Information-Greedy Global Optimization

Philipp Hennig and Christian J. Schuler
Max Planck Institute for Intelligent Systems
Department of Empirical Inference
Spemannstraße 38, 72070 Tübingen, Germany
[phennig|cschuler]@tuebingen.mpg.de

Abstract

Optimization is about inferring the location of the optimum of a function. An information-optimal optimizer should thus aim to collapse its belief about the location of the optimum towards a point-distribution, as fast as possible. But the state of the art rarely addresses this inference problem. Instead, it usually relies on some heuristic predicting function optima, then evaluates at the maximum of the heuristic. The reason there are no truly probabilistic optimizers yet is that they are intractable in several ways. In this paper, we present tractable approximations for each of these issues, and arrive at a flexible global optimizer for functions under Gaussian process priors, which performs well in comparison to a state of the art Gaussian process optimizer.

1 Introduction

Global optmization is a wide field, re-discovered several times in different communities. While the resulting algorithms vary widely in their motivation, characteristic behaviour and performance, they can usually be reduced to one of two general classes: Stochastic exploration (as in simulated annealing, genetic algorithms, etc.), or heuristic exploration. The latter class uses some model for the function (parametric, as in most local convex optimizers, or nonparametric, as in Gaussian Process optimization), then defines a heuristic on this model (such as the minimum of the model, or the "Expected Improvement" and "Probability of Improvement" heuristics and their variants, on Gaussian beliefs [1, 2, 3]). Neither of these two general classes are information efficient. Stochastic explorers do not have a notion of guided learning at all. Heuristic explorers have such a notion but, from an information-theoretic viewpoint, it is non-optimal in two ways: First, the heuristic is at best an approximation to, but not the probability distribution over the location of the optimum. Second, the heuristic optimizer tries to optimize the heuristic, not information gain about the heuristic. Instead of trying to *learn most* about the location of the optimum, these algorithms try to evaluate where some approximation thinks the optimum is *most likely to be*.

There is a reason why there are no information-optimal function optimizers around: They are intractable, in three ways:

- Even given a proper probabilistic model $p(f)$ for the function $f : x \mapsto f(x)$ in question, the probability distribution $p_{\min} = p(x = \arg\min_{x'} f(x'))$ is usually intractable, even for relatively simple priors, such as multivariate Gaussians, on finite discrete domains. This problem is compounded when the domain of f is continuous.
- Optimizing for information gain requires a model for how the belief $p_{\min}$ *changes* as a result of evaluation. This model is usually itself probabilistic, and not always tractable.
- Optimizing a number of subsequent evaluation points, whether under the probabilistic description of $p_{\min}$ or any heuristic, is an exponentially hard dynamic programming problem.

In an effort to find approximate answers to these challenges, we propose the following approach:

- Following other authors [e.g. 1, 2, 3], we adopt a Gaussian process prior for f. We then construct an approximation on $p_{\min}$, by representing the belief with finitely many samples and constructing a belief over the minimum using Expectation Propagation (EP) [4]. The samples for this discretisation should be chosen efficiently, but the most efficient sampling distribution is intractable, so we use a heuristic.
- We use analytic properties of the Gaussian process belief and the EP approximation to construct a first-order probabilistic prediction for the change of $p_{\min}$ as a function of the next evaluation location.
- Finally, we choose the next evaluation point by maximizing the expected change of an information measure on the continuous but bounded domain of f: relative entropy between the uniform distribution and $p_{\min}$. This is a greedy approach, so it essentially ignores the third point above, but there is evidence that greedy approaches work well in such settings.

We also present empirical results suggesting that this approach considerably improves on the closest competitor – Expected improvement search [1], which is among the best global optimizers [2].

For simplicity, we will assume that the function f to be optimized is defined on a bounded region of a Hilbert space, and known up to a specific Gaussian process prior induced by a positive-definite kernel k. A general purpose implementation should arguably have a more general prior than this, but our derivations can be extended straightforwardly to priors defined as mixtures of Gaussian processes as gained from sampling the hyper-parameters of the Gaussian process model (see [3]).

2 Methods

2.1 Approximating the Belief over the Location of the Optimum

Assume for the moment that the optimization problem is discrete: There is a finite number of locations $\{x_i\}_{i=1,\dots,I}$, we are given a multivariate Gaussian prior $\mathcal{N}(\boldsymbol{f}(\boldsymbol{x}); \boldsymbol{\mu}, \boldsymbol{\Sigma})$ over the values of the function f on the domain, and asked to infer the belief over the identity of the minimum $p_{\min}$. In other words, what is the probability that $f(x_i) < f(x_j)\ \forall\, j \neq i$?

$$p_{\min}(x_i) = \int \mathcal{N}(\boldsymbol{f}(\boldsymbol{x}); \boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{j \neq i}^{I} \theta[f(x_j) - f(x_i)]\, \mathrm{d}\boldsymbol{f} \tag{1}$$

where θ is Heaviside's step function. This is an integral of a multivariate normal distribution, over a piecewise linear, half-open, convex integration region. Such integrals are known to be intractable, but can be approximated well with Expectation Propagation [4], with cost $\mathcal{O}(I^4)$.

However, in our setting the belief over f is an infinite-dimensional Gaussian process. Assuming the kernel is continuous, the true function is continuous, and $p_{\min}$ can be approximated arbitrarily well with a finite number of discrete representer points $\{x_n\}_{n=1,\dots,N}$. But, given that EP is costly, how should we choose the representer points such that the discrete representation comes as close to the true distribution as possible? A naïve approach would be to put x_n on a regular grid. This is intractable in high-dimensional domains, and also not efficient. The optimal choice of representer points is given by the natural metric of – the measure defined by – $p_{\min}$. Unfortunately, this being the distribution we are trying to construct from those representers, it is not available at this point in the algorithm. So we choose to sample $\{x_n\}$ from some heuristic that should be close to $p_{\min}$. Here the literature comes to our help: The expected improvement heuristic [1] $\iota(x; \mu, \Sigma; \eta)$, which is a measure relative to a current best guess function value η, is given by (Φ, ϕ are standard Gaussian cdf and pdf, respectively)

$$\iota(x) = Z^{-1} \left\{ [\eta - \mu(x)] \Phi\left(\frac{\eta - \mu(x)}{\sqrt{\Sigma(x,x)}} \right) + \sqrt{\Sigma(x,x)} \phi\left(\frac{\eta - \mu(x)}{\sqrt{\Sigma(x,x)}} \right) \right\}. \tag{2}$$

It is computationally cheap and known to contain much information about $p_{\min}$ [2]. The normalization constant Z is intractable, but irrelevant for our problem. So we approximate $p_{\min}(x)$ on the domain

of f by a staircase function defined using a set of samples $\{x_i\} \sim \iota$ and the discrete $p_{\min}(x_i)$ constructed by EP on them:

$$p_{\min}(x) \approx p_{\min}(x_c)\iota(x_c)N \qquad \text{where } x_c = \arg\min_{x_i}\{||x_i - x||\}_i \tag{3}$$

2.2 Information Measure on Continuous Domains

It is well known that the concept of entropy does not generalise trivially from probability distributions to probability densities. In continuous domains, consistent measures of uncertainty and information are provided by *relative entropies* [e.g. 5] with respect to a base measure. We choose the uniform measure $p_0(x_{\min}) = |I|^{-1}$ on the bounded search domain I as this measure, providing an uninformative prior belief on the location of the minimum (Other measures are possible, but complicate EP inference somewhat). In other words, the algorithm tries to maximize KL-divergence between its belief and uniform uncertainty. With this, the uncertainty over $p_{\min}(x)$ is approximated by our samples as (note that the step size in the staircase function is $\delta x_i \approx (\iota N)^{-1}$)

$$\mathsf{H}[p_{\min}(x)] = -\int p_{\min}(x)\log[p_{\min}(x)|I|]\,\mathrm{d}x \approx -\sum_i p_{\min}(x_i)\log[p_{\min}(x_i)\iota(x)|I|N] \tag{4}$$

This formulation offers a consistent extension of the concept of uncertainty (as measured by Shannon Entropy) from discrete to continuous domains.

2.3 Predicting Information Gain

One of the many convenient aspects of Gaussian process uncertainty is that the predicted *change* of the belief after evaluation at location x_e is itself a sample from a Gaussian process, with a covariance function given by an *innovation function* $L(x^*; x_e)$ (details can be found in [6, in press]). The change to mean and covariance functions are

$$\delta\mu(x^*)|_{x_e} = L(x^*, x_e)\omega \quad \text{where } \omega \sim \mathcal{N}(0,1) \text{ and } \delta\Sigma(x^*, x_*)|_{x_e} = -L(x^*; x_e)L(x_*; x_e) \tag{5}$$

The EP approximation to $p_{\min}$ also provides analytic derivatives with respect to μ and Σ [7]. So we can construct a first-order prediction of the change of relative entropy as (using Itō's lemma, and the sum convention)

$$\mathsf{E}[\delta\mathsf{H}]|_{x_e} \approx \int \mathsf{H}\left[p_{\min,0} + \frac{\partial p_{\min}}{\partial\Sigma_i}\delta\Sigma^i_{x_e} + \frac{1}{2}\frac{\partial^2 p_{\min,0}}{\partial\mu_i\partial\mu_j}\delta\mu^i_{x_e}\delta\mu^j_{x_e} + \frac{\partial p_{\min,0}}{\partial\mu_i}\delta\mu^i_{x_e}\cdot\omega\right]\phi(\omega)\,\mathrm{d}\omega. \tag{6}$$

The expectation over the Gaussian uncertainty can most simply be taken by Monte Carlo integration using a small number of samples. The resulting function is differentiable if the kernel is differentiable, and can thus be efficiently optimized locally. So our algorithm boils down to choosing the next evaluation point as the one locally minimizing $\mathsf{E}[\delta\mathsf{H}]|_{x_e}$. Note that our choice of logarithmic loss function (relative entropy) for this decision problem has no bearing on computational complexity – should one prefer a different loss, it is trivial to use it instead.

3 Experiments

Figure 1 provides intuition for the new paradigm on a simple example of a 1-dimensional function, given 3 observations and a rational quadratic kernel. See caption for details, note the structural difference between the *local* expected improvement heuristic and the *global* view afforded by $p_{\min}$ and the information theoretic description. Figure 2 compares the performance of the new algorithm, *Entropy Search* to the expected improvement heuristic for global search [1, 2], on two standard test functions (Branin's and Goldstein's & Price's functions).

4 Conclusion

We have outlined information-theoretic desiderata for global optimization, pointed out that they are analytically challenging, then suggested a number of approximations to address each challenge. The result is an information-greedy global optimization algorithm for functions defined on Hilbert spaces, and known up to Gaussian process priors. Empirical evidence suggests this treatment leads to considerable performance increase over one of the best known competitors.

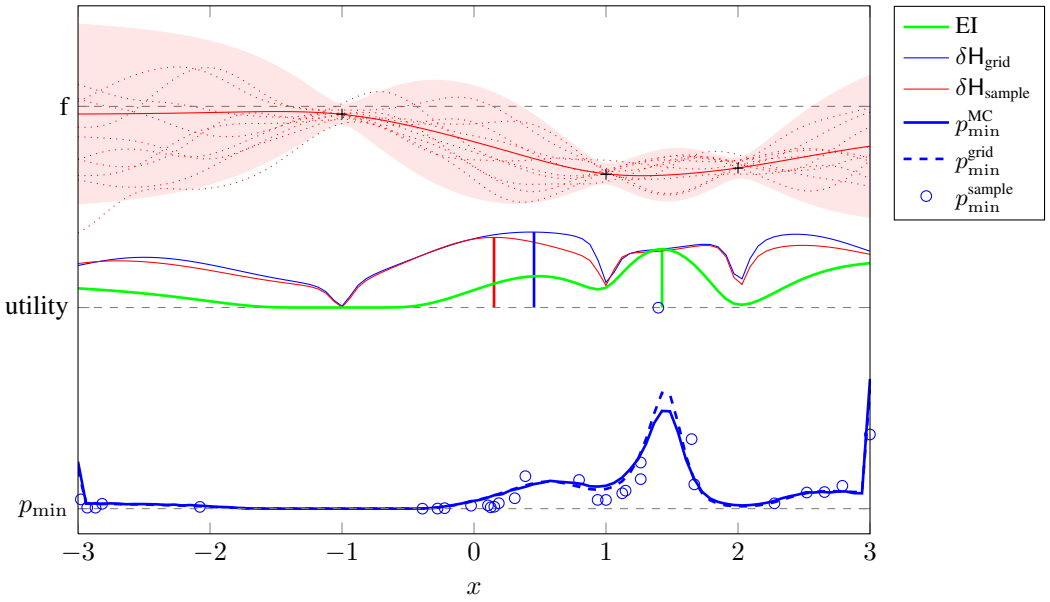

Figure 1: Overview of problem and algorithm. **Top:** 3 observed datapoints, GP belief with samples, mean, marginal variance. **Middle:** Expected Improvement heuristic in green, next evaluation point at vertical mark. Predicted change of Entropy, from a regular grid of 100 points (blue) or 30 samples as described in the text (red), with next evaluation points marked. Note that the information-greedy algorithm takes nonlocal structure into account, like the extend of the uncertain region around the evaluation point, while the EI heuristic does not. **Bottom:** Belief over location of minimum, from MC samples (solid), from EP on regular grid (dashed, mostly overlapping solid line), and 30 samples. Note the nontrivial edge-effects, not taken into account by the EI heuristic.

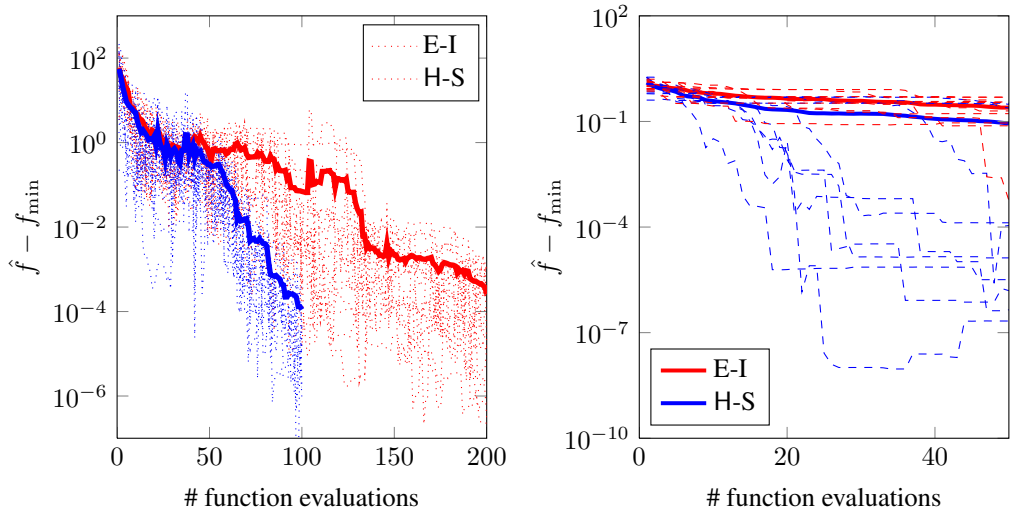

Figure 2: Performance on two test environments (value of function at optimizer's best guess, minus true minimum). Ten individual experiments as thin lines, means as thick lines. E-I: Expected Improvement. H-S: Entropy search (this work). **Left:** Branin's function. The H-searcher is between 1.5 and 2 times faster. **Right:** Goldstein-Price function. The EI heuristic does not typically discover the minimum within the 50 alloted evaluations, while the H-Searcher usually does.

References

[1] 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.

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

[3] M.A. Osborne, R. Garnett, and S.J. Roberts. Gaussian processes for global optimization. In *3rd International Conference on Learning and Intelligent Optimization (LION3)*, 2009.

[4] T.P. Minka. Expectation Propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann.

[5] E.T. Jaynes and G.L. Bretthorst. *Probability Theory: the Logic of Science*. Cambridge University Press, 2003.

[6] P. Hennig. Optimal reinforcement learning for Gaussian systems. In *Advances in Neural Information Processing Systems*, 2011.

[7] M. Seeger. Expectation propagation for exponential families. Technical report, U.C. Berkeley, 2008.