Predictive Entropy Search for Bayesian Optimization with Unknown Constraints

José Miguel Hernández-Lobato[†*], Michael A. Gelbart[†*], Matthew W. Hoffman[‡], Ryan P. Adams[†], Zoubin Ghahramani[‡]

Harvard University[†], University of Cambridge[‡], Authors contributed equally[*]

Constrained Bayesian optimization

Problem: we are interested in solving

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \quad \text{s.t.} \quad c_1(\mathbf{x}) \geq 0, \ldots, c_K(\mathbf{x}) \geq 0 \qquad (1)$$

where

- the objective f and constraints c_i are evaluated via expensive, black-box queries,
- we sequentially select inputs $\mathbf{x}_t$ and
- we observe outputs $\mathbf{y}_t = [f(\mathbf{x}_t), c_1(\mathbf{x}_t)\epsilon_t^1, \ldots, c_K(\mathbf{x}_t)]^{\text{T}}$

A Bayesian approach: Given observations $\mathcal{D}_t = (\mathbf{x}_{1:t}, \mathbf{y}_{1:t})$ we use Gaussian processes (GPs) to construct Bayesian posteriors over the unknown functions f and c_i. These posteriors are then used to select $\mathbf{x}_{t+1}$ by maximizing an acquisition function $\alpha(\mathbf{x})$ which takes the information gained about the constrained optimizer into account.

The Challenge: to construct an acquisition function that can be used in all scenarios, even when the objective and constraints may be evaluated independently (decoupled).

Predictive entropy search with constraints

We take the information-based approach, and maximize the expected information gain about the location of the global constrained optimizer $\mathbf{x}_\star$:

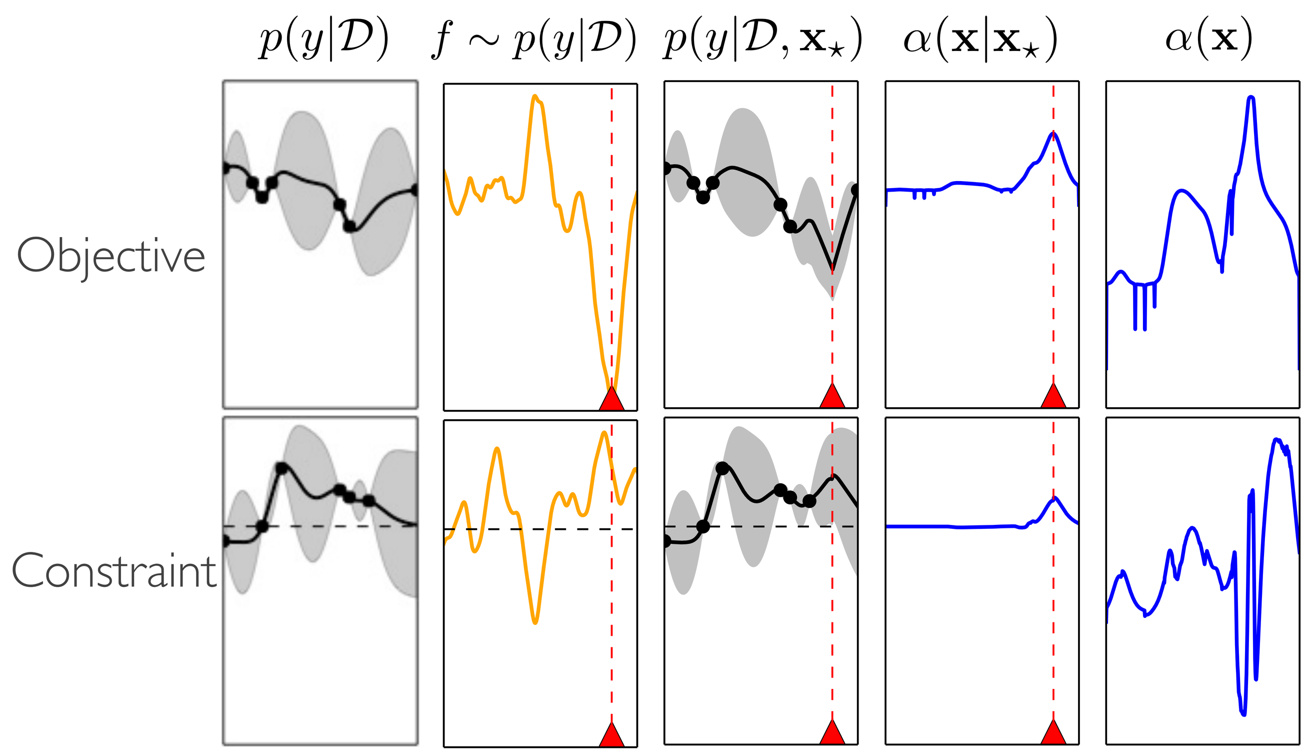
$$\alpha(\mathbf{x}) = \text{H}\left[\mathbf{x}_\star|\mathcal{D}\right] - \mathbb{E}_{\mathbf{y}}\left\{\text{H}\left[\mathbf{x}_\star|\mathcal{D}\cup(\mathbf{x},\mathbf{y})\right]\right\} \qquad (2)$$

Following Hernández-Lobato et al. (2014) we rewrite this as:

$$\alpha(\mathbf{x}) = \text{H}\left[\mathbf{y}|\mathcal{D}\right] - \mathbb{E}_{\mathbf{x}_\star}\left\{\text{H}\left[\mathbf{y}|\mathcal{D},\mathbf{x},\mathbf{x}_\star\right]\right\} \qquad (3)$$

- The term $\text{H}[\mathbf{y}|\mathcal{D}]$ is the entropy of a product of independent Gaussians, and is computable in closed form.
- The second term involves an expectation which we approximate by averaging over samples of $\mathbf{x}_\star$ drawn by approximate Thompson sampling.
- The second term involves an entropy which we compute by approximating the conditioned predictive distribution (CPD) $p(\mathbf{y}|\mathcal{D},\mathbf{x},\mathbf{x}_\star)$ with a Gaussian using Expectation Propagation (EP). This is described below.

Visualizing the PESC approximation



This figure shows (from left to right) the posterior predictive distribution, a sample of $\mathbf{x}_\star$, the PESC approximation to the CPD, the acquisition function for the one shown sample of $\mathbf{x}_\star$, and the acquisition function averaged of 100 samples of $\mathbf{x}_\star$.

Approximating the conditioned predictive distribution (CDP)

First, let $\Psi(\mathbf{x})$ denote the condition that any point $\mathbf{x} \neq \mathbf{x}_\star$ must be sub-optimal if the constraints are satisfied at $\mathbf{x}$:

$$\Psi(\mathbf{x}) = \left(\prod_{k=1}^K \Theta\left[c_k(\mathbf{x})\right]\right)\Theta[f(\mathbf{x}) - f(\mathbf{x}_\star)] + \left(1 - \prod_{k=1}^K \Theta\left[c_k(\mathbf{x})\right]\right)$$

Next let $\mathbf{z} = [f(\mathbf{x}), c_1(\mathbf{x}), \ldots, c_K(\mathbf{x})]^{\text{T}}$ denote the value of the objective and constraint functions at some test input $\mathbf{x}$. The distribution of these latent values can be written as

$$p(\mathbf{z}|\mathcal{D},\mathbf{x},\mathbf{x}_\star) \propto \int \delta[z_0 - f(\mathbf{x})]\left[\prod_{k=1}^K \delta[z_k - c_k(\mathbf{x})]\right]\left[\prod_{k=1}^K \Theta\left[c_k(\mathbf{x}_\star)\right]\right]$$
$$\left[\prod_{\mathbf{x}'\neq\mathbf{x}} \Psi(\mathbf{x}')\right]\Psi(\mathbf{x})p(f, c_1, \ldots, c_K|\mathcal{D})\,df\,dc_1\ldots dc_K$$

However, we will approximate this by considering only the finite points observed in our dataset. We first approximate the factors that do not depend on $\mathbf{x}$ as

$$q_1(\mathbf{f},\mathbf{c}_1,\ldots,\mathbf{c}_K) = \left[\prod_{k=1}^K \Theta[c_{k0}]\right]\left[\prod_{n=1}^N \Psi(\mathbf{x}_n)\right]p(\mathbf{f},\mathbf{c}_1,\ldots,\mathbf{c}_K|\mathcal{D})$$

and approximate this resulting distribution with EP

$$q_2(\mathbf{f},\mathbf{c}_1,\ldots,\mathbf{c}_K) = \mathcal{N}(\mathbf{f}|\mathbf{m}_0,\mathbf{V}_0)\prod_{k=1}^K \mathcal{N}(\mathbf{c}_k|\mathbf{m}_k,\mathbf{V}_k)\,,$$

Plugging this into the earlier equation, our full approximation is then

$$p(\mathbf{z}|\mathcal{D},\mathbf{x},\mathbf{x}_\star) \approx Z_2^{-1}\int p(\mathbf{z}|\mathbf{f},\mathbf{c}_1,\ldots,\mathbf{c}_K)\Psi(\mathbf{x})q_2(\mathbf{f},\mathbf{c}_1,\ldots,\mathbf{c}_K)\,d\mathbf{f}\,d\mathbf{c}_1\cdots d\mathbf{c}_K\,,$$

All variables other than $z_0 = f(\mathbf{x})$ are integrated out and a final approximation can be employed approximating this distribution as a Gaussian. The full acquisition function is,

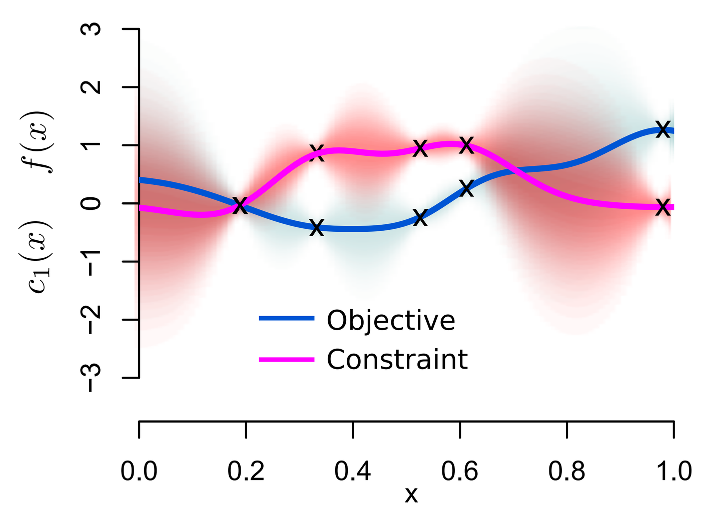
$$\alpha(\mathbf{x}) = \left\{\log v_f(\mathbf{x}) + \sum_{k=1}^K \log v_k(\mathbf{x})\right\} - \frac{1}{M}\sum_{m=1}^M\left\{\log \hat{v}_f\left(\mathbf{x}|\mathbf{x}_\star^{(m)}\right) + \sum_{k=1}^K \log \hat{v}_k\left(\mathbf{x}|\mathbf{x}_\star^{(m)}\right)\right\}$$

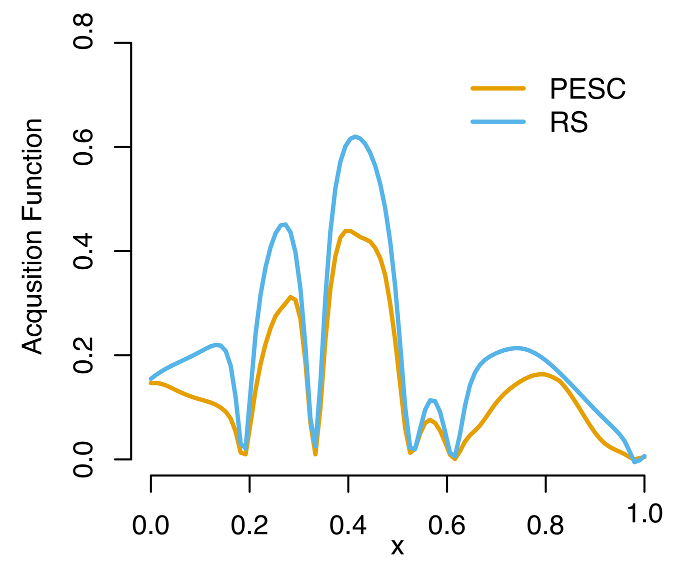
where v_f and v_k are the posterior predictive variances and $\hat{v}_f$ and $\hat{v}_k$ are these variances conditioned on $\mathbf{x}_\star$. Hyperparameters are easily marginalized out using MCMC.

Accuracy of the PESC approximations

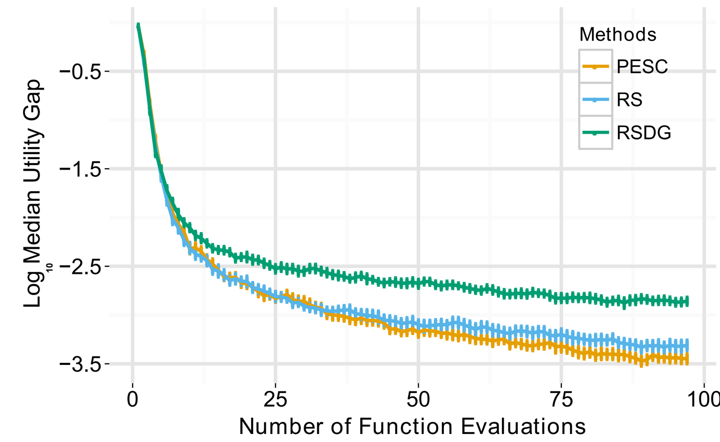
We compare the PESC approximation with ground truth computed using rejection sampling (RS) on a dense grid.

Shown below on the **left** is the posterior for a 1d objective with a single constraint. The plot on the **right** compares the accuracy of PESC to the *ground-truth* RS.



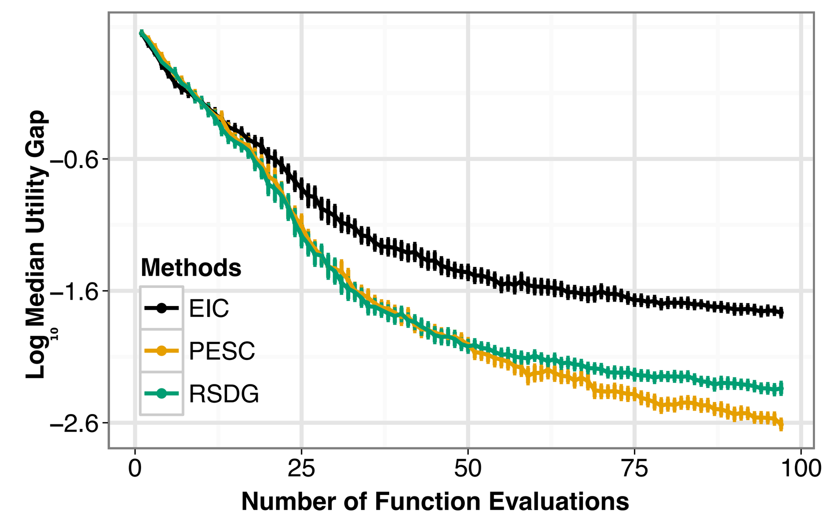


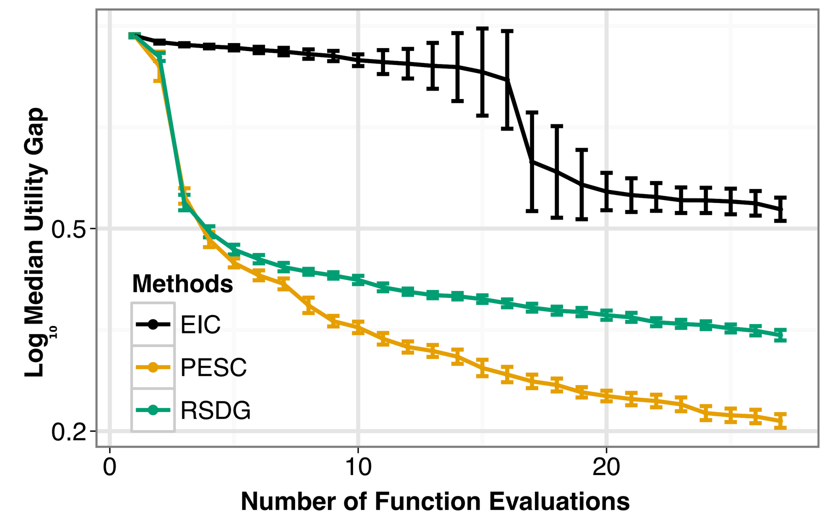
Shown to the **right** is the resulting performance on this problem. Also shown is a method which uses *rejection sampling on a dynamic grid (RSDG)* which samples grid points from $p(\mathbf{x}_\star)$. This scales better to higher dimensions than RS because the number of grid points is fixed.



Results on synthetic functions

Below we extend these experiments to 2-dimensional (**left**) and 8-dimensional (**right**) synthetic problems and compare against *expected improvement with constraints (EIC)*.



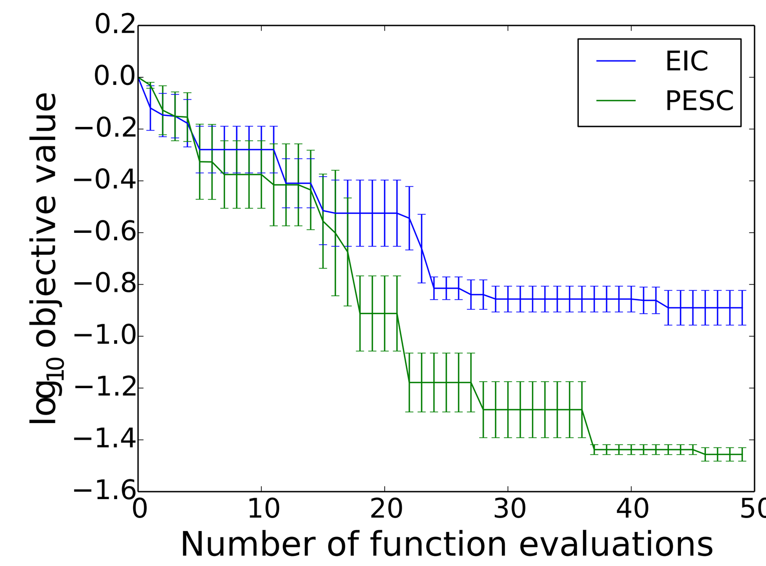


Constrained hyperparameter optimization

We studied the performance of PESC on two constrained hyperparameter optimization tasks.

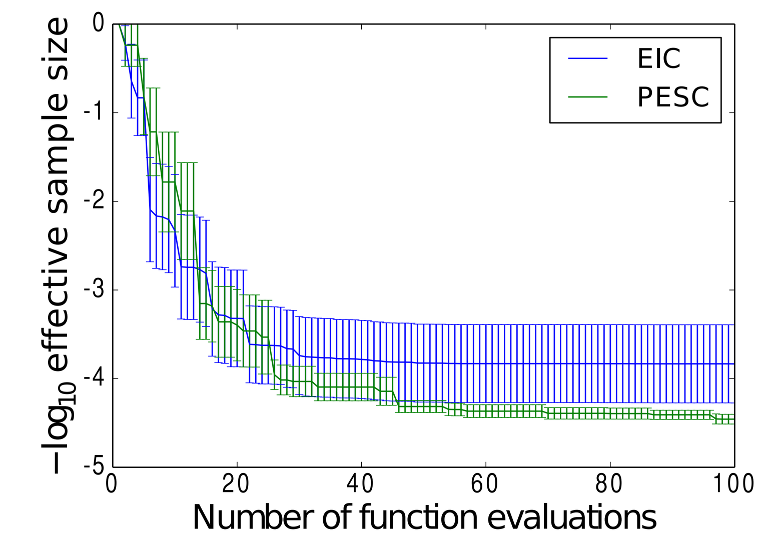
Tuning a fast neural network

We tune the hyperparameters of a three-hidden-layer neural network subject to the constraint that the prediction time must not exceed 2 ms. The search space consists of 12 parameters including parameters for learning rates, momentum, dropout, regularization, number of units, and activation function.



Tuning Hamiltonian MCMC

We optimize the number of effective samples produced by HMC limited to 5 minutes of computation time, subject to convergence diagnostics and a non-divergence constraint. Parameters include the integration step size and number of steps, fraction of time spent in burn-in, and an HMC mass parameter.



Related work

Gelbart, Michael A., Snoek, Jasper, and Adams, Ryan P. Bayesian optimization with unknown constraints. In *UAI*, 2014.

Gramacy, Robert B. and Lee, Herbert K. H. Optimization under unknown constraints. *Bayesian Statistics*, 9, 2011.

Gramacy, Robert B., Gray, Genetha A., Digabel, Sebastien Le, Lee, Herbert K. H., Ranjan, Pritam, Wells, Garth, and Wild, Stefan M. Modeling an augmented Lagrangian for improved blackbox constrained optimization, 2014. arXiv:1403.4890v2 [stat.CO].

Hernández-Lobato, J. M, Hoffman, M. W., and Ghahramani, Z. Predictive entropy search for efficient global optimization of black-box functions. In *NIPS*. 2014.

Picheny, Victor. A stepwise uncertainty reduction approach to constrained global optimization. In *AISTATS*, 2014.