Predictive Entropy Search for Bayesian Optimization with Unknown Constraints

José Miguel Hernández-Lobato*

joint work with
*Michael A. Gelbart**, *Matt W. Hoffman, Ryan P. Adams and Zoubin Ghahramani*.

July 8, 2015,

* Authors contributed equally

HARVARD
INTELLIGENT
PROBABILISTIC
SYSTEMS

Bayesian optimization with Unknown Constraints

- We aim to solve **black-box** constrained optimization problems:

$$\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \quad \text{s.t.} \quad c_1(\mathbf{x}) \geq 0, \ldots, c_K(\mathbf{x}) \geq 0 \,.$$

Coupled evaluations!

$\mathbf{x}_t$ query inputs → Black-box, f → y_f noisy outputs

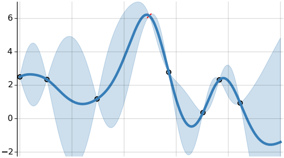
$\mathbf{x}_t$ query inputs → Black-box, c_1 → y_1 noisy outputs

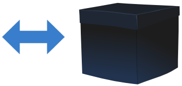
...

$\mathbf{x}_t$ query inputs → Black-box, c_K → y_K noisy outputs

- Queries are **very expensive** (time, economic cost, etc.).

Gaussian Process Predictions





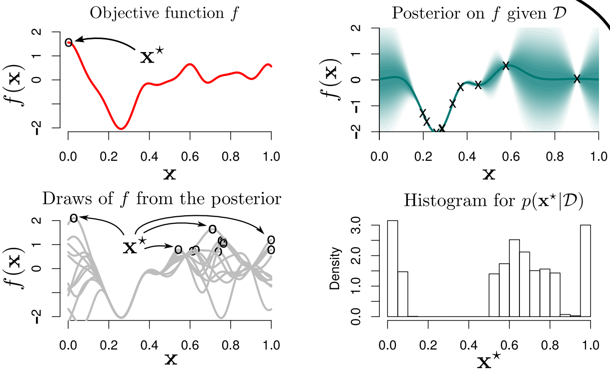
Entropy Search (ES) in the Unconstrained Case

Let $\mathbf{x}_\star$ be the **global optimum**. Entropy search (ES) maximizes the expected reduction in the **entropy** of the **posterior on $\mathbf{x}_\star$**.

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

How much we know about $\mathbf{x}_\star$ now.

How much we will know about $\mathbf{x}_\star$ after collecting y at $\mathbf{x}$.



Computing (ES) is **very difficult in practice**!

Villemonteix et al. [2009], Hennig and Schuler [2012]

Predictive Entropy Search with Constraints (PESC)

We can swap y and $\mathbf{x}_\star$ to obtain a new reformulation which we call **Predictive Entropy Search** (PES) (**Hernández-Lobato et al. [2014]**):

$$\alpha_t(\mathbf{x}) = \mathrm{H}\left[\mathbf{x}_\star \middle| \mathcal{D}_t\right] - \mathbb{E}_y\left[\mathrm{H}\left[\mathbf{x}_\star \middle| \mathcal{D}_t \cup \{\mathbf{x}, y\}\right] \middle| \mathcal{D}_t, \mathbf{x}\right] \equiv \mathrm{MI}(y, \mathbf{x}_\star) \qquad \text{(ES)}$$

$$\alpha_t(\mathbf{x}) = \mathrm{H}\left[y \middle| \mathcal{D}_t, \mathbf{x}\right] - \underbrace{\mathbb{E}_{\mathbf{x}_\star}}_{❶}\left[\underbrace{\mathrm{H}\left[y \middle| \mathcal{D}_t, \mathbf{x}, \mathbf{x}_\star\right]}_{❷} \middle| \mathcal{D}_t, \mathbf{x}\right] \equiv \mathrm{MI}(\mathbf{x}_\star, y) \qquad \text{(PES)}$$

❶ Approximated by sampling from $p(\mathbf{x}_\star | \mathcal{D}_t)$ (**Thompson sampling**).

❷ Approximated with **expectation propagation** (Minka [2001]).

The PES acquisition function is the same in the constrained case:

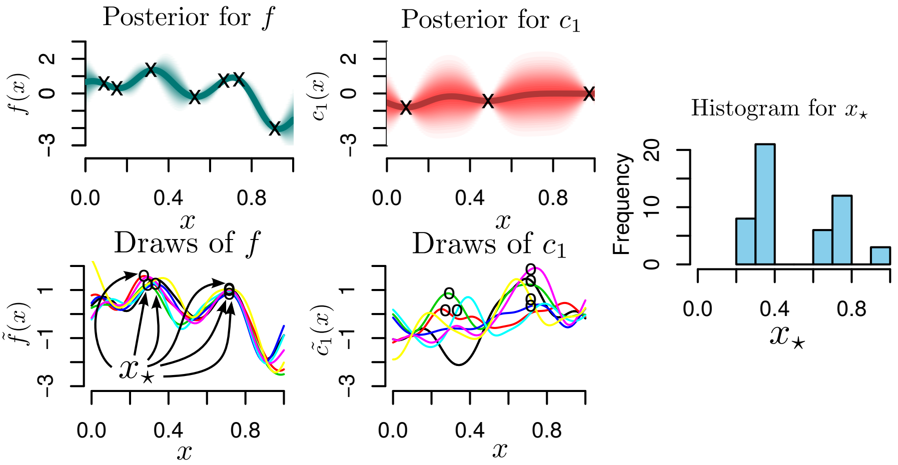
$$\alpha_t(\mathbf{x}) = \mathrm{H}\left[\mathbf{y} \middle| \mathcal{D}_t, \mathbf{x}\right] - \underbrace{\mathbb{E}_{\mathbf{x}_\star}}_{❶}\left[\underbrace{\mathrm{H}\left[\mathbf{y} \middle| \mathcal{D}_t, \mathbf{x}, \mathbf{x}_\star\right]}_{❷} \middle| \mathcal{D}_t, \mathbf{x}\right], \qquad \text{(PESC)}$$

with $\mathbf{y} = (y_f, y_1, \ldots, y_K)^\mathsf{T}$.

$$\alpha_t(\mathbf{x}) = \mathrm{H}\left[\mathbf{y} \middle| \mathcal{D}_t, \mathbf{x}\right] - \underset{1}{\mathbb{E}_{\mathbf{x}_\star}} \left[\mathrm{H}\left[\mathbf{y} \middle| \mathcal{D}_t, \mathbf{x}, \mathbf{x}_\star\right] \middle| \mathcal{D}_t, \mathbf{x} \right], \qquad \text{(PESC)}$$

Step 1: Sampling the Optimum $\mathbf{x}_\star$ (I)

We sample $\tilde{f} \sim p(f|\mathcal{D}_t)$ and $\tilde{c}_1 \sim p(c_1|\mathcal{D}_t), \ldots, \tilde{c}_K \sim p(c_1|\mathcal{D}_t)$ and return $\arg\max_{\mathbf{x}} \tilde{f}(\mathbf{x})$ s.t. $\tilde{c}_1(\mathbf{x}) \geq 0, \ldots \tilde{c}_K(\mathbf{x}) \geq 0$.

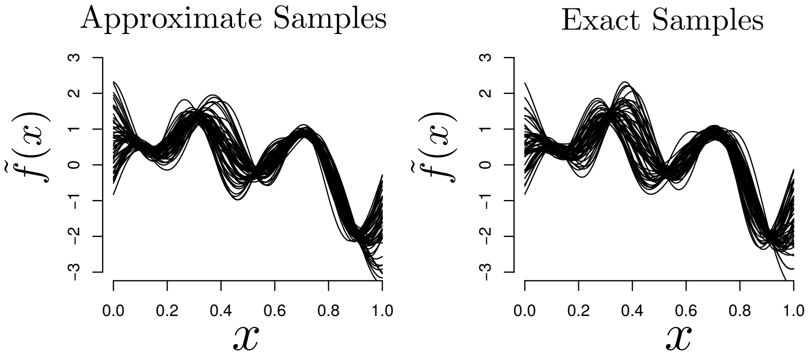


Step 1: Sampling the Optimum $\mathbf{x}_\star$ (II)

However, $\tilde{f}$ and $\tilde{c}_1, \ldots, \tilde{c}_K$ are an infinite dimensional objects!

Instead we use $\tilde{f}(\cdot) \approx \phi(\cdot)^\mathsf{T}\boldsymbol{\theta}$ where $\phi(\mathbf{x}) = \sqrt{2\alpha/m}\cos(\mathbf{W}\mathbf{x} + \mathbf{b})$.

Bochner's theorem shows that when $m \to \infty$ the approximation is exact.



We choose $m = 500$.

Bochner [1959]

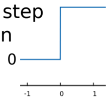
$$\alpha_t(\mathbf{x}) = \mathsf{H}\left[\mathbf{y}\middle|\mathcal{D}_t, \mathbf{x}\right] - \mathbb{E}_{\mathbf{x}_\star}\left[\underbrace{\mathsf{H}\left[\mathbf{y}\middle|\mathcal{D}_t, \mathbf{x}, \mathbf{x}_\star\right]}_{❷}\middle|\mathcal{D}_t, \mathbf{x}\right], \qquad \text{(PESC)}$$

Step 2: Approximating $p(\mathbf{y}|\mathcal{D}_t, \mathbf{x}, \mathbf{x}_\star)$

$$\Psi(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \text{ is a better solution than } \mathbf{x}_\star . \\ 1 & \text{otherwise.} \end{cases}$$

Heaviside step function

$\Theta(\mathbf{x}) \rightarrow$



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

Constraints satisfied

x is not optimal

Constraints not satisfied

Let $\mathbf{z} = [f(\mathbf{x}), c_1(\mathbf{x}), \ldots, c_K(\mathbf{x})]^{\mathsf{T}}$, then

$$p(\mathbf{z}|\mathcal{D}, \mathbf{x}, \mathbf{x}_\star) \propto \int \delta[z_0 - f(\mathbf{x})] \left[\prod_{k=1}^{K} \delta[z_k - c_k(\mathbf{x})]\right] \left[\prod_{k=1}^{K} \Theta\left[c_k(\mathbf{x}_\star)\right]\right]$$

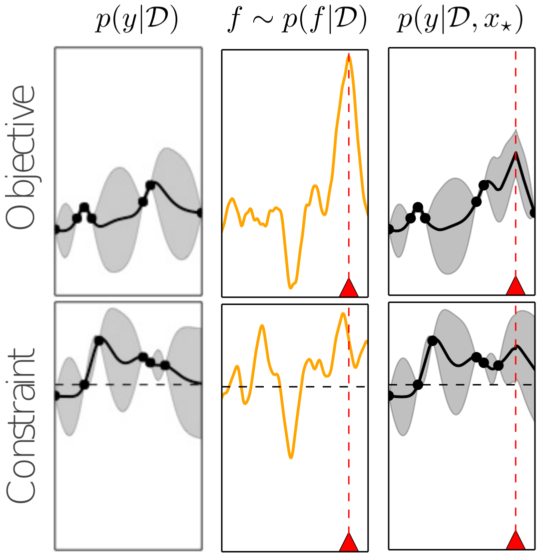
$$\left[\prod_{\mathbf{x}' \neq \mathbf{x}_\star} \Psi(\mathbf{x}')\right] p(f, c_1, \ldots, c_K|\mathcal{D})\, df\, dc_1 \ldots dc_k$$

No other point is a better solution than $\mathbf{x}_\star$

$\mathbf{x}_\star$ must be feasible

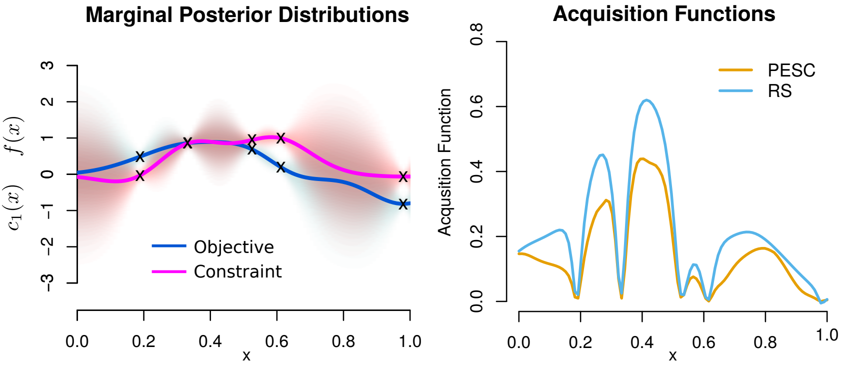
We find a **Gaussian** approximation using **expectation propagation**.

Visualizing the Approximation to $p(\mathbf{y}|\mathcal{D}_t, \mathbf{x}, \mathbf{x}_\star)$



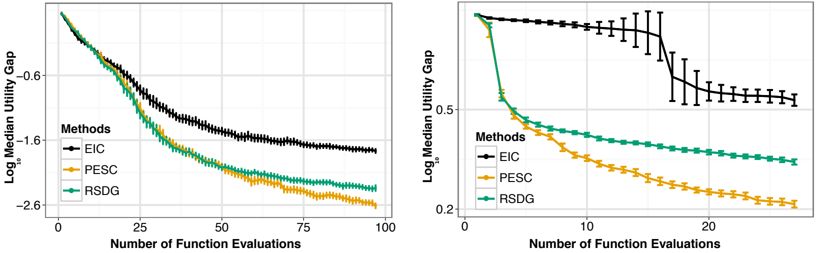
Accuracy of the PESC Approximation to $\alpha(\mathbf{x})$

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



Results on Synthetic Functions

Below we show experiments with 2-dimensional (**left**) and 8-dimensional (**right**) synthetic problems.



Baseline: **expected improvement with constraints** (EIC):

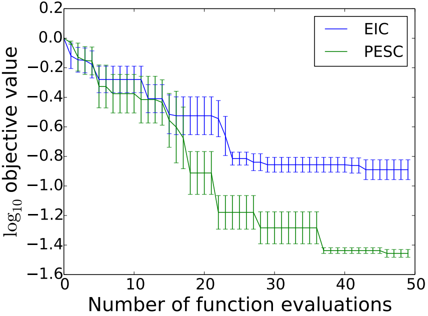
$$\alpha_t(\mathbf{x}) = \mathbb{E}\Big[\max\Big(0, f(\mathbf{x}) - f(\mathbf{x}_+)\Big)\Big|\mathcal{D}_t\Big]\Big[\prod_{k=1}^{K} p(c_k(\mathbf{x}) \geq 0)\Big]$$

Baseline: **rejection sampling on a dynamic grid** (RSDG).

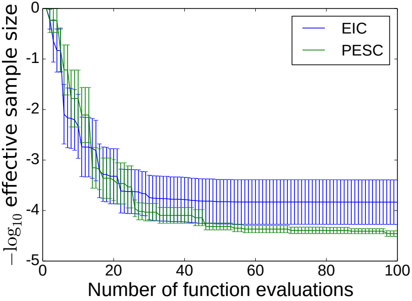
Gelbart et al. [2014], Schonlau et al. [1998]

Experimental Results with Real-world Data

Optimizing a neural network validation error on MNIST when constrained to make predictions in under 2ms.

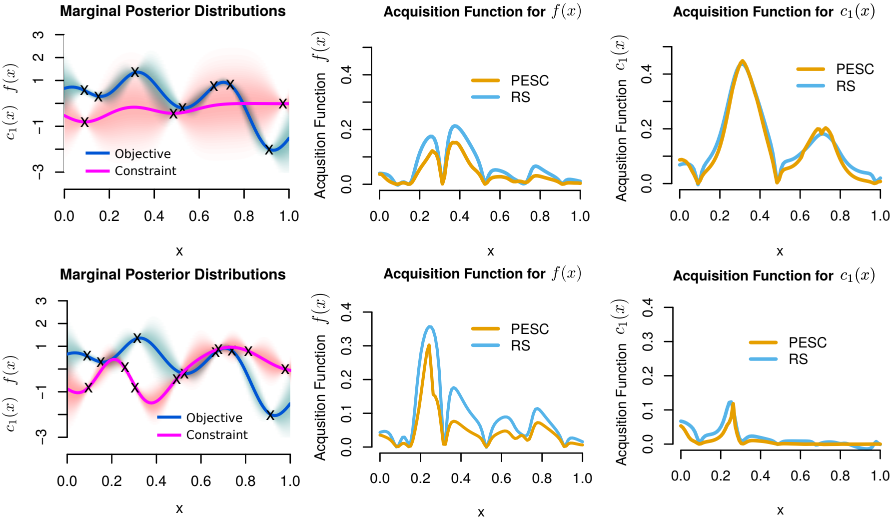


Optimizing the effective sample size of HMC on logistic regression when constrained to pass convergence diagnostics.



PESC in a Decoupled Evaluation Setting

The PESC acquisition function is **additive** across f and $c_1, \ldots, c_K$.



Summary

- EI can lead to **pathologies** when used with constraints.
  - Computing EI requires a current **best solution**, which may not exist.
  - EI fails when the objective and the constraints are **decoupled**.
- Information-based methods like PESC do not have these problems.
- PESC achieves **state-of-the-art** results in the coupled scenario.
- PESC can easily be applied to the **decoupled** case.
  - The acquisition function for PESC is **additive**!
  - Exhaustive evaluation in the decoupled case in a forthcoming paper!

PESC is implemented within **spearmint** and it is available at

`https://github.com/HIPS/Spearmint/tree/PESC`.

Thanks!

Thank you for your attention!

References I

S. Bochner. *Lectures on Fourier integrals*. Number 42. Princeton University Press, 1959.

M. A. Gelbart, J. Snoek, and R. P. Adams. Bayesian optimization with unknown constraints. In *UAI*, 2014.

P. Hennig and C. J. Schuler. Entropy search for information-efficient global optimization. *JMLR*, 13, 2012.

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 25*. Curran Associates, Inc., 2014.

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

References II

M. Schonlau, W. J. Welch, and D. R. Jones. Global versus local search in constrained optimization of computer models. *Lecture Notes-Monograph Series*, pages 11–25, 1998.

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.