Fast Information-theoretic Bayesian Optimisation

Binxin Ru
University of Oxford
Oxford OX1 3PJ
robin@robots.ox.ac.uk

Mark Mcleod
University of Oxford
Oxford OX1 3PJ
mark.mcleod@magd.ox.ac.uk

Michael Osborne
University of Oxford
Oxford OX1 3PJ
mosb@robots.ox.ac.uk

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; limit the choice of kernels available to model the objective; and introduce often-prohibitive computational overhead. We develop a fast information-theoretic Bayesian optimisation method, FITBO, that circumvents 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

Bayesian optimisation is a powerful tool to tackle optimisation challenges [1] whose objective functions are unknown, nonconvex and expensive to evaluate [10]. A core step in Bayesian optimisation is to definite an acquisition function which uses the available observations effectively to recommend the next query location [10]. There are many types of acquisition functions such as Probability of Improvement (PI) [8], Expected Improvement (EI) [7] and Gaussian Process Upper Confidence Bound (UCB) [11]. 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 $\mathbf{x}_*$ [10]. Such methods have demonstrated impressive empirical performance and tend to outperform traditional methods in tasks with highly multimodal and noisy latent functions [4].

One popular information-based acquisition function is Predictive Entropy Search (PES) [12, 3, 4]. However, PES is very slow to evaluate and faces serious constraints on kernel choices. Moreover, PES deals with the input space, thus less efficient in higher dimensional problems [13]. The more recent methods such as Output-space Predictive Entropy Search (OPES) [5] and Max-value Entropy Search (MES) [13] improve on PES by focusing on the information content in output space. However, all current entropy search methods, being it dealing with minimiser or minimum value, need two separate sampling processes: 1) sampling hyperparameters for marginalisation and 2) sampling the global minimum 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 latent function [4], which introduces some kernel constraints.

In view of the limitations of the existing methods, we propose a fast information-theoretic Bayesian optimisation technique, FITBO. Inspired by the Bayesian quadrature work in [2], the creative contribution of our technique is to approximate any black-box function in a parabolic form:

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

$f(\mathbf{x}) = \eta + {}^1\!/\!_2 g(\mathbf{x})^2$. The global minimum is then 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. First, our approach circumvents the need to sample the global minimiser/minimum, thus saving much sampling effort and speeding up the evaluation of acquisition function tremendously. Second, our approach faces fewer constraints on the choice of appropriate kernel functions for the Gaussian process prior. Third, similar to MES [13], our approach works on information in the output space, thus 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 $p(\mathbf{x}_*|D_n)$ [3]. The acquisition function for such techniques has the form [3] [4]:

$$\alpha_{ES}(\mathbf{x}|D_n) = H[p(\mathbf{x}_*|D_n)] - \mathbb{E}_{p(y|D_n,\mathbf{x})}\Big[H\big[p\big(\mathbf{x}_*|D_n \cup \{(\mathbf{x},y)\}\big)\big]\Big]. \tag{1}$$

PES makes use of the symmetry of mutual information and turns the function (1) to the form:

$$\alpha_{PES}(\mathbf{x}|D_n) = H[p(y|D_n,\mathbf{x})] - \mathbb{E}_{p(\mathbf{x}_*|D_n)}\Big[H\big[p(y|D_n,\mathbf{x},\mathbf{x}_*)\big]\Big], \tag{2}$$

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

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 [13]. In other words, FITBO aims to select the next query point which minimises the entropy of the global minimum:

$$\alpha_{FITBO}(\mathbf{x}|D_n) = H[p(y|D_n,\mathbf{x})] - \mathbb{E}_{p(f_*|D_n)}\Big[H\big[p(y|D_n,\mathbf{x},f_*)\big]\Big]. \tag{3}$$

This idea of changing entropy computation from the input space to the output space is also shared by [5] and [13]. Hence, the acquisition function of the FITBO method is very similar to those of OPES [5] and MES [13]. However, our novel contribution is to express the unknown objective function in a parabolic form: $f(\mathbf{x}) = \eta + {}^1\!/\!_2 g(\mathbf{x})^2$, thus representing the global minimum f_* explicitly by a hyperparameter η. FITBO acquisition function can then be reformulated as:

$$\begin{aligned}\alpha_{FITBO}(\mathbf{x}|D_n) &= H[p(y|D_n,\mathbf{x})] - \mathbb{E}_{p(\eta|D_n)}\Big[H\big[p(y|D_n,\mathbf{x},\eta)\big]\Big] \\ &= H\Big[\int p(y|D_n,\mathbf{x},\eta)p(\eta|D_n)\mathrm{d}\eta\Big] - \int p(\eta|D_n)H\big[p(y|D_n,\mathbf{x},\eta)\big]\mathrm{d}\eta. \end{aligned}\tag{4}$$

The intractable integral terms can be approximated via Monte Carlo method [4]. 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 approximation as described in Section 2.1. Then, the first term in the above FITBO acquisition function becomes an entropy of a Gaussian mixture which can be approximated as described in Section 2.2. The second term can be computed analytically because the entropy of a Gaussian has the closed form: $H[p(y|D_n,\mathbf{x},\eta)] = 0.5\log\big[2\pi e\big(v_f(\mathbf{x}|D_n,\eta) + \sigma_n^2\big)\big]$ 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 Approximation and Predictive Posterior Distribution

The warped sequential active Bayesian integration method [2] uses a square-root transformation on the integrand to ensure non-negativity. Inspired by this work, we creatively express any unknown objective function $f(\mathbf{x})$ in the parabolic form:

$$f(\mathbf{x}) = \eta + {}^1\!/\!_2 g(\mathbf{x})^2, \tag{5}$$

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 $g(\mathbf{x})$ 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: $p(g|D_g, \mathbf{x}, \eta) = \mathcal{GP}\big(g; m_g(\cdot), K_g(\cdot,\cdot)\big)$ where $m_g(\mathbf{x}) = K(\mathbf{x}, \mathbf{X}_n)K(\mathbf{X}_n, \mathbf{X}_n)^{-1}\mathbf{g}_n$, $K_g(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}, \mathbf{x}') - K(\mathbf{x}, \mathbf{X}_n)K(\mathbf{X}_n, \mathbf{X}_n)^{-1}K(\mathbf{X}_n, \mathbf{x}')$.

Due to the parabolic transformation, the distribution for any f is now a non-central χ^2 distribution, 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 resort to the linearisation technique proposed in [2]. We perform a local linearisation of the parabolic transformation $h(g) = \eta + {}^1\!/_2 g^2$ 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 that is linear in g:

$$f(\mathbf{x}) \approx [\eta + {}^1\!/_2 m_g(\mathbf{x})^2] + m_g(\mathbf{x})[g(\mathbf{x}) - m_g(\mathbf{x})] = \eta - {}^1\!/_2 m_g(\mathbf{x})^2 + m_g(\mathbf{x})g(\mathbf{x}). \tag{6}$$

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

$$p(f|D_f, \mathbf{x}, \eta) = \mathcal{GP}\big(f; m_f(\cdot), K_f(\cdot,\cdot)\big) \tag{7}$$

where $m_f(\mathbf{x}) = \eta + {}^1\!/_2 m_g(\mathbf{x})^2$, $K_f(\mathbf{x}, \mathbf{x}') = m_g(\mathbf{x})K_g(\mathbf{x}, \mathbf{x}')m_g(\mathbf{x}')$.

However, in real world situation, 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 [9]. 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 (7) becomes:

$$p(y|D_n, \mathbf{x}, \eta) = \mathcal{GP}\big(y; m_f(\cdot), K_f(\cdot,\cdot) + \sigma_n^2\delta(\cdot,\cdot)\big). \tag{8}$$

2.2 Approximation for the Entropy of A Gaussian Mixture

The entropy of our univariate Gaussian mixture is intractable and can be estimated via a number of methods: the Taylor expansion proposed in [6], 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 Supplementary Material).

Alternatively, we can approximate using moment matching. 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: $\mathbb{E}[z] = \sum_j^M \frac{1}{M}m_j$, $Var(z) = \sum_j^M \frac{1}{M}(K_j + m_j^2) - \mathbb{E}[z]^2$. By fitting a Gaussian to the Gaussian mixture, the first entropy term can be approximated with an analytical expression: $H[p(z)] \approx 0.5 \log\big[2\pi e\big(Var(z) + \sigma_n^2\big)\big]$. We will compare numerical integration (FITBO) and moment-matching approaches (FITBO-MM) in our experiments in Section 3.

3 Experiments

We conduct two sets of experiments to test the empirical performance of FITBO and compare it with other popular acquisition functions. The first set of experiments measure and compare the runtime of evaluating different acquisition functions $\alpha_n(\mathbf{x}|D_n)$. The runtime measured excludes the time taken for sampling hyperparameters and optimising the acquisition functions. For the experiments, we take 10 initial observation data from a N-D test function to sample a set of M hyperparameters $\boldsymbol{\psi} = \{\boldsymbol{\theta}_i, \eta_i|i = 1, \dots, M\}$ from $\log p(\boldsymbol{\psi}|D_n)$ and use the set of hyperparameters to evaluate all acquisition functions. Finally, we compute the mean and standard deviation of the runtime taken for evaluating various acquisition functions at one test point. Figure 1 shows that FITBO is significantly faster to evaluate than PES and MES and the moment matching technique manages to further enhance the speed of FITBO by a large margin, making FITBO-MM faster than EI and comparable with PI and GP-UCB. 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 cause significant increase in the sampling burden. 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.

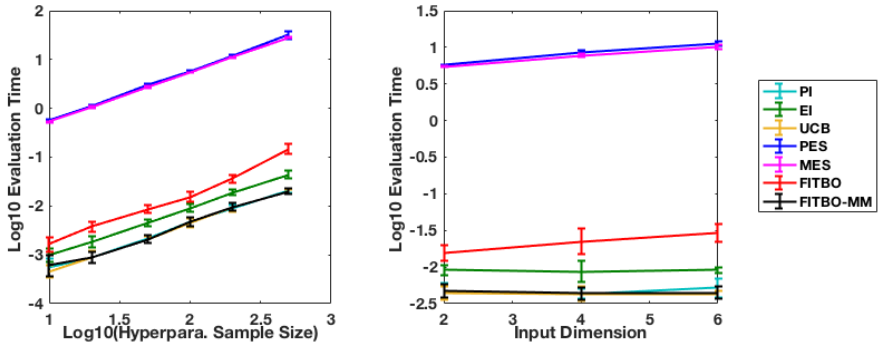
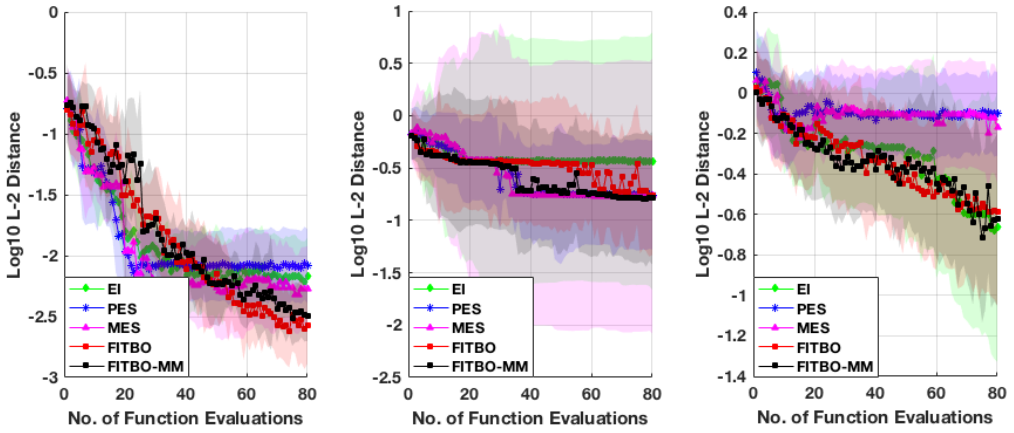


Figure 1: The runtime of evaluating PI, EI, UCB, PES, MES, FITBO and FITBO-MM. The left plot shows the runtime for using different hyperparameter sample sizes (M = 10, 20, 50, 100, 200, 500). The right plot shows the runtime for test point data of different dimensions (N=2,4,6).

The second set of experiments perform optimisation tasks on three benchmark functions. In all tests, we set the observation noise to $\sigma_n^2 = 10^{-3}$ and resample all the hyperparameters after each function evaluation.The results in Figure 2 show that FITBO and FITBO-MM outperform the other two information-theoretical approaches in the problems of Brain 2D and Hartman 6D while performing comparably well as PES and MES in the case of Eggholder 2D.



(a) Branin 2D (b) Eggholder 2D (c) Hartman 6D

Figure 2: Optimisation results for benchmark test functions in terms of the median Euclidean distance between the predicted global minimiser $\hat{\mathbf{x}}_n$ and the true global minimiser $\mathbf{x}_*$: $\|L\|_2 = \|\hat{\mathbf{x}}_n - \mathbf{x}_*\|$.

4 Conclusion

We have proposed a novel information-theoretic approach for Bayesian optimisation, FITBO. With the creative use of the parabolic approximation and the hyperparameter η, FITBO enjoys the merits of less sampling effort 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 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 competitive alternative to existing Bayesian optimisation approaches.

References

[1] 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.

[2] Tom Gunter, Michael A Osborne, Roman Garnett, Philipp Hennig, and Stephen J Roberts. Sampling for inference in probabilisktic models with fast Bayesian quadrature. In *Advances in neural information processing systems*, pages 2789–2797, 2014.

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

[4] José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in neural information processing systems*, pages 918–926, 2014.

[5] Matthew W. Hoffman and Zoubin Ghahramani. Output-space predictive entropy search for flexible global optimization. In *the NIPS workshop on Bayesian optimization*, 2015.

[6] Marco F Huber, Tim Bailey, Hugh Durrant-Whyte, and Uwe 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.

[7] 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.

[8] Harold J Kushner. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering*, 86(1):97–106, 1964.

[9] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.

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

[11] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

[12] Julien Villemonteix, Emmanuel Vazquez, and Eric Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44(4):509–534, 2009.

[13] Zi Wang and Stefanie Jegelka. Max-value entropy search for efficient Bayesian optimization. *arXiv preprint arXiv:1703.01968*, 2017.