Output-Space Predictive Entropy Search for Flexible Global Optimization

Matthew W. Hoffman, Zoubin Ghahramani
University of Cambridge

1 Introduction

Recently, a great deal of interest has been paid in the field of Bayesian optimization to methods which take an information-theoretic approach to the design of acquisition functions. Such techniques equate the value of potential query points directly with their information content. The first of these techniques, developed separately as *an informational approach to global optimization (IAGO)* [9] or *entropy search (ES)* [3], considers the latent maximizer as a random variable and selects the point which results in the greatest reduction in posterior entropy. The entropy reduction cannot, though, be computed in closed form—as a result these earlier works relied on approximations that are often unwieldy in practice. A more recent formulation due to [4] instead rewrites ES as the mutual information between the latent maximizer and the next observation. This approach, known as *predictive entropy search (PES)*, greatly simplifies the required approximations and allows for further extensions of the optimizer [5].

In this work we build upon the strategy employed by PES, however whereas this earlier approach considers the information content of the latent maximizer, we instead maximize the information gained about the *maximum value*. This algorithm is referred to as *output-space predictive entropy search (OPES)* due to its computation of entropy in the outputs rather than the inputs of our model. We will show that rewriting the strategy in this manner leads to further simplifications of the necessary approximations and alleviates a number of irregularities evident in PES. In particular we show that OPES can be applied to problems whose inputs are the union of disjoint and differently-dimensioned spaces where PES does not apply. Finally, we further describe an additional extension that naturally allows for non-conjugate likelihoods that would otherwise be difficult within the PES framework.

2 Output-space Predictive Entropy Search

The prior and hence posterior process used in this work define a distribution over latent functions f, given as a zero-mean Gaussian process (GP) [8]. As a result the unknown maximizer $\mathbf{x}_\star = \arg\max_{\mathbf{x}} f(\mathbf{x})$ is stochastic, with distribution $p(\mathbf{x}_\star|\mathcal{D})$. The acquisition strategy employed by *entropy search (ES)* selects at every iteration the point which minimizes the expected entropy of $\mathbf{x}_\star$ after observing $y_{\mathbf{x}}$. The *Predictive Entropy Search (PES)* acquisition function alternatively attacks the point which maximally reduces the entropy, a quantity that is equivalent in terms of its maximizer to that of ES. However this quantity can also be formally written as the mutual information between $\mathbf{x}_\star$ and $y_{\mathbf{x}}$, allowing one to rewrite this acquisition as

$$\alpha_{\text{PES}}(\mathbf{x}|\mathcal{D}) = \text{H}[y_{\mathbf{x}}|\mathcal{D}] - \mathbb{E}_{\mathbf{x}_\star}\Big[\text{H}\big[y_{\mathbf{x}}\big|\mathcal{D}, \mathbf{x}_\star\big]\Big|\mathcal{D}\Big]. \tag{1}$$

This formulation, previously applied in the context of active learning [6], leads to an approximation that is both simpler than ES and more accurate [4].

However, both of these approaches take as their starting point the entropy of the latent maximizer $\mathrm{H}[\mathbf{x}_\star|\mathcal{D}]$. Alternatively, in this work we propose to select points which minimize the entropy of the latent maximum value, i.e. $\mathrm{H}[f_\star|\mathcal{D}]$ where $f_\star = f(\mathbf{x}_\star)$. Similar to the definition of PES the order of the information gain can again be swapped and written as

$$\alpha_{\mathrm{OPES}}(\mathbf{x}|\mathcal{D}) = \mathrm{H}[y_\mathbf{x}|\mathcal{D}] - \mathbb{E}_{f_\star}\Big[\mathrm{H}\big[y_\mathbf{x}\big|\mathcal{D}, f_\star\big]\Big|\mathcal{D}\Big]. \tag{2}$$

We will refer to this approach as *output-space predictive entropy search (OPES)*. Notationally, this alternative acquisition function replaces the location $\mathbf{x}_\star$ from (1) with its value $f_\star$. In practice, however, this modification significantly extends the applicability of the method. The acquisition function can then be evaluated in the following way:

1. the first term $\mathrm{H}[y_\mathbf{x}|\mathcal{D}]$ is the entropy of a Gaussian random variable and as a result has a closed form expression;
2. the expectation for the second term can be approximated with Monte Carlo via samples from $p(f_\star|\mathcal{D})$ using a similar strategy as in [4];
3. finally, although the entropy of the second term has no closed form expression we can use expectation propagation (EP) to approximate $p(f_\mathbf{x}|\mathcal{D}, f_\star)$ with a Gaussian [7] and then compute the resulting quantity in closed form; this is described in Section 2.1.

2.1 Approximating the conditional predictive distribution

Note that due to space restrictions this section is very abbreviated. A longer version containing the full derivations is available.

In order to approximate $p(f_\mathbf{x}|\mathcal{D}, f_\star)$ we will first compute the distribution over the latent values $f(\mathbf{x}_i)$ conditioned on the observations and the constraining maximizer $f_\star$. Let $\mathbf{f}$ and $\mathbf{y}$ denote vectors containing the latent function values and observations respectively, evaluated at all previous query points $\mathbf{x}_{1:n}$. The distribution of the latent values $\mathbf{f}$ can be written as $p(\mathbf{f}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, where these values follow standard GP inference. Here we are also assuming that each observation y_i is subject to iid Gaussian noise with variance σ^2, although this assumption can be relaxed later. Next we introduce the constraint that each latent value $f_i \leq f_\star$ is bounded by the given maximum value $f_\star$. Conditioning on this constraint can be written as

$$p(\mathbf{f}|\mathbf{y}, f_\star) \propto \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \prod_{i=1}^{n} \Theta(f_\star - f_i) \approx \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\, \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) \propto \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1),$$

where $\boldsymbol{\mu}_1 = \boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}})$ and $\boldsymbol{\Sigma}_1 = (\boldsymbol{\Sigma}_0^{-1} + \tilde{\boldsymbol{\Sigma}}^{-1})^{-1}$. Here Θ is the Heaviside step function which is one when its arguments are zero or greater and is zero otherwise. The additional Gaussian terms $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ are approximate factors computed with EP where the covariance term will be diagonal.

Next, given the constrained distribution over the latent values $\mathbf{f}$ one can compute predictions at any point $\mathbf{x}$. The resulting distribution can be written as

$$\begin{aligned} p_0(f_\mathbf{x}|\mathbf{y}, f_\star) &= \int p(f_\mathbf{x}|\mathbf{f})\, p(\mathbf{f}|\mathbf{y}, f_\star)\, \mathrm{d}\mathbf{f} \\ &\approx \int \mathcal{N}(f_\mathbf{x}|\mathbf{k}_\mathbf{x}^\mathsf{T}\mathbf{K}^{-1}\mathbf{f}, k_{\mathbf{xx}} - \mathbf{k}_\mathbf{x}^\mathsf{T}\mathbf{K}^{-1}\mathbf{k}_\mathbf{x})\mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\, \mathrm{d}\mathbf{f} = \mathcal{N}(\bar{m}, \bar{v}) \end{aligned}$$

We have written this distribution as p_0 due the fact that we haven't yet included the final constraint, i.e. that $f_\star \geq f_\mathbf{x}$. In order to obtain numerically stability and avoid numerous inversions of the kernel matrix $\mathbf{K}$ the final mean and variance can be simplified as,

$$\bar{m} = \mathbf{k}_\mathbf{x}^\mathsf{T}(\mathbf{K} + \boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}(\sigma^{-2}\mathbf{y} + \tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}), \tag{3}$$

$$\bar{v} = k_{\mathbf{xx}} - \mathbf{k}_\mathbf{x}^\mathsf{T}(\mathbf{K} + \boldsymbol{\Omega})^{-1}\mathbf{k}_\mathbf{x}, \tag{4}$$

where $\boldsymbol{\Omega} = \sigma^2(\sigma^2\tilde{\boldsymbol{\Sigma}}^{-1} + \mathbf{I})^{-1}$. We can next incorporate the constraint that $f_\star \geq f_\mathbf{x}$ for the arbitrary query point $\mathbf{x}$. This can again be done using a step-function factor and written as

$$p(f_\mathbf{x}|\mathbf{y}, f_\star) \approx \mathcal{N}(\bar{m}, \bar{v})\, \Theta(f_\star - f_\mathbf{x}) \approx \mathcal{N}(\hat{m}, \hat{v})$$

The final approximation involves moment-matching to arrive at

$$\hat{m}(\mathbf{x}|\mathcal{D}, f_\star) = \bar{m}_\mathbf{x} - \sqrt{\bar{v}_\mathbf{x}} r, \tag{5}$$
$$\hat{v}(\mathbf{x}|\mathcal{D}, f_\star) = \bar{v}_\mathbf{x} - \bar{v}_\mathbf{x} r(r + \alpha), \tag{6}$$

where $\alpha = (f_\star - \bar{m}_\mathbf{x})/\sqrt{\bar{v}_\mathbf{x}}$ and $r = \phi(\alpha)/\Phi(\alpha)$. We have also explicitly written these quantities as a function of the test input and data. An example of this procedure is shown in Figure 1 constrasted with that of PES.

The OPES acquisition function. Using the results of this section, the acquisition function corresponding to the OPES strategy can be written as

$$\alpha_{\text{OPES}}(\mathbf{x}|\mathcal{D}) = \frac{1}{2}\Big[\log\left(v(\mathbf{x}|\mathcal{D}) + \sigma^2\right) - \frac{1}{M}\sum_{i=1}^{M} \log\left(\hat{v}(\mathbf{x}|\mathcal{D}, f_\star^{(i)}) + \sigma^2\right)\Big] \tag{7}$$

where $\{f_\star^{(i)}\}$ is the set of function maxima sampled as described earlier. OPES can also be trivially extended to sample over GP hyperparameters by adding an extra summation and selecting a single $f_\star^{(i)}$ for each such sample.

Comparison with PES. The approximation used by PES requiers that $\mathbf{x}_\star^{(i)}$ is constrained to be a global maximum. This same difficulty is faced by OPES and both algorithms reduce this to requiring that the maximum is better than previous observations as well as any individual test points. When considering the latent observations, however, since PES only constrains on the maximizer location it then holds that both $f_\star$ and $\mathbf{f}$ are latent variables. As a result PES instead ignores the latent $\mathbf{f}$ and introduces a soft-constraint that the maximum observation $\max_i y_i$ is less than the latent $f_\star$. This can prove problematic precisely when used for non-conjugate likelihoods for which this soft-constraint is inappropriate.

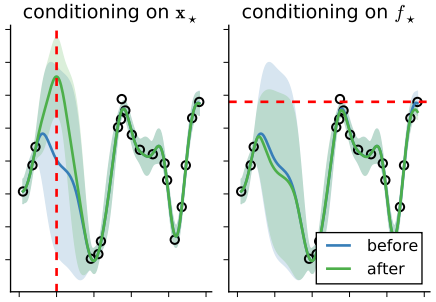

Figure 1: Illustration of OPES and PES.

An extension to non-conjugate likelihoods. In introducing OPES we initially assumed that the observations y_i are each subject to iid Gaussian noise. This allows the predictions before constraining on $f_\star$ to be made in closed form. However, we can also consider a Bernoulli likelihood model with $y_i \in \{-1, +1\}$ such that $p(y_i|f_i) = \sigma(y_i f_i)$ for a sigmoid function σ, typically logistic or probit. This is quite natural for optimization problems where the outcome is e.g. success $(+1)$ or failure (-1) and the goal is the find the design $\mathbf{x}$ which maximizes probability of success. Additionally, many of the problems considered under the guise of bandit or contextual-bandit problems are of this form. Under this likelihood the predictive distribution can no longer be computed in closed form, but it is possible to approximate this distribution using either Laplace or EP. In this setting, the constrained predictions described in Section 2.1 can be written as a Normal distribution with mean and variance

$$\bar{m} = \mathbf{k}_\mathbf{x}^\mathsf{T}(\mathbf{K} + \boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\rho} + \tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}),$$
$$\bar{v} = k_{\mathbf{x}\mathbf{x}} - \mathbf{k}_\mathbf{x}^\mathsf{T}(\mathbf{K} + \boldsymbol{\Omega})^{-1}\mathbf{k}_\mathbf{x}.$$

where $\boldsymbol{\Omega} = (\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Lambda} + \mathbf{I})^{-1}\boldsymbol{\Lambda}$ and $(\boldsymbol{\rho}, \boldsymbol{\Lambda})$ are due to the non-conjugate approximation. As a sanity check we can also see that that the conjugate model is recovered exactly when $\boldsymbol{\Lambda} = \sigma^2\mathbf{I}$ and $\boldsymbol{\rho} = \mathbf{y}$. This leaves us with unconstrained and constrained approximations $q(f_\mathbf{x}|\mathcal{D})$ and $q(f_\mathbf{x}|\mathcal{D}, f_\star)$. Finally the entropy for both distributions q this is given simply by $-\pi_\mathbf{x}\log\pi_\mathbf{x} - (1-\pi_\mathbf{x})\log(1-\pi_\mathbf{x})$ where $\pi_\mathbf{x} = \int \sigma(f_\mathbf{x})q(f_\mathbf{x})\,\mathrm{d}f_\mathbf{x}$ is the marginal probability of observing a $+1$ at input $\mathbf{x}$. This integral can be computed in closed form for the probit and approximated for the logistic [see 8, chap. 3]. Finally, although we have given as an example the Bernoulli likelihood this technique applies much more broadly, [e.g. 1, 2].

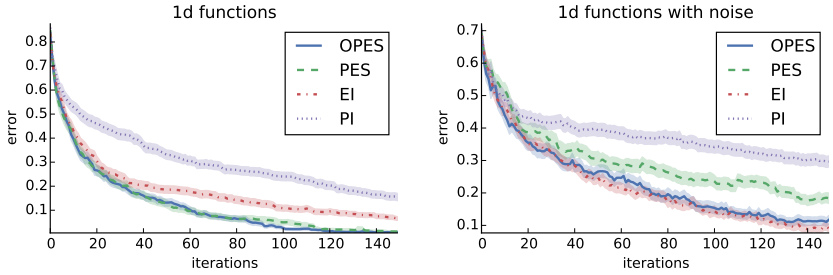

Figure 2: Average results for random, 1-dimensional functions.

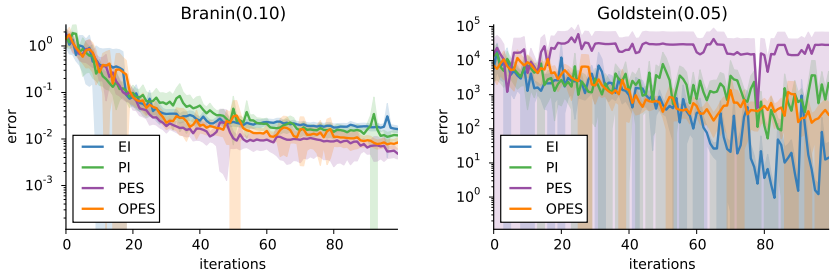

Figure 3: Average results for global optimization benchmark functions.

3 Results

In this section we include preliminary results which compare the performance of OPES and PES; also included are the EI and PI strategies as baselines. Previous work has already shown PES to outperform the ES strategy, although these approaches are equivalent up to approximation errors, so we do not include this in our experiments.

Our first experiment, shown in Figure 2 compares each acquisition strategy for random 1-dimensional functions sampled from a zero- mean GP prior. Here we see that PES and OPES both perform quite well, however we notice that OPES does perform better with higher noise. The next set of experiments, in Figure **??** shows the results on two global optimization benchmark functions. And finally in Figure 4 we show results for random functions defined over a union of disjoint sets of differing dimensionalities (for which PES does not apply). For the setting of random functions we see rather conclusive evidence as to the performance gain of information-based methods. Results are somewhat less conclusive for the benchmark functions, however we see that these methods are competitive (and give better performance for Branin).

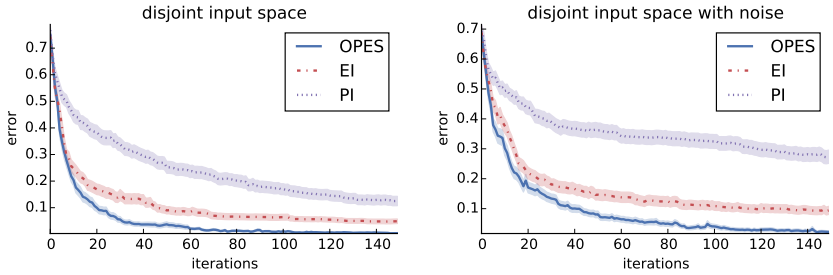

Figure 4: Average results for random functions defined over an input space which is the union of 1- and 2-dimensional values.

References

[1] W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *the Journal of Machine Learning Research*, pages 1019–1041, 2005.

[2] P. Groot and P. Lucas. Gaussian process regression with censored data using expectation propagation. In *the European Workshop on Probabilistic Graphical Models*, 2012.

[3] P. Hennig and C. J. Schuler. Entropy search for information-efficient global optimization. *the Journal of Machine Learning Research*, 13:1809–1837, 2012.

[4] 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*, 2014.

[5] J. M. Hernández-Lobato, M. Gelbart, M. W. Hoffman, R. P. Adams, and Z. Ghahramani. Predictive entropy search for Bayesian optimization with unknown constraints. In *the International Conference on Machine Learning*, 2015.

[6] N. Houlsby, J. M. Hernández-Lobato, F. Huszar, and Z. Ghahramani. Collaborative Gaussian processes for preference learning. In *Advances in Neural Information Processing Systems*, pages 2096–2104, 2012.

[7] T. P. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

[8] C. E. Rasmussen and C. K. Williams. *Gaussian processes for machine learning*. The MIT Press, 2006.

[9] 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.