An Entropy Search Portfolio for Bayesian Optimization

Bobak Shahriari[*] bshahr@cs.ubc.ca
Ziyu Wang[†] ziyu.wang@cs.ox.ac.uk
Matthew W. Hoffman[‡] mwh30@cam.ac.uk
Alexandre Bouchard-Côté[*] bouchard@stat.ubc.ca
Nando de Freitas[†§°] nando@cs.ox.ac.uk

[*]University of British Columbia, Canada [†]University of Oxford, United Kingdom
[‡]University of Cambridge, United Kingdom [§]Canadian Institute for Advanced Research
[°]Google DeepMind

Abstract

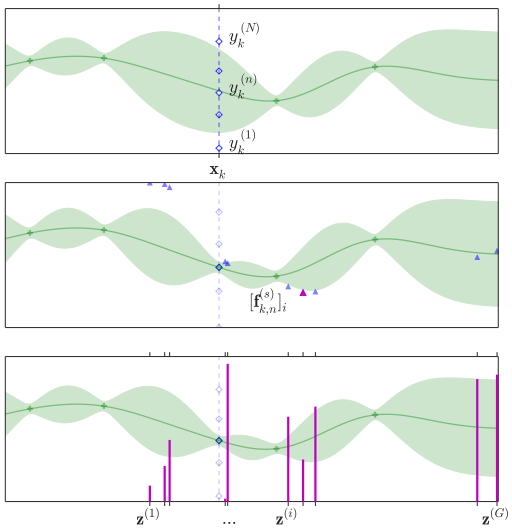
Portfolio methods provide an effective, principled way of combining a collection of acquisition functions in the context of Bayesian optimization. We introduce a novel approach to this problem motivated by an information theoretic consideration. We show that our method outperforms existing portfolio methods on several real and synthetic problems, including geostatistical datasets and simulated control tasks. We also demonstrate that as well as outperforming other portfolio methods, our proposed method is robust to the inclusion of poor acquisition functions.

1 Introduction

We are interested in finding a global minimizer $\mathbf{x}_\star = \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$ of a function f over some bounded domain, typically $\mathcal{X} \subset \mathbb{R}^d$. We further assume that $f(\mathbf{x})$ can only be evaluated via a series of queries $\mathbf{x}_t$ to a black-box that provides noisy outputs y_t from some set, typically $\mathcal{Y} \subseteq \mathbb{R}$. For this work we assume $y_t \sim \mathcal{N}(f(\mathbf{x}_t), \sigma^2)$, however, our framework can be extended to other non-Gaussian likelihoods. In this setting, we describe a sequential search algorithm that, after t iterations, proposes to evaluate f at some location $\mathbf{x}_{t+1}$ given by an acquisition strategy $\alpha(\mathcal{D}_t)$ where $\mathcal{D}_t = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_t, y_t)\}$ is the history of previous observations. Finally, after T iterations the algorithm must make a final recommendation $\tilde{\mathbf{x}}_T$, i.e. its best estimate for the optimum.

No single acquisition strategy known today provides better performance over all problem instances. In fact, we have empirically observed that the preferred strategy may change at various stages of the sequential optimization process due to varying trade-offs of exploration and exploitation. To address this issue, Hoffman *et al.* propose the use of a portfolio containing multiple acquisition strategies, and selecting between them using a meta-criterion [5].

Our contribution is a novel meta-criterion which we call the Entropy Search Portfolio (ESP). By viewing the location of the minimizer $\mathbf{x}_\star$ as a random variable, information-theoretic approaches, such as the work of [8, 3, 4], aim to select a new query point which minimizes posterior entropy of $\mathbf{x}_\star$ at each iteration. We use this same strategy to select between points suggested by acquisition functions in a portfolio. We provide empirical evidence that our approach results in performance gains not only over previous portfolio strategies but also over the fundamental strategies that make up the portfolio. We also show that ESP exhibits increased robustness with respect to poorly performing acquisition strategies.



Algorithm 1 Entropy Search Portfolio

Require: candidates $\{\mathbf{x}_k\}$, observations $\mathcal{D}$

1: $\mathbf{z}^{(i)} \sim p(\mathbf{x}_\star|\mathcal{D}), \quad i = 1, \ldots, G$
2: **for** $k = 1 : K$ **do**
3: **for** $n = 1 : N$ **do**
4: $y_k^{(n)} \sim p(y|\mathbf{x}_k, \mathcal{D})$
5: $\tilde{\mathcal{D}}_k^{(n)} = \mathcal{D} \cup \{(\mathbf{x}_k, y_k^{(n)})\}$
6: $\mathbf{f}_{kn}^{(s)} \sim p(\mathbf{f}|\tilde{\mathcal{D}}_k^{(n)})$ for $s = 1 : S$
7: $\hat{p}_{ikn} = \frac{1}{S} \sum_s \mathbb{I}\big[i = \arg\min_j [\mathbf{f}_{kn}^{(s)}]_j\big]$
8: **end for**
9: $u_k = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{G} \hat{p}_{ikn} \log \hat{p}_{ikn}$
10: **end for**
11: $k_\star = \arg\max_k u_k$
12: **return** $\mathbf{x}_{k_\star}$

Figure 1: Visualization of the Entropy Search Portfolio. **Top panel:** for each candidate $\mathbf{x}_k$ (blue dashed line) we draw N hallucinations (blue diamonds) from the conditional. In practice this is done with quasi-Monte-Carlo to reduce variance. **Middle panel:** for each hallucination $y_k^{(n)}$ we augment the GP model (green line and shaded area) and sample it S times at the discrete points $\mathbf{z}^{(i)}$ to obtain the $\mathbf{f}_{kn}^{(s)}$ (blue triangles) for $s = 1, \ldots, S$. We find the minimizer of each vector $\mathbf{f}_{kn}^{(s)}$ (magenta triangle). **Bottom panel:** finally, we bin the S minimizers into a discrete empirical distribution $\hat{p}$ depicted here as a magenta histogram.

2 Entropy search over portfolios

Portfolios are collections of base strategies $\mathcal{A} = \{\alpha_k\}_{k=1}^K$. Each strategy is an expert which recommends a candidate point $\mathbf{x}_k = \alpha_k(\mathcal{D})$ to be selected at iteration t. Our task is to select the most promising candidate according to some meta-criterion. In particular, our approach uses a probabilistic model of the location of the *unknown* global minimizer $\mathbf{x}_\star$.

Given data $\mathcal{D}$, let $\mathbb{P}(\,\mathrm{d}\mathbf{x}_\star|\mathcal{D}) = \mathbb{P}(\arg\min f(\mathbf{x}) \in \,\mathrm{d}\mathbf{x}_\star|\mathcal{D})$ denote the posterior over minimizer locations, with density $p(\mathbf{x}_\star|\mathcal{D})$. This distribution is induced by our GP posterior. We propose our meta- criterion $u(\mathbf{x}_k|\mathcal{D}) = \mathbb{E}_{p(y_k|\mathcal{D},\mathbf{x}_k)}[H[p(\mathbf{x}_\star|\tilde{\mathcal{D}}_k)]]$, where $\tilde{\mathcal{D}} = \mathcal{D} \cup \{(\mathbf{x}_k, y_k)\}$ contains one additional fixed $\mathbf{x}_k$ and a corresponding random y_k simulated from the posterior predictive distribution $p(y_k|\mathcal{D}, \mathbf{x}_k)$. This corresponds to the expected entropy of the distribution of the minimizer after selecting candidate input $\mathbf{x}_k$. In other words, the candidate selected by this criterion is the one that results in the greatest decrease in uncertainty about the location of the minimizer.

The expectation with respect to the predictive distribution can easily be approximated via sampling and Monte Carlo integration. Meanwhile, the entropy estimation is more involved due to the difficulty of evaluating $p(\mathbf{x}_\star|\tilde{\mathcal{D}}_k)$. For continuous densities p the differential entropy can be written as $H[p(\mathbf{x})] = -\int p(\mathbf{x}) \log p(\mathbf{x}) \,\mathrm{d}\mathbf{x}$. Instead we approximate this density with a discrete distribution $\hat{p}$ restricted to a finite set of *representer points* denoted $\{\mathbf{z}^{(i)}\}_{i=1}^G$ sampled directly from $p(\mathbf{x}_\star|\tilde{\mathcal{D}}_k)$, corresponding to Line 1 in Algorithm 1. Exactly how to do this is non-trivial and we give an outline of the steps at the end of the section.

With this discretized distribution $\hat{p}$ and the Monte Carlo integration of the outer expectation, we approximate our meta-criterion as $u(\mathbf{x}_k|\mathcal{D}) \approx \frac{1}{N} \sum H\big[\hat{p}(\mathbf{x}_\star|\tilde{\mathcal{D}}_k^{(n)})\big]$, where $\tilde{\mathcal{D}}_k^{(n)} = \mathcal{D} \cup \{(\mathbf{x}_k, y_k^{(n)}\}$ with $y_k^{(n)} \sim p(\cdot|\mathbf{x}_k, \mathcal{D})$ and where H now represents the discrete entropy

$$H\big[\hat{p}(\mathbf{x}_\star|\tilde{\mathcal{D}}_k^{(n)})\big] = -\sum_{i=1}^G \hat{p}(\mathbf{x}_\star = \mathbf{z}^{(i)}|\tilde{\mathcal{D}}_k^{(n)}) \log \hat{p}(\mathbf{x}_\star = \mathbf{z}^{(i)}|\tilde{\mathcal{D}}_k^{(n)}). \tag{1}$$

We are left with the problem of computing $\hat{p}(\mathbf{x}_\star = \mathbf{z}^{(i)}|\tilde{\mathcal{D}}_k^{(n)})$. Recall that $\hat{p}$ is the probability distribution over minimizers of a GP-distributed random function where the minimizers are restricted to a discrete and finite set. This can be sampled exactly as follows. Let the random variable $[\mathbf{f}_{kn}]_i = f(\mathbf{z}^{(i)})$, $i = 1, \ldots, G$, be a vector of latent function values evaluated at the representer points and conditioned on data $\tilde{\mathcal{D}}_k^{(n)}$. This vector simply has a Gaussian distribution

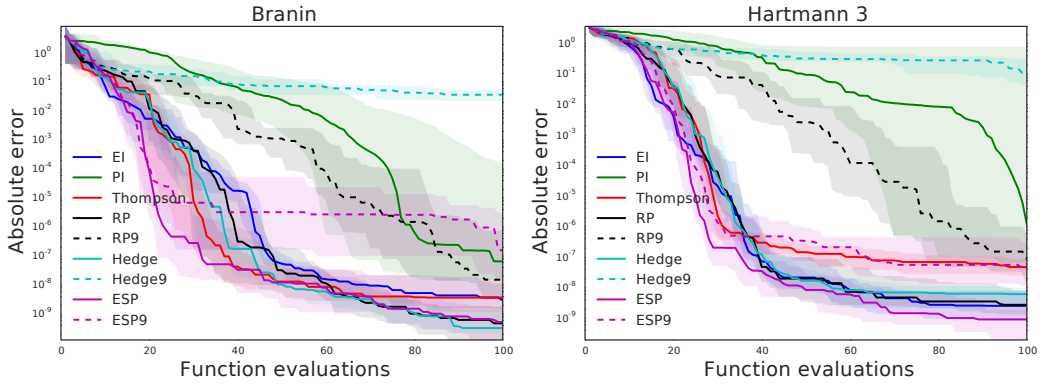


Figure 2: Absolute error of the best observation for the Branin and Hartmann 3 synthetic functions. The 9 additional random experts in RP9, GPHedge9, and ESP9 affect the RP and GPHedge methods much more dramatically than ESP.

and as a result we can produce S samples $\mathbf{f}_{kn}^{(s)} \sim p(\cdot|\tilde{\mathcal{D}}_k^{(n)})$ from the resulting GP posterior. The probabilities necessary to compute the entropy can then be approximated by the relative counts $\hat{p}_{ikn} = \frac{1}{S}\sum_s \mathbb{I}\big[i = \arg\min_j [\mathbf{f}_{kn}^{(s)}]_j\big]$. In other words, $\hat{p}_{ikn}$ represents the number of times representer point $\mathbf{z}^{(i)}$ was the minimizer in S samples from the posterior GP with data $\mathcal{D} \cup \{(\mathbf{x}_k, y_k^{(n)}\}$. Finally, by combining these ideas we can express our entropy-based meta-criterion

$$u(\mathbf{x}_k|\mathcal{D}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{G} \hat{p}_{ikn} \log \hat{p}_{ikn}. \tag{2}$$

Pseudocode computing this quantity is given in Algorithm 1 and we provide a corresponding visualization in Figure 1.

Let us now address the issue of producing samples from the posterior over global minima $p(\mathbf{x}_\star|\mathcal{D})$. In the discrete and finite setting, this can be done via probability matching by repeating the following generative process: i) draw a sample from the posterior distribution $p(\mathbf{f}|\mathcal{D})$ and ii) return the index of the maximum element in the sampled vector. We employ this same approach over our continuous domain. To avoid constructing an infinite-dimensional object representing the function f, we sample and optimize an analytic approximation to f which uses Bochner's lemma [1, 7].

We omit the details here but intuitively we sample a function by first sampling a fixed number of frequencies from the spectral density of the GP kernel and then the corresponding coefficients for those frequencies. We sample the coefficients from a specific distribution such that the weighted sum of those Fourier basis functions approximates one sample from the posterior GP. We then have a fixed function to maximize and obtain an approximate sample from $p(\mathbf{x}_\star|\mathcal{D})$. In this paper we used this process in two ways: first to obtain an approximation to Thompson sampling in the continuous domain (as was also done in [4]), and second to sample our representer points $\mathbf{z}^{(i)}$.

3 Experiments

We compare ESP against three well-known Bayesian optimization acquisition functions, namely EI, PI, and Thompson. For EI we used the implementation available in the *spearmint* package[1], while the latter two were implemented in the same framework. All three methods were included in the portfolios. We also compare ESP against the GPHedge portfolio method [5] and an approach which selects between different base strategies uniformly at random labeled RP.

We begin with two synthetic functions commonly used for benchmarking global optimization methods: Branin and Hartmann 3 [6]. Figure 2 reports the observed performance measured in absolute error on a logarithmic scale. As expected, ESP makes the best use of its base experts and outperforms all other methods in both examples. It is also interesting to note that each of the two examples favour their own base strategy. Thompson is a clear winner on Branin, while EI is the more attractive option on Hartmann 3. This observation motivates the use of portfolios of acquisition functions.

[1] https://github.com/JasperSnoek/spearmint

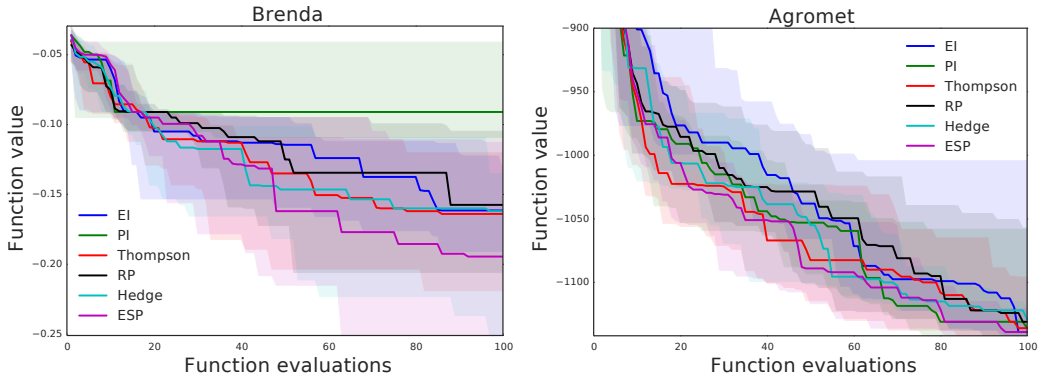


Figure 3: Best observed evaluation on mining datasets Brenda and Agromet. ESP outperforms the other portfolio methods while RP performs worst.

In these synthetic experiments we also demonstrate the robustness of ESP with respect to to the inclusion of poor base strategies. We do so by adding 9 random experts to each portfolio (we denote these ESP9, RP9, etc.). These so-called random experts select a point uniformly at random in the bounding box $\mathcal{X}$. We expect this sort of random search to be comparable to the other base methods in the initial stage of the optimization and eventually provide too much exploration and not enough exploitation. Note however that a few random experts in a portfolio could actually be beneficial in providing a constant source of purely exploratory candidates; precisely how many is an interesting question we do not discuss in the present work. Nevertheless, for the dimensionality and difficulty of these examples, we propose 9 additional random experts as being too many and indeed we observe empirically that they substantially deteriorate performance for all portfolios.

We observe that, especially on Hartmann 3, ESP is virtually unaffected until it reaches 5 digits of accuracy. Meanwhile, the progress made by RP is hindered by the random experts which it selects as frequently as the legitimate acquisition functions. Significantly worse is the performance of GPHedge which, due to the initial success of the random experts, favours these until the horizon is reached. Note in contrast that ESP does not rely on any expert's past performance, which makes it robust to lucky guesses and time-varying expert performances.

The next set of experiments were carried out on two datasets from the geostatistics community, referred to here as Brenda and Agromet [2]. Since these datasets consist of a finite sets of points, we transformed each of them into a function that we can query at arbitrary points via nearest neighbour interpolation. This produces a jagged piecewise constant function, which is outside the smoothness class of our surrogate models and hence a relatively difficult problem. Brenda is a dataset of 1,856 three-dimensional observations while Agromet is a dataset of 18,188 two-dimensional observations. Results on these functions are shown in Figure 3.

We note that PI, which has so far been an under-achiever, is among the better strategies on Agromet, while EI is the worst. This is further motivation for the use of portfolios. On both of these examples, RP performs poorly whereas GPHedge fares somewhat better. We can see that ESP performs particularly well on Brenda. On Agromet, ESP outperforms the other portfolio methods and is competitive with the best acquisition function—Thompson in the initial exploration phase, followed by PI after around 60 evaluations.

4 Conclusion

In this work we revisited the use of portfolios for Bayesian optimization. We introduced a novel, information-theoretic meta-criterion ESP which can indeed provide performance matching or exceeding that of its component experts. This is particularly important since we show in our experiments that the best acquisition function varies between problem instances and horizons considered. We have also shown that ESP has robust behavior across functions of different dimensionality even when the members of its portfolio do not exhibit this behavior. Further, ESP is more robust to poorly performing experts than other portfolio mechanisms.

References

[1] S. Bochner. *Lectures on Fourier integrals*. Princeton University Press, 1959.

[2] Isobel Clark and William V Harper. *Practical Geostatistics 2000: Case Studies*. Ecosse North America, 2008.

[3] P. Hennig and C.J. Schuler. Entropy search for information-efficient global optimization. *The Journal of Machine Learning Research*, 98888:1809–1837, 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] M. W. Hoffman, E. Brochu, and N. de Freitas. Portfolio allocation for Bayesian optimization. In *Uncertainty in Artificial Intelligence*, pages 327–336, 2011.

[6] D. Lizotte. *Practical Bayesian Optimization*. PhD thesis, University of Alberta, Canada, 2008.

[7] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2007.

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