# Predictive Entropy Search for Bayesian Optimization with Unknown Constraints

**José Miguel Hernández-Lobato**
Harvard University
jmh@seas.harvard.edu

**Michael A. Gelbart**
Harvard University
mgelbart@seas.harvard.edu

**Matthew W. Hoffman**
University of Cambridge
mwh30@cam.ac.uk

**Ryan P. Adams**
Harvard University
rpa@seas.harvard.edu

**Zoubin Ghahramani**
University of Cambridge
zoubin@eng.cam.ac.uk

## Abstract

Unknown constraints arise in many types of black-box optimization problems. They may arise due to unpredictable system failures, search space bounds that are unknown *a priori*, or simply as a means of trading off different objectives. Several methods have been proposed recently for performing Bayesian optimization with constraints. These methods are based on the expected improvement (EI) heuristic. However, EI can lead to several pathologies when used with constraints. For example, computing EI requires a current best solution, which may not exist if the data collected so far does not satisfy the constraints. Furthermore, in the case of decoupled constraints, i.e., when one can independently evaluate the objective or the constraints, using EI leads to a pathology that prevents exploration. By contrast, information-based approaches do not have these problems. In this paper, we present a new information-based method called Predictive Entropy Search with Constraints (PESC). We show that PESC compares favorably to EI-based approaches on synthetic data. This is a promising direction towards a unified solution for constrained Bayesian optimization.

## 1 Introduction

We are interested in finding the global maximum $\mathbf{x}_\star$ of an objective function $f(\mathbf{x})$ over some bounded domain, typically $\mathcal{X} \subset \mathbb{R}^d$, subject to the non-negativity of a series of constraint functions $c_1, \ldots, c_K$. This can be formalized as

$$\max \quad f(\mathbf{x}) \quad \text{s.t.} \quad c_1(\mathbf{x}) \geq 0, \ldots, c_K(\mathbf{x}) \geq 0\,. \tag{1}$$

However, $f$ and $c_1, \ldots, c_K$ are unknown and can only be evaluated via expensive queries to black-boxes that provide noisy outputs of the form $y_i^f \sim \mathcal{N}(f(\mathbf{x}_i), \sigma_f^2)$ for $f$ and $y_i^k \sim \mathcal{N}(c_k(\mathbf{x}_i), \sigma_k^2)$ for $c_k$ and $k = 1, \ldots, K$. We seek to find a solution to (1) with as few queries as possible.

In this we work we extend predictive entropy search (PES) [6] to solve (1), an approach that we call PES with constraints (PESC). PESC is a sequential optimization method which after $n$ evaluations of $f$ and $c_1, \ldots, c_K$, proposes to evaluate these functions at the location $\mathbf{x}_{n+1}$ which approximately maximizes the expected information gain about the constrained maximizer $\mathbf{x}_\star$. We compute this information gain by conditioning on the objective data $\mathcal{D}_n^f = \{(\mathbf{x}_1, y_i^f), \ldots, (\mathbf{x}_n, y_n^f)\}$ and constraint data $\mathcal{D}_n^1, \ldots, \mathcal{D}_n^K$, where $\mathcal{D}_n^k = \{(\mathbf{x}_1, y_i^k), \ldots, (\mathbf{x}_n, y_n^k)\}$ for $k = 1, \ldots, K$. PESC models this data assuming that $f$ and $c_1, \ldots, c_K$ follow independent Gaussian process (GP) priors [12].

While previous approaches to the problem of Bayesian optimization with unknown constraints have been proposed, most are variants of expected improvement (EI) [10, 8]. Initially proposed by [13] one method of extending EI to the constrained setting considers the expected *feasible* improvement, where the constraints are given as above; such approaches have recently been independently developed in [3, 2, 14]. Alternatively [4] consider the integrated change in improvement under additional points with respect to expected feasibility and [11] consider the probability of improvement under a similar measure. In the next section we describe the pathologies that arise when applying EI to constrained problems—providing much of the motivation for this work.

## 2 Expected improvement with constraints (EIC)

Because improvement can only occur if all the constraints are satisfied, approaches which compute the feasible EI are obtained by weighting the original expected improvement with the posterior probability of the constraints being satisfied. The associated EIC acquisition function is given by

$$\alpha(\mathbf{x}) = \mathrm{EI}(\mathbf{x}|\eta, \mathcal{D}_f) \textstyle\prod_{k=1}^K p(c_k(\mathbf{x}) \geq 0|\mathcal{D}_k)\,, \tag{2}$$

where $\mathrm{EI}(\mathbf{x}|\eta, \mathcal{D}_f) = \int \max(0, f(\mathbf{x}) - \eta) p(f(\mathbf{x})|\mathcal{D}_f) df(\mathbf{x})$ is the expected improvement over $\eta$ at $\mathbf{x}$ given the collected data $\mathcal{D}_f$. The constant $\eta$ represents the expected value of $f$ at the best solution found so far. In the unconstrained case, this is usually the maximum of the posterior mean for $f$ [1]. In the constrained case, $\eta$ is the largest value of the posterior mean for $f$ such that all the constraints are satisfied at the corresponding location. However, since information about the constraints is only available through noisy measurements, we can never be certain that the constraints will be satisfied at any particular location. To avoid this problem, Gelbart et al. consider in [3] a location $\mathbf{x}$ to be feasible only if all the constraints are satisfied at $\mathbf{x}$ with high posterior probability, that is, if

$$\forall k \in \{1, \ldots, K\},\ p(c_k(\mathbf{x}) \geq 0|\mathcal{D}_k) \geq 1 - \delta_k\,, \tag{3}$$

where the $\delta_k$ are small positive numbers. This is called a *probabilistic constraint*. Under this new formulation, $\eta$ is the the largest value of the posterior mean for $f$ such that (3) is satisfied at the corresponding location. However, when no point in the search space is feasible, $\eta$ does not exist and EI cannot be computed. In this case, Gelbart et al. ignore the factor $\mathrm{EI}(\mathbf{x}|\eta, \mathcal{D}_f)$ in (2) and only consider the probability of the constraints being satisfied. The resulting acquisition function focuses only on searching for a feasible location and ignores learning about $f$.

Note that with probabilistic constraints, EIC is not the true expected improvement of the best feasible solution. This is clear because the EIC does not depend on $\delta$. Computing the true expected improvement does not work because of the following pathology: observing a single noisy constraint satisfaction is typically insufficient to push the posterior probability of satisfaction above $1 - \delta$; thus a myopic strategy like EI will see zero potential improvement from a single evaluation.

Furthermore, Gelbart et al. identify a pathology with EI when one can individually evaluate the objective or the constraints (called *decoupled* constraints): the best solution $\mathbf{x}_*$ must have a high objective value and high (non-negative) constraint values. This is a conjunction ("AND") of several conditions. If one is only evaluating the objective, or a single constraint, this conjunction cannot be satisfied by that single observation. Thus, the observed $\mathbf{x}$ cannot become the new best as a result of a decoupled observation and so the expected improvement will be zero. Therefore EI is not sensible in this decoupled setting. Gelbart et al. circumvent this pathology by treating decoupling as a special case and using a two-stage acquisition function: first, $\mathbf{x}$ is chosen with EIC, and then the task (whether to evaluate the objective or one of the constraints) is chosen with Entropy Search (ES) [5] given $\mathbf{x}$. This approach does not take full advantage of the available information in the way a joint decision of $\mathbf{x}$ and the task would.

Our new method PESC does not have the problems mentioned above. First, the PESC acquisition function does not depend on the current best feasible solution; and, second, PESC naturally separates the contribution of each function evaluation (objective or constraint) in its acquisition function. Our experiments with synthetic data show the improved performance of PESC compared to the method proposed by Gelbart et al. in [3] in the basic setting of joint evaluations, where EI can be applied.

## 3 Predictive entropy search with constraints

We want to maximize our information about the location $\mathbf{x}_\star$, the constrained global maximum, whose posterior distribution is $p(\mathbf{x}_\star|\mathcal{D}_n^f, \mathcal{D}_n^1, \ldots, \mathcal{D}_n^K)$. Therefore, we collect data at the $\mathbf{x}_{n+1}$ that maximizes the expected reduction in the differential entropy $\mathrm{H}(\cdot)$ of the posterior on $\mathbf{x}_\star$. We follow [7, 6] and rewrite this acquisition function as the mutual information between $y^f, y^1, \ldots, y^K$ and $\mathbf{x}_\star$ given the collected data $\mathcal{D}_n^f, \mathcal{D}_n^1, \ldots, \mathcal{D}_n^K$, that is,

$$
\begin{aligned}
\alpha(\mathbf{x}) = \mathrm{H}\left[p(y^f, y^1, \ldots, y^K | \mathcal{D}_n^f, \mathcal{D}_n^1, \ldots, \mathcal{D}_n^K, \mathbf{x})\right] - \\
\mathbf{E}\left\{\mathrm{H}\left[p(y^f, y^1, \ldots, y^K | \mathcal{D}_n^f, \mathcal{D}_n^1, \ldots, \mathcal{D}_n^K, \mathbf{x}, \mathbf{x}_\star)\right]\right\}\,, \qquad (4)
\end{aligned}
$$

where the expectation is now with respect to the posterior on $\mathbf{x}_\star$, that is, $p(\mathbf{x}_\star | \mathcal{D}_n^f, \mathcal{D}_n^1, \ldots, \mathcal{D}_n^K)$, and $p(y^f, y^1, \ldots, y^K | \mathcal{D}_n^f, \mathcal{D}_n^1, \ldots, \mathcal{D}_n^K, \mathbf{x}, \mathbf{x}_\star)$ is the posterior predictive distribution for $y^f, y^1, \ldots, y^K$ given $\mathcal{D}_n^f, \mathcal{D}_n^1, \ldots, \mathcal{D}_n^K$ and conditioned to the location $\mathbf{x}_\star$ of the global solution to the constrained optimization problem. We call this distribution the *conditioned predictive distribution* (CPD).

The first term on the right-hand side of (4) is straightforward to compute: it is the entropy of the predictive distribution of independent GPs. This is one half of the sum of the log predictive variances plus $\frac{1}{2}(K+1)\log 2\pi e$. However, the second term has to be approximated. For this, we first approximate the expectation by averaging over samples of $\mathbf{x}_\star$ approximately drawn from $p(\mathbf{x}_\star | \mathcal{D}_n^f, \mathcal{D}_n^1, \ldots, \mathcal{D}_n^K)$ using the method described in [6]. To sample $\mathbf{x}_\star$, we first approximately draw $f$ and $c_1, \ldots, c_K$ from their GP posteriors using a finite parameterization of these functions. Then we solve a constrained optimization problem using the sampled functions. The solution to this problem is the sample of $\mathbf{x}_\star$. For each value of $\mathbf{x}_\star$ generated by this procedure, we approximate the CPD $p(y^f, y^1, \ldots, y^K | \mathcal{D}_n^f, \mathcal{D}_n^1, \ldots, \mathcal{D}_n^K, \mathbf{x}, \mathbf{x}_\star)$ as described in the next section.

### 3.1 Approximating the conditioned predictive distribution

We can approximate the CPD by first approximating the noise free version of the CPD (NFCPD), that is, the posterior predictive distribution for $\bar{f} = f(\mathbf{x})$, $\bar{c}_1 = c_1(\mathbf{x}), \ldots, \bar{c}_K = c_K(\mathbf{x})$ given $\mathcal{D}_n^f, \mathcal{D}_n^1, \ldots, \mathcal{D}_n^K$ and $\mathbf{x}_\star$, and then convolving that approximation with additive Gaussian noise on $\bar{f}, \bar{c}_1, \ldots, \bar{c}_K$ with variance $\sigma_f^2, \sigma_1^2, \ldots, \sigma_K^2$. The NFCPD can be informally written as

$$
\begin{aligned}
p(\bar{f}, \bar{c}_1, \ldots, \bar{c}_K | \mathcal{D}_n^f, \mathcal{D}_n^1, \ldots, \mathcal{D}_n^K, \mathbf{x}, \mathbf{x}_\star) = Z^{-1} \textstyle\int \delta[\bar{f} - f(\mathbf{x})]\delta[\bar{c}_1 - c_1(\mathbf{x})] \cdots \delta[\bar{c}_K - c_K(\mathbf{x})] \\
\left[\textstyle\prod_{\mathbf{x}' \neq \mathbf{x}_\star} \left(\left\{\prod_{k=1}^K \Theta[c_k(\mathbf{x}')]\right\} \Theta[f(\mathbf{x}_\star) - f(\mathbf{x}')] + \left\{1 - \prod_{k=1}^K \Theta[c_k(\mathbf{x}')]\right\}\right)\right] \\
\Theta[c_1(\mathbf{x}_\star)] \cdots \Theta[c_K(\mathbf{x}_\star)] p(f | \mathcal{D}_n^f) p(c_1 | \mathcal{D}_n^1) \cdots p(c_K | \mathcal{D}_n^K) df\, dc_1 \, \ldots \, dc_K\,, \qquad (5)
\end{aligned}
$$

where the integral above marginalizes out the infinite dimensional vectors $f$, $c_1, \ldots, c_K$ encoding the objective and the constraint functions. These vectors are sampled from the posteriors $p(f | \mathcal{D}_n^f)$, $p(c_1 | \mathcal{D}_n^1), \ldots, p(c_K | \mathcal{D}_n^K)$, which are infinite dimensional multivariate Gaussian distributions. The Dirac deltas in the first line of (5) project these infinite dimensional vectors to their corresponding values at $\mathbf{x}$. The Heaviside step functions (denoted by $\Theta$) in the bottom line of (5) guarantee that $\mathbf{x}_\star$ is a feasible solution. The infinite product in the middle line of (5) guarantees that $\mathbf{x}_\star$ is the global solution. This factor takes value one when $f(\mathbf{x}')$ is smaller than $f(\mathbf{x}_\star)$ for all $\mathbf{x}' \neq \mathbf{x}_\star$ such that all the constraints are satisfied at $\mathbf{x}'$ and zero otherwise. Finally, $Z$ is a normalization constant.

We find a Gaussian approximation to (5) in several steps. We first approximate the infinite product in (5) with a finite dimensional one only over the locations at which we have collected data, that is, $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Let $\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K$ be the $(n+1)$-dimensional vectors containing the concatenation of the evaluations of $f, c_1, \ldots, c_K$ at $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and $\mathbf{x}_\star$. We can then obtain a finite dimensional approximation to the factors in the second and third lines of (5) as

$$
\begin{aligned}
q_1(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K) = \left[\textstyle\prod_{i=1}^n \left(\left\{\prod_{k=1}^K \Theta[c_{k,i}]\right\} \Theta[f_{n+1} - f_i] + \left\{1 - \prod_{k=1}^K \Theta[c_{k,i}]\right\}\right)\right] \\
\Theta[c_{1,n+1}] \cdots \Theta[c_{K,n+1}] p(\mathbf{f} | \mathcal{D}_n^f) p(\mathbf{c}_1 | \mathcal{D}_n^1) \cdots p(\mathbf{c}_K | \mathcal{D}_n^K)\,, \qquad (6)
\end{aligned}
$$

where $p(\mathbf{f} | \mathcal{D}_n^f) p(\mathbf{c}_1 | \mathcal{D}_n^1) \cdots p(\mathbf{c}_K | \mathcal{D}_n^K)$ are the GP predictive distributions on $\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K$ given the collected data. Because (6) is not tractable, we approximate the normalized version of $q_1$ with a product of Gaussians using expectation propagation (EP) [9]. In particular, we obtain

$$
Z_{q_1}^{-1} q_1(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K) \approx q_2(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K) = \mathcal{N}(\mathbf{f} | \mathbf{m}_f, \mathbf{V}_f) \textstyle\prod_{k=1}^K \mathcal{N}(\mathbf{c}_k | \mathbf{m}_k, \mathbf{V}_k)\,, \qquad (7)
$$

where $Z_{q_1}$ is the normalization constant of $q_1$ and $\mathbf{m}_f, \mathbf{m}_1, \ldots, \mathbf{m}_K$ and $\mathbf{V}_f, \mathbf{V}_1, \ldots, \mathbf{V}_K$ are mean vectors and covariance matrices determined by EP. Details will be given in a supplementary material that will be attached to the final version of this work. Given $q_2$, we can approximate (5) by

$$
\begin{aligned}
p(\bar{f}, \bar{c}_1, \ldots, \bar{c}_K | \mathcal{D}_n^f, \mathcal{D}_n^1, \ldots, \mathcal{D}_n^K, \mathbf{x}, \mathbf{x}_\star) \approx Z_2^{-1} \textstyle\int p(\bar{f} | \mathbf{f}) p(\bar{c}_1 | \mathbf{c}_1) \cdots p(\bar{c}_K | \mathbf{c}_K) \\
\left[\left\{\textstyle\prod_{k=1}^K \Theta[\bar{c}_k]\right\} \Theta[f_{n+1} - \bar{f}] + \left\{1 - \textstyle\prod_{k=1}^K \Theta[\bar{c}_k]\right\}\right]
\end{aligned}
$$

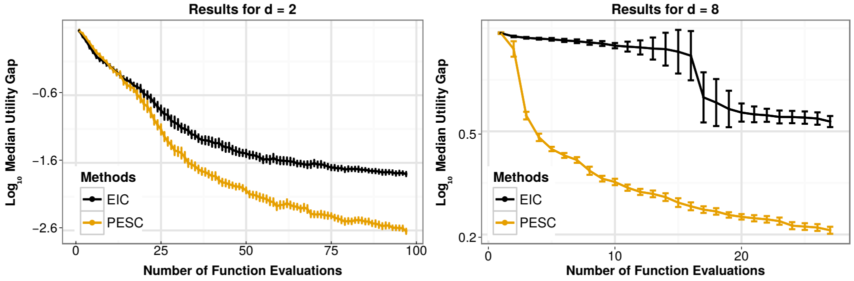

Figure 1: Median utility gap for PESC and EIC with $d = 2$ (left) and $d = 8$ (right).

$$q_2(\mathbf{f}, \mathbf{c}_1, \dots, \mathbf{c}_K)\, d\mathbf{f}\, d\mathbf{c}_1 \cdots d\mathbf{c}_K\,, \qquad (8)$$

where $Z_2$ is a normalization constant and $p(\bar{f}|\mathbf{f}), p(\bar{c}_1|\mathbf{c}_1), \dots, p(\bar{c}_K|\mathbf{c}_K)$ are Gaussian conditional distributions given by the GP priors on $f, c_1, \dots, c_K$. These conditional distributions approximate the deltas in the first line of (5). Note that, in the second line of (8), we have introduced one of the factors forming the infinite product in (5). This is the factor for which $\mathbf{x}'$ corresponds to the location $\mathbf{x}$ at which we are making predictions and it guarantees that $\bar{f} = f(\mathbf{x})$ is smaller than $f_{n+1} = f(\mathbf{x}_\star)$ when all the constraints are satisfied at $\mathbf{x}$, that is, $\bar{c}_k = c_k(\mathbf{x}) \geq 0$ for all $k$. Note that the product of $p(\bar{f}|\mathbf{f}), p(\bar{c}_1|\mathbf{c}_1), \dots, p(\bar{c}_K|\mathbf{c}_K)$ and $q_2$ in (8) results in $K + 1$ joint multivariate Gaussian distributions, the first one for $\bar{f}$ and $\mathbf{f}$ and the $K$ remaining ones for each of the $\bar{c}_k$ and $\mathbf{c}_k$. After marginalizing out $f_1, \dots, f_n$ and $\mathbf{c}_1, \dots, \mathbf{c}_K$ we can rewrite (8) as

$$p(\bar{f}, \bar{c}_1, \dots, \bar{c}_K | \mathcal{D}_n^f, \mathcal{D}_n^1, \dots, \mathcal{D}_n^K, \mathbf{x}, \mathbf{x}_\star) \approx Z_2^{-1} \int \left[ \left\{ \prod_{k=1}^K \Theta[\bar{c}_k] \right\} \Theta[f_{n+1} - \bar{f}] + \left\{ 1 - \prod_{k=1}^K \Theta[\bar{c}_k] \right\} \right]$$
$$\mathcal{N}([\bar{f}, f_{n+1}] | \mathbf{m}'_f, \mathbf{V}'_f) \prod_{k=1}^K \mathcal{N}(\bar{c}_k | m'_k, v'_k)\, df_{n+1}\,. \qquad (9)$$

Details for $\mathbf{m}'_f$, $\mathbf{V}'_f$, $m'_1, \dots, m_K$, $v'_1, \dots, v'_K$ will be included in the supplementary material.

### 3.2 Approximating the entropy of the CPD

The normalization constant $Z_2$ in (9) can be computed analytically. This allows us to obtain the marginal variances of the right-hand-side of (9) by computing the gradient of $\log Z_2$ with respect to $\mathbf{m}'_f$, $\mathbf{V}'_f$, $m'_1, \dots, m_K$, $v'_1, \dots, v'_K$ using formula 5.13 in [9]. These expressions will be included in the supplementary material. If we assume independence in the NFCPD (5), we can then approximate the entropy in the CPD by performing the following operations. First, we add the noise variances $\sigma_f^2, \sigma_1^2, \dots, \sigma_K^2$ to the marginal variances of the right-hand-side of (9) and second, assuming Gaussianity, we sum one half of the logarithm of the resulting variances and finally add $\frac{(K+1)}{2} \log(2\pi e)$.

## 4 Experiments

We evaluate the performance of predictive entropy search with constraints (PESC) in experiments with synthetic functions following the same experimental set up as in [5, 6]. The search space is the unit hypercube of dimension $d$, and the ground truth objective $f$ is a sample from a zero-mean GP with a squared exponential covariance function of unit amplitude and length scale $\ell = 0.1$ in each dimension. We represent the function $f$ by first sampling from the GP prior on a grid of 1000 points and then defining $f$ as the resulting GP posterior mean. We use a single constraint function $c_1$ whose ground truth is sampled in the same way as $f$. The evaluations for $f$ and $c_1$ are contaminated with i.i.d. Gaussian noise with variance $\sigma_f^2 = \sigma_1^2 = 0.01$. We compare PESC and EIC with $\delta = 0.05$ using GP hyperparameters that are matched to the ground truth. In both methods we make recommendations by finding the location with highest posterior mean for $f$ such that $c_1 \geq 0$ with probability at least $1 - \delta$. For each recommendation at $\mathbf{x}$, we compute the utility gap $|f(\mathbf{x}) - f(\mathbf{x}_\star)|$, where $\mathbf{x}_\star$ is the true solution of the optimization problem and we treat a recommendation that violates the constraint as the worst possible objective function value. PESC and EIC are initialized with the same 3 random points drawn using latin hyper-cube sampling.

Figure 1 shows the median of the utility gap for each method across 500 different samples of $f$ and $c_1$ for dimensionalities $d = 2$ (left) and $d = 8$ (right). We report the median because the empirical distribution of the utility gap is heavy-tailed and therefore the median is more representative of the bulk data location than the mean. Overall, PESC performs significantly better than EIC.

## 5 Future work and conclusion

Two main lines of future work are needed. First, a more general form of PESC will be derived which encompasses the constraint scenarios described earlier (e.g. decoupled constraints). Second, real-world experiments will be performed to show the utility of PESC beyond synthetic problems. This paper shows that PESC is a promising algorithm and has the potential to be a unified framework for constrained Bayesian optimization that is both theoretically appealing and effective in practice.

**Acknowledgements** J.M.H.L acknowledges support from the Rafael del Pino Foundation.

## References

[1] E. Brochu, V. M. Cora, and N. de Freitas. A tutorial on Bayesian optimization of expensive cost functions, 2010. arXiv:1012.2599 [cs.LG].

[2] J. R. Gardner, M. J. Kusner, Z. E. Xu, K. Q. Weinberger, and J. P. Cunningham. Bayesian optimization with inequality constraints. In *ICML*, 2014.

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

[4] R. B. Gramacy and H. K. H. Lee. Optimization under unknown constraints, 2010. arXiv:1004.4027 [stat.ME].

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

[6] 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.

[7] N. Houlsby, J. M. Hernández-Lobato, F. Huszar, and Z. Ghahramani. Collaborative gaussian processes for preference learning. In *Advances in Neural Information Processing Systems 25*, pages 2096–2104. Curran Associates, Inc., 2012.

[8] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

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

[10] J. Mockus, V. Tiesis, and A. Zilinskas. The application of Bayesian methods for seeking the extremum. *Towards Global Optimization*, 2, 1978.

[11] V. Picheny. A stepwise uncertainty reduction approach to constrained global optimization. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 787–795, 2014.

[12] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[13] 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.

[14] J. Snoek. *Bayesian Optimization and Semiparametric Models with Applications to Assistive Technology*. PhD thesis, University of Toronto, Toronto, Canada, 2013.