Practical Bayesian Optimization for Variable Cost Objectives

Mark McLeod MARKM@ROBOTS.OX.AC.UK
Department of Engineering Science, University of Oxford

Michael A. Osborne MOSB@ROBOTS.OX.AC.UK
Stephen J. Roberts SJROB@ROBOTS.OX.AC.UK
Oxford-Man Institute of Quantitative Finance
Department of Engineering Science, University of Oxford

Abstract

We propose a novel Bayesian Optimization approach for black-box functions with an environmental variable whose value determines the tradeoff between evaluation cost and the fidelity of the evaluations. Further, we use a novel approach to sampling support points, allowing faster construction of the acquisition function. This allows us to achieve optimization with lower overheads than previous approaches and is implemented for a more general class of problem. We show this approach to be effective on synthetic and real world benchmark problems.

1. Introduction

In the context of Bayesian optimization, we wish to find the minimum of a function $x_* = \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$. This function is considered to be multimodal and costly to evaluate, either due to the computation required, or because an expensive physical experiment must be undertaken. We therefore wish to make as few evaluations as possible and are willing to expend non-trivial computation time on determining the next point to evaluate. There is a significant body of literature addressing this problem, with applications including tuning the hyperparameters of many machine learning algorithms (Snoek et al., 2012; Hernández-Lobato et al., 2014; Klein et al., 2015), sensor set selection (Garnett et al., 2010) and tuning the gait parameters of bipedal (Peters & Deisenroth, 2014), quadrupedal (Lizotte et al., 2007) and snake (Tesch et al., 2011) robots.

A common approach to the problem of choosing the next step is to construct a model of the objective using all previous observations. In the above applications, the model used is a Gaussian Process (GP) (Rasmussen, 2006), which provides a normally distributed estimate of the objective function at any point given training data. We then define an acquisition function, $\alpha(x)$, which quantifies how useful evaluation at x is expected to be. Since this acquisition function can be evaluated on the GP estimate rather than the expensive objective we can perform a global search to find its maximum value and choose that point as our next evaluation of the objective. We provide a brief overview and propose an alternative metric for evaluating the performance of Bayesian Optimization methods in §2.

In this work we consider an extension of the expensive optimization problem by considering an additional variable, s, an environmental variable (Klein et al., 2015). The choice of this variable allows the objective to be evaluated with reduced accuracy at a lower cost. By carefully selecting a value for s at each step, we wish to further reduce the cost of finding the minimum of the full cost objective. Previous approaches to the environmental variable setting (Klein et al., 2015; Swersky et al., 2013) are based on the Entropy search method (Hennig & Schuler, 2012). We adapt Predictive Entropy Search (Hernández-Lobato et al., 2014) to form a novel acquisition function over the environmental variable in §3, and show a novel method of selecting sample points to reduce overheads which is applicable to both Entropy Search and Predictive Entropy Search in §4. Our computationally efficient approach allows the practical optimization of objectives that have hitherto proved infeasible. We show results for common optimization benchmarks and real world applications in §5.

2. Bayesian Optimization

2.1. Gaussian Processes

Gaussian Processes are a standard tool for making inference of a function value with uncertainty given a set of observations. A good introductory text is (Rasmussen, 2006). The model is characterized by a kernel function

$k(x_1, x_2 \mid \theta)$ where θ are hyperparameters. Common choices are the squared exponential kernel $A\exp(-0.5r^2)$, where $r = \sqrt{\sum_{d\in D}\frac{(x_{2d}-x_{1d})^2}{h_d}}$, which models infinitely smooth functions, and the Matérn 3/2 and 5/2 kernels $A(1+\sqrt{3}r)\exp(-\sqrt{3}r)$ and $A(1+\sqrt{5}r+\frac{5}{3}r^2)\exp(-\sqrt{5}r)$ giving once and twice differentiable functions respectively. In the experiments below we have used the Matérn 5/2 kernel which is a common choice in Bayesian optimization.

Ideally we would marginalize over the hyperparameters A and h to obtain posterior estimates

$$m(x \mid D) = \int m(x \mid D, \lambda)p(\lambda)\mathrm{d}\lambda, \tag{1}$$

given some prior $p(\lambda)$ to obtain the full posterior mean function.

However, this is not usually an analytic function so cannot be achieved exactly. Instead, we use slice sampling (Neal, 2003) of the hyperparameters to approximate the posterior as

$$m(x \mid D) = \frac{1}{K}\sum_{k=0}^{K} m(x \mid D, \lambda_k) \tag{2}$$

where the K draws of the hyperparameter values have been made by slice sampling of their posterior likelihood given the data observed so far.

2.2. Acquisition Functions

A selection of acquisition functions are available, Probability of improvement over the current best observation (Lizotte, 2008; Kushner, 1964), the Expectation of Improvement (Jones et al., 1998; Močkus, 1975), and a Lower Confidence Bound (Srinivas et al., 2009) on the GP are common and simple choices. We consider the Entropy Search (ES) acquisition function proposed by (Hennig & Schuler, 2012), specifically the Predictive Entropy Search (PES) acquisition of (Hernández-Lobato et al., 2014), which is a fast approximation to ES.

In Entropy Search, the optimization is viewed not as finding locations of progressively lower values of the objective, but as gaining knowledge about the location of the global minimum. Specifically, prior belief about the location of the global minimum is represented as a probability distribution, $p(x_*)$, the probability that $x_* = \mathrm{argmax}_x f(x)$. We desire to maximize the relative entropy (KL-divergence) of this distribution from the uniform distribution. This occurs when $p(x_*)$ is a delta located at x_{min}. Therefore the ES acquisition function selects points to produce greedy maximization of the mutual information between x_* and y,

$$\begin{aligned} x_{n+1} =& \mathrm{argmax}_x\big(H\big[p(x_* \mid D_n)\big] - \\ & \mathbb{E}_{x_*}\big[H[p(x_* \mid D_n, x, y)]\big]\big). \end{aligned} \tag{3}$$

2.3. Predictive Entropy Search

The procedure for implementing ES requires considerable computation to achieve a good approximation to the ideal acquisition function. PES seeks a fast approximation to the ES acquisition. This is achieved by noting that the mutual information between the location x_* and the next observed values y_{n+1} is given by

$$\begin{aligned} & I[x_*, y_{n+1} \mid D_n, x_{n+1}] \\ =& H[x_* \mid D_n] - \mathbb{E}_{y_{n+1}}\big[H[x_* \mid D_n, x_{n+1}, y_{n+1})]\big] \\ =& H[y \mid D_n, x] - \mathbb{E}_{x_*}\big[H[y \mid D_n, x, x_*]\big] \end{aligned} \tag{4}$$

That is, the information gained about the location of x_{min} by evaluating at x_{n+1} is equal to the information gained about the value of $f(x_{n+1})$ given the true location of the global minimum. The acquisition function, $\alpha(x)$ is the expected information gain about the value at x_{n+1} given a true observation of the global minimum.

$$\begin{aligned} \alpha(x_{n+1}) =& \Delta H \\ =& H[y_{n+1} \mid x_{n+1}, D_n] - H[y_{n+1} \mid D_n, x_{n+1}, x_*] \end{aligned} \tag{5}$$

To implement this variation a draw x_d is from the distribution over the location of minimum given the current model, and at this location the minimizing conditions,

$$f(x_d) \leq \min(Y_n), \tag{6}$$

$$\frac{\partial f(x_d)}{\partial x_i} = 0 \quad \forall i \qquad \text{and} \tag{7}$$

$$\frac{\partial^2 f(x_d)}{\partial x_i \partial x_j} \begin{cases} = 0 & i \neq j \\ \geq 0 & i = j \end{cases} \tag{8}$$

are imposed. Expectation Propagation (Minka, 2001) is used to achieve a Gaussian approximation to the inequalities. The change in entropy of y at a candidate x_{n+1}, averaged over draws of the minimum, is a good approximation to the ideal objective. We use a bespoke GP package[1] to accommodate observation and inference of first and second derivatives.

2.4. Evaluation of Performance

Since the point determined by the acquisition function is different from the posterior minimum the sequence of points evaluated does not represent the best guess for the global minimizer at each step. When using acquisition functions based on the values observed, such as expected improvement, these points do still provide good results. However, when using entropy-based methods, we find the points evaluated tend to be far from the posterior minimum. We therefore propose a greedy evaluation at the posterior

[1] To be released.

minimum as the final step of optimization. In the experiments below, we perform this evaluation offline at each step and report the immediate regret (IR) that would have been returned if that step had been the last, before continuing according the regular optimization policy. The difference in performance by taking this approach is illustrated in Figure 1. We argue that all information theoretic means of Bayesian Optimization should be evaluated according to this metric.

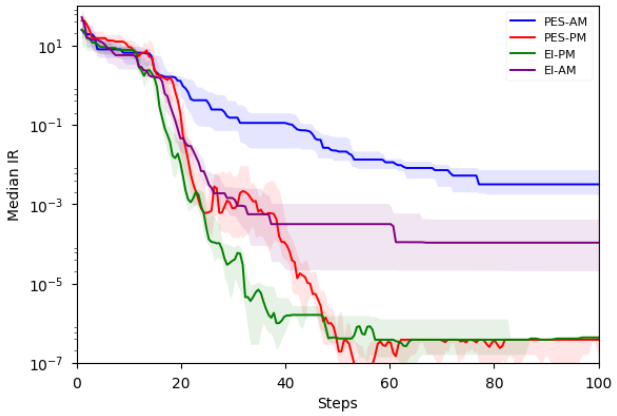

Figure 1. Performance on the Branin function of EI and PES with the minimum observed value (purple and blue respectively) and the posterior minimum value (green and red respectively) being reported. The median (solid line) and first to third quartiles (shaded) of ten runs are shown. Performance is similar initially but reporting the posterior minimum rather than the minimum observed value reaches a substantial lower errors, particularly with PES.

3. Environmental Variables

We now consider an extension of the classic Bayesian Optimization setting. At each evaluation of the objective function we specify an additional parameter, named by (Klein et al., 2015) as the environmental variable. This parameter allows a tradeoff between accurate but more expensive and cheap but poor evaluations. We define the environmental variable as $0 \leq s \leq 1$ where $s = 0$ yields the true objective and cost is expected to decrease towards $s = 1$. We desire an acquisition function that will provide the most cost efficient information gain at each step.

3.1. Augmented space

If variation of the environmental variable causes change in the mean of the response then, as in (Klein et al., 2015), we model the black box as a $d + 1$ dimensional function in $\mathbb{R}^d \times s \in [0, 1]$ such that $s = 0$ corresponds to the true objective. This is the case in many scenarios. For example, in optimizing the parameters of a classifier, the performance is expected to be worse with fewer training data, or, in fitting the hyperparameters of a Gaussian process, the likelihood on a subset of the data is not expected to have the same value as on the full dataset. We then wish to define an acquisition function over the full space such that we learn about the minimum in the $s = 0$ plane. Rather than the Entropy search method as in (Klein et al., 2015), we make use of Predictive Entropy search with some modifications as a faster method of evaluating the expected change in entropy. We model the response of the objective as a GP over the augmented $d + 1$ dimensional space, with a Matérn 5/2 kernel, but rather than using a spectral decomposition to draw from $p(x_*)$ we use sampling in the $s = 0$ plane to draw support points, and draw from the posterior on these points. The method is detailed in Section 4.

To optimize a wide range of objectives we are not able to guarantee that the minimum of the objective (for $s = 0$) is the global minimum in the augmented space spanning all s. For this reason, we do not include the global minimum constraint used in the original specification of PES (Hernández-Lobato et al., 2014) since the values reported at reduced cost may be lower as well as higher than the full cost result. For the same reason we use the $D + 1$ dimensional Matérn 5/2 product kernel to model the objective without any further assumptions on behaviour due to the environmental variable.

3.2. Overhead adjustment

In this section we present a novel acquisition function over environmental variables. In the usual setting for Bayesian Optimization the overhead computational cost of optimizing the acquisition function is considered negligible compared to evaluations of the true objective. When the option of a larger number of less expensive evaluations is made available, this may no longer be the case, particularly considering the poor scaling of Gaussian Processes (typically $\mathcal{O}(n^3)$) as additional points are added. (Klein et al., 2015) use an acquisition function of the form

$$\alpha = \frac{\Delta H}{c(x, s) + c_{\text{over}}}, \tag{9}$$

where c_{over} is the time for the previous step to choose the next point to evaluate and $c(x, s)$ is the GP posterior mean of evaluation cost conditioned on the evaluations observed so far using MAP hyperparameters of the Matérn 5/2 kernel. Since the overhead grows substantially over the course of optimization, we prefer to use an estimate of the average overhead between the current and final steps.

We model the overhead as growing according to a power plus constant rule

$$\hat{c}_{\text{over}}(n \mid \theta) = \theta_0 + \theta_1 n^{\theta_2} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \theta_3^2) \tag{10}$$

with independent Gamma priors on θ_i. Given a remaining optimization budget B, our modified acquisition function is

$$\frac{\Delta H}{c(x,s) + \frac{1}{N}\sum_{i=0}^{N} \hat{c}_{\text{over}}(i \mid \hat{\theta})} \tag{11}$$

where N is the greatest value such that

$$B \geq \sum_{i=0}^{N} \hat{c}_{\text{over}}(i \mid \hat{\theta}) + c_{\text{evaluation}} \quad . \tag{12}$$

Here $\hat{\theta}$ is the maximum-a-posteriori estimate given the overheads observed so far and the current step is considered to be step zero.

This change causes the algorithm to prefer slightly more expensive evaluations than otherwise, particularly when a large number of evaluations remain, which we find to improve performance.

4. Fast Draws from Pmin

We now present a novel sampling strategy for PES that renders our method computationally efficient. The original formulation of PES in (Hernández-Lobato et al., 2014) makes use of Bochner's theorem to obtain approximate draws from the Gaussian process which can easily be minimized to obtain a draw from the posterior distribution of the global minimizer. This method is only applicable to a stationary kernel. Since we wish to be able to use non-stationary kernels in general we prefer the alternate method of generating draws proposed in (Hennig & Schuler, 2012) of drawing support points from some distribution $q(x)$, which is similar to $p(x_*)$, then making draws from the GP posterior to provide samples of $p(x_*)$. This process does not place any requirements on the kernel used. As noted by (Hennig & Schuler, 2012) any q with non-zero support over the search domain may be used, with more samples of q being required to obtain good results if it is not similar to p.

Slice sampling over the EI of the GP posterior is the suggested method of drawing support points in (Hennig & Schuler, 2012). However, the evaluation of EI requires an $O(n^2)$ inference and is performed many times for each point produced by slice sampling. This is further increased by the practice of discarding a burn-in period and subsampling the resulting sequence. This represented a large portion of the runtime in our implementation. Noting that any q can be used, we seek an alternative from which we can draw points with lower computational overhead.

To achieve a fast approximation for p we note that the probability of a point being a stationary is equal to the probability

$$p_{\text{stat}} = \prod_{i=0}^{D} P\left(\frac{\partial f}{\partial x_i} = 0\right) \tag{13}$$

by definition. Further, the local minimum, x_i^l, of the posterior mean are local maxima of p_{stat}. We can find these points easily by performing local searches on the posterior mean from random start points. Given a candidate local minima x_i^c we infer the GP mean and covariance of all elements of the gradient, g_μ, Σ_g, and Hessian, H_μ, Σ_H, at that point. We then make a second order Taylor approximation centered at the candidate minimum

$$f = \frac{1}{2} z^T H z + z^T g + c, \tag{14}$$

where $z = x - x_i^l$. We make the further assumption that $H_\mu \gg \Sigma_H$ for all elements of H, therefore the Hessian is treated as constant $H = H_\mu$. The local minimum under this model is located at

$$\begin{aligned} \frac{\partial f}{\partial z} &= \underline{0} \\ z &= H^{-1} g \end{aligned} \tag{15}$$

Since g is Normally distributed, and zero at x_i^c by definition, the posterior distribution for the local minimum under the Taylor approximation is

$$p(x^l) = \mathcal{N}\left(x^c, H^{-1}\Sigma_g H^{-T}\right) \tag{16}$$

We define $q(x)$, the distribution we use to draw support points, as

$$q = Z\left[1 + \sum_{i}^{m} \mathbb{I}(x \in D)\mathcal{N}\left(x_i^c, H_i^{-1}\Sigma_{gi}H_i^{-T}\right)\right] \tag{17}$$

where there are m local minima and an additional uniform component is included to retain capacity for exploration. The distribution is clipped within the search domain.

Following the initial step of locating the local minima we can make draws from this q trivially. If desired this approximation could be improved by marginalizing over the Hessian by taking draws from H at each minimum and calculating $p(x^l \mid H)$ for each draw. The improvement in computational cost by using this method is shown in Figure 2

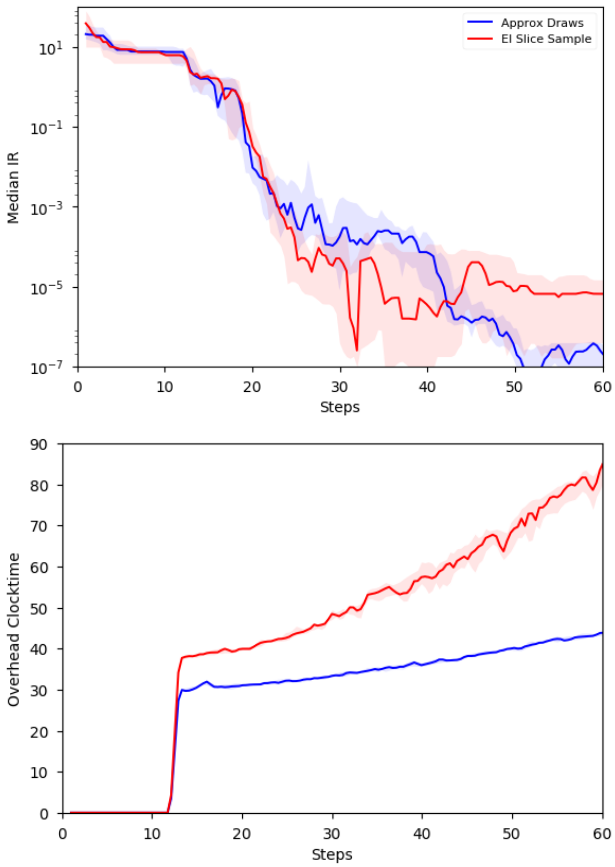

Figure 2. Overhead of optimizing the acquisition function and immediate regret for an offset Branin objective with quadratic cost. Slice sampling under EI (red) and our sum of Gaussians approximation (blue) are shown. Performance remains similar while overhead costs are significantly reduced.

For initialization we follow (Klein et al., 2015) by evaluating a set of values of the environmental variable for each random draw from x rather than each evaluation being independent. We choose $s = 0.5, 0.75, 0.875$ and use twenty total evaluations in the initialization dataset.

5. Experiments

We compare our method (EnvPES) to Expected Improvement (EI), Predictive Entropy Search (PES) and FABOLAS. Our GP package is used for PES and EI, the implementation of FABOLAS is provided in the RoBO package[2]. We modify these implementations to return the posterior mean minimizer as in §2.4, and further modify FABOLAS to use the Matérn 5/2 kernel over the environmental variable rather than the parametric form used in the original since the monotonic assumption does not necessarily hold in our experiments. We show that we are able to match or exceed the performance of existing methods over a selection of objectives.

[2]https://github.com/automl/RoBO/

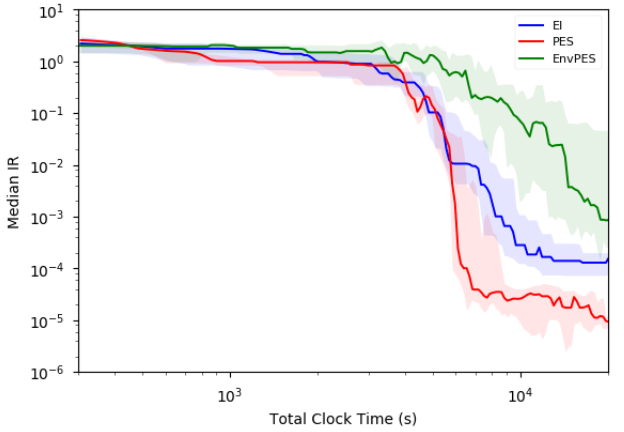

Figure 3. Performance of EnvPES (green), PES (red) and Expected Improvement (blue) on draws from the Matérn kernel. The median (solid line) and first to third quartiles (shaded) of optimizing ten draws are shown. The objective and cost function had $l_c = 1$ and $l_{ev} = 0.4$ which induces poor performance as the cost of evaluation does not reduce much and evaluations at increasing s are only loosely linked to the true objective. Evaluations cannot be made as cheaply compared to the full cost as the previous example and convey little information about the true objective. EnvPES is able to perform as well as other methods since it needs to learn a more complex model.

5.1. In model test

To illustrate expected performance we use draws from the Matérn 5/2 kernel as objective function. The characteristic lengthscale is set to 0.3 and the search domain is $[-1, 1]^2$. We simulate the environmental variable as an additional dimension of the objective over $[0, 1]$ with characteristic length l_{ev} and the cost as an exponential $\exp(-l_c s)$. We show results for advantageous and adversarial values of l_c and l_{ev} in Figures 4 and 3 respectively. This shows the potential gains of making use of the environmental variable in a suitable scenario, while retaining reasonable performance otherwise.

5.2. Off model test

5.2.1. Common Synthetic Functions

We test our method on a selection of common objectives. The results are shown in Figure 5. Following the approach of (Swersky et al., 2013) we use a linear shift of the objective for the lower cost evaluations, in this case the shift is continuous rather than discrete. The cost imposed is of quadratic form rising from two minutes as the cheapest available up to thirty minutes for the full objective. Per-

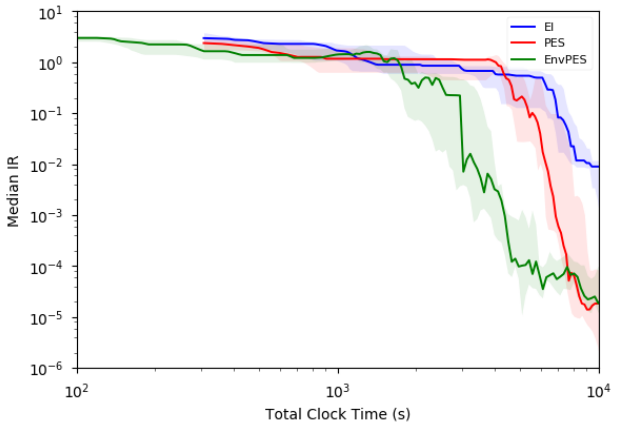

Figure 4. Performance of EnvPES (green), PES (red) and Expected Improvement (blue) on draws from the Matérn kernel. The median (solid line) and first to third quartiles (shaded) of optimizing ten draws are shown. The objective and cost function had $l_c = 3$ and $l_{ev} = 1.5$ which allows very good performance as the cost decays quickly while the approximate objective remains very similar to the true objective. Substantial improvement is therefore available using less expensive evaluations.

formance of EnvPES matches or exceeds the other methods shown on each test. Considering only the cost of evaluation EnvPES and FABOLAS have similar performance, including the time to optimize the acquisition the lower overheads of EnvPES allow better performance.

5.2.2. MNIST CLASSIFER HYPERPARAMETERS

Finding the best parameters for a classifier is another common problem in machine learning. We optimize the hyperparameters of an SVM classifier on the the MNIST dataset, We allow the dataset size to vary from 100 to 100000, which incurs a cost of around five minutes on the full dataset using a standard laptop. As shown in Figure 6 we are able to achieve superior performance to the existing methods. Both our method and FABOLAS are able to achieve low values faster than methods not making use of the environmental variable. However, due to the high overhead cost FABOLAS is then slow to make further improvement.

5.2.3. GP KERNEL PARAMETER FITTING

Fitting the hyperparameters of a Gaussian Process is a common machine learning problem, with evaluation cost that scales cubically with the number of datapoints. We train a GP with a Matérn 5/2 kernel on freely available half hourly time series data for UK electricity demand for 2015[3]. Evaluation of this objective with the full dataset again typically incurs a cost of around ten minutes. EnvPES and FABOLAS are able to evaluate the log-likelihood of random subsets down to 2% of the full dataset. EI and PES are only able to use the full dataset. As shown in Figure 7 we are able to achieve performance similar to the methods that do not make use of the environmental variable, in contrast to FABOLAS which incurs costs an order of magnitude greater than the true objective in selecting the next point to evaluate.

[3] www2.nationalgrid.com/UK/Industry-information/Electricity-transmission-operational-data/Data-explorer

6. Conclusion

We have proposed a novel acquisition function based on Predictive Entropy Search for use in variable cost Bayesian Optimization. We further introduce a novel sampling strategy applicable to both ES and PES which makes our implementation more computationally efficient. We have also proposed an alternative method for evaluating the performance of Bayesian Optimization methods. Bringing these together we demonstrate a practical Bayesian Optimization algorithm for variable cost methods and have shown that we are able to match or exceed the performance of existing methods on a selection of synthetic and real world applications.

Acknowledgements

Mark McLeod is supported by an EPSRC DTA studentship.

References

Garnett, Roman, Osborne, Michael A, and Roberts, Stephen J. Bayesian optimization for sensor set selection. In *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pp. 209–219. ACM, 2010.

Hennig, Philipp and Schuler, Christian J. Entropy search for information-efficient global optimization. *The Journal of Machine Learning Research*, 13(1):1809–1837, 2012.

Hernández-Lobato, José Miguel, Hoffman, Matthew W, and Ghahramani, Zoubin. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems*, pp. 918–926, 2014.

Jones, Donald R, Schonlau, Matthias, and Welch, William J. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13 (4):455–492, 1998.

Klein, Aaron, Bartels, Simon, Falkner, Stefan, Hennig, Philipp, and Hutter, Frank. Towards efficient Bayesian

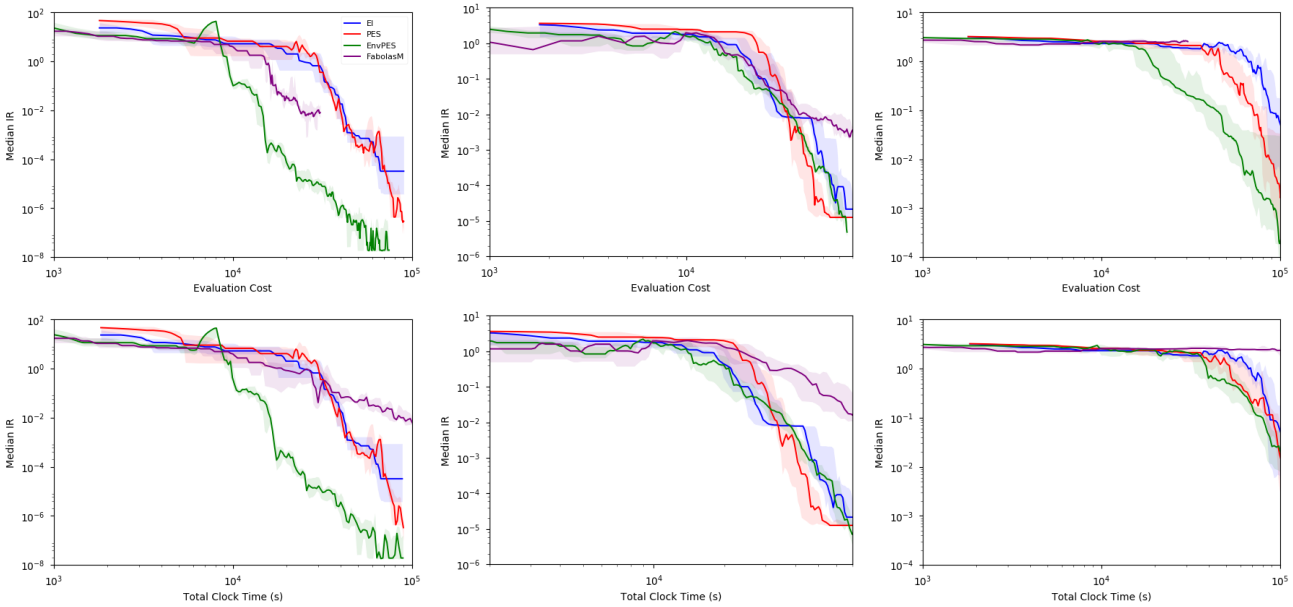

Figure 5. Performance of EnvPES (green), PES (red) and Expected Improvement (blue) and FABOLAS (purple) on the Branin (left), Hartman 3D (middle) and Hartman 6D (right) test functions with less expensive evaluations available under a linear shift from the true objective. The performance is shown against evaluation time for the objective (top) and including overhead due to the acquisition function (bottom). The median (solid line) and first to third quartiles (shaded) of eight runs are shown.

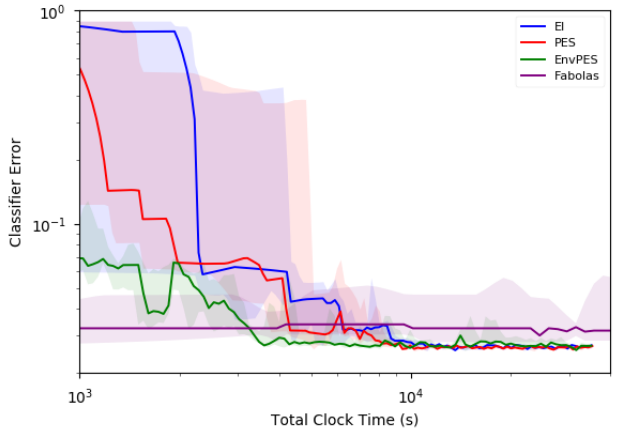

Figure 6. Performance of EnvPES (green), PES (red) and Expected Improvement (blue) and FABOLAS (purple) finding the best hyperparameters for a support vector machine classifying the MNIST dataset. The median (solid line) and first to third quartiles (shaded) of seven runs are shown. Here we have used the original form of FABOLAS.

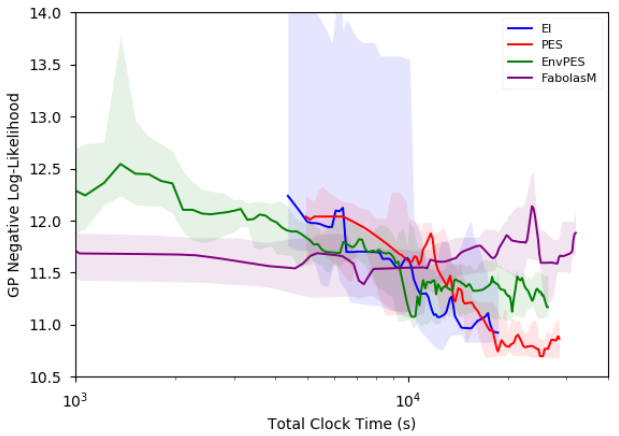

Figure 7. Performance of EnvPES (green), PES (red) and Expected Improvement (blue) and FABOLAS (purple) minimizing the negative log-likelihood of kernel hyperparameters for a Gaussian Process on UK power data. The median (solid line) and first to third quartiles (shaded) of eight runs are shown.

optimization for big data. In *NIPS 2015 workshop on Bayesian Optimization (BayesOpt 2015)*, 2015.

Kushner, Harold J. 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.

Lizotte, Daniel J, Wang, Tao, Bowling, Michael H, and Schuurmans, Dale. Automatic gait optimization with Gaussian process regression. In *IJCAI*, volume 7, pp. 944–949, 2007.

Lizotte, Daniel James. *Practical bayesian optimization*. University of Alberta, 2008.

Minka, Thomas P. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pp. 362–369. Morgan Kaufmann Publishers Inc., 2001.

Močkus, J. On Bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference*, pp. 400–404. Springer, 1975.

Neal, Radford M. Slice sampling. *Annals of statistics*, pp. 705–741, 2003.

Peters, Jan and Deisenroth, Marc Peter. Bayesian gait optimization for bipedal locomotion. In *Learning and Intelligent Optimization: 8th International Conference, Lion 8, Gainesville, FL, USA, February 16-21, 2014. Revised Selected Papers*, volume 8426, pp. 274. Springer, 2014.

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

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

Srinivas, Niranjan, Kakade, Sham M, Seeger, Matthias, and Krause, Andreas. Gaussian process bandits without regret: An experimental design approach. Technical report, 2009.

Swersky, Kevin, Snoek, Jasper, and Adams, Ryan P. Multi-task Bayesian optimization. In *Advances in neural information processing systems*, pp. 2004–2012, 2013.

Tesch, Matthew, Schneider, Jeff, and Choset, Howie. Using response surfaces and expected improvement to optimize snake robot gait parameters. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pp. 1069–1074. IEEE, 2011.