Extending Expected Improvement for High-dimensional Stochastic Optimization of Expensive Black-Box Functions

Piyush Pandita
School of Mechanical Engineering
Purdue University
West Lafayette, Indiana 47907
Email: ppandit@purdue.edu

Ilias Bilionis[*]
School of Mechanical Engineering
Purdue University
West Lafayette, Indiana 47907
Email: ibilion@purdue.edu

Jitesh Panchal
School of Mechanical Engineering
Purdue University
West Lafayette, Indiana 47907
Email: panchal@purdue.edu

Design optimization under uncertainty is notoriously difficult when the objective function is expensive to evaluate. State-of-the-art techniques, e.g, stochastic optimization or sampling average approximation, fail to learn exploitable patterns from collected data and require an excessive number of objective function evaluations. There is a need for techniques that alleviate the high cost of information acquisition and select sequential simulations optimally. In the field of deterministic single-objective unconstrained global optimization, the Bayesian global optimization (BGO) approach has been relatively successful in addressing the information acquisition problem. BGO builds a probabilistic surrogate of the expensive objective function and uses it to define an information acquisition function (IAF) whose role is to quantify the merit of making new objective evaluations. Specifically, BGO iterates between making the observations with the largest expected IAF and rebuilding the probabilistic surrogate, until a convergence criterion is met. In this work, we extend the expected improvement (EI) IAF to the case of design optimization under uncertainty wherein the EI policy is reformulated to filter out parametric and measurement uncertainties. To increase the robustness of our approach in the low sample regime, we employ a fully Bayesian interpretation of Gaussian processes by constructing a particle approximation of the posterior of its hyperparameters using adaptive Markov chain Monte Carlo. We verify and validate our approach by solving two synthetic optimization problems under uncertainty and demonstrate it by solving the oil-well-placement problem with uncertainties in the permeability field and the oil price time series.

1 Introduction

The majority of stochastic optimization techniques are based on Monte Carlo sampling, e.g., stochastic gradient descent [1], sample average approximation [2], and random search [3]. Unfortunately, the advantages offered by these techniques can be best leveraged [4] only when a large number of objective evaluations is possible. Therefore, their applicability to engineering design/optimization problems involving expensive physics-based models or even experimentally measured objectives is severely limited.

Bayesian global optimization (BGO) has been successfully applied to the field of single-objective unconstrained optimization. [5, 6, 7, 8, 9, 10, 11]. BGO builds a probabilistic surrogate of the expensive objective function and uses it to define an information acquisition function (IAF). The role of the IAF is to quantify the merit of making new objective evaluations. Given an IAF, BGO iterates between making the observation with the largest expected IAF and rebuilding the probabilistic surrogate until a convergence criterion is met. The most commonly used IAFs are the expected improvement (EI) [12], resulting in a version of BGO known as efficient global optimization (EGO), and the probability of improvement (PoI) [8]. The operations research literature has developed the concept of knowledge gradient (KG) [13,14,15,16], which is essentially a generalization of the EI, and the machine learning community has been experimenting with the expected information gain (EIG) [17, 18, 19].

BGO is not able to deal with stochastic optimization in a satisfactorily robust way. In this work, we propose a natural modification of the EI IAF, which is able to filter out the effect of noise in the objective and, thus, enable stochastic optimization strategies under an information acquisition budget.

---
[*]Corresponding author

We will be referring to our version of EI as the Extended EI (EEI). Our approach does not suffer from the curse of dimensionality in the stochastic space, since it represents both parametric and measurement noise in an equal footing and does not explicitly try to learn the map between the uncertain parameters and the objective. However, we observed that naive applications of our strategy fail to converge in the regime of low samples and high noise. To deal with this problem, we had to retain the full epistemic uncertainty of the underlying objective surrogate. This epistemic uncertainty corresponds to the fact that the parameters of the surrogate cannot be determined exactly due to limited data and/or increased noise. Ignoring this uncertainty by picking specific parameter values, e.g., by maximizing the marginal likelihood, typically yields an overconfident, but wrong, surrogate. This is a known problem in sequential information acquisition literature, first mentioned by MacKay in [20]. To avoid this issue, we had to explicitly characterize the posterior distribution of the surrogate parameters by adaptive Markov chain Monte Carlo sampling. Remarkably, by keeping the full epistemic uncertainty induced by the limited objective evaluations, we are able to characterize our state of knowledge about the location of the optimum and the optimal value.

The outline of the paper is as follows. We start Sec. 2 by providing the mathematical definition of the stochastic optimization problem that is being studied. In Sec. 2.1, we introduce Gaussian process regression (GPR) which is used to construct a probabilistic surrogate of the map between the design variables and the objective. In Sec. 2.2, we show how the epistemic uncertainty on the location of the optimum and the optimal value can be quantified. In Sec. 2.3, we derive our extension to EI suitable for stochastic optimization. Our numerical results are presented in Sec. 3. In particular, in Sec. 3.1 and 3.2, we validate our approach using two synthetic stochastic optimization problems with known optimal solutions and we experiment with various levels of Gaussian noise, as well as heteroscedastic, i.e., input dependent, noise. In Sec. 3.3, we apply our methodology to solve the oil-well placement problem with uncertainties in soil permeability and the oil price timeseries. Our conclusions are presented in Sec. 4.

2 Methodology

We are interested in the following design optimization problem under uncertainty:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \mathbb{E}_{\boldsymbol{\xi}}\left[V(\mathbf{x};\boldsymbol{\xi})\right], \tag{1}$$

where $V(\mathbf{x};\boldsymbol{\xi})$ is the *objective function* depending on a set of *design parameters* $\mathbf{x}$ and *stochastic parameters* $\boldsymbol{\xi}$. The operator $\mathbb{E}_{\boldsymbol{\xi}}[\cdot]$ denotes the expectation over $\boldsymbol{\xi}$, i.e.,

$$\mathbb{E}_{\boldsymbol{\xi}}\left[V(\mathbf{x};\boldsymbol{\xi})\right] = \int V(\mathbf{x};\boldsymbol{\xi})p(\boldsymbol{\xi})d\boldsymbol{\xi}, \tag{2}$$

where $p(\boldsymbol{\xi})$ is the probability density function (PDF) of $\boldsymbol{\xi}$. We will develop a methodology for the solution of Eq. (1) that addresses the following challenges:

1. The objective is expensive to evaluate.
2. It is not possible to compute the gradient of the objective with respect to $\mathbf{x}$.
3. The stochastic parameters $\boldsymbol{\xi}$ are either not observed directly, or they are so high-dimensional that learning the dependence of the objective with respect to them is impossible.

Before we get to the specifics of our methodology, it is worth clarifying a few things about the data collection process. We assume that we can choose to evaluate the objective at any design point $\mathbf{x}$ we wish. We envision this evaluation to take place as follows. Behind the scenes, a random variable $\boldsymbol{\xi}$ is sampled from the, unknown, PDF $p(\boldsymbol{\xi})$, and the function $y = V(\mathbf{x};\boldsymbol{\xi})$ is evaluated. We only see y and not $\boldsymbol{\xi}$. In this way, we can obtain an *initial* data set consisting of observed design points,

$$\mathbf{x}_{1:n} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}, \tag{3}$$

and the corresponding observed noisy objective evaluations,

$$\mathbf{y}_{1:n} = \{y_1, \cdots, y_n\}. \tag{4}$$

What can be said about the solution of Eq. (1) using only the observed data $\mathbf{x}_{1:n}$ and $\mathbf{y}_{1:n}$? In the language of probability theory [21], we would like to characterize the probability of a design being optimal conditional on the observations, and similarly for the optimal objective value. Here probability corresponds to a state of belief and not to something random. The uncertainty encoded in this probability is epistemic and it is induced by the fact that inference is based on just n observations. We will answer this question by making no discounts on the Bayesian nature of Gaussian process surrogates, see Sec. 2.1 and Sec. 2.2.

Where should we evaluate the objective next? Of course, looking for an optimal information acquisition policy is a futile task since the problem is mathematically equivalent to a non-linear stochastic dynamic programming problem [22, 23]. As in standard BGO, we will rely on a suboptimal one-step-look-ahead strategy that makes use of an information acquisition function, albeit we will extend the EI information acquisition function so that it can cope robustly with noise, see Sec. 2.3.

2.1 Gaussian process regression

Gaussian process regression [24] is the Bayesian interpretation of classical Kriging [25, 26]. It is a powerful non-linear and non-parametric regression technique that has the added benefit of being able to quantify the epistemic uncertainties induced by limited data. We will use it to learn the function that corresponds to the expectation of the objective $f(\cdot) = \mathbb{E}_{\boldsymbol{\xi}}[V(\cdot;\boldsymbol{\xi})]$ from the observed data $\mathbf{x}_{1:n}$ and $\mathbf{y}_{1:n}$.

2.1.1 Expressing prior beliefs

A GP defines a probability measure on the space of meta-models, here $f(\cdot)$, which can be used to encode our prior beliefs about the response, e.g., lengthscales, regularity, before we see any data. Mathematically, we write:

$$p(f(\cdot)|\boldsymbol{\psi}) = \mathrm{GP}(f(\cdot)|m(\cdot;\boldsymbol{\psi}),k(\cdot,\cdot;\boldsymbol{\psi})), \qquad (5)$$

where $m(\cdot;\boldsymbol{\psi})$ and $k(\cdot,\cdot;\boldsymbol{\psi})$ are the mean and covariance functions of the GP, respectively, and $\boldsymbol{\psi}$ is a vector including all the hyperparameters of the model. Following the hierarchical Bayes framework, one would also have to specify a prior on the hyperparameters, $p(\boldsymbol{\psi})$.

Note that information about the mean can actually be encoded in the covariance function. Thus, without loss of generality, in this work we take $m(\cdot;\boldsymbol{\psi})$ to be identically equal to zero. In our numerical examples, we will use the squared exponential (SE) covariance:

$$k(\mathbf{x},\mathbf{x}';\boldsymbol{\psi}) = s^2 \exp\left\{ -\frac{1}{2}\sum_{i=1}^{d}\frac{(x_i - x_i')^2}{\ell_i^2} \right\}, \qquad (6)$$

where d is the dimensionality of the design space, $s > 0$ and $\ell_i > 0$ can be interpreted as the signal strength of the response and the lengthscale along input dimension i, respectively, and $\boldsymbol{\psi} = \{s,\ell_1,\ldots,\ell_d\}$. Finishing, we assume that all the hyperparameters are a priori independent:

$$p(\boldsymbol{\psi}) = p(s)\prod_{i=1}^{d}p(\ell_i), \qquad (7)$$

where

$$p(s) \propto \frac{1}{s} \qquad (8)$$

is the Jeffreys' prior [27], and

$$p(\ell_i) \propto \frac{1}{1+\ell_i^2} \qquad (9)$$

is a log-logistic prior [28].

2.1.2 Modeling the measurement process

To ensure analytical tractability, we assume that the measurement noise is Gaussian with unknown variance σ^2. Note that this could easily be relaxed to a student-t noise, which is more robust to outliers. The more general case of heteroscedastic, i.e., input-dependent, noise is an open research problem and beyond the scope of the current work. Note, however, that in our numerical examples we observe that our approach is robust to modest heteroscedasticity levels.

Mathematically, the likelihood of the data is:

$$p(\mathbf{y}_{1:n}|\mathbf{x}_{1:n},\boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{y}_{1:n}\big|\mathbf{f}_{1:n},\mathbf{K}_n(\boldsymbol{\psi})+\sigma^2\mathbf{I}_n\right), \qquad (10)$$

where $\mathcal{N}(\cdot|\mu,\Sigma)$ is the PDF of a multivariate normal random variable with mean μ and covariance matrix Σ, $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ is the identity matrix, $\mathbf{K}_n(\boldsymbol{\psi}) \in \mathbb{R}^{n\times n}$ is the covariance matrix,

$$\mathbf{K}_n(\boldsymbol{\psi}) = \begin{pmatrix} k(\mathbf{x}_1,\mathbf{x}_1;\boldsymbol{\psi}) & \ldots & k(\mathbf{x}_1,\mathbf{x}_n;\boldsymbol{\psi}) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_n,\mathbf{x}_1;\boldsymbol{\psi}) & \ldots & k(\mathbf{x}_n,\mathbf{x}_n;\boldsymbol{\psi}) \end{pmatrix}, \qquad (11)$$

and, for notational convenience, we have defined $\boldsymbol{\theta} = \{\boldsymbol{\psi},\sigma\}$. Finally, we need to assign a prior to σ. We assume that σ is a priori independent of all the variables in $\boldsymbol{\psi}$ and set:

$$p(\sigma) \propto \frac{1}{\sigma}. \qquad (12)$$

2.1.3 Posterior state of knowledge

Bayes rule combines our prior beliefs with the likelihood of the data and yields a posterior probability measure on the space of meta-models. Conditioned on the hyperparameters $\boldsymbol{\theta}$, this measure is also a Gaussian process,

$$p(f(\cdot)|\mathbf{x}_{1:n},\mathbf{y}_{1:n},\boldsymbol{\theta}) = \mathrm{GP}\left(f(\cdot)\big|m_n(\mathbf{x};\boldsymbol{\theta}),k_n(\mathbf{x},\mathbf{x}';\boldsymbol{\theta})\right), \qquad (13)$$

albeit with posterior mean and covariance functions,

$$m_n(\mathbf{x};\boldsymbol{\theta}) = (\mathbf{k}_n(\mathbf{x};\boldsymbol{\psi}))^T \left(\mathbf{K}_n(\boldsymbol{\psi})+\sigma^2\mathbf{I}_n\right)^{-1}\mathbf{y}_{1:n}, \qquad (14)$$

and

$$
\begin{aligned}
k_n(\mathbf{x},\mathbf{x}';\boldsymbol{\theta}) = {}& k(\mathbf{x},\mathbf{x}';\boldsymbol{\psi}) \\
& - (\mathbf{k}_n(\mathbf{x};\boldsymbol{\psi}))^T \left(\mathbf{K}_n(\boldsymbol{\psi})+\sigma^2\mathbf{I}_N\right)^{-1}\mathbf{k}_n(\mathbf{x}';\boldsymbol{\psi})
\end{aligned}
$$
$$(15)$$

respectively, where $\mathbf{k}_n(\mathbf{x};\boldsymbol{\psi}) = (k(\mathbf{x},\mathbf{x}_1;\boldsymbol{\psi}),\ldots,k(\mathbf{x},\mathbf{x}_n;\boldsymbol{\psi}))^T$, and $\mathbf{A}^T$ is the transpose of $\mathbf{A}$. Restricting our attention to a specific design point $\mathbf{x}$, we can derive from Eq. (13) the *point-predictive probability density* conditioned on the hyperparameters $\boldsymbol{\theta}$:

$$p(f(\mathbf{x})|\mathbf{x}_{1:n},\mathbf{y}_{1:n},\boldsymbol{\theta}) = \mathcal{N}\left(f(\mathbf{x})\big|m_n(\mathbf{x};\boldsymbol{\theta}),\sigma_n^2(\mathbf{x};\boldsymbol{\theta})\right), \qquad (16)$$

where $\sigma_n^2(\mathbf{x};\boldsymbol{\theta}) = k_n(\mathbf{x},\mathbf{x};\boldsymbol{\theta})$.

To complete the characterization of the posterior state of knowledge, we need to express our updated beliefs about the hyperparameters $\boldsymbol{\theta}$. By a standard application of the Bayes rule, we get:

$$p(\boldsymbol{\theta}|\mathbf{x}_{1:n},\mathbf{y}_{1:n}) \propto p(\mathbf{y}_{1:n}|\mathbf{x}_{1:n},\boldsymbol{\theta})p(\boldsymbol{\theta}), \qquad (17)$$

where $p(\boldsymbol{\theta}) = p(\boldsymbol{\psi})p(\sigma)$. Unfortunately, Eq. (17) cannot be computed analytically. Thus, we characterize it by a *particle approximation* consisting of N samples, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ obtained by adaptive Markov chain Monte Carlo (MCMC) [29]. Formally, we write:

$$p(\boldsymbol{\theta}|\mathbf{x}_{1:n}, \mathbf{y}_{1:n}) \approx \frac{1}{N} \sum_{i=1}^{N} \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_i), \qquad (18)$$

where $\delta(\cdot)$ is Dirac's delta function. In our numerical results, we use $N = 90$ and the samples are generated as follows: 1) We obtain a starting point for the MCMC chain by maximizing the log of the posterior Eq. (17); 2) We burn $10,000$ MCMC steps during which the MCMC proposal parameters are tuned; and 3) We perform another $90,000$ MCMC steps and record $\boldsymbol{\theta}$ every $1,000$ steps.

2.2 Epistemic uncertainty on the solution of a stochastic optimization problem

Now, we are in a position to quantify the epistemic uncertainty in the solution of Eq. (1) induced by the limited number of acquired data. Let $Q[\cdot]$ be any operator acting on functions $f(\cdot)$. Examples of such operators, are the minimum of $f(\cdot)$, $Q_{\min}[f(\cdot)] = \min_{\mathbf{x}} f(\mathbf{x})$, or the location of the minimum, $Q_{\arg\min}[f(\cdot)] = \arg\min_{\mathbf{x}} f(\mathbf{x})$. Conditioned on $\mathbf{x}_{1:n}$ and $\mathbf{y}_{1:n}$ our state of knowledge about the value of any operator $Q[\cdot]$ is

$$p(Q|\mathbf{x}_{1:n}, \mathbf{y}_{1:n}) = (\int (\int \delta(Q - Q[f(\cdot)]) \, p(f(\cdot)|\mathbf{x}_{1:n}, \mathbf{y}_{1:n}, \boldsymbol{\theta})$$
$$df(\cdot))p(\boldsymbol{\theta}|\mathbf{x}_{1:n}, \mathbf{y}_{1:n})d\boldsymbol{\theta}), \qquad (19)$$

By sampling M functions, $f_1(\cdot), \ldots, f_M(\cdot)$ from Eq. (13) and using Eq. (18), we get the particle approximation:

$$p(Q|\mathbf{x}_{1:n}, \mathbf{y}_{1:n}) \approx \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \delta(Q - Q[f_i(\cdot)]). \qquad (20)$$

Our derivation is straightforward and uses only the product and sum rules of probability theory. The implementation, however, is rather technical. For more details see the publications of Bilionis in the subject, [30, 31, 32, 33, 34]. In our numerical examples we use $M = 100$.

2.3 Extended expected improvement function

The classic definition of expected improvement, see [12], relies on the observed minimum $\tilde{y}_n = \min_{1 \le i \le n} y_i$. Unfortunately, this definition breaks down when y_i is noisy. To get a viable alternative, we have to filter out this noise. To this end, let us define the *observed filtered minimum* conditioned on $\boldsymbol{\theta}$:

$$\tilde{m}_n(\boldsymbol{\theta}) = \min_{1 \le i \le n} m_n(\mathbf{x}_i; \boldsymbol{\theta}), \qquad (21)$$

where $m_n(\mathbf{x}; \boldsymbol{\theta})$ is the posterior mean of Eq. (14). Using $\tilde{m}_n(\boldsymbol{\theta})$, the improvement we would get if we observed $f(\mathbf{x})$ at design point $\mathbf{x}$ is:

$$I(\mathbf{x}, f(\mathbf{x}); \boldsymbol{\theta}) = \max\{0, \tilde{m}_n(\boldsymbol{\theta}) - f(\mathbf{x})\}. \qquad (22)$$

This is identical to the improvement function formulated in Sequential kriging optimization (SKO) [35]. However, the EEI retains the full epistemic uncertainty unlike SKO, which relies on a point estimate to the hyper-parameters. Since we don't know $f(\mathbf{x})$ or $\boldsymbol{\theta}$, we have to take their expectation over our posterior state of knowledge, see Sec. 2.1.3,

$$\mathrm{EEI}_n(\mathbf{x}) = (\int \int I(\mathbf{x}, f(\mathbf{x}); \boldsymbol{\theta}) p(f(\mathbf{x})|\mathbf{x}_{1:n}, \mathbf{y}_{1:n}, \boldsymbol{\theta}) df(\mathbf{x})$$
$$p(\boldsymbol{\theta}|\mathbf{x}_{1:n}, \mathbf{y}_{1:n})d\boldsymbol{\theta}), \qquad (23)$$

where $p(f(\mathbf{x})|\mathbf{x}_{1:n}, \mathbf{y}_{1:n}, \boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\mathbf{x}_{1:n}, \mathbf{y}_{1:n})$ are given in Eq. (16) and Eq. (17), respectively. The inner integral can be carried out analytically in exactly the same way as one derives the classic expected improvement. To evaluate the outer integral, we have to employ the particle approximation to $p(\boldsymbol{\theta}|\mathbf{x}_{1:n}, \mathbf{y}_{1:n})$ given in Eq. (18). The end result is:

$$\mathrm{EEI}_n(\mathbf{x}) \approx \frac{1}{N} \sum_{i=1}^{N} [\sigma_n(\mathbf{x}; \boldsymbol{\theta}_i) \phi \left(\frac{\tilde{m}_n(\boldsymbol{\theta}_i) - m_n(\mathbf{x}; \boldsymbol{\theta}_i)}{\sigma_n(\mathbf{x}; \boldsymbol{\theta}_i)} \right)$$
$$+ (\tilde{m}_n(\boldsymbol{\theta}_i) - m_n(\mathbf{x}; \boldsymbol{\theta}_i)) \Phi \left(\frac{\tilde{m}_n(\boldsymbol{\theta}_i) - m_n(\mathbf{x}; \boldsymbol{\theta}_i)}{\sigma_n(\mathbf{x}; \boldsymbol{\theta}_i)} \right)]. \qquad (24)$$

Algorithm 1 demonstrates how the derived information acquisition criterion can be used in a modified version of BGO to obtain an approximation to Eq. (1). Note that instead of attempting to maximize $\mathrm{EEI}_n(\mathbf{x})$ over $\mathbf{x}$ exactly, we just search for the most informative point among a set of n_d randomly generated test points. In our numerical examples we use $n_d = 1,000$ test points following a latin hypercube design [36].

3 Numerical Results

We validate our approach, see Sec. 3.1 and 3.2, using two synthetic stochastic optimization problems with known optimal solutions. To assess the robustness of the methodology, we experiment with various levels of Gaussian noise, as well as heteroscedastic, i.e., input dependent, noise. In Sec. 3.3, we solve the oil-well placement problem with uncertainties in soil permeability and the oil price timeseries. Note that all the parameters required by our method, e.g., covariance function, priors of hyperparameters, MCMC steps, have already been introduced in the previous paragraphs and they are the same for all examples. The only thing that we vary is the initial number of observations n.

Algorithm 1 The Bayesian global optimization algorithm with the Extended expected improvement function

---

Require: Observed inputs $\mathbf{x}_{1:n}$, observed outputs $\mathbf{y}_{1:n}$, number of candidate points tested for maximum EEI at each iteration n_d, maximum number of allowed iterations S, EEI tolerance ε.

1: $s \leftarrow 0$.
2: **while** $s < S$ **do**
3:     Construct the particle approximation to the posterior of θ, Eq. (18).
4:     Generate a set of candidate test points $\hat{\mathbf{x}}_{1:n_d}$, e.g., via a latin hypercube design [36].
5:     Compute EEI on all of the candidate points $\hat{\mathbf{x}}_{1:n_d}$ using Eq. (24).
6:     Find the candidate point $\hat{\mathbf{x}}_j$ that exhibits the maximum EEI.
7:     **if** $\text{EEI}_{n+s}(\mathbf{x}_j) < \varepsilon$ **then**
8:         Break.
9:     **end if**
10:     Evaluate the objective at $\hat{\mathbf{x}}_j$ measuring $\hat{y}$.
11:     $\mathbf{x}_{1:n+s+1} \leftarrow \mathbf{x}_{1:n+s} \cup \{\hat{\mathbf{x}}_j\}$.
12:     $\mathbf{y}_{1:n+s+1} \leftarrow \mathbf{y}_{1:n+s} \cup \{\hat{y}\}$.
13:     $s \leftarrow s+1$.
14: **end while**

3.1 One-dimensional synthetic example

Consider the one-dimensional synthetic objective:

$$V(x,\xi) = 4\left(1 - \sin\left(6x + 8e^{6x-7}\right)\right) + s(x)\xi, \quad (25)$$

for $x \in [0,1]$, where ξ is a standard normal and for the noise standard deviation, $s(x)$, we will experiment with $s(x) = 0.01, 0.1, 1$, and the heteroscedastic $s(x) = \left(\frac{x-3}{3}\right)^2$. Here, $\mathbb{E}_\xi[V(x,\xi)]$ is analytically available and it is quite trivial to find that this function has two minima exhibiting the same objective value.

Fig. 1 (a) and (b) visualize the posterior state of knowledge along with the EEI (dashed purple line) as a function of x and the epistemic uncertainty on the location of the optimal design, respectively, for $s(x) = 0.01$ when $n = 5$. In Fig. 1 (a), the solid blue line is the median of the predictive distribution of the GP and the shaded blue area corresponds to a 95% prediction interval. Fig. 2 (a) and (b) depict the maximum EEI and the evolution of the 95% predictive bounds for the optimal objective value (PBOO), respectively, as a function of the iteration number. Fig. 3 (a) and (b) show the evolution of the PBOO for $(s(x) = 0.01)$ and $(s(x) = 0.1)$ respectively and Fig. 3 (a) and (b) show the evolution of the PBOO for $(s(x) = 1)$ and $(s(x) = \left(\frac{x-3}{3}\right)^2)$ respectively.

As expected, the larger the noise the more iterations are needed for convergence. In general, we have observed that the method is robust to noise as soon as the initial number of observations is not too low. For example, the case $s(x) = 1$ fails to converge to the truth, if one starts from less than five initial observations.
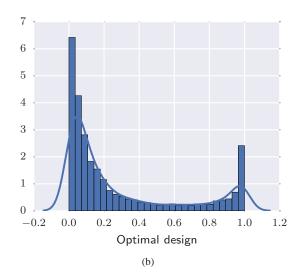

(a)

(b)

Fig. 1. One-dimensional synthetic example $(s(x) = 0.1, n = 5)$. Subfigure (a) depicts our initial state of knowledge about the true expected objective (dotted red line) conditioned on $n = 5$ noisy observations (black crosses). Subfigure (b), shows a histogram of the predictive distribution of the optimal design x^*.

3.2 Two-dimensional synthetic example

Consider the two-dimensional function [37]:

$$V(\mathbf{x};\xi) = 2 + \frac{(x_2 - x_1^2)^2}{100} + (1 - x_1)^2 + 2(2 - x_2)^2 \quad (26)$$
$$+ 7\sin(0.5x_2)\sin(0.7x_1 x_2) + s(\mathbf{x})\xi,$$

for $\mathbf{x} \in [0,5]^2$, ξ a standard normal, and $s(\mathbf{x}) = 0.01, 0.1, 1$, or the heteroscedastic $s(\mathbf{x}) = \left(\frac{x_2-x_1}{3}\right)^2$. As before, the expectation over ξ is analytically available. It can easily be verified that the objective exhibits three minima two of which are suboptimal.

Fig. 5 (a) and (b) show the PBOO for $(s(\mathbf{x}) = 0.01)$ and $(s(\mathbf{x}) = 0.1)$ and Fig. 6 (a) and (b) show the PBOO for $(s(\mathbf{x}) = 1)$ and $(s(\mathbf{x}) = \left(\frac{x_2-x_1}{3}\right)^2)$, respectively, as a function
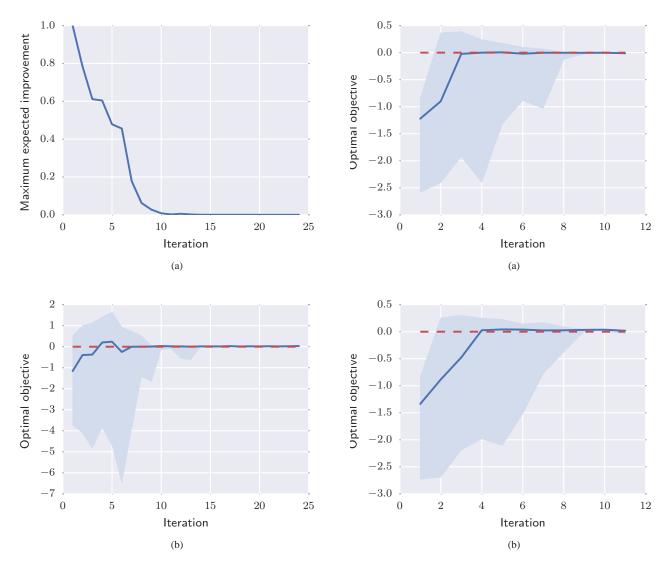
(a)

(b)

Fig. 2. One-dimensional synthetic example ($s(x) = 0.1, n = 5$). The dashed red line in Subfigure (b) marks the real optimal value.

(a)

(b)

Fig. 3. One-dimensional synthetic example ($n = 10$).

of the number of iterations. As before, the larger the noise the more iterations are required for convergence. The observed spikes are caused by the limited data used to build the surrogate. In particular, the model is "fooled" to believe that the noise is smaller than it actually is and, as a result, it becomes more certain about the solution of the optimization problem. As more observations are gathered though, the model is self-corrected. This is a manifestation of the well known S-curve effect of information acquisition [22, Ch. 5.2]. The existence of this effect means, however, that one needs to be very careful in choosing the stopping criterion.

3.3 Oil well placement problem

During secondary oil production, water (potentially enhanced with chemicals or gas) is injected into the reservoir through an *injection* well. The injected fluid pushes the oil out of the *production* well. The *oil well placement problem* (OWPP) involves the specification of the number and loca-

tion of the injection and production wells, the operating pressures, the production schedule, etc., that maximize the net present value (NPV) of the investment. This problem is of extreme importance for the oil industry and an active area of research. Several sources of uncertainty influence the NPV, the most important of which are the time evolution of the oil price (aleatoric uncertainty) and the uncertainty about the underground geophysical parameters (epistemic uncertainty).

We consider an idealized 2D oil reservoir over the spatial domain $\Omega = [0, 356.76] \times [0, 670.56]$ (measured in meters). The four-dimensional design variable $\mathbf{x} = (x_1, x_2, x_3, x_4)$ specifies the location of the injection well (x_1, x_2), in which we pump water (w), and the production well (x_3, x_4), out of which comes oil (o) and water. Letting $\mathbf{x}_s \in \Omega$ denote a spatial location, we assume that the permeability of the ground is an isotropic tensor,

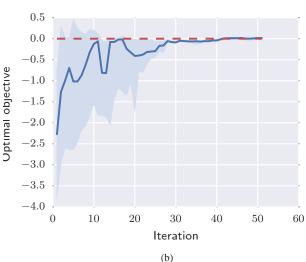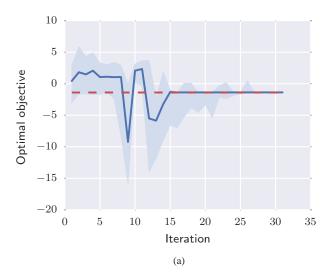$$\mathbf{C}(\mathbf{x}_s; \boldsymbol{\xi}_c) = e^{g(\mathbf{x}_s; \boldsymbol{\xi}_c)} c(\mathbf{x}_s) \mathbf{I}_3, \qquad (27)$$

Fig. 4. One-dimensional synthetic example ($n = 10$).

Fig. 5. Two-dimensional synthetic example ($n = 20$).

where $c(\mathbf{x}_s)$ is the geometric mean (assumed to be the first layer of the x-component of the SPE10 reservoir model permeability tensor [38]), $g(\mathbf{x}_s; \boldsymbol{\xi}_c)$ is the truncated, at $13,200$ terms, Karhunen-Loève expansion of a random field with exponential covariance function of lengthscale $\ell = 10$ meters and variance 10, see [39], and $\boldsymbol{\xi}_c$ is a $(13,200)$-dimensional vector of standard normal random variables. Four samples of the permeability field are depicted in Fig. 7.

Given the well locations $\mathbf{x}$ and the stochastic variables $\boldsymbol{\xi}_c$, we solve a coupled system of time-dependent partial differential equations (PDEs) describing the two-phase immiscible flow of water and oil through the reservoir. The solution is based on a finite volume scheme with a 60×220 regular grid. The form of the PDEs, the required boundary and initial conditions, as well as the details of the finite volume discretization are discussed in [40]. The parameters of the model that remain constant are as follows. The water injection rate is 9.35 m^3/day, the connate water saturation is $s_{\text{wc}} = 0.2$, the irreducible oil saturation is $s_{\text{or}} = 0.2$, the water viscosity is set to $\mu_w = 3 \times 10^{-4}$ Pa·s, the oil viscosity to
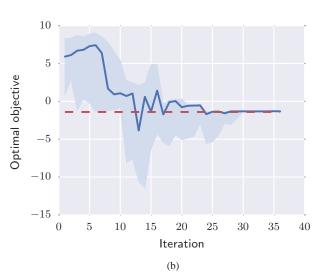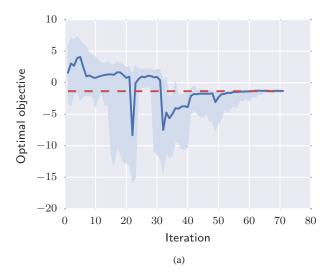
$\mu_o = 3 \times 10^{-3}$ Pa·s, the soil porosity is 10^{-3}, the timestep used is $\delta t = 0.1$ days, and operations last $T = 2,000$ days. From the solution of the PDE system, we obtain the oil and water extraction rates $q_o(t; \mathbf{x}, \boldsymbol{\xi}_c)$ and $q_w(t; \mathbf{x}, \boldsymbol{\xi}_c)$, respectively, where t is the time in days and the units of these quantities are in m^3/day.

The oil price is modeled on a daily basis as $S_{o,t} = S_{o,0}e^{W_t}$, where $S_{o,0} = \$560.8/\text{m}^3$, and W_t is a random walk with a drift:

$$W_{t+1} = W_t + \mu + \alpha \xi_{o,t}, \tag{28}$$

where the $\mu = 10^{-8}$, $\alpha = 10^{-3}$, and $\xi_{o,t}$ are independent standard normal random variables. Fig. 8 visualizes four samples from the oil price model. Since the process runs for $T = 2,000$ days, we can think of $S_{o,t}$ as a function of the $2,000$ independent identically distributed random variables $\boldsymbol{\xi}_o = \{\boldsymbol{\xi}_{o,1}, \ldots, \boldsymbol{\xi}_{o,T}\}$, i.e., $S_{o,t} = S_{o,t}(\boldsymbol{\xi}_o)$. For simplicity, we take the cost of disposing contaminated water is constant over time $S_{w,t}^- = \$0.30/\text{m}^3$. Assuming a discount rate
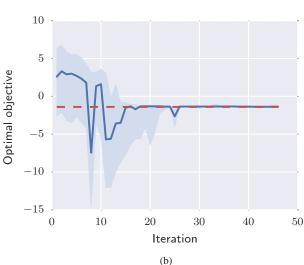
(a)

(b)

Fig. 6. Two-dimensional synthetic example ($n = 20$).

Fig. 7. OWPP: Samples from the stochastic permeability model (in logarithmic scale) defined in Eq. (27).

$r = 10\%$ and risk neutrality, our objective is to maximize the NPV of the investment. Equivalently, we wish to minimize:

$$V(\mathbf{x};\boldsymbol{\xi}) = 10^{-6} \sum_{t=1}^{2{,}000\,\text{days}} \left[S_{w,t} q(t;\mathbf{x},\boldsymbol{\xi}_c) - S_{o,t}(\boldsymbol{\xi}_o) q_o(t;\mathbf{x},\boldsymbol{\xi}_c) \right]$$
$$(1+r)^{-t/365\,\text{days}}, \quad (29)$$

where $\boldsymbol{\xi} = \{\boldsymbol{\xi}_c, \boldsymbol{\xi}_o\}$, and the units are in million dollars.

Fig. 9 (a) shows the evolution of the PBOO as a function of the iterations of our algorithm for the case of $n = 20$ initial observations. Note that in this case, we do not actually know what the optimal value of the objective is. In subfigures (b) and (c) of the same figure, we visualize the initial set of observed well pairs and the well pairs selected for simulation by our algorithm (where the blue 'x' stands for the injection well, the red 'o' for the production well) respectively. Our algorithm quickly realizes the wells that are two close together
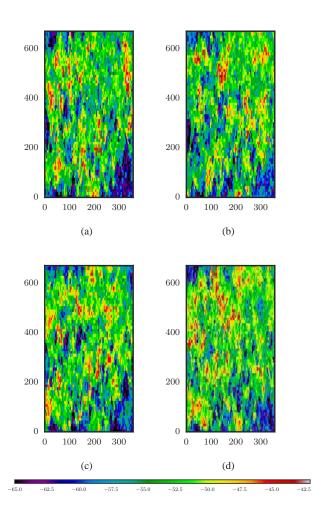
are suboptimal and that it seems to favor wells that are located at the bottom right and top right corners. Note that the noise in this case is moderate, albeit heteroscedastic.

4 Conclusions

We constructed an extension to the expected improvement which makes possible the application of Bayesian global optimization to stochastic optimization problems. In addition, we have shown how the epistemic uncertainty induced by the limited number of simulations can be quantified, by deriving predictive probability distributions for the location of the optimum as well as the optimal value of the problem. We have validated our approach with two synthetic examples with known solution and various noise levels, and we applied it to the challenging oil well placement problem. The method offers a viable alternative to the sampling average approximation when the cost of simulations is significant. We observe that our approach is robust to moderate noise heteroscedasticity. There remain several open research questions. In our opinion, the most important direction would be to construct surrogates that explicitly model het-
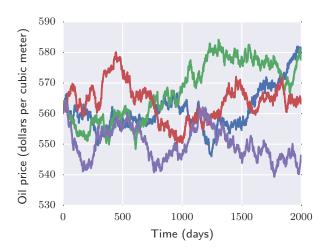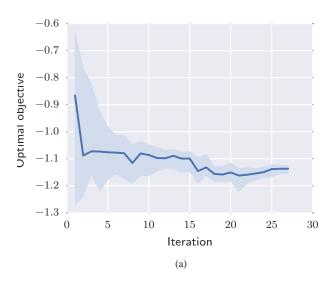
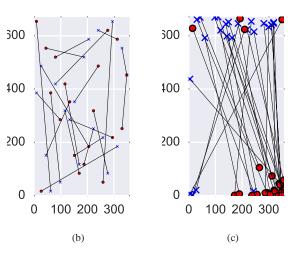Fig. 8.   OWPP: Samples from the stochastic oil price model.

(a)

(b)

(c)

Fig. 9.   OWPP ($n = 20$).

eroscedasticity and use them to extend the present methodology to robust stochastic optimization and, subsequently, to multi-objective stochastic optimization.

References

[1] Bottou, L., 2010. "Large-scale machine learning with stochastic gradient descent". In *Proceedings of COMPSTAT'2010*. Springer, pp. 177–186.

[2] Kleywegt, A. J., Shapiro, A., and Homem-de Mello, T., 2002. "The sample average approximation method for stochastic discrete optimization". *SIAM Journal on Optimization,* **12**(2), pp. 479–502.

[3] Heyman, D. P., and Sobel, M. J., 2003. *Stochastic Models in Operations Research: Stochastic Optimization*, Vol. 2. Courier Corporation.

[4] Zinkevich, M., Weimer, M., Li, L., and Smola, A. J., 2010. "Parallelized stochastic gradient descent". In *Advances in neural information processing systems,* pp. 2595–2603.

[5] Torn, A., and Zilinskas, A., 1987. *Global Optimization*. Springer.

[6] Mockus, J., 1994. "Application of bayesian approach to numerical methods of global and stochastic optimization". *Journal of Global Optimization,* **4**(4), pp. 347–365.

[7] Locatelli, M., 1997. "Bayesian algorithms for one-dimensional global optimization". *Journal of Global Optimization,* **10**(1), pp. 57–76.

[8] Jones, D. R., 2001. "A taxonomy of global optimization methods based on response surfaces". *Journal of global optimization,* **21**(4), pp. 345–383.

[9] Lizotte, D., 2008. "Practical bayesian optization". Thesis.

[10] Benassi, R., Bect, J., and Vazquez, E., 2011. *Robust Gaussian process-based global optimization using a fully Bayesian expected improvement criterion.* Springer, pp. 176–190.

[11] Bull, A. D., 2011. "Convergence rates of efficient global optimization algorithms". *Journal of Machine Learning Research,* **12**, pp. 2879–2904.

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

[13] Frazier, P. I., Powell, W. B., and Dayanik, S., 2008. "A knowledge-gradient policy for sequential information collection". *SIAM Journal on Control and Optimization,* **47**(5), pp. 2410–2439.

[14] Frazier, P., Powell, W., and Dayanik, S., 2009. "The knowledge-gradient policy for correlated normal beliefs". *Informs Journal on Computing,* **21**(4), pp. 599–613.

[15] Negoescu, D. M., Frazier, P. I., and Powell, W. B., 2011. "The knowledge-gradient algorithm for sequencing experiments in drug discovery". *Informs Journal on Computing,* **23**(3), pp. 346–363.

[16] Scott, W., Frazier, P., and Powell, W., 2011. "The correlated knowledge gradient for simulation optimization of continuous parameters using gaussian process regression". *SIAM Journal on Optimization,* **21**(3), pp. 996–1026.

[17] Villemonteix, J., Vazquez, E., and Walter, E., 2009. "An informational approach to the global optimization of expensive-to-evaluate functions". *Journal of Global Optimization,* **44**(4), pp. 509–534.

[18] Hennig, P., and Schuler, C. J., 2012. "Entropy search for infromation-efficient global optimization". *Journal of Machine Learning Research,* **13**, pp. 1809–1837.

[19] Hernadez-Lobato, J. M., Hoffman, M., and Ghahramani, Z. "Predictive entropy search for efficient global optimization of black-box functions". In Advances in Neural Information Processing Systems.

[20] MacKay, D. J. C., 1992. "Information-based objective functions for active data selection". *Neural Computation,* **4**(4), pp. 590–604.

[21] Jaynes, E. T., 2003. *Probability Theory: The Logic of Science*. Cambridge.

[22] Powell, W. B., and Ryzhov, I. O., 2012. *Optimal learning*, Vol. 841. John Wiley & Sons.

[23] Bertsekas, D., 2007. *Dynamic Programming and Optimal Control*, 4th ed. Athena Scientific.

[24] Rasmussen, C. E., and Williams, C. K. I., 2006. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, MA.

[25] Cressie, N., 1990. "The origins of kriging". *Mathematical geology,* **22**(3), pp. 239–252.

[26] Smith, T. E., and Dearmon, J., 2014. "Gaussian process regression and bayesian model averaging: An alternative approach to modeling spatial phenomena".

[27] Jeffreys, H., 1946. "An invariant form for the prior probability in estimation problems". *Proceedings of the Royal Society of London Series a-Mathematical and Physical Sciences,* **186**(1007), pp. 453–461.

[28] Conti, S., and O'Hagan, A., 2010. "Bayesian emulation of complex multi-output and dynamic computer models". *Journal of Statistical Planning and Inference,* **140**(3), pp. 640–651.

[29] Haario, H., Laine, M., Mira, A., and Saksman, E., 2006. "Dram: Efficient adaptive mcmc". *Statistics and Computing,* **16**(4), pp. 339–354.

[30] Bilionis, I., and Zabaras, N., 2012. "Multi-output local gaussian process regression: Applications to uncertainty quantification". *Journal of Computational Physics,* **231**(17), pp. 5718–5746.

[31] Bilionis, I., and Zabaras, N., 2012. "Multidimensional adaptive relevance vector machines for uncertainty quantification". *Siam Journal on Scientific Computing,* **34**(6), pp. B881–B908.

[32] Bilionis, I., Zabaras, N., Konomi, B. A., and Lin, G., 2013. "Multi-output separable gaussian process: Towards an efficient, fully bayesian paradigm for uncertainty quantification". *Journal of Computational Physics,* **241**, pp. 212–239.

[33] Bilionis, I., and Zabaras, N., 2014. "Solution of inverse problems with limited forward solver evaluations: a bayesian perspective". *Inverse Problems,* **30**(1). 278BA Times Cited:0 Cited References Count:32.

[34] Chen, P., Zabaras, N., and Bilionis, I., 2015. "Uncertainty propagation using infinite mixture of gaussian processes and variational bayesian inference". *Journal of Computational Physics,* **284**, pp. 291–333.

[35] Huang, D., Allen, T. T., Notz, W. I., and Zeng, N., 2006. "Global optimization of stochastic black-box systems via sequential kriging meta-models". *Journal of global optimization,* **34**(3), pp. 441–466.

[36] Mckay, M. D., Beckman, R. J., and Conover, W. J., 2000. "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code". *Technometrics,* **42**(1), pp. 55–61.

[37] Sasena, M. J., 2002. "Flexibility and efficiency enhancements for constrained global design optimization with kriging approximations". PhD thesis, General Motors.

[38] Christie, M., and Blunt, M., 2001. "Tenth spe comparative solution project: A comparison of upscaling techniques". *SPE Reservoir Evaluation & Engineering,* **4**(04), pp. 308–317.

[39] Ghanem, R. G., and Spanos, P. D., 2003. *Stochastic finite elements: a spectral approach*. Courier Corporation.

[40] Bilionis, I., and Zabaras, N., 2014. "Solution of inverse problems with limited forward solver evaluations: a bayesian perspective". *Inverse Problems,* **30**(1), p. 015004.