
Fast Information-theoretic Bayesian Optimisation

Binxin Ru¹ Mark McLeod¹ Diego Granziol¹ Michael A. Osborne^{1,2}

Abstract

Information-theoretic Bayesian optimisation techniques have demonstrated state-of-the-art performance in tackling important global optimisation problems. However, current information-theoretic approaches require many approximations in implementation, introduce often-prohibitive computational overhead and limit the choice of kernels available to model the objective. We develop a fast information-theoretic Bayesian Optimisation method, FITBO, that avoids the need for sampling the global minimiser, thus significantly reducing computational overhead. Moreover, in comparison with existing approaches, our method faces fewer constraints on kernel choice and enjoys the merits of dealing with the output space. We demonstrate empirically that FITBO inherits the performance associated with information-theoretic Bayesian optimisation, while being even faster than simpler Bayesian optimisation approaches, such as Expected Improvement.

1. Introduction

Optimisation problems arise in numerous fields ranging from science and engineering to economics and management (Brochu et al., 2010). In classical optimisation tasks, the objective function is usually known and cheap to evaluate (Hennig and Schuler, 2012). However, in many situations, we face another type of tasks for which the above assumptions do not apply. For example, in the cases of clinical trials, financial investments or constructing a sensor network, it is very costly to draw a sample from the latent function underlying the real-world processes (Brochu et al., 2010). The objective functions in such type

of problems are generally non-convex and their closed-form expressions and derivatives are unknown (Shahriari et al., 2016). Bayesian optimisation is a powerful tool to tackle such optimisation challenges (Brochu et al., 2010).

A core step in Bayesian optimisation is to define an acquisition function which uses the available observations effectively to recommend the next query location (Shahriari et al., 2016). There are many types of acquisition functions such as Probability of Improvement (PI) (Kushner, 1964), Expected Improvement (EI) (Moćkus et al., 1978; Jones et al., 1998) and Gaussian Process Upper Confidence Bound (GP-UCB) (Srinivas et al., 2009). The most recent type is based on information theory and offers a new perspective to efficiently select the sequence of sampling locations based on entropy of the distribution over the unknown minimiser x_* (Shahriari et al., 2016). The information-theoretic approaches guide our evaluations to locations where we can maximise our learning about the unknown minimum rather than to locations where we expect to obtain lower function values (Hennig and Schuler, 2012). Such methods have demonstrated impressive empirical performance and tend to outperform traditional methods in tasks with highly multimodal and noisy latent functions.

One popular information-based acquisition function is Predictive Entropy Search (PES) (Villemon-teix et al., 2009; Hennig and Schuler, 2012; Hernández-Lobato et al., 2014). However, it is very slow to evaluate in comparison with traditional methods like EI, PI and GP-UCB and faces serious constraints in its application. For example, the implementation of PES requires the first and second partial derivatives as well as the spectral density of the Gaussian process kernel function (Hernández-Lobato et al., 2014; Requeima, 2016). This limits our kernel choices. Moreover, PES deals with the input space, thus less efficient in higher dimensional problems (Wang and Jegelka, 2017). The more recent methods such as Output-space Predictive Entropy Search (OPES) (Hoffman and Ghahramani, 2015) and Max-value Entropy Search (MES) (Wang and Jegelka, 2017) improve on PES by focusing on the information content in output space instead of input space. However, current entropy search methods, whether dealing with the minimiser or the minimum value, all involve two separate sampling processes : 1) sampling

¹Department of Engineering Science, University of Oxford, Oxford, UK ²Mind Foundry Ltd., Oxford. Correspondence to: Binxin Ru <robin@robots.ox.ac.uk>, Mark McLeod <mark.mcleod@magd.ox.ac.uk>, Diego Granziol <diego@robots.ox.ac.uk>, Michael A. Osborne <mosb@robots.ox.ac.uk>.

hyperparameters for marginalisation and 2) sampling the global minimum/minimiser for entropy computation. The second sampling process not only contributes significantly to the computational burden of these information-based acquisition functions but also requires the construction of a good approximation for the objective function based on Bochner’s theorem (Hernández-Lobato et al., 2014), which limits the kernel choices to the stationary ones (Bochner, 1959).

In view of the limitations of the existing methods, we propose a fast information-theoretic Bayesian optimisation technique (FITBO). Inspired by the Bayesian integration work in (Gunter et al., 2014), the creative contribution of our technique is to approximate any black-box function in a parabolic form: $f(\mathbf{x}) = \eta + 1/2g(\mathbf{x})^2$. The global minimum is explicitly represented by a hyperparameter η , which can be sampled together with other hyperparameters. As a result, our approach has the following three major advantages:

1. Our approach reduces the expensive process of sampling the global minimum/minimiser to the much more efficient process of sampling one additional hyperparameter, thus overcoming the speed bottleneck of information-theoretic approaches.
2. Our approach faces fewer constraints on the choice of appropriate kernel functions for the Gaussian process prior.
3. Similar to MES (Wang and Jegelka, 2017), our approach works on information in the output space and thus is more efficient in high dimensional problems.

2. Fast Information-theoretic Bayesian Optimisation

Information-theoretic techniques aim to reduce the uncertainty about the unknown global minimiser \mathbf{x}_* by selecting a query point that leads to the largest reduction in entropy of the distribution $p(\mathbf{x}_*|D_n)$ (Hennig and Schuler, 2012). The acquisition function for such techniques has the form (Hennig and Schuler, 2012; Hernández-Lobato et al., 2014):

$$\alpha_{\text{ES}}(\mathbf{x}|D_n) = H[p(\mathbf{x}_*|D_n)] - \mathbb{E}_{p(y|D_n, \mathbf{x})} \left[H[p(\mathbf{x}_*|D_n \cup (\mathbf{x}, y))] \right]. \quad (1)$$

PES makes use of the symmetry of mutual information and arrives at the following equivalent acquisition function:

$$\alpha_{\text{PES}}(\mathbf{x}|D_n) = H[p(y|D_n, \mathbf{x})] - \mathbb{E}_{p(\mathbf{x}_*|D_n)} \left[H[p(y|D_n, \mathbf{x}, \mathbf{x}_*)] \right], \quad (2)$$

where $p(y|D_n, \mathbf{x}, \mathbf{x}_*)$ is the predictive posterior distribution for y conditioned on the observed data D_n , the test location \mathbf{x} and the global minimiser \mathbf{x}_* of the objective function.

FITBO harnesses the same information-theoretic thinking but measures the entropy about the latent global minimum $f_* = f(\mathbf{x}_*)$ instead of that of the global minimiser \mathbf{x}_* . Thus, the acquisition function of FITBO method is the mutual information between the function minimum f_* and the next query point (Wang and Jegelka, 2017). In other words, FITBO aims to select the next query point which minimises the entropy of the global minimum:

$$\alpha_{\text{FITBO}}(\mathbf{x}|D_n) = H[p(y|D_n, \mathbf{x})] - \mathbb{E}_{p(f_*|D_n)} \left[H[p(y|D_n, \mathbf{x}, f_*)] \right]. \quad (3)$$

This idea of changing entropy computation from the input space to the output space is also shared by Hoffman and Ghahramani (2015) and Wang and Jegelka (2017). Hence, the acquisition function of the FITBO method is very similar to those of OPES (Hoffman and Ghahramani, 2015) and MES (Wang and Jegelka, 2017).

However, our novel contribution is to express the unknown objective function in a parabolic form $f(\mathbf{x}) = \eta + 1/2g(\mathbf{x})^2$, thus representing the global minimum f_* by a hyperparameter η and circumventing the laborious process of sampling the global minimum. FITBO acquisition function can then be reformulated as:

$$\alpha_{\text{FITBO}}(\mathbf{x}|D_n) = H[p(y|D_n, \mathbf{x})] - \mathbb{E}_{p(\eta|D_n)} \left[H[p(y|D_n, \mathbf{x}, \eta)] \right] = H \left[\int p(y|D_n, \mathbf{x}, \eta) p(\eta|D_n) d\eta \right] - \int p(\eta|D_n) H[p(y|D_n, \mathbf{x}, \eta)] d\eta. \quad (4)$$

The intractable integral terms can be approximated by drawing M samples of η from the posterior distribution $p(\eta|D_n)$ and using a Monte Carlo method (Hernández-Lobato et al., 2014). The predictive posterior distribution $p(y|D_n, \mathbf{x}, \eta)$ can be turned into a neat Gaussian form by applying a local linearisation technique on our parabolic transformation as described in Section 2.1. Thus, the first term in the above FITBO acquisition function is an entropy of a Gaussian mixture, which is intractable and demands approximation as described in Section 2.3. The second term is the expected entropy of a one-dimensional Gaussian distribution and can be computed analytically because the entropy of a Gaussian has the closed form: $H[p(y|D_n, \mathbf{x}, \eta)] = 0.5 \log [2\pi e (v_f(\mathbf{x}|D_n, \eta) + \sigma_n^2)]$ where the variance $v_f(\mathbf{x}|D_n, \eta) = K_f(\mathbf{x}, \mathbf{x}')$ and σ_n^2 is the variance of observation noise.

2.1. Parabolic Transformation and Predictive Posterior Distribution

Gunter et al. (2014) use a square-root transformation on the integrand in their warped sequential active Bayesian integration method to ensure non-negativity. Inspired by this work, we creatively express any unknown objective function $f(\mathbf{x})$ in the parabolic form:

$$f(\mathbf{x}) = \eta + 1/2g(\mathbf{x})^2, \quad (5)$$

where η is the global minimum of the objective function. Given the noise-free observation data $D_f = \{(\mathbf{x}_i, f_i) | i = 1, \dots, n\} = \{\mathbf{X}_n, \mathbf{f}_n\}$, the observation data on g is $D_g = \{(\mathbf{x}_i, g_i) | i = 1, \dots, n\} = \{\mathbf{X}_n, \mathbf{g}_n\}$ where $g_i = \sqrt{2(f_i - \eta)}$.

We impose a zero-mean Gaussian process prior on $g(\mathbf{x})$, $g \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$, so that the posterior distribution for g conditioned on the observation data D_g and the test point \mathbf{x} also follows a Gaussian process:

$$p(g|D_g, \mathbf{x}, \eta) = \mathcal{GP}(g; m_g(\cdot), K_g(\cdot, \cdot)) \quad (6)$$

where

$$m_g(\mathbf{x}) = K(\mathbf{x}, \mathbf{X}_n)K(\mathbf{X}_n, \mathbf{X}_n)^{-1}\mathbf{g}_n,$$

$$K_g(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}, \mathbf{x}') - K(\mathbf{x}, \mathbf{X}_n)K(\mathbf{X}_n, \mathbf{X}_n)^{-1}K(\mathbf{X}_n, \mathbf{x}').$$

The parabolic transformation causes the distribution for any f to become a non-central χ^2 process, making the analysis intractable. In order to tackle this problem and obtain a posterior distribution $p(f|D_f, \mathbf{x}, \eta)$ that is also Gaussian, we employ a linearisation technique (Gunter et al., 2014).

We perform a local linearisation of the parabolic transformation $h(g) = \eta + 1/2g^2$ around g_0 and obtain $f \approx h(g_0) + h'(g_0)(g - g_0)$ where the gradient $h'(g) = g$. By setting g_0 to the mode of the posterior distribution $p(g|D_g, \mathbf{x}, \eta)$ (i.e. $g_0 = m_g$), we obtain an expression for f which is linear in g :

$$\begin{aligned} f(\mathbf{x}) &\approx [\eta + 1/2m_g(\mathbf{x})^2] + m_g(\mathbf{x})[g(\mathbf{x}) - m_g(\mathbf{x})] \\ &= \eta - 1/2m_g(\mathbf{x})^2 + m_g(\mathbf{x})g(\mathbf{x}). \end{aligned} \quad (7)$$

Since the affine transformation of a Gaussian process remains Gaussian, the predictive posterior distribution for f now has a closed form:

$$p(f|D_f, \mathbf{x}, \eta) = \mathcal{GP}(f; m_f(\cdot), K_f(\cdot, \cdot)) \quad (8)$$

where

$$\begin{aligned} m_f(\mathbf{x}) &= \eta + 1/2m_g(\mathbf{x})^2 \\ K_f(\mathbf{x}, \mathbf{x}') &= m_g(\mathbf{x})K_g(\mathbf{x}, \mathbf{x}')m_g(\mathbf{x}'). \end{aligned}$$

However, in real world situations, we do not have access to the true function values but only noisy observations of the function, $y(\mathbf{x}) = f(\mathbf{x}) + \epsilon$, where ϵ is assumed to be an independently and identically distributed Gaussian noise with variance σ_n^2 (Rasmussen and Williams, 2006). Given noisy observation data $D_n = \{(\mathbf{x}_i, y_i) | i = 1, \dots, n\} = \{\mathbf{X}_n, \mathbf{y}_n\}$, the predictive posterior distribution (8) becomes:

$$p(y|D_n, \mathbf{x}, \eta) = \mathcal{GP}(y; m_f(\cdot), K_f(\cdot, \cdot) + \sigma_n^2\delta(\cdot, \cdot)). \quad (9)$$

2.2. Hyperparameter Treatment

Hyperparameters are the free parameters, such as output scale and characteristic length scales in the kernel function for the Gaussian processes as well as noise variance. We use θ to represent a vector of hyperparameters that includes all the kernel parameters and the noise variance. Recall that we introduce a new hyperparameter η in our model to represent the global minimum. To ensure that η is not greater than the minimum observation y_{min} , we assume that $\log(y_{min} - \eta)$ follows a broad normal distribution. Thus the prior for η has the form:

$$p(\eta) = \frac{1}{(y_{min} - \eta)} \mathcal{N}(\log(y_{min} - \eta); \mu, \sigma^2). \quad (10)$$

The most popular approach to hyperparameter treatment is to learn hyperparameter values via maximum likelihood estimation (MLE) or maximum a posterior estimation (MAP). However, both MLE and MAP are not desirable as they give point estimates and ignore our uncertainty about the hyperparameters (Hernández-Lobato et al., 2014). In a fully Bayesian treatment of the hyperparameters, we should consider all possible hyperparameter values. This can be done by marginalising the terms in the acquisition function with respect to the posterior $p(\psi|D_n)$ where $\psi = \{\theta, \eta\}$:

$$\begin{aligned} \alpha_{\text{FITBO}}(\mathbf{x}|D_n) &= H \left[\int p(y|D_n, \mathbf{x}, \psi)p(\psi|D_n)d\psi \right] \\ &\quad - \int p(\psi|D_n)H[p(y|D_n, \mathbf{x}, \psi)]d\psi. \end{aligned}$$

Since complete marginalisation over hyperparameters is analytically intractable, the integral can be approximated using the Monte Carlo method (Hoffman and Ghahramani, 2015; Snoek et al., 2012), leading to the final expression:

$$\begin{aligned} \alpha_{\text{FITBO}}(\mathbf{x}|D_n) &= H \left[\frac{1}{M} \sum_j p(y|D_n, \mathbf{x}, \theta^{(j)}, \eta^{(j)}) \right] \\ &\quad - \frac{1}{2M} \sum_j \log [2\pi e(v_f(\mathbf{x}|D, \theta^{(j)}, \eta^{(j)}) + \sigma_n^2)]. \end{aligned} \quad (11)$$

2.3. Approximation for the Gaussian Mixture Entropy

The entropy of a Gaussian mixture is intractable and can be estimated via a number of methods: the Taylor expansion proposed in (Huber et al., 2008), numerical integration and Monte Carlo integration. Of these three, our experimentation revealed that numerical integration (in particular, an adaptive Simpson’s method) was clearly the most performant for our application (see the supplementary material). Note that our Gaussian mixture is univariate.

A faster alternative is to approximate the first entropy term by matching the first two moments of a Gaussian mixture. The mean and variance of a univariate Gaussian mixture model $p(z) = \sum_j^M \frac{1}{M} \mathcal{N}(z|m_j, K_j)$ have the analytical form:

$$\mathbb{E}[z] = \sum_j^M \frac{1}{M} m_j \quad (12)$$

$$\text{Var}(z) = \sum_j^M \frac{1}{M} (K_j + m_j^2) - \mathbb{E}[z]^2. \quad (13)$$

By fitting a Gaussian to the Gaussian mixture, we can obtain a closed-form upper bound for the first entropy term: $H[p(z)] \approx 0.5 \log [2\pi e (\text{Var}(z) + \sigma_\pi^2)]$, thus further enhancing the computational speed of FITBO approach. However, the moment-matching approach results in a looser approximation than numerical integration (shown in the supplementary material) and we will compare both approaches in our experiments in Section 3.

2.4. The Algorithm

The procedures of computing the acquisition function of FITBO are summarised by Algorithm 1. Figure 1 illustrates the sampling behaviour of FITBO method for a simple 1D Bayesian optimisation problem. The optimisation process is started with 3 initial observation data. As more samples are taken, the mean of the posterior distribution for the objective function gradually resembles the objective function and the distribution of η converges to the global minimum.

3. Experiments

We conduct a series of experiments to test the empirical performance of FITBO and compare it with other popular acquisition functions. In this section, FITBO denotes the version using numerical integration to estimate the entropy of the Gaussian mixture while FITBO-MM denotes the version using moment matching. In all experiments, we adopt a zero-mean Gaussian process prior with the squared exponential kernel function and use the elliptical slice sampler (Murray et al., 2010) for sampling hyperparameters θ and η . For the implementation of EI, PI, GP-UCB, MES

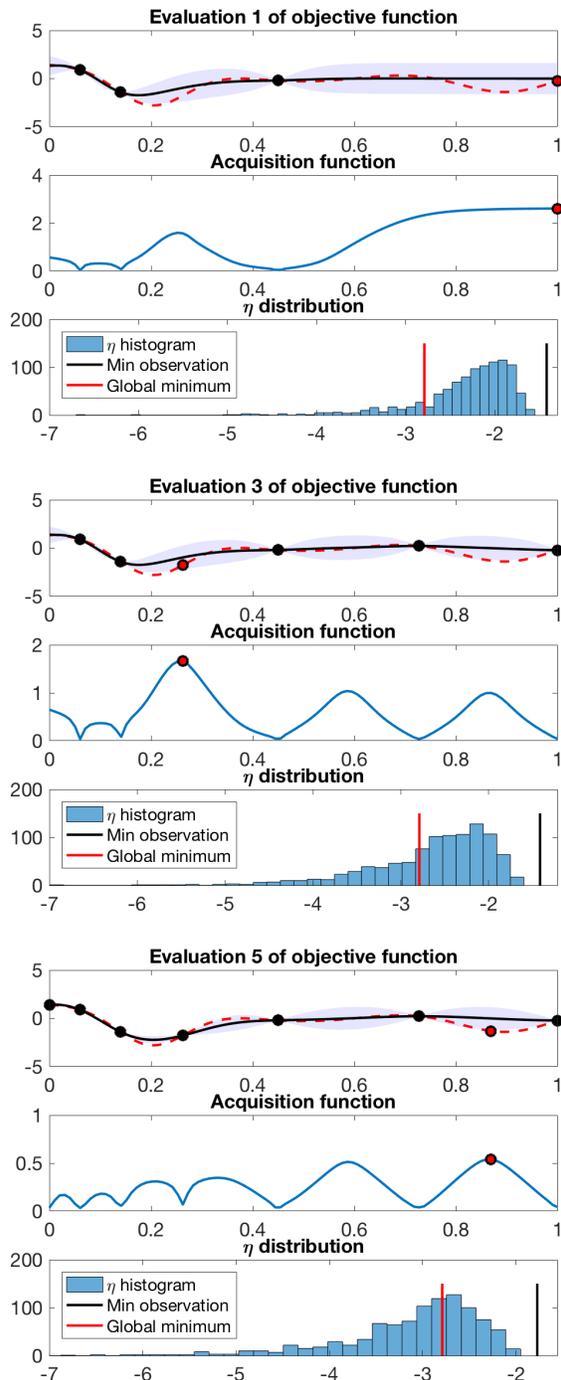


Figure 1. Bayesian optimisation for a 1D objective function using FITBO method at the 1st, 3rd, 5th evaluations. In each subfigure, the top plot shows the objective function (red dotted line), the posterior mean (black solid line) and the 95% confidence interval (blue shaded area) estimated by the Gaussian process model as well as the observation points (black dot) and the next query point (red dot). The middle plot shows the acquisition function. The bottom plot is the histogram of η samples as well as its relation to the minimum observation (black vertical line) and the true global minimum (red vertical line).

Algorithm 1 FITBO acquisition function

- 1: **Input:** a test input \mathbf{x} ; noisy observation data
 $D_n = \{(\mathbf{x}_i, y_i) | i = 1, \dots, n\}$
- 2: Sample hyperparameters and η from $p(\boldsymbol{\psi} | D_n)$:
 $\boldsymbol{\Psi} = \{\boldsymbol{\theta}^{(j)}, \eta^{(j)} | j = 1, \dots, M\}$
- 3: **for** $j = 1, \dots, M$ **do**
- 4: Use $f(\mathbf{x}) = \eta + 1/2g(\mathbf{x})^2$ to approximate
 $p(f | D_n, \mathbf{x}, \boldsymbol{\theta}^{(j)}, \eta^{(j)}) = \mathcal{GP}(m_f(\cdot), K_f(\cdot, \cdot))$
- 5: Compute $p(y | D_n, \mathbf{x}, \boldsymbol{\theta}^{(j)}, \eta^{(j)})$
- 6: Compute $H[p(y | D_n, \mathbf{x}, \boldsymbol{\theta}^{(j)}, \eta^{(j)})]$
- 7: **end for**
- 8: Estimate the entropy of the Gaussian mixture :
 $E_1(\mathbf{x} | D_n) = H\left[\frac{1}{M} \sum_j^M p(y | D_n, \mathbf{x}, \boldsymbol{\theta}^{(j)}, \eta^{(j)})\right]$
- 9: Compute the entropy expectation:
 $E_2(\mathbf{x} | D_n) = \frac{1}{M} \sum_j^M H[p(y | D_n, \mathbf{x}, \boldsymbol{\theta}^{(j)}, \eta^{(j)})] =$
 $\frac{1}{2M} \sum_j^M \log [2\pi e (v_f(\mathbf{x} | D_n, \boldsymbol{\theta}^{(j)}, \eta^{(j)}) + \sigma_n^2)]$
- 10: **return** $\alpha_n(\mathbf{x} | D_n) = E_1(\mathbf{x} | D_n) - E_2(\mathbf{x} | D_n)$

and PES, we use the open source Matlab code by Wang and Jegelka (2017) and Hernández-Lobato et al. (2014). Our Matlab code for FITBO will be available at <https://github.com/rubinxin/FITBO>. We use the type of MES method that samples the global minimum $f(\mathbf{x}_*)$ from an approximated posterior function $\tilde{f}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^T \tilde{\mathbf{a}}$ where $\boldsymbol{\phi}(\mathbf{x})$ is an m -dimensional feature vector and $\tilde{\mathbf{a}}$ is a Gaussian weight vector (Wang and Jegelka, 2017). This is also the minimiser sampling strategy adopted by PES (Hernández-Lobato et al., 2014). The computational complexity of sampling $\tilde{\mathbf{a}}$ from its posterior distribution $p(\tilde{\mathbf{a}} | D_n)$ is $\mathcal{O}(n^2m)$ when $n < m$ (Hernández-Lobato et al., 2014). Minimising $\tilde{f}(\mathbf{x})$ to within ζ accuracy using any grid search or branch and bound optimiser requires $\mathcal{O}(\zeta^{-d})$ calls to $\tilde{f}(\mathbf{x})$ for d -dimensional input data (Kandasamy et al., 2015). For both PES and MES, we apply their fastest versions which draw only 1 minimum or minimiser sample to estimate the acquisition function.

3.1. Runtime Tests

The first set of experiments measure and compare the runtime of evaluating the acquisition functions $\alpha_n(\mathbf{x} | D_n)$ for methods including GP-UCB, PI, EI, PES, MES, FITBO and FITBO-MM. All the timing tests were performed exclusively on a 2.3 GHz Intel Core i5. The runtime measured excludes the time taken for sampling hyperparameters as well as optimising the acquisition functions. The methodology of the tests can be summarised as follows:

1. Generate 10 initial observation data from a d -dimensional test function and sample a set of M hyperparameters $\boldsymbol{\Psi} = \{\boldsymbol{\theta}^{(j)}, \eta^{(j)} | j = 1, \dots, M\}$ from the log posterior distribution $\log \tilde{p}(\boldsymbol{\psi} | D_n)$ using the

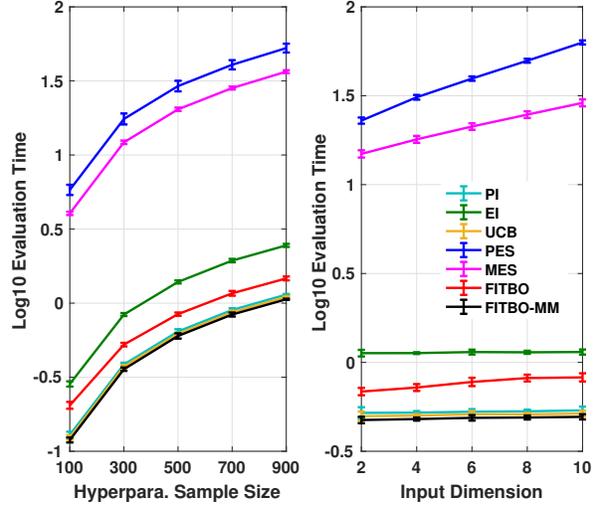


Figure 2. The runtime of evaluating 7 different acquisition functions (PI, EI, GP-UCB, PES, MES, FITBO and FITBO-MM) at 100 test inputs. The left plot shows the runtime of evaluating the acquisition functions using M hyperparameter samples for 2D input data and M tested are 100, 300, 500, 700, 900. The right plot shows the runtime of evaluating the acquisition functions using 400 hyperparameter samples for input data of dimension d where d are 2, 4, 6, 8, 10. The y-axes are the evaluation runtime expressed in the logarithm to the base 10.

elliptical slice sampler.

2. Use this set of hyperparameters to evaluate all acquisition functions at 100 test points.
3. Repeat the procedures 1 and 2 for 100 different initializations and compute the mean and standard deviation of the runtime taken for evaluating various acquisition functions.

We did not include the time for sampling η alone into the runtime of evaluating FITBO and FITBO-MM because η is sampled jointly with other hyperparameters and does not add to the overall sampling burden significantly. In fact, we have tested that sampling η by the elliptical slice sampler adds 0.09 seconds on average when drawing 2 000 samples and 0.93 seconds when drawing 20 000 samples. Note further that we will limit all methods to a fixed number of hyperparameter samples in both runtime tests and performance experiments: this will impart a slight performance penalty to our method, which must sample from a hyperparameter space of one additional dimension.

The above tests are repeated for different hyperparameter sample sizes $M = 100, 300, 500, 700, 900$ and input data of different dimensions $d = 2, 4, 6, 8, 10$. The results are presented graphically in Figure 2 with the

Table 1. Runtime of evaluating PI, GP-UCB and FITBO-MM at 100 2D inputs using M hyperparameter samples (Unit: Second).

M	PI	GP-UCB	FITBO-MM
100	0.1293 (± 0.006)	0.1238 (± 0.005)	0.1193 (± 0.005)
300	0.3856 (± 0.011)	0.3731 (± 0.010)	0.3582 (± 0.009)
500	0.6442 (± 0.025)	0.6205 (± 0.012)	0.6011 (± 0.027)
700	0.8990 (± 0.026)	0.8670 (± 0.026)	0.8382 (± 0.028)
900	1.1426 (± 0.011)	1.1025 (± 0.014)	1.0618 (± 0.010)

Table 2. Runtime of evaluating PI, GP-UCB and FITBO-MM for 100 test inputs of dimension d with $M = 400$ (Unit: Second).

d	PI	GP-UCB	FITBO-MM
2	0.5217 (± 0.047)	0.4991 (± 0.034)	0.4745 (± 0.021)
4	0.5215 (± 0.011)	0.5020 (± 0.010)	0.4800 (± 0.010)
6	0.5281 (± 0.019)	0.5112 (± 0.023)	0.4879 (± 0.019)
8	0.5307 (± 0.011)	0.5102 (± 0.010)	0.4899 (± 0.013)
10	0.5378 (± 0.029)	0.5159 (± 0.019)	0.4942 (± 0.017)

evaluation runtime being expressed in the logarithm to the base 10 and the exact numerical results for methods that are very close in runtime are presented in Tables 1 and 2.

Figure 2 shows that FITBO is significantly faster to evaluate than PES and MES for various hyperparameter sample sizes used and for problems of different input dimensions. Moreover, FITBO even gains a clear speed advantage over EI. The moment matching technique manages to further enhance the speed of FITBO, making FITBO-MM comparable with, if not slightly faster than, simple algorithms like PI and GP-UCB. In addition, we notice that the runtime of evaluating FITBO-MM, EI, PI and GP-UCB tend to remain constant regardless of the input dimensions while the runtime for PES and MES tends to increase with input dimensions at a rate of 10^d . Thus, our

approach is more efficient and applicable in dealing with high-dimensional problems.

3.2. Tests with Benchmark Functions

We perform optimisation tasks on three challenging benchmark functions: Branin (defined in $[0, 1]^2$), Eggholder (defined in $[0, 1]^2$) and Hartmann (defined in $[0, 1]^6$). In all tests, we set the observation noise to $\sigma_n^2 = 10^{-3}$ and re-sample all the hyperparameters after each function evaluation. In evaluating the optimisation performance of various Bayesian optimisation methods, we use the two common metrics adopted by Hennig and Schuler (2012). The first metric is Immediate regret (IR) which is defined as:

$$IR = |f(\mathbf{x}_*) - f(\hat{\mathbf{x}}_n)| \quad (14)$$

where \mathbf{x}_* is the location of true global minimum and $\hat{\mathbf{x}}_n$ is the best guess recommended by a Bayesian optimiser after n iterations, which corresponds to the minimiser of the posterior mean. The second metric is the Euclidean distance of an optimiser’s recommendation $\hat{\mathbf{x}}_n$ from the true global minimiser \mathbf{x}_* , which is defined as:

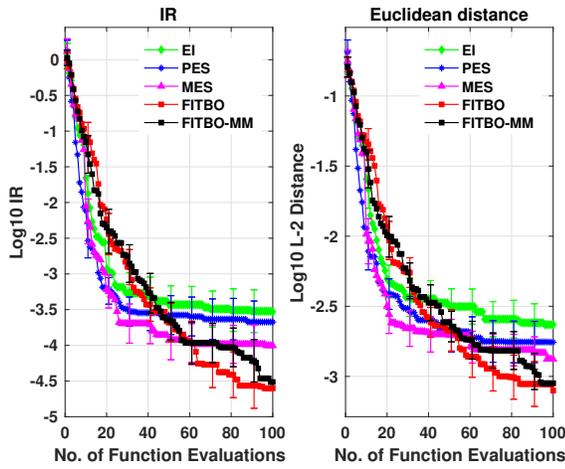
$$\|L\|_2 = \|\mathbf{x}_* - \hat{\mathbf{x}}_n\|. \quad (15)$$

We compute the median IR and the median $\|L\|_2$ over 40 random initialisations. At each initialisation, all Bayesian optimisation algorithms start from 3 random observation data for Branin-2D and Eggholder-2D problems and from 9 random observation data for Hartmann-6D problem.

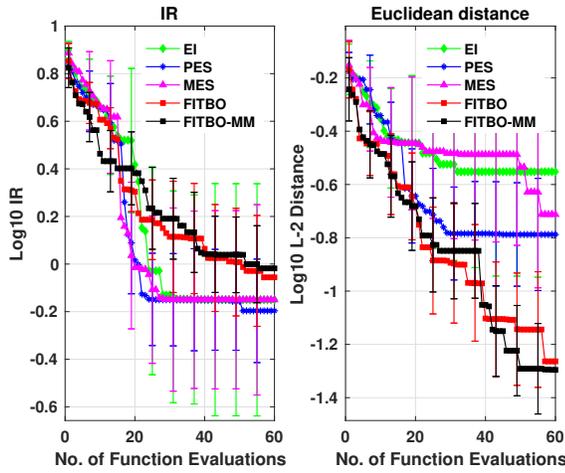
The results are presented in Figure 3. The plots on the left show the median IR achieved by each approach as more evaluation steps are taken. The plots on the right show the median $\|L\|_2$ between each optimiser’s recommended global minimiser and the true global minimiser. The error bars indicate one standard deviation.

In the case of Branin-2D, FITBO and FITBO-MM lose out to other methods initially but surpass other methods after 50 evaluations. One interesting point we would like to illustrate through the Branin problem is the fundamentally different mechanisms behind information-based approaches like FITBO and improvement-based approaches like EI. As shown in Figure 4, FITBO is much more explorative compared to EI in taking new evaluations because FITBO selects the query points that maximise the information gain about the minimiser instead of those that lead to an improvement over the best function value observed. FITBO successfully finds all three global minimisers but EI quickly concentrates its searches into regions of low function values, missing out one of the global minimisers.

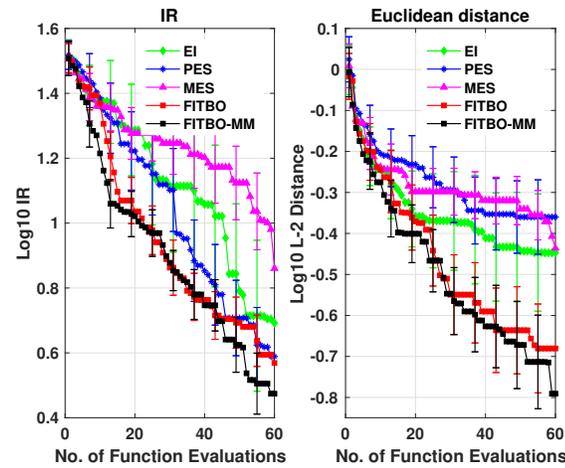
In the case of Eggholder-2D which is more complicated and multimodal, FITBO and FITBO-MM perform



(a) Branin-2D



(b) Eggholder-2D



(c) Hartmann-6D

Figure 3. Optimisation performance of EI, PES, MES, FITBO and FITBO-MM for three benchmark test functions.

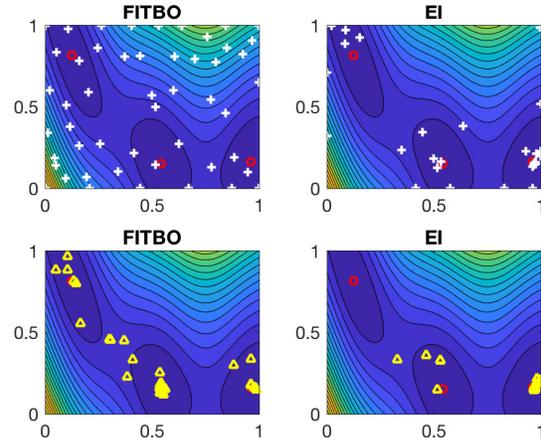


Figure 4. Evaluations taken by FITBO and EI in the Branin-2D problem. The white crosses in the top plots indicate the first 50 query points recommended by the two algorithms. The yellow triangles in the bottom plots indicate the guesses of the global minimiser recommended by the algorithms (i.e. \hat{x}_n) after each evaluation. FITBO, which is more explorative in taking evaluations, successfully identifies all three global minimisers (red circle) but EI misses out one of the global minimisers.

not as well as other methods in finding lower function values but outperform all competitors in locating the global minimiser by a large margin. One reason is that the function value near the global minimiser of Eggholder-2D rises sharply. Thus, although FITBO and FITBO-MM are able to better identify the location of the true global minimum, they return higher function values than other methods that are trapped in locations of good local minima.

As for a higher dimensional problem, Hartmann-6D, FITBO and FITBO-MM outperform all other methods in finding both the lower function value and the location of the global minimum. In all three tasks, FITBO-MM, despite using a looser upper bound of the Gaussian mixture entropy, still manages to demonstrate similar, sometimes better, results compared with FITBO. This shows that the performance of our information-theoretic approach is robust to slightly worse approximation of the Gaussian mixture entropy.

3.3. Test with Real-world Problems

Finally, we experiment with a series of real-world optimisation problems. The first problem (Boston) returns the L2 validation loss of a 1-hidden layer neural network (Wang and Jegelka, 2017) on the Boston housing dataset (Bache and Lichman, 2013). The dataset is randomly partitioned into train/validation/test sets and the neural network is trained with Levenberg-Marquardt optimisation. The 2 variables tuned with Bayesian optimisation are the number

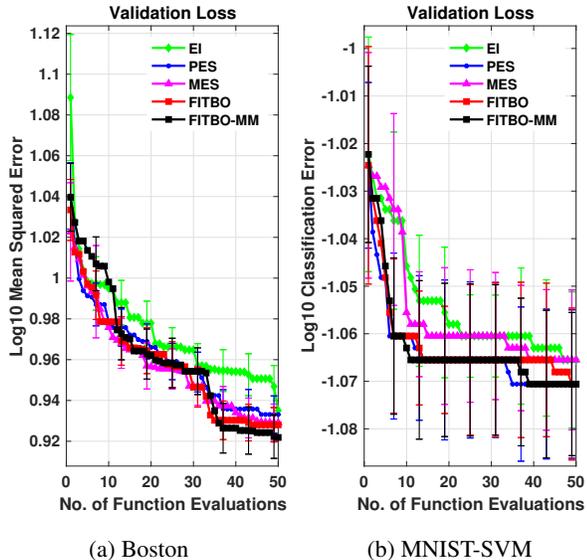


Figure 5. Performance on tuning hyperparameters for (a) training a neural network on the Boston housing dataset and (b) training an SVM on the MNIST dataset.

of neurons and the damping factor μ .

The second problem (MNIST-SVM) outputs the classification error of a support vector machine (SVM) classifier on the validation set of the MNIST dataset (LeCun et al., 1998). The SVM classifier adopts a radial basis kernel and the 2 variables to optimise are the kernel scale parameter and the box constraint.

The third problem (Cancer) returns the cross-entropy loss of a 1-hidden layer neural network (Wang and Jegelka, 2017) on the validation set of the breast cancer dataset (Bache and Lichman, 2013). This neural network is trained with the scaled conjugate gradient method and we use Bayesian optimisation methods to tune the number of neurons, the damping factor μ , the μ -increase factor and the μ -decrease factor.

We initialise all Bayesian optimisation algorithms with 3 random observation data and set the observation noise to $\sigma_n^2 = 10^{-3}$. All experiments are repeated 40 times. In each case, the ground truth is unknown but our aim is to minimise the validation loss. Thus, the corresponding loss functions are used to compare the performance of various Bayesian optimisation algorithms.

Figure 5 shows the median of the best validation losses achieved by all Bayesian optimisation algorithms after n iterations for the Boston and MNIST-SVM problems. Our FITBO and FITBO-MM perform competitively well compared to their information-theoretic counterparts and

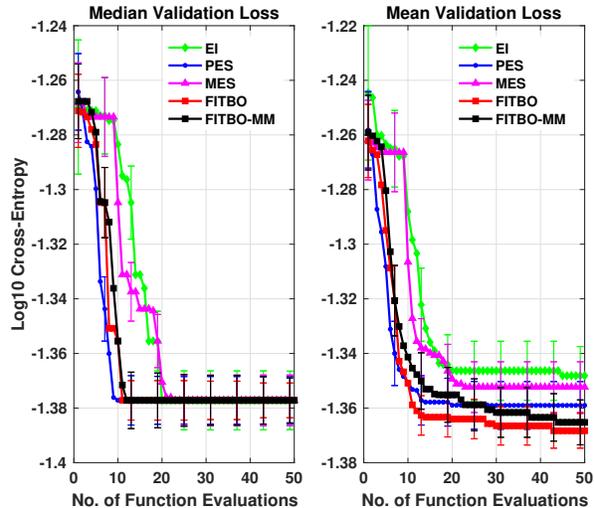


Figure 6. Performance on tuning hyperparameters for training a classification neural network on the breast cancer dataset.

all information-theoretic methods outperform EI in these real-world applications.

As for the Cancer problem (Figure 6), FITBO and FITBO-MM converge to the stable median value of the validation loss at a much faster speed than MES and EI and are almost on par with PES. By examining the mean validation loss shown in the right plot of Figure 6, it is evident that both FITBO and FITBO-MM demonstrate better performance than all other methods on average with FITBO gaining a slight advantage over FITBO-MM. Moreover, the comparable performance of FITBO and FITBO-MM in all three real-world tasks re-affirmed the robustness of our approach to entropy approximation as our moment matching technique, while improving the speed of the algorithm, does not really compromise the performance.

4. Conclusion

We have proposed a novel information-theoretic approach for Bayesian optimisation, FITBO. With the creative use of the parabolic transformation and the hyperparameter η , FITBO enjoys the merits of less sampling effort, more flexible kernel choices and much simpler implementation in comparison with other information-based methods like PES and MES. As a result, its computational speed outperforms current information-based methods by a large margin and even exceeds EI to be on par with PI and GP-UCB. While requiring much lower runtime, it still manages to achieve satisfactory optimisation performance which is as good as or even better than PES and MES in a variety of tasks. Therefore, FITBO approach offers a very efficient and competitive alternative to existing Bayesian optimisation approaches.

Acknowledgements

We wish to thank Roman Garnett and Tom Gunter for the insightful discussions and Zi Wang for sharing the Matlab implementation of EI, PI, GP-UCB, MES and PES. We would also like to thank Favour Mandanji Nyikosa, Logan Graham, Arno Blaas and Olga Isupova for their helpful comments about improving the paper.

References

- K. Bache and M. Lichman. UCI machine learning repository. 2013.
- S. Bochner. *Lectures on Fourier Integrals: With an Author's Suppl. on Monotonic Functions, Stieltjes Integrals and Harmonic Analysis. Transl. from the Orig. by Morris Tennenbaum and Harry Pollard.* University Press, 1959.
- E. Brochu, V. M. Cora, and N. 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.
- T. Gunter, M. A. Osborne, R. Garnett, P. Hennig, and S. J. Roberts. Sampling for inference in probabilistic models with fast Bayesian quadrature. In *Advances in neural information processing systems*, pages 2789–2797, 2014.
- P. Hennig and C. J. Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(Jun):1809–1837, 2012.
- J. 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.
- M. W. Hoffman and Z. Ghahramani. Output-space predictive entropy search for flexible global optimization. In *the NIPS workshop on Bayesian optimization*, 2015.
- M. F. Huber, T. Bailey, H. Durrant-Whyte, and U. D. Hanebeck. On entropy approximation for Gaussian mixture random vectors. In *Multisensor Fusion and Integration for Intelligent Systems, 2008. MFI 2008. IEEE International Conference on*, pages 181–188. IEEE, 2008.
- 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.
- K. Kandasamy, J. Schneider, and B. Póczos. High dimensional Bayesian optimisation and bandits via additive models. In *International Conference on Machine Learning*, pages 295–304, 2015.
- 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.
- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- J. Moćkus, V. Tiesis, and A. Žilinskas. Toward global optimization, volume 2, chapter the application of Bayesian methods for seeking the extremum, 1978.
- I. Murray, R. Prescott Adams, and D. J. MacKay. Elliptical slice sampling. 2010.
- C. E. Rasmussen and C. K. Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.
- J. R. Requeima. Integrated predictive entropy search for Bayesian optimization. 2016.
- B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.
- J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.
- N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.
- J. Villemonteix, E. Vazquez, and E. Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44(4):509–534, 2009. URL <http://www.springerlink.com/index/T670U067V47922VK.pdf>.
- Z. Wang and S. Jegelka. Max-value entropy search for efficient Bayesian optimization. *arXiv preprint arXiv:1703.01968*, 2017.