}$$

Combining these bounds and noting that the M term cancels out, we get

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

Using the bound of $\tau(-z_{t-1}(x_t))$ in Lemma 9 from [14] and setting $C \triangleq \log\left[\frac{1}{2\pi\kappa^2}\right]$ we can simplify this to

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

We now look at the sum of the regret,

$$R_t = \sum_{t=1}^{T} r_t \leq \sum_{t=1}^{T} L_t + \sum_{t=1}^{T} U_t$$

Using the Cauchy-Schwartz inequality that $(a + b + c) \leq 3(a^2 + b^2 + c^2)$, that $\beta_T \geq \beta_t, \forall t \leq T$, and Lemma 7 from [14]) we can bound $\sum_{t=1}^{T} L_t$ with the following

$$\begin{aligned}\sum_{t=1}^{T} L_t &\leq \sum_{t}^{T} \sigma_{t-1}^2(x_t) 3(\beta_t + 1 + C) \\ &\leq 3(\beta_T + 1 + C)\sum_{t}^{T} \sigma_{t-1}^2(x_t) \leq \frac{6(\beta_T + 1 + C)\gamma_T}{\log(1 + \sigma^{-2})}\end{aligned}$$

Using the Cauchy-Schwartz inequality again we get

$$\sum_{t=1}^{T} L_t \leq \sqrt{T}\sqrt{\sum_{t=1}^{T} L_t} \leq \sqrt{\frac{6T(\beta_T + 1 + C)\gamma_T}{\log(1 + \sigma^{-2})}} \tag{7}$$

We can use Lemma 7 [14] and the Cauchy-Schwartz inequality on $\sum_{t=1}^{T} U_t$ as well to obtain a similar result:

$$\sum_{t=1}^{T} U_t \leq \beta_T \sum_{t=1}^{T} \sigma_{t-1}(x^*) \leq \sqrt{\frac{2T\beta_T\gamma_T}{\log(1 + \sigma^{-2})}} \tag{8}$$

Combining Equations (7) and (8) gives us our regret bound:

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

The function of the maximum information gain, $\sqrt{T \times \gamma_T}$, will usually dominate this expression as $\beta_T \sim \mathcal{O}\left((\log T)^2\right)$. It is kernel dependent but for the squared exponential kernel used in this paper it is $\gamma_T \sim \mathcal{O}\left((\log T)^{d+1}\right)$. This means that our regret bound for this kernel is $R_T \sim \mathcal{O}\left(\sqrt{T \times (\log T)^{d+1}}\right)$, which vanishes in the limit of $\lim_{T\to\infty} \frac{R_T}{T} = 0$. We note that we achieve the similar form to the one in [14]. $\square$

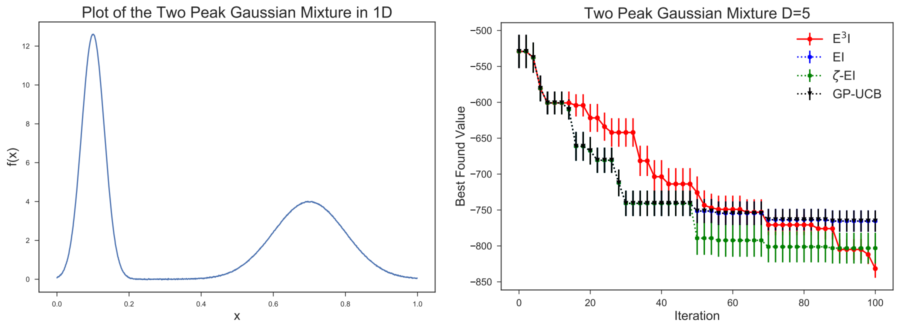

Fig. 1. A plot of the 1D Gaussian mixture function for illustration (left) and the performance of various methods on the 5D version of the same Gaussian mixture function (right). The higher dimensional function was used as little exploration is required in 1D. Lower is better. Note that GP-UCB and EI both get stuck on the initial lower value peak. ζ-EI manages to find the larger peak, but fails to exploit it. This suggests that it may be over-exploring due to an imperfect choice of ζ. On the contrary, it can be seen that E^{3}I is able to obtain a superior result through better late-stage exploitation.

4 Experiments

In this section we outline and discuss our experimental results. We apply our method to synthetic, benchmark and real world functions. We also test some other important properties of our method, such as the dependence on M and the convergence of the Thompson samples. **The code used for this paper can be found at `https://github.com/jmaberk/BO_with_E3I`.**

4.1 Experimental Setup

We performed several experiments comparing E^{3}I with the standard EI, ζ-EI with $\zeta = 0.01$ [2], and GP-UCB. In our algorithm, we scaled the inputs to be in the range $[0, 1]$ in all dimensions and standardized the sampled function values to have zero mean and a standard deviation of 1. This guarantees that our kernel magnitude will be scaled correctly for all functions. After this scaling, we use a square exponential kernel.

We ran our experiment 10 times per function with $d + 1$ random initial points, where d is the number of input dimensions. The experiment was stopped after $T = 20d$ iterations. As we used a simple multi-start L-BFGS-B optimizer, we minimized the negative of all functions instead of maximizing them. As such, lower results are better.

4.2 Synthetic Multi-Peak Function

As we expect our method to have higher exploration than EI, we will test it on functions which require exploration for better performance. In particular, we consider multi-peak functions. Methods with poor exploration can get stuck on sup-optimal peaks, significantly reducing performance. As such, we chose to use a two-peak Gaussian mixture function to verify the high exploration of our method. One peak was chosen to be wide ($\mathcal{N}(0.7, 0.01)$) while the other peak was chosen to be narrow but taller ($\mathcal{N}(0.1, 0.001)$) so that acquisition functions with poor exploration will tend to get stuck on the wider, smaller peak more often and hence not perform as well. We applied our suite of acquisition functions to the problem and have summarised our results in **Fig. 1**.

It is evident that our method is both better able to find the narrow peak faster, and that the performance gap increases as the number of dimensions increases.

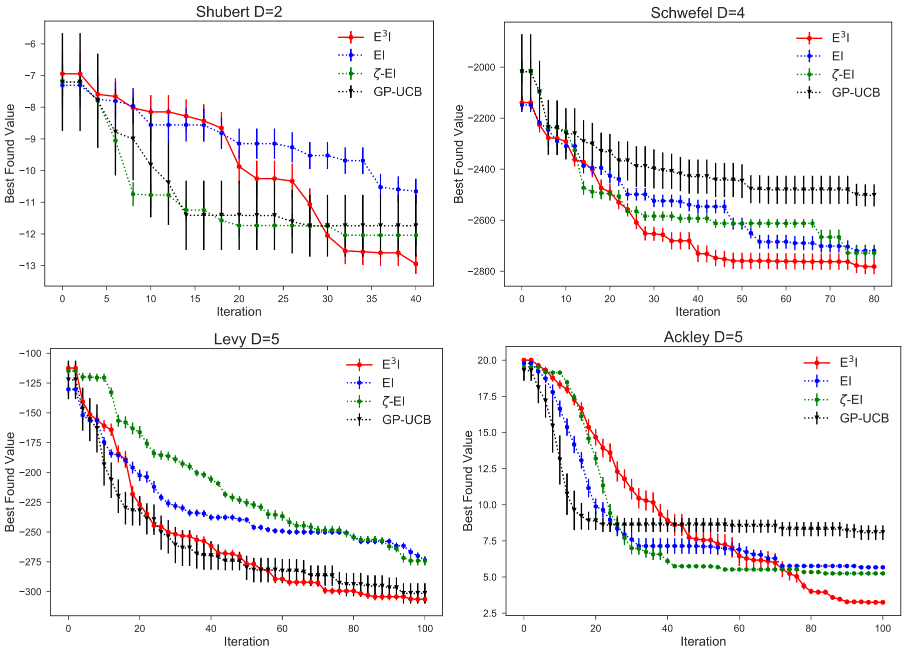

Fig. 2. Performance of various methods on a range of multi-peak benchmark functions. Lower is better. Note that our method generally seems to converge slowly at the early stages when other methods are exploiting, but it can beat the exploitative methods by finding better peaks. This is because E^{3}I tends to explore in the early stages and then tends to exploit later to hit the optimum.

4.3 Benchmark Functions

We also tested our method on several common multi-peak benchmark functions. These include the Levy (5D), Schwefel (4D), Shubert (2D), and Ackley (5D) functions[1]. The results for these are displayed in **Fig. 2**.

[1] All benchmark functions use the recommended parameters from https://www.sfu.ca/ ssurjano/optimization.html

These results show that our method is able to find a better optima more quickly than the other methods in these multi-peak test functions. GP-UCB also does very well with the Levy and Shubert functions, while the ζ-EI does fairly well on all experiments except the Levy function. This variance in performance is unsurprising, as both methods have parameters that control their level of exploration. If these are not suited to the problem, they can be detrimental to the algorithm's performance. Our methods increased exploration is automatically adjusted through the Thompson samples and therefore does not face this issue. As such, even in cases where these methods performed well, E^{3}I was able to show improvements over them.

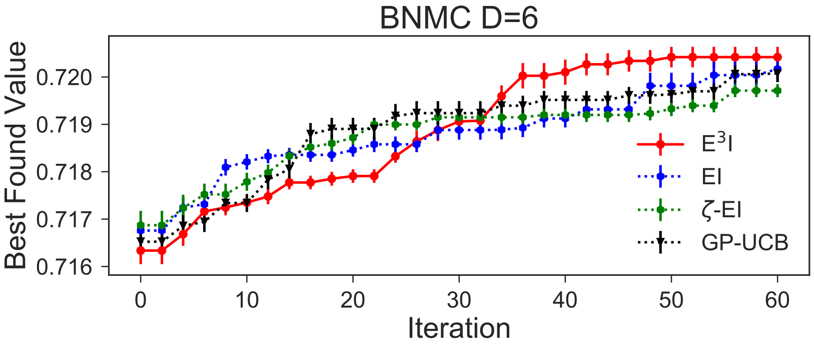

Fig. 3. The results for hyperparameter tuning a BNMC experiment. Note that, again, our algorithm performs poorly in the early stages while it explores for new optima but is able to find a better optima sooner than the other methods.

4.4 Machine Learning Hyperparameter Tuning

Finally, we tested our method on a real-world application; the determination of optimal hyperparameters for a machine learning algorithm called *Bayesian Nonparametric Multi-label Classification* (BNMC) [12]. This algorithm is used to efficiently classify multi-labelled data by exploiting the correlation between the multiple labels and features. Its performance is dependant on six hyperparameters which we can tune with Bayesian optimization. These are the Dirichlet concentration parameters for both the feature and label, the learning rates for both SVI and SGD, the truncation threshold and the stick-breaking parameter. The data the algorithm was used on, called SceneData, consisted of 1196 test and 1211 training samples, each with 294 features. The results of this with our suite of acquisition functions is given in **Fig. 3** with the F1 score used as the performance measure.

4.5 Sensitivity Analysis with respect to the Number of Optima Samples, M

One of our key assumptions is that we can approximate the expectation over the distribution of Thompson samples as its sample mean. As $M \to \infty$, this will be true.

However, using a very large M will increase computational costs. As such, we wish to find a value for M that will not compromise our results while also not being too expensive. To determine this, we ran the same experiment on both a 2D Shubert function and a 2D Schwefel function for a range of M values from 1 to 200. The results of which are summarised in **Fig. 4**. From these results, we can see that $M = 100$ seems to be an appropriate number of Thompson samples.

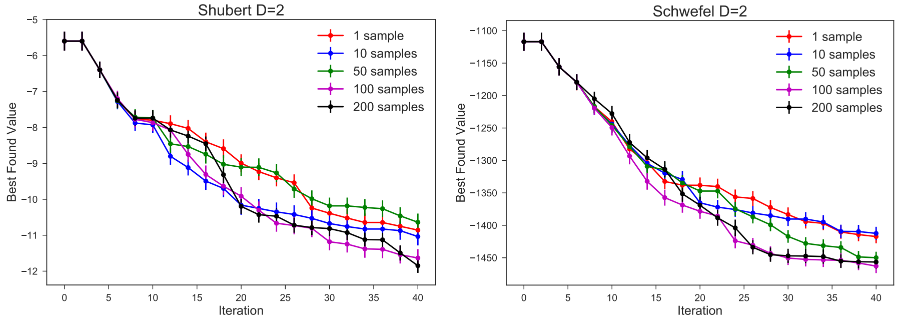

Fig. 4. Performance of E^{3}I on two functions with various number of Thompson samples, M. We can see that increasing M noticeably improves the results until $M = 50$, after which time little improvement is seen.

4.6 Computational Considerations

While our method has competitive performance with other methods, it has a considerable computational cost. Each Thompson sample requires both the inversion of a $V \times V$ matrix and a global optimization step which may increase exponentially with the dimension, d. The overall cost scales significantly with both the number of data points and the number of input dimensions, making it $\mathcal{O}(MNV^2)$ where N is the number of observations [13]. To give this some context, the average time per iteration for the 2D Shubert function earlier was 0.33 seconds with EI and 26 seconds for E^{3}I. Moving up to 4D with the Schwefel function, these become 1.4 and 81 seconds respectively. These may seem high, but they are negligible when compared to the costs associated with sampling in many of the areas that Bayesian optimization is applied to.

One way to potentially reduce computational costs is to use a method by Wang et al. [21] to find the maxima of Thompson samples by sampling a Gumbel distribution. However, this method makes several assumptions that may lead to inaccurate results and as such is left for future work.

4.7 Empirical Convergence Analysis of Thompson Samples

One of the key assumptions of our method is that the Thompson sample functions will converge to the true function as T increases. This convergence was experimentally tested and the results are shown in **Fig. 5**. It is evident that the sample mean,

$\bar{g^*} = \frac{1}{M}\sum_{m=1}^{M} g_m^*$ is converging to f^* and that the sample standard deviation, calculated with $\sigma(g^*) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(g_m^* - \bar{g}^*)^2}$, is reducing with the number of iterations. These suggest that the Thompson samples are converging properly.

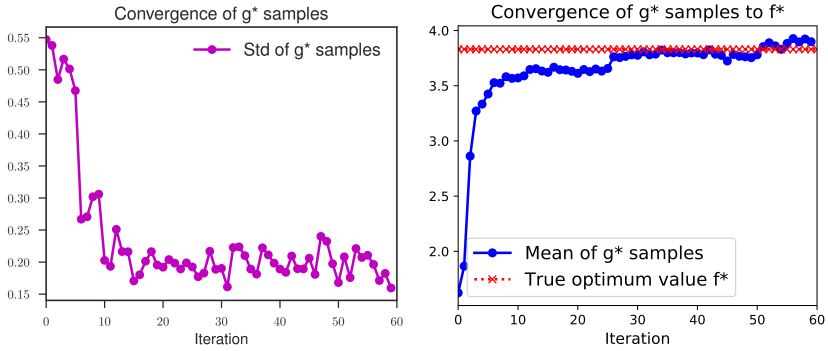

Fig. 5. The reduction of standard deviation between the Thompson sample function maxima (left) and the convergence of $\bar{g}^*$ to $f^* = \max_x(f(x))$ (right). We can see that as more samples are taken, the inter-sample variance is reducing and their mean is approaching f^*. This suggests that they are properly converging as the space is explored.

5 Conclusion

We have proposed a new approach for balancing exploration and exploitation in Bayesian optimization. Our approach makes use of Thompson sampling to guide the level of exploration. This results in the E^{3}I acquisition function.

Our method has been shown to perform better than competing methods on both several multi-peak test functions and on hyperparameter tuning for a BNMC experiment. We also show that it has a sub-linear regret bound.

The most important next step in improving E^{3}I is to resolve some of its computational issues. Beyond this, the effects of similar distribution-based approaches should be explored on other acquisition functions besides EI.

6 Acknowledgements

This research was supported by an Australian Government Research Training Program (RTP) Scholarship awarded to Mr Berk, and was partially funded by the Australian Government through the Australian Research Council (ARC). Prof Venkatesh is the recipient of an ARC Australian Laureate Fellowship (FL170100006).

Appendix A: E^{3}I Derivation

In this section, we provide the analytical derivation of E^{3}I, described in Equation (6). In particular, we make use of the improvement function over the perceived optima sample generated from Thompson sampling, $I(x) = \max(f(x) - g^*, 0)$.
We wish to find the PDF of $I(x)$ so that later we can take its expectation. As we are modeling the system with a Gaussian process, we assume that $f(x) \sim \mathcal{N}(\mu(x), \sigma(x))$. This means that $f(x)$ has the PDF

$$p(f(x)) = \frac{1}{\sqrt{2\pi}\sigma(x)} \exp\left(\frac{-(f(x) - \mu(x))^2}{2\sigma^2(x)}\right) \tag{9}$$

Now that we have a PDF for $f(x)$, we can use it to find the PDF of $I(x)$ with the distribution function technique. Let us look at the CDF of $I(x)$ for $f(x) > g^*$ with the substitution $f(x) = I(x) + g^* \, \forall f(x) > g^*$:

$$CDF_{I(x)}(a) = \int_0^a \frac{1}{\sqrt{2\pi}\sigma(x)} \exp\left(\frac{-(I(x) + g^* - \mu(x))^2}{2\sigma^2(x)}\right) dI \tag{10}$$

Taking the partial derivative with respect to $I(x)$ this gives us its the PDF:

$$p(I(x)) = \frac{1}{\sqrt{2\pi}\sigma(x)} \exp\left(\frac{-(I(x) + g^* - \mu(x))^2}{2\sigma^2(x)}\right) \tag{11}$$

Now that we have the PDF, can take its expectation to derive our acquisition function:

$$\alpha^{E^3I}(x) = \mathbb{E}_{g^*}\left[\int_0^\infty \frac{I(x)}{\sqrt{2\pi}\sigma(x)} \exp\left(\frac{-(I(x) + g^* - \mu(x))^2}{2\sigma^2(x)}\right) dI(x)\right] \tag{12}$$

Unfortunately, g^* does not have a tractable algebraic expression. To circumvent this, we approximate the expectation over g^* with the sample mean. Assuming that we have M samples of g^* our acquisition function becomes

$$\alpha^{E^3I}(x) = \sum_{m=1}^{M} \int_0^\infty \frac{I_m(x)}{\sqrt{2\pi}\sigma(x)} \exp\left(\frac{-(I_m(x) + g_m^* - \mu(x))^2}{2\sigma^2(x)}\right) dI_m(x) \tag{13}$$

As each g_m^* is now a constant, the expression inside the summation is now functionally the same expression as found in this stage of the derivation of EI. This means that E^{3}I can be expressed as a sum of standard expected improvement acquisition functions with $z = \frac{\mu(x) - g_m^*}{\sigma(x)}$:

$$\alpha^{E^3I}(x) = \begin{cases} \frac{1}{M}\sum_{m=1}^{M}\left[(\mu(x) - g_m^*)\Phi(z) + \sigma(x)\phi(z)\right], & \text{if } \sigma(x) > 0 \\ 0 & \text{if } \sigma(x) = 0 \end{cases} \tag{14}$$

References

1. S. Bochner. *Lectures on Fourier Integrals: With an Author's Supplement on Monotonic Functions, Stieltjes Integrals and Harmonic Analysis; Translated from the Original German by Morris Tenenbaum and Harry Pollard*. Princeton University Press, 1959.
2. E. Brochu, V. M. Cora, and Nando de Freitas. A tutorial on Bayesian optimisation of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arxiv.org*, 2010.
3. A. D. Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(Oct):2879–2904, 2011.
4. J. González, J. Longworth, D. C James, and N. D Lawrence. Bayesian optimization for synthetic gene design. *arXiv preprint arXiv:1505.01627*, 2015.
5. José M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in neural information processing systems*, pages 918–926, 2014.
6. A. Jalali, J. Azimi, X. Fern, and R. Zhang. A lipschitz exploration-exploitation scheme for bayesian optimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 210–224. Springer, 2013.
7. D. R Jones, M. Schonlau, and W. J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.
8. H. J Kushner. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering*, 86(1):97–106, 1964.
9. C. Li, D. R.ín de Celis Leal, S. Rana, S. Gupta, A. Sutti, S. Greenhill, T. Slezak, M. Height, and S. Venkatesh. Rapid Bayesian optimisation for synthesis of short polymer fiber materials. *Scientific reports*, 7, 2017.
10. D. J. Lizotte. *Practical Bayesian Optimization*. PhD thesis, University of Alberta, 2008.
11. J. Mockus. Application of Bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*, 4(4):347–365, 1994.
12. V. Nguyen, S. Gupta, S. Rana, C. Li, and S. Venkatesh. A Bayesian nonparametric approach for multi-label classification. In *Asian Conference on Machine Learning*, pages 254–269, 2016.
13. V. Nguyen, S. Gupta, S. Rana, C. Li, and S. Venkatesh. Predictive variance reduction search. In *NIPS Workshop on Bayesian Optimization*, 12 2017.
14. V. Nguyen, S. Gupta, S. Rana, C. Li, and S. Venkatesh. Regret for expected improvement over the best-observed value and stopping condition. In *Asian Conference on Machine Learning*, pages 279–294, 2017.
15. C.E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
16. I. O Ryzhov. On the convergence rates of expected improvement methods. *Operations Research*, 64(6):1515–1528, 2016.
17. J. Snoek, H. Larochelle, and R. P Adams. Practical Bayesian optimization of machine learning algorithms. In *NIPS*, pages 2951–2959, 2012.
18. N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1015–1022, 2010.
19. M. Turgeon, C. Lustig, and W. H. Meck. Cognitive aging and time perception: roles of Bayesian optimization and degeneracy. *Frontiers in aging neuroscience*, 8, 2016.
20. Z. Wang and N. de Freitas. Theoretical analysis of Bayesian optimisation with unknown gaussian process hyper-parameters. *NIPS Workshop on Bayesian Optimization*, 2014.
21. Z. Wang and S. Jegelka. Max-value entropy search for efficient bayesian optimization. In *International Conference on Machine Learning*, pages 3627–3635, 2017.