Predictive Entropy Search for Multi-objective Bayesian Optimization

Daniel Hernández-Lobato[1], José Miguel Hernández-Lobato[2],
Amar Shah[3] and Ryan P. Adams [2]

[1]Universidad Autónoma de Madrid. [2]Harvard University. [3]University of Cambridge.

1. Introduction to Multi-objective Bayesian Optimization

We are interested in solving the **problem**:

$$\min_{\mathbf{x}\in\mathcal{X}} \quad f_1(\mathbf{x}),\ldots,f_K(\mathbf{x})$$

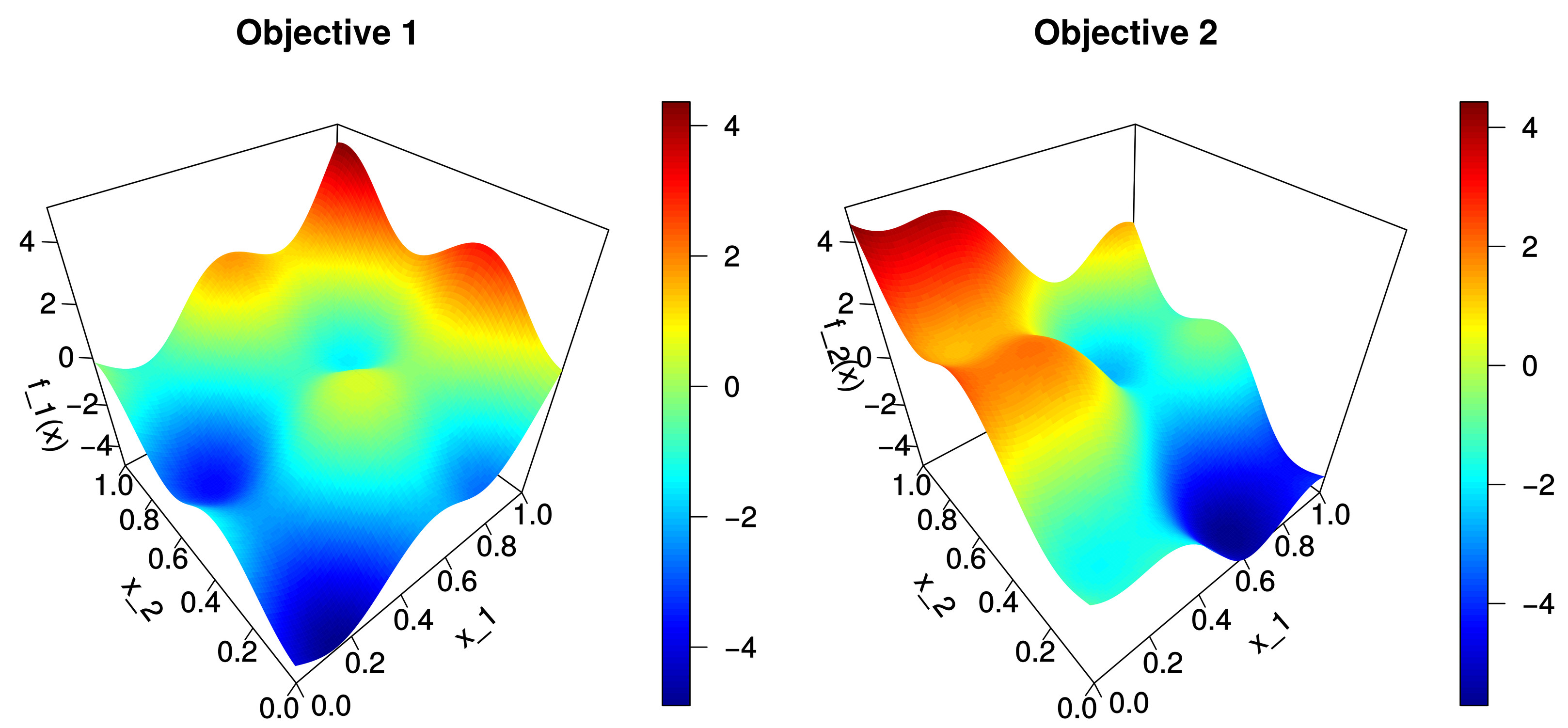
where $\mathcal{X}$ is the domain of each $f_k(\cdot)$. For this, we assume:

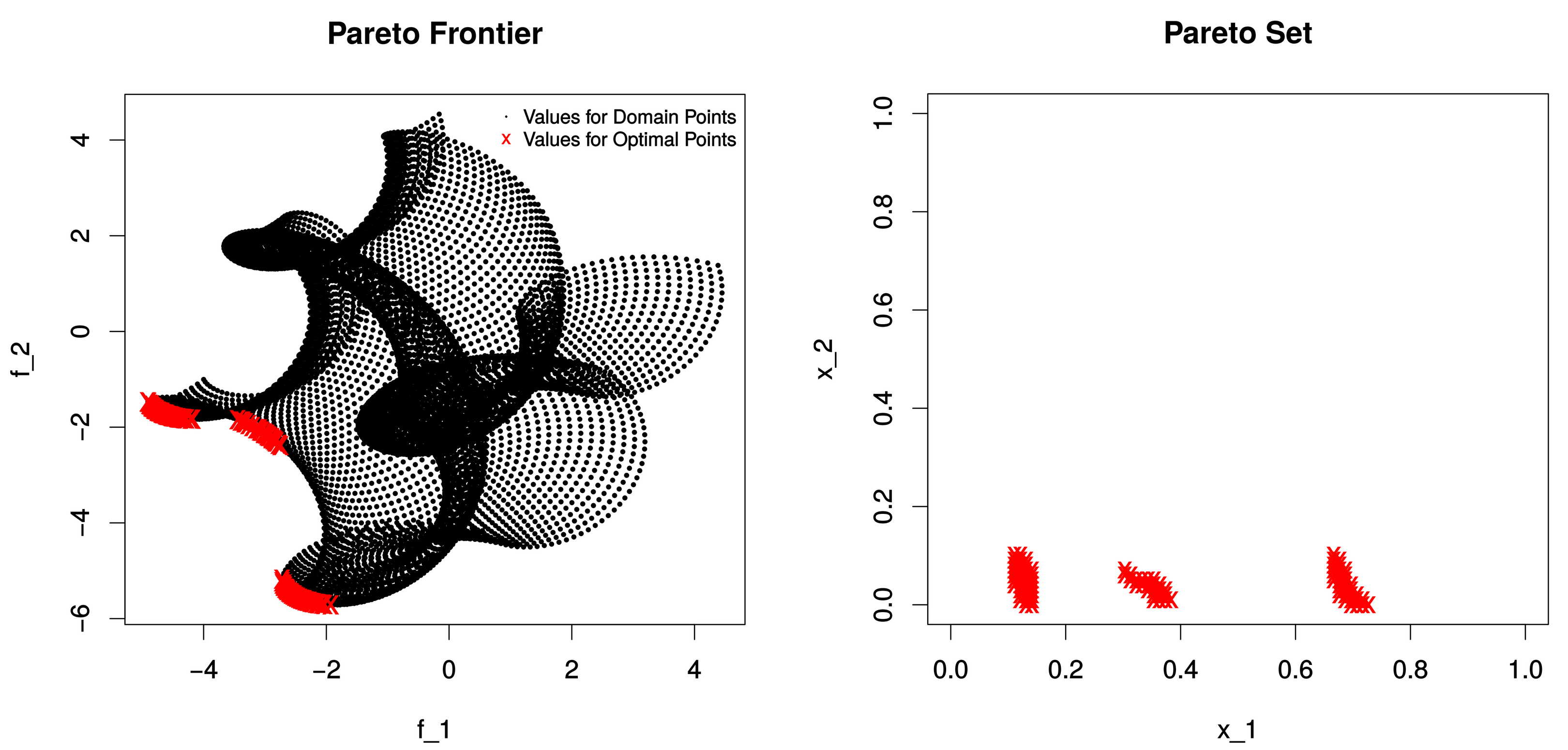
- Each $f_k(\cdot)$ is evaluated via expensive black-box queries.
- We select $\mathbf{x}$ and we observe output $\mathbf{y}=(f_1(\mathbf{x}),\ldots,f_K(\mathbf{x}))^{\text{T}}$
- The evaluations may be **contaminated** with Gaussian noise ϵ.

Most times the objectives are conflicting and there is no common minimizer of all $f_k(\cdot)$. The goal is to find the **Pareto set** $\mathcal{X}^\star$ of non-dominated points.

$$\forall \mathbf{x}^\star \in \mathcal{X}^\star, \forall \mathbf{x}\in\mathcal{X} \quad \exists k \quad \text{s.t.} \quad f_k(\mathbf{x}^\star)\leq f_k(\mathbf{x})$$

Example multi-objective optimization problem:





In practice, we would like to identify the Pareto set $\mathcal{X}^\star$ with the **smallest number of evaluations** of the objective functions $f_1,\ldots,f_K$.

2. Predictive Entropy Search for Multi-objective Optimization (MPES)

Given some observations in the form of a dataset $\mathcal{D}=\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^N$ we should to choose on which location $\mathbf{x}_{N+1}$ evaluate next the objectives. For that, we model each objective $f_k(\cdot)$ using a **Gaussian process**.

We choose the next point $\mathbf{x}_{N+1}$ as the one that maximizes the reduction on the entropy of the Pareto set $\mathcal{X}^\star$. The **acquisition function** employed is:

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

Computing the entropy of the Pareto set, $H(\mathcal{X}^\star)$, is very challenging!

The acquisition function is equal to $I(\mathcal{X}^\star,\mathbf{y})$ which is a symmetric quantity, *i.e.*, $I(\mathcal{X}^\star,\mathbf{y})=I(\mathbf{y},\mathcal{X}^\star)$. This allows to **rewrite** the acquisition function:

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

$H(\mathbf{y}|\mathcal{D},\mathbf{x})$ is the entropy of the predictive distribution of each GP at $\mathbf{x}$:

$$H(\mathbf{y}|\mathcal{D},\mathbf{x}) = \frac{K}{2}\log(2\pi e) + \sum_{k=1}^K \frac{1}{2}\log(v_k(\mathbf{x})^{\text{PD}}),$$

The expectation w.r.t. $\mathcal{X}^\star$ is approximated by Monte Carlo. We sample from the GP posteriors, and optimize the samples using an **evolutionary strategy**.

To compute $H(\mathbf{y}|\mathcal{X}^\star,\mathcal{D},\mathbf{x})$ we use **expectation propagation (EP)**:

$$p(\mathbf{y}|\mathcal{D},\mathcal{X}^\star,\mathbf{x}) \propto \int p(\mathbf{y}|\mathbf{f},\mathbf{x})p(\mathcal{X}^\star|\mathbf{f})p(\mathbf{f}|\mathcal{D})d\mathbf{f},$$

where the only non-Gaussian factor is

$$p(\mathcal{X}^\star|\mathbf{f}) \propto \prod_{\mathbf{x}^\star\in\mathcal{X}^\star}\prod_{\mathbf{x}'\in\mathcal{X}}\left[1-\prod_{k=1}^K\Theta\left(f_k(\mathbf{x}')-f_k(\mathbf{x}^\star)\right)\right] = \prod_{\substack{\mathbf{x}^\star\in\mathcal{X}^\star\\ \mathbf{x}'\in\mathcal{X}}}\Psi(\mathbf{x}',\mathbf{x}^\star),$$

where $\Theta(\cdot)$ is a step function and we set $\mathcal{X}=\{\mathbf{x}_n\}_{n=1}^N\cup\{\mathbf{x}\}\cup\mathcal{X}^\star$.

EP approximates each factor $\Psi(\mathbf{x}',\mathbf{x}^\star)$ with a product of K Gaussians:

$$\Psi(\mathbf{x}',\mathbf{x}^\star)\approx\prod_{k=1}^K\tilde{\mathcal{N}}\left((f_k(\mathbf{x}'),f_k(\mathbf{x}^\star))^{\text{T}}|\tilde{\mathbf{m}},\tilde{\Sigma}\right)$$

The approximate acquisition function is hence:

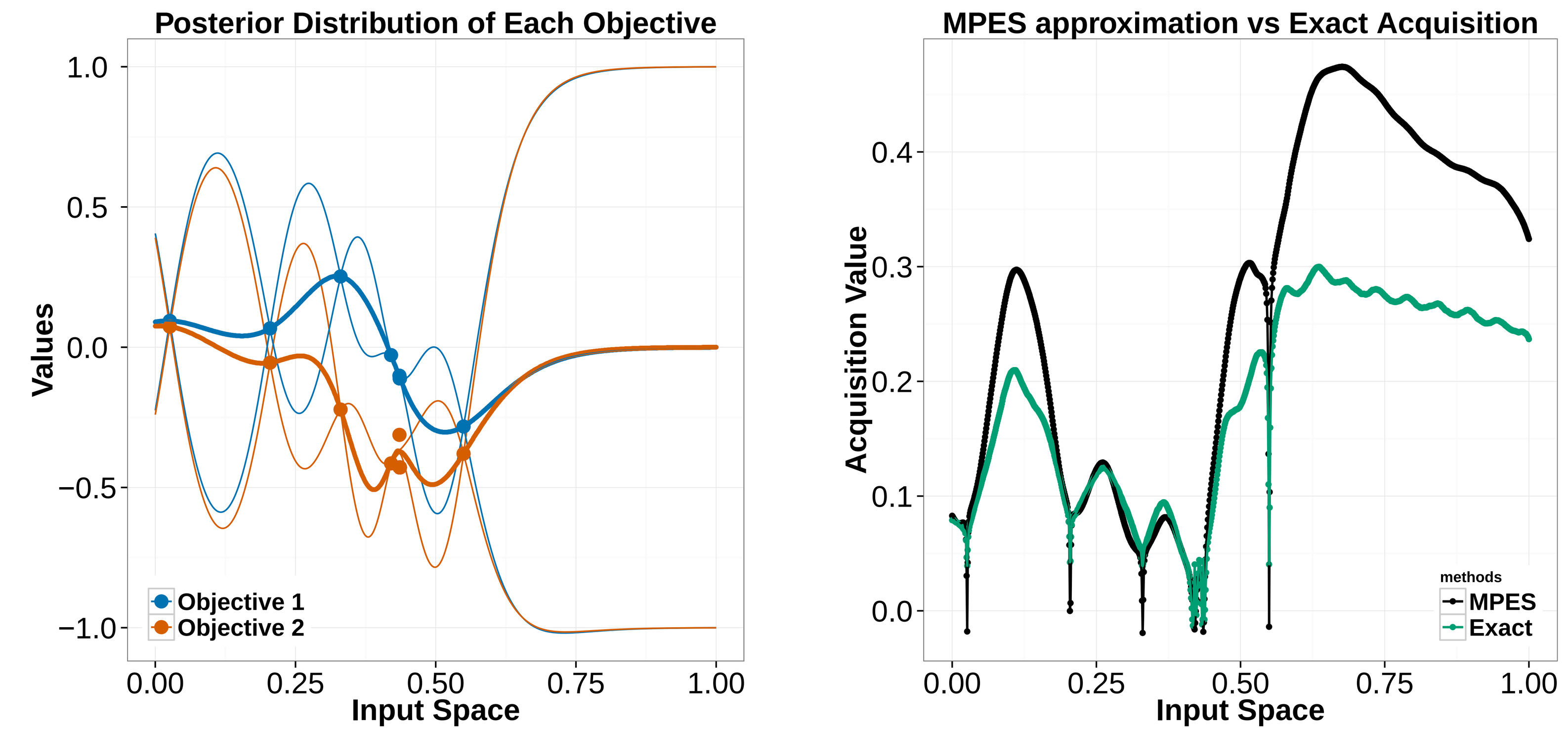
$$\alpha(\mathbf{x}) \approx \sum_{k=1}^K\frac{1}{2}\log v_k^{\text{PD}}(\mathbf{x}) - \frac{1}{S}\sum_{s=1}^S\sum_{k=1}^K\frac{1}{2}\log v_k^{\text{CPD}}(\mathbf{x}|\mathcal{X}^\star_{(s)}).$$

This acquisition function allows for a **decoupled evaluation scenario**!

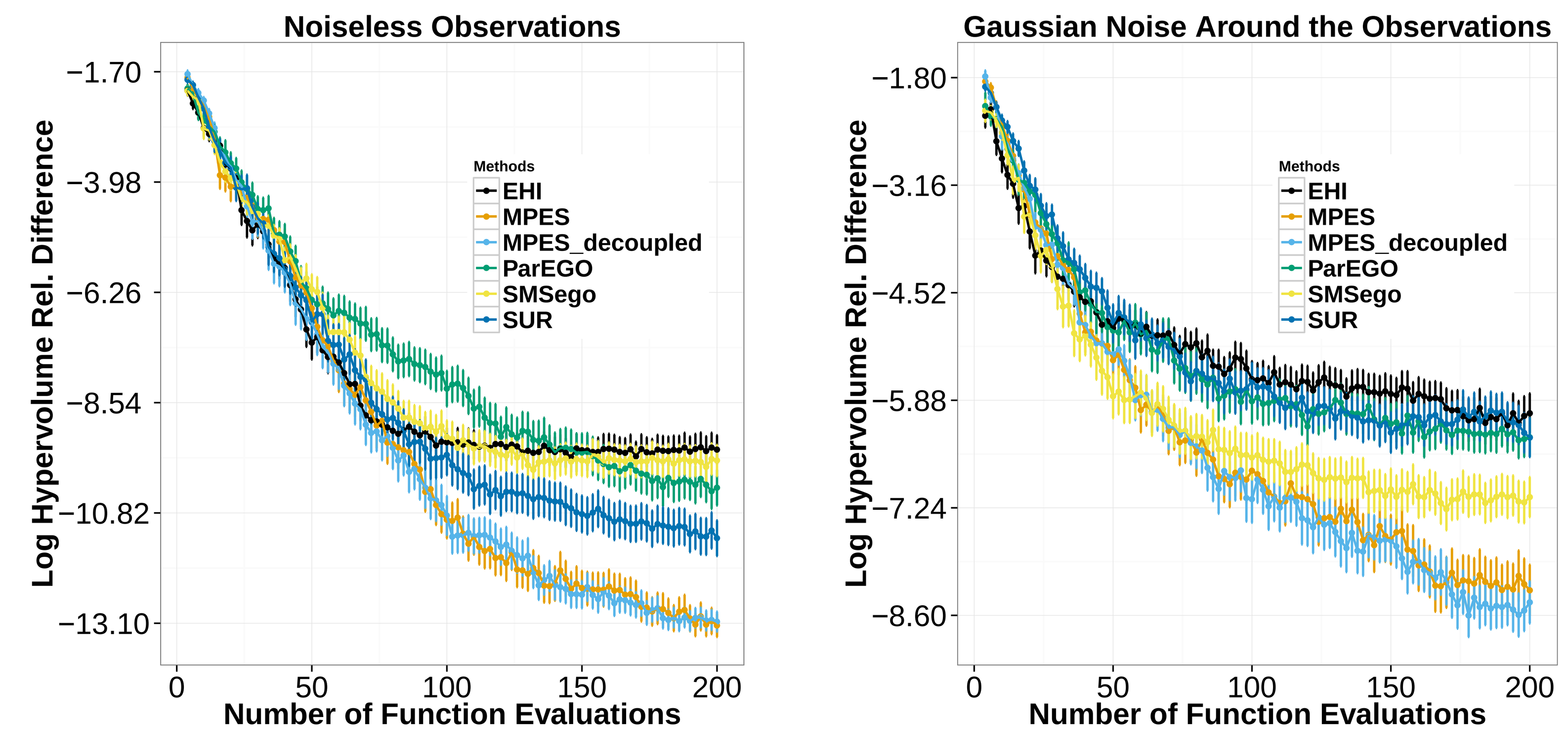
3. Experimental Evaluation

We implemented MPES in the software for BO **Spearmint**.

Accuracy of the MPES approximation to the acquisition function:



We carry out experiments involving problems with 2 objectives and 3 dimensions that are obtained by sampling the functions from the GP prior.



We compare the results of MPES with other strategies from the literature for multi-objective Bayesian Optimization: ParEGO, SMSego, SUR and EHI.

4. Conclusions

- We have described MPES, a successful strategy to carry out Bayesian Optimization of very expensive-to-evaluate black-box functions.
- MPES chooses the next location on which to evaluate the objectives as the one that is expected to minimize the entropy of the Pareto set.
- Our experiments show that MPES has better performance than most of the already known strategies for multi-objective Bayesian optimization.
- MPES allows for a decoupled evaluation scenario in which different objectives are evaluated at different input locations at each iteration.

http://dhnzl.org/ daniel.hernandez@uam.es