Predictive Entropy Search for Bayesian Optimization with Unknown Constraints

José Miguel Hernández-Lobato

joint work with

July 8, 2015,

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

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

Queries are very expensive (time, economic cost, etc.).
Let \( x_\star \) be the **global optimum**. Entropy search (ES) maximizes the expected reduction in the entropy of the posterior on \( x_\star \).

\[
\alpha_t(x) = H[x_\star | \mathcal{D}_t] - \mathbb{E}_y \left[ H[x_\star | \mathcal{D}_t \cup \{x, y\}] \right] \quad \text{(ES)}
\]

How much we know about \( x_\star \) now.

How much we will know about \( x_\star \) after collecting \( y \) at \( x \).

Computing (ES) is very difficult in practice!
We can swap $y$ and $x_\star$ to obtain a new reformulation which we call **Predictive Entropy Search** (PES) ([Hernández-Lobato et al. 2014](#)):

\[
\alpha_t(x) = H[x_\star | D_t] - \mathbb{E}_y \left[ H[x_\star | D_t \cup \{x, y\}] \big| D_t, x \right] \equiv \text{MI}(y, x_\star) \quad \text{(ES)}
\]

\[
\alpha_t(x) = H[y | D_t, x] - \mathbb{E}_{x_\star} \left[ H[y | D_t, x, x_\star] \big| D_t, x \right] \equiv \text{MI}(x_\star, y) \quad \text{(PES)}
\]

**1** Approximated by sampling from $p(x_\star | D_t)$ (**Thompson sampling**).

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

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

\[
\alpha_t(x) = H[y | D_t, x] - \mathbb{E}_{x_\star} \left[ H[y | D_t, x, x_\star] \big| D_t, x \right], \quad \text{(PESC)}
\]

with $y = (y_f, y_1, \ldots, y_K)^T$. 

3/17
\[ \alpha_t(x) = H[y|D_t, x] - \mathbb{E}_{x_\star}[H[y|D_t, x, x_\star]|D_t, x], \]  
(PESC)
Step 1: Sampling the Optimum $x_*$

We sample $\tilde{f} \sim p(f|D_t)$ and $\tilde{c}_1 \sim p(c_1|D_t), \ldots, \tilde{c}_K \sim p(c_1|D_t)$ and return $\arg \max_x \tilde{f}(x)$ s.t. $\tilde{c}_1(x) \geq 0, \ldots, \tilde{c}_K(x) \geq 0$. 
Step 1: Sampling the Optimum $x_*$

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

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

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

We choose $m = 500$. [Bochner [1959]]
\[ \alpha_t(x) = H[y|\mathcal{D}_t, x] - \mathbb{E}_{x_\star} \left[ H[y|\mathcal{D}_t, x, x_\star] \bigg| \mathcal{D}_t, x \right], \] (PESC)
Step 2: Approximating $p(y|D_t, x, x_\star)$

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

Constraints satisfied

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

$x$ is not optimal

Constraints not satisfied

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

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

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

No other point is a better solution than $x_\star$

$x_\star$ must be feasible

We find a Gaussian approximation using expectation propagation.
Visualizing the Approximation to $p(y|\mathcal{D}_t, x, x_*)$
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):

\[ \alpha_t(x) = \mathbb{E} \left[ \max \left( 0, f(x) - f(x_+) \right) \middle| D_t \right] \left[ \prod_{k=1}^{K} p(c_k(x) \geq 0) \right] \]

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.
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.
Thank you for your attention!


