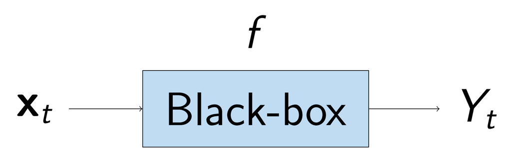
Predictive Entropy Search for Efficient Global Optimization of Black-box Functions

José Miguel Hernández-Lobato, Matthew W. Hoffman, Zoubin Ghahramani
University of Cambridge

UNIVERSITY OF CAMBRIDGE

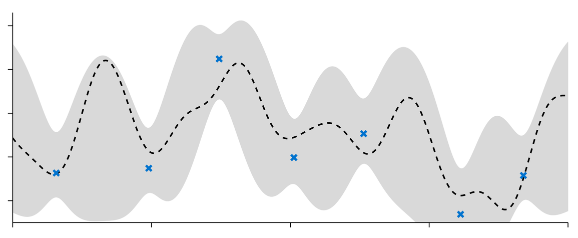
Bayesian black-box optimization

We are interested in solving black-box optimization problems of the form $\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$ by sequentially querying points $\mathbf{x}_t$ and observing Y_t.



Here black-box means:

- we may only be able to observe the function value, i.e. no gradients
- our observations may be corrupted by noise, i.e. we observe Y_t, but $f(\mathbf{x}_t) = \mathbb{E}[Y_t|\mathbf{x}_t]$



Given some function to optimize (the dashed line) we can make observations Y_t and construct a Bayesian posterior over the (unknown!) function.

This posterior can then be used to guide our search, i.e. the selection of inputs $\mathbf{x}_t$.

A framework for Bayesian optimization

The following pseudocode outlines an algorithmic framework for Bayesian optimization:

```
for t = 1, ..., T do
    select point x_t = arg max_x α_{t-1}(x)        Exploration     select the next point to query by maximizing some acquisition function
    observe y_t ~ p(·|x_t)
    update posterior p(f|D_t)                      Prediction      posterior model for prediction; generally use a GP
end for
return x̃_T                                         Recommendation  return final recommendation; can maximize the posterior mean
```

The acquisition function used to explore the function should in some sense try to gain as much information about the optimizer location as possible.

Predictive Entropy Search

A common active learning approach is to select points which maximize the expected reduction in posterior entropy about some predictor. Applying this to optimization as in [2, 3], let $\mathbf{x}_\star$ be the unknown optimizer and write its information gain:

$$\alpha_t(\mathbf{x}) = \mathrm{H}\left[\mathbf{x}_\star|\mathcal{D}_t\right] - \mathbb{E}_{P(y|\mathcal{D}_t,\mathbf{x})}\Big[\mathrm{H}\left[\mathbf{x}_\star|\mathcal{D}_t \cup \{(\mathbf{x},y)\}\right]\Big] \tag{ES}$$

But:
- the distribution $P(\mathbf{x}_\star|\dots)$ has no closed form expression; and
- this expensive (and rough) approximation must be done for every $(\mathbf{x}, y)$

Note, however, that the information gain is symmetric. Changing the order of these arguments we can write the acquisition as

$$\alpha_t(\mathbf{x}) = \mathrm{H}\left[y|\mathcal{D}_t,\mathbf{x}\right] - \mathbb{E}_{p(\mathbf{x}_\star|\mathcal{D}_t)}\Big[\mathrm{H}\left[y|\mathcal{D}_t,\mathbf{x},\mathbf{x}_\star\right]\Big] \tag{PES}$$

which we call Predictive Entropy Search.

This requires:
- sampling from the distribution over maximizers $\mathbf{x}_\star$, and
- computing the predictive entropy conditioned on this maximizer.

Sampling optima

To sample an optimum we need only sample $f \sim p(\cdot|\mathcal{D}_t)$ and return $\arg\max_\mathbf{x} f(\mathbf{x})$.

- if our set of possibly query points is discrete this is Thompson sampling;
- however, f is an infinite dimensional object!

Instead we will approximately sample a $f(\cdot) = \phi(\cdot)^\mathsf{T}\boldsymbol{\theta}$ where ϕ consist of random features and $\boldsymbol{\theta}$ are sampled from the resulting approximate posterior.

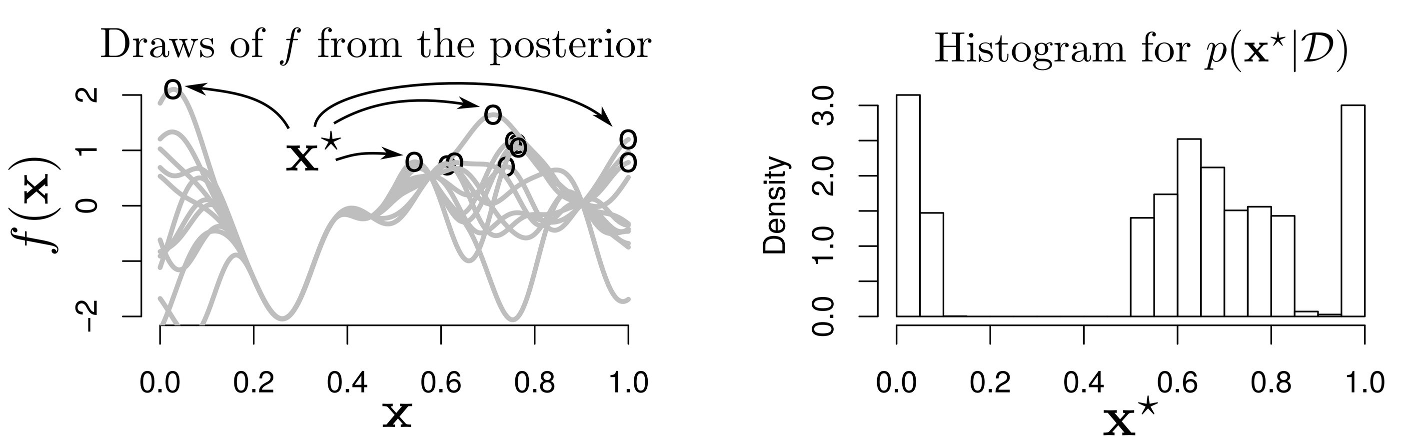
Via Bochner's theorem [1] a shift-invariant kernel can be written as the Fourier transform of its spectral density. Treating this as a probability distribution $p(\mathbf{w})$ we can write:

$$k(\mathbf{x},\mathbf{x}') = \alpha\ \mathbb{E}_{p(\mathbf{w})}[e^{-i\mathbf{w}^\mathsf{T}(\mathbf{x}-\mathbf{x}')}] = 2\alpha\ \mathbb{E}_{p(\mathbf{w},b)}[\cos(\mathbf{w}^\mathsf{T}\mathbf{x}+b)\cos(\mathbf{w}^\mathsf{T}\mathbf{x}'+b)] \tag{1}$$

where b is uniformly distributed in $[0, 2\pi]$. Letting

$$\phi(\mathbf{x}) = \sqrt{2\alpha/m}\cos(\mathbf{W}\mathbf{x}+\mathbf{b})$$

be the m-dimensional feature map we can approximate the GP posterior with a simple linear-Gaussian model; can sample $\boldsymbol{\theta}$ directly.



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

[2] P. Hennig and C. J. Schuler. *Entropy search for information-efficient global optimization. Journal of Machine Learning Research*, 13, 2012.

[3] 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.

Approximating the entropy

To construct $\alpha_t(\mathbf{x})$ we can approximate the fact that $\mathbf{x}_\star$ is a maximum with the following constraints:

- $\nabla f(\mathbf{x}_\star) = 0$;
- $\mathrm{upper}[\nabla^2 f(\mathbf{x}_\star)] = 0$;
- $\mathrm{diag}[\nabla^2 f(\mathbf{x}_\star)] < 0$.
- $f(\mathbf{x}_\star) > \max_t f(\mathbf{x}_t)$.
- $f(\mathbf{x}_\star) > f(\mathbf{x})$.

The first two constraints can be incorporated exactly, producing $\mathcal{N}(\mathbf{z}|\mathbf{m}_0,\mathbf{V}_0)$ where $\mathbf{z}$ are the latent values for the maximizer and Hessian. The second set of constraints can be approximated as

$$\mathcal{N}(\mathbf{z}|\mathbf{m}_0,\mathbf{V}_0) \times \underbrace{\Phi_{\sigma^2}(f(\mathbf{x}_\star) - y_{\max})}_{\text{maximizer constraint}} \times \underbrace{\prod_{i=1}^{d} \mathbb{I}\left([\nabla^2 f(\mathbf{x}_\star)]_{ii} \le 0\right)}_{\text{Hessian constraints}}$$

Using Expectation Propagation we can approximate this density with a single multivariate Gaussian, leading to a Gaussian distribution over the latent $f(\mathbf{x}_\star)$. This need only be computed once per iteration.

Finally, given any $\mathbf{x}$ we can compute the joint $p(f(\mathbf{x}), f(\mathbf{x}_\star))$. An additional factor can be incorporated requiring $f(\mathbf{x}) < f(\mathbf{x}_\star)$ and approximated using a single step of EP.

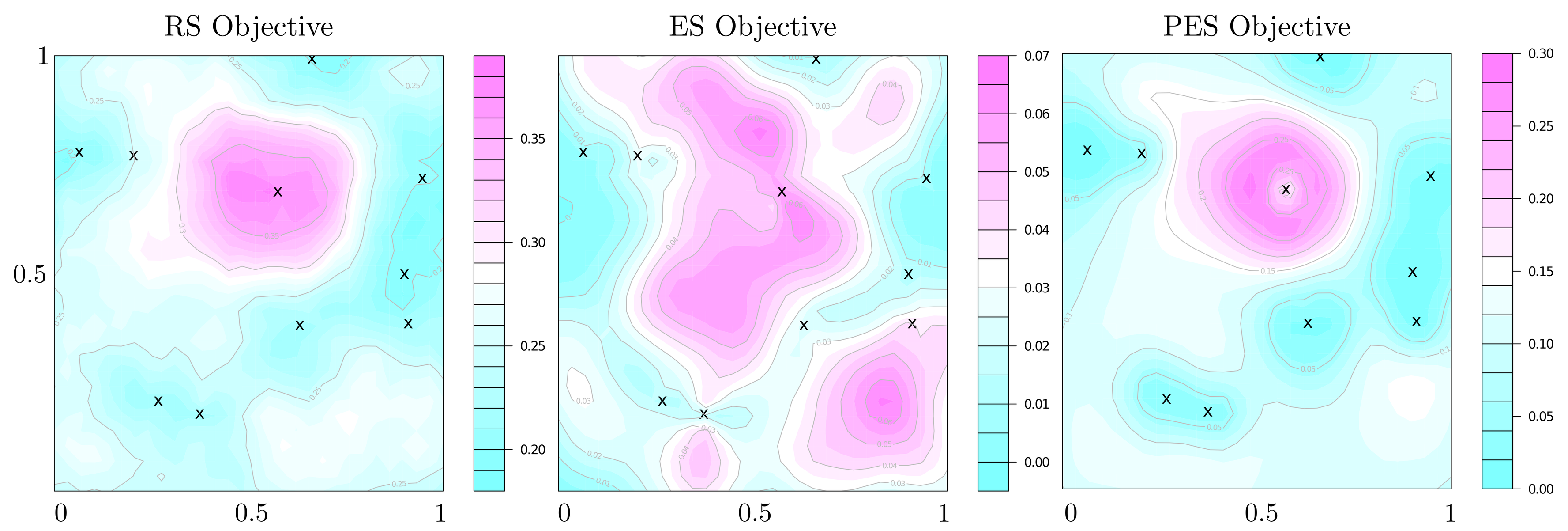
Taking entropies of these Gaussian distributions the acquisition function is:

$$\alpha_n(\mathbf{x}) = 0.5\log[v_n(\mathbf{x}) + \sigma^2] - 0.5\log[v_n(\mathbf{x}|\mathbf{x}_\star) + \sigma^2]$$

where $v_n(\mathbf{x})$ and $v_n(\mathbf{x}|\mathbf{x}_\star)$ are the unconditioned and conditioned variances.

Accuracy of the PES acquisition

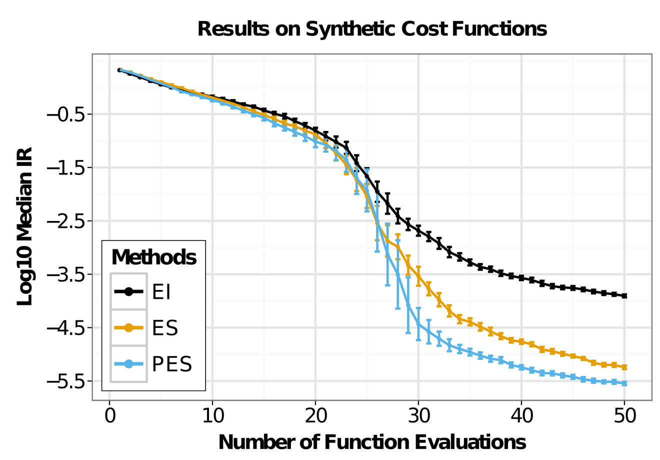
The following compares a fine-grained random sampling (RS) scheme to compute the ground truth objective with ES and PES. We see PES provides a much better approximation.

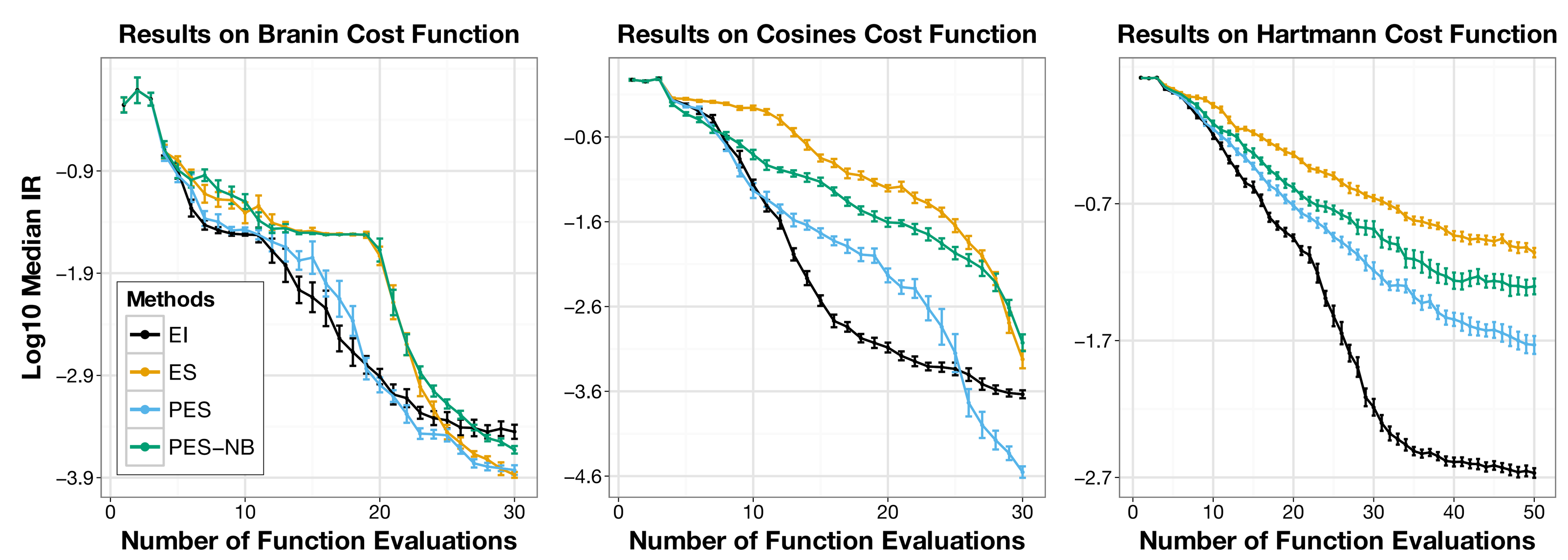


Performance of PES

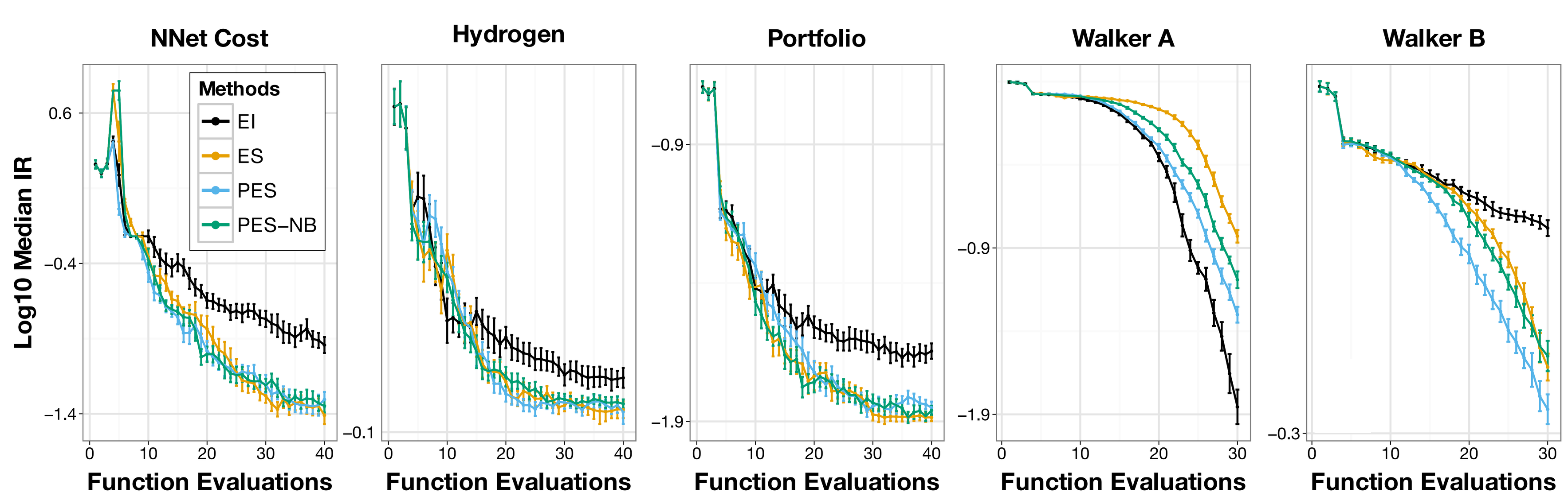
The *Right* plot shows the performance on a function randomly sampled from the given GP prior.

The *Bottom* plots compare the performance on various global optimization test problems.





We further consider several real-world experiments. (NNet) the hyperparameters of a neural network; (Hydrogen) the hydrogen production of a particular bacteria; (Portfolio) the Sharpe ratio of 1-year simulated returns; and the speed of a bipedal robot under (Walker A) noiseless and (Walker B) noisy observations.



Note also: the Walker-A experiment had less noise, and hence here EI showed better performance, due to its more exploitative behavior.