Bayesian Optimization Under Uncertainty

Justin J. Beland
University of Toronto
justin.beland@mail.utoronto.ca

Prasanth B. Nair
University of Toronto
pbn@utias.utoronto.ca

Abstract

We consider the problem of robust optimization, where it is sought to design a system such that it sustains a specified measure of performance under uncertainty. This problem is challenging since modeling a complex system under uncertainty can be expensive and for most real-world problems robust optimization will not be computationally viable. In this paper, we propose a Bayesian methodology to efficiently solve a class of robust optimization problems that arise in engineering design under uncertainty. The central idea is to use Gaussian process models of loss functions (or robustness metrics) together with appropriate acquisition functions to guide the search for a robust optimal solution. Numerical studies on a test problem are presented to demonstrate the efficacy of the proposed approach.

1 Introduction

Consider a scalar output of an expensive computer simulation $f(\boldsymbol{x}_1, \boldsymbol{x}_2+\boldsymbol{\delta}, \boldsymbol{\xi})$, where $\boldsymbol{x}_1 \in \mathcal{X}_1 \subset \mathbb{R}^{d_1}$ and $\boldsymbol{x}_2 \in \mathcal{X}_2 \subset \mathbb{R}^{d_2}$ can be precisely controlled (control factors) while $\boldsymbol{\delta} \in \mathcal{Y} \subset \mathbb{R}^{d_2}$ and $\boldsymbol{\xi} \in \mathcal{Z} \subset \mathbb{R}^{d_{\boldsymbol{\xi}}}$ are random variables (noise factors) with the specified joint probability density function $p(\boldsymbol{\delta}, \boldsymbol{\xi})$. Now, suppose that we seek the minima of f subject to the set of inequality constraints, $c_j(\boldsymbol{x}_1, \boldsymbol{x}_2 + \boldsymbol{\delta}, \boldsymbol{\xi}) \leq 0,\ j = 1, \ldots, d_{\boldsymbol{c}}$, by varying the control factors. This problem can be posed as a robust optimization problem where we seek to minimize some measure of loss such that the optimum is least sensitive to the noise factors $\boldsymbol{\delta}$ and $\boldsymbol{\xi}$. The robust optimization problem that we consider is of the form

$$\boldsymbol{x}^\star = \underset{\boldsymbol{x} \in \{\boldsymbol{x}_1, \boldsymbol{x}_2\}}{\arg\min}\ \mathcal{J}(\boldsymbol{x}) \qquad \text{s.t.} \qquad \Pr[c_j \leq 0] \ \geq\ 1 - \eta,\ j = 1, \ldots, d_{\boldsymbol{c}}, \tag{1}$$

where $\mathcal{J} : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$ denotes a loss function (or a robustness metric) [1] and $\eta \in [0, 1]$ is a user defined parameter that controls the probability of constraint satisfaction. Optimization problems of this form are encountered in many areas such as the design of circuits, aircraft and automotive components [2, 3].

The primary focus of the present work is to develop efficient Bayesian optimization (BO) methods for solving (1). It is well known that Bayesian methods are well suited for locating the minima of complex optimization problems on a limited computational budget, particularly when gradient information is not available and the underlying function is corrupted by noise [4–7]. To the best of our knowledge, BO methods have not been formulated for robust optimization problems of the form considered here. The key challenge is that we don't have access to observations of $\mathcal{J}$ due to computational resource limitations, instead we only have the ability to query f and $c_j,\ j = 1, \ldots, d_{\boldsymbol{c}}$. In this paper, we present a methodology to estimate loss functions that are of interest in optimization under uncertainty using Gaussian process (GP) models [8] conditioned on observations of f and $c_j,\ j = 1, \ldots, d_{\boldsymbol{c}}$. Subsequently, we propose acquisition functions that can be used to iteratively converge to the minima of (1). Finally, we present numerical studies to demonstrate the performance of the proposed algorithm.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2 Bayesian Optimization Under Uncertainty

In this section we outline the proposed BO strategy to solve (1). To simplify our notation, we introduce two new variables: $\widetilde{\boldsymbol{x}} = \{\boldsymbol{x}_1, \boldsymbol{x}_2 + \boldsymbol{\delta}, \boldsymbol{\xi}\}$ defined over the product space $\widetilde{\mathcal{X}} \subset \mathbb{R}^{d_1+d_2+d_\xi}$ and $\boldsymbol{\zeta} = \{\boldsymbol{\delta}, \boldsymbol{\xi}\}$ defined over the product space $\mathcal{Q} = \mathcal{Y} \times \mathcal{Z}$. In addition, we will use the notation $\widetilde{\boldsymbol{x}}^i$ to denote the ith observation of $\widetilde{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{x}}^{1:t}$ to denote t observations. Lastly, we denote the set of optimization variables (or control factors) $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ by $\boldsymbol{x} \in \mathcal{X}$ where $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$. Note that $\mathcal{Y}$ and $\mathcal{Z}$ are the image spaces of $\boldsymbol{\delta}$ and $\boldsymbol{\xi}$ respectively.

In the BO under uncertainty framework, the loss function $\mathcal{J}$ is specified such that the noise factors $\boldsymbol{\zeta}$ are integrated out and decisions can be made entirely in the decision space $\mathcal{X}$. One possibility is to use Bayes risk [9] as the loss function, i.e.

$$\mathcal{J}(\boldsymbol{x}) = \int_{\mathcal{Q}} f(\widetilde{\boldsymbol{x}}) p(\boldsymbol{\zeta}) d\boldsymbol{\zeta}. \tag{2}$$

Here, the loss function is the first statistical moment of f over the space $\mathcal{Q}$ given a setting for the control factors $\boldsymbol{x}$. Alternatively, one may define $\mathcal{J}$ as the second-order statistical moment given by $\int_{\mathcal{Q}} f(\widetilde{\boldsymbol{x}})^2 p(\boldsymbol{\zeta}) d\boldsymbol{\zeta}$. This measure of loss simultaneously minimizes Bayes risk and the variance of f [10]. The proposed framework for BO under uncertainty can accommodate a wide variety of alternative robustness metrics, such as the aggregate of the mean and the variance [11], the minimax principle [3] and horsetail matching [12].

Next, we specify a zero mean GP prior over the function f with a covariance function $k_f^{\text{pr}} : \widetilde{\mathcal{X}} \times \widetilde{\mathcal{X}} \to \mathbb{R}$. In other words, we approximate f as a function of $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{\xi}$. Note that it is not necessary to explicitly model the dependence of f on $\boldsymbol{\delta}$ since this noise factor can be interpreted to be a perturbation to $\boldsymbol{x}_2$. Now, suppose that we have gathered the t observations $\boldsymbol{y}^{1:t} = f(\widetilde{\boldsymbol{x}}^{1:t}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \nu)$ denotes measurement noise. Subsequently, by conditioning the prior on the t observations we obtain the posterior distribution $f \sim \mathcal{GP}(\mu_f^{\text{pos}}, k_f^{\text{pos}})$, where $\mu_f^{\text{pos}}(\widetilde{\boldsymbol{x}}) = \boldsymbol{k}(\widetilde{\boldsymbol{x}})^\mathsf{T}(\boldsymbol{K} + \nu \boldsymbol{I})^{-1} \boldsymbol{y}^{1:t}$ is the posterior mean and $k_f^{\text{pos}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}') = k_f^{\text{pr}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}') - \boldsymbol{k}(\widetilde{\boldsymbol{x}})^\mathsf{T}(\boldsymbol{K} + \nu \boldsymbol{I})^{-1} \boldsymbol{k}(\widetilde{\boldsymbol{x}}')$ denotes the posterior covariance. The elements of the covariance matrix $\boldsymbol{K} \in \mathbb{R}^{t \times t}$ are given by $K_{pq} = k_f^{\text{pr}}(\widetilde{\boldsymbol{x}}^p, \widetilde{\boldsymbol{x}}^q) + \nu \delta_{pq}$ where δ_{pq} denotes the Kronecker delta and $\boldsymbol{k}(\widetilde{\boldsymbol{x}}) = [k_f^{\text{pr}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}^1), \ldots, k_f^{\text{pr}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}^t)]^\mathsf{T}$ is the vector of cross covariances. The hyperparameters in the prior covariance function and the noise variance can be estimated using an empirical or fully Bayesian approach [8].

Consider the case when the loss function $\mathcal{J}$ is defined as Bayes risk (see (2)). Since the loss function is a linear operator applied to f it follows that,

$$\mathcal{J} \sim \mathcal{GP}(\mu_{\mathcal{J}}^{\text{pos}}, k_{\mathcal{J}}^{\text{pos}}), \tag{3}$$

where

$$\mu_{\mathcal{J}}^{\text{pos}}(\boldsymbol{x}) = \int_{\mathcal{Q}} \mu_f^{\text{pos}}(\widetilde{\boldsymbol{x}}) p(\boldsymbol{\zeta}) d\boldsymbol{\zeta} = \boldsymbol{z}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{K} + \nu \boldsymbol{I})^{-1} \boldsymbol{y}^{1:t}, \tag{4}$$

$$k_{\mathcal{J}}^{\text{pos}}(\boldsymbol{x}, \boldsymbol{x}') = \int_{\mathcal{Q}} \int_{\mathcal{Q}} k_f^{\text{pos}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}') p(\boldsymbol{\zeta}, \boldsymbol{\zeta}') d\boldsymbol{\zeta} d\boldsymbol{\zeta}' = z(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{z}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{K} + \nu \boldsymbol{I})^{-1} \boldsymbol{z}(\boldsymbol{x}'), \tag{5}$$

$\boldsymbol{z}(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}^t$ with its ith component defined as $z_i(\boldsymbol{x}) = \int_{\mathcal{Q}} k_f^{\text{pr}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}^i) p(\boldsymbol{\zeta}) d\boldsymbol{\zeta}$, and $z : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is given by

$$z(\boldsymbol{x}, \boldsymbol{x}') = \int_{\mathcal{Q}} \int_{\mathcal{Q}} k_f^{\text{pr}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}') p(\boldsymbol{\zeta}, \boldsymbol{\zeta}') d\boldsymbol{\zeta} d\boldsymbol{\zeta}'. \tag{6}$$

The integrals that appear in (4) - (6) can be evaluated analytically provided that the covariance function $k_f^{\text{pr}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}')$ and the joint distribution $p(\boldsymbol{\zeta})$ are separable with respect to their input arguments. When this is not feasible, a multivariate sparse quadrature scheme can be used to approximate these integrals [13].

We note that if $\mathcal{J}$ is not a linear operator applied to f then the estimator for the loss function is no longer a GP, for example, if $\mathcal{J}(\boldsymbol{x}) = \int_{\mathcal{Q}} f(\widetilde{\boldsymbol{x}})^2 p(\boldsymbol{\zeta}) d\boldsymbol{\zeta}$. In this specific case, we can apply GP inferencing to the integrand f^2 so that the estimator for $\mathcal{J}$ is a GP. For a more general nonlinear loss function, we can construct a Gaussian approximation of $\mathcal{J}$ using a sampling procedure or a sparse quadrature scheme.

To guide the search towards the minimizer $\boldsymbol{x}^\star$ while ensuring a high probability of constraint satisfaction, we require a strategy to identify the point $\boldsymbol{x}^{t+1} = \{\boldsymbol{x}_1^{t+1}, \boldsymbol{x}_2^{t+1}\} \in \mathcal{X}$, where we should next query the deterministic computer model. However, this information alone is not sufficient to evaluate f and $c_j,\ j = 1, \ldots, d_{\boldsymbol{c}}$, as we must also select a setting for $\boldsymbol{\xi}^{t+1} \in \mathcal{Z}$. We then require the acquisition function $\alpha_{\boldsymbol{x}}^{\boldsymbol{c}} : \mathcal{X} \to \mathbb{R}$ to guide the optimization and $\alpha_{\boldsymbol{\xi}}^{\boldsymbol{c}} : \mathcal{Z} \to \mathbb{R}$ to identify an appropriate setting for $\boldsymbol{\xi}^{t+1}$.

Locating the query point $\boldsymbol{x}^{t+1}$ is accomplished by maximizing $\alpha_{\boldsymbol{x}}^{\boldsymbol{c}}(\boldsymbol{x})$. To avoid sampling far away from the feasible region we select $\boldsymbol{x}^{t+1}$ such that $\Pr[c_j(\boldsymbol{x}_1^{t+1}, \boldsymbol{x}_2^{t+1} + \boldsymbol{\delta}, \boldsymbol{\xi}) \leq 0],\ j = 1, \ldots, d_{\boldsymbol{c}}$ is high before querying the computer model [14]. The proposed acquisition function can be written as

$$\alpha_{\boldsymbol{x}}^{\boldsymbol{c}}(\boldsymbol{x}) = \alpha_{\boldsymbol{x}}(\boldsymbol{x}) \prod_{j=1}^{d_{\boldsymbol{c}}} \Pr[c_j(\boldsymbol{x}_1, \boldsymbol{x}_2 + \boldsymbol{\delta}, \boldsymbol{\xi}) \leq 0], \tag{7}$$

where one example of $\alpha_{\boldsymbol{x}} : \mathcal{X} \to \mathbb{R}$ is the probability of improvement (POI) criterion [15]. This means $\alpha_{\boldsymbol{x}} = \Pr[\mathcal{J} \leq \mathcal{J}^\dagger]$, where $\mathcal{J}^\dagger$ denotes a target for the loss function. Another possibility is the expected improvement (EI) criterion [16], which is defined as $\mathbb{E}[\max(0, \mathcal{J}^\dagger - \mathcal{J})]$. It is also possible to use the lower confidence bound (LCB) [17] defined as $\mu_{\mathcal{J}}^{\text{pos}}(\boldsymbol{x}) - \beta k_{\mathcal{J}}^{\text{pos}}(\boldsymbol{x}, \boldsymbol{x})$ with $\beta \in \mathbb{R}^+$. The LCB is well suited to problems where the goal is to minimize regret but may suggest new points that do not explore beyond a local minima. It is worth noting that it is also possible to use information gain metrics as candidate acquisition functions for use in the BO under uncertainty framework [18–20]. If we have GP models for all of the constraints then the product terms appearing in (7) can be efficiently approximated using the ideas presented in [21, 22].

Finally, we consider the case when $\boldsymbol{\xi}^{t+1}$ is selected such that the predictive capability of the GP models for the objective and constraint functions are improved. One way to proceed further would be to assume that $\boldsymbol{x}^{t+1}$ is fixed and define the acquisition function as $\alpha_{\boldsymbol{\xi}}^{\boldsymbol{c}}(\boldsymbol{\xi}) = k_f^{\text{pos}}(\{\boldsymbol{x}^{t+1}, \boldsymbol{\xi}\}, \{\boldsymbol{x}^{t+1}, \boldsymbol{\xi}\})$. By maximizing this acquisition function we select the location in the image space $\mathcal{Z}$ where the model is least accurate. An alternative would be to express $\alpha_{\boldsymbol{\xi}}^{\boldsymbol{c}}(\boldsymbol{\xi})$ as the aggregate of both the variance of the objective function and all of the constraints. Another option would be to integrate over the control variables from the GP posterior variance and maximize $\alpha_{\boldsymbol{\xi}}^{\boldsymbol{c}}(\boldsymbol{\xi}) = \int_{\mathcal{X}} k_f^{\text{pos}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}) d\boldsymbol{x}$ to obtain the setting for $\boldsymbol{\xi}^{t+1}$. Again, it would be possible to amalgamate some combination of the variance of each constraint $c_j,\ j = 1, \ldots, d_{\boldsymbol{c}}$ and integrate over $\mathcal{X}$. With the new points selected, we query the objective and constraint functions and then augment the dataset $\mathcal{D}^{1:t} = \{\widetilde{\boldsymbol{x}}^{1:t}, \boldsymbol{y}^{1:t}, \boldsymbol{c}_y^{1:t}\}$ with the new observations $\{\widetilde{\boldsymbol{x}}^{t+1}, y^{t+1}, \boldsymbol{c}_y^{t+1}\}$ to obtain $\mathcal{D}^{t+1}$, where $\widetilde{\boldsymbol{x}}^i$ contains $\{\boldsymbol{x}^i, \boldsymbol{\xi}^i\}$ and $\boldsymbol{c}_y^i \in \mathbb{R}^{d_{\boldsymbol{c}}}$ denotes the ith vector of noise corrupted constraint observations. The key steps of the proposed methodology are outlined in Algorithm 1.

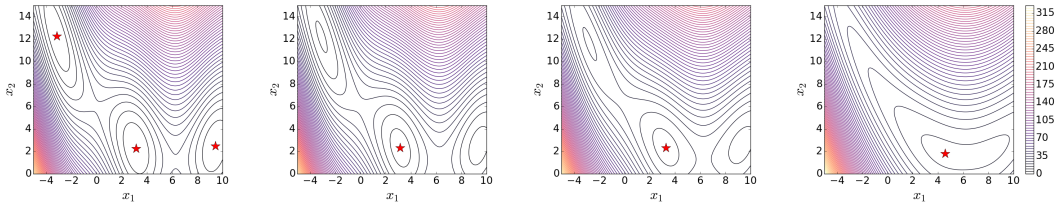
Algorithm 1: Bayesian Optimization Under Uncertainty

$\mathcal{D}^{1:t} = \{\widetilde{\boldsymbol{x}}^{1:t}, \boldsymbol{y}^{1:t}, \boldsymbol{c}_y^{1:t}\}$ `// Initialize training dataset with` t `samples`
while $cost \leq budget$ **do**
- $f \sim \mathcal{GP}(\mu_f^{\text{pos}}, k_f^{\text{pos}}) \xleftarrow[\text{on } \mathcal{D}^{1:t}]{\text{conditioning}} \mathcal{GP}(\mu_f^{\text{pr}}, k_f^{\text{pr}})$
- $\mathcal{J} \sim \mathcal{GP}(\mu_{\mathcal{J}}^{\text{pos}}, k_{\mathcal{J}}^{\text{pos}})$
- $c_j \sim \mathcal{GP}(\mu_{c_j}^{\text{pos}}, k_{c_j}^{\text{pos}}) \xleftarrow[\text{on } \mathcal{D}^{1:t}]{\text{conditioning}} \mathcal{GP}(\mu_{c_j}^{\text{pr}}, k_{c_j}^{\text{pr}}),\ j = 1, \ldots, d_{\boldsymbol{c}}$
- $\boldsymbol{x}^{t+1} = \arg\max_{\boldsymbol{x}} \alpha_{\boldsymbol{x}}^{\boldsymbol{c}}(\boldsymbol{x}) \qquad \boldsymbol{\xi}^{t+1} = \arg\max_{\boldsymbol{\xi}} \alpha_{\boldsymbol{\xi}}^{\boldsymbol{c}}(\boldsymbol{\xi})$
- $\mathcal{D}^{1:t+1} \leftarrow \mathcal{D}^{1:t} \cup \{\widetilde{\boldsymbol{x}}^{t+1}, y^{t+1}, \boldsymbol{c}_y^{t+1}\}$
- $t \leftarrow t + 1$

3 Numerical Studies

We present numerical studies involving the minimization of the Branin function [23] under uncertainty. In particular, we rewrite the Branin function as $f(\boldsymbol{x} + \boldsymbol{\delta})$, where $\boldsymbol{x} \in [-5, 10] \times [0, 15]$ and $\boldsymbol{\delta}$ is a vector of uniformly distributed random variables defined over the interval $[-\boldsymbol{\delta}_b, \boldsymbol{\delta}_b]$. For this study, we choose Bayes risk as the loss function. Since $\mathcal{J}$ is inexpensive to evaluate precisely using a quadrature scheme for this particular problem, we illustrate the contour plots of Bayes risk while varying $\boldsymbol{\delta}_b \in \mathbb{R}^2$ as shown in

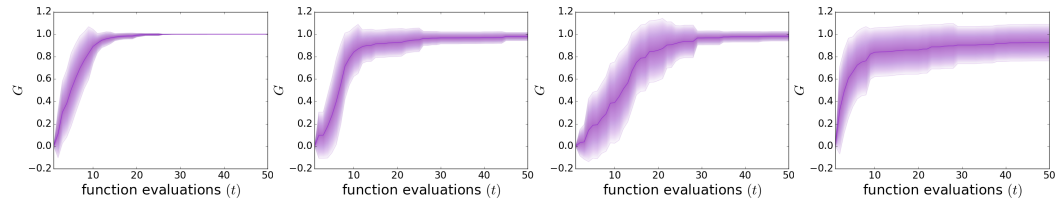
Figure 1. For the case when $\boldsymbol{\delta}_b$ is a vector of zeros as shown in Figure 1a we recover the original Branin function with three global minima. In the scenario where $\boldsymbol{\delta}_b > \mathbf{0}$ there exists a single global minima as illustrated in Figures 1b, 1c and 1d.



(a) $\boldsymbol{\delta}_b = [0, 0]^\mathsf{T}$ (b) $\boldsymbol{\delta}_b = [1, 1]^\mathsf{T}$ (c) $\boldsymbol{\delta}_b = [2, 2]^\mathsf{T}$ (d) $\boldsymbol{\delta}_b = [3, 3]^\mathsf{T}$

Figure 1: Bayes risk for the Branin function where red stars indicate global minima.

Now, suppose that we seek to minimize $\mathcal{J}$ using the methodology presented in Section 2. If the prior distribution $p(f)$ is defined by a zero mean function and the squared exponential covariance function [8] then we can derive closed form expressions for $\mu_{\mathcal{J}}^{\text{pos}}$ and $k_{\mathcal{J}}^{\text{pos}}$. We study the performance of POI, EI and UCB using the gap metric $G = (\mathcal{J}(\boldsymbol{x}^t) - \mathcal{J}(\boldsymbol{x}^\dagger))/(\mathcal{J}(\boldsymbol{x}^t) - \mathcal{J}(\boldsymbol{x}^\star))$, where $\mathcal{J}(\boldsymbol{x}^t)$ denotes the Bayes risk evaluated at the first query point and $\mathcal{J}(\boldsymbol{x}^\dagger)$ is the minimum observed value in $\mathcal{J}(\boldsymbol{x}^{1:t})$ [24]. A comparison of the results obtained for $\boldsymbol{\delta}_b = [1, 1]^\mathsf{T}$ are shown in Figure 2. We initialize the dataset $\mathcal{D}^{1:10}$ with 10 random query points and carry out 25 independent runs. Additionally, in Figure 2d,



(a) $\alpha_{\boldsymbol{x}}^{\mathbf{c}}$: POI (b) $\alpha_{\boldsymbol{x}}^{\mathbf{c}}$: EI (c) $\alpha_{\boldsymbol{x}}^{\mathbf{c}}$: LCB (d) $\alpha_{\boldsymbol{x}}^{\mathbf{c}}$: random

Figure 2: Convergence of the gap metric G for various acquisition functions.

we use a random number generator as the acquisition function for comparison. For this setting of $p(\boldsymbol{\delta})$, the POI acquisition function performs consistently well while the EI criterion explores more in the earlier iterations, thus requiring additional evaluations of f to reach the global minima. When compared with the randomized acquisition function, all methods have lower uncertainty in the later iterations.

4 Conclusions and Future Work

We propose an efficient framework for solving constrained optimization problems where the objective or the constraint functions are sensitive to uncertainty. We first specify a loss function such that decisions can be made entirely in the space of the control factors. However, if evaluating the underlying objective and constraint functions are expensive, then solving the robust optimization problem can be computationally demanding. It was shown that by specifying a GP prior over the objective function, estimating the loss function becomes tractable and in some cases the mean and covariance functions can be expressed analytically. Similarly, using GP models for the constraints, the probability of constraint satisfaction can be efficiently approximated. Finally, we showed that update points used to query the expensive objective and constraint functions can be selected by maximizing specified acquisition functions. The focus of ongoing work is on the formulation of alternative acquisition functions to identify multiple query points in parallel as well as the use of gradient observations to accelerate convergence.

Acknowledgments: This research is funded by an NSERC Discovery Grant and the Canada Research Chairs program.

References

[1] S. M. Göhler, T. Eifler, and T. J. Howard, "Robustness metrics: Consolidating the multiple approaches to quantify robustness," *Journal of Mechanical Design*, vol. 138, no. 11, p. 111 407, 2016.

[2] M. S. Phadke, *Quality engineering using robust design*. Prentice Hall PTR, 1995.

[3] H.-G. Beyer and B. Sendhoff, "Robust optimization–a comprehensive survey," *Computer methods in applied mechanics and engineering*, vol. 196, no. 33, pp. 3190–3218, 2007.

[4] J. Mockus, "Application of Bayesian approach to numerical methods of global and stochastic optimization," *Journal of Global Optimization*, vol. 4, no. 4, pp. 347–365, 1994.

[5] D. R. Jones, "A taxonomy of global optimization methods based on response surfaces," *Journal of global optimization*, vol. 21, no. 4, pp. 345–383, 2001.

[6] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.

[7] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: A review of bayesian optimization," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148–175, 2016.

[8] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. MIT press Cambridge, 2006, vol. 1.

[9] J. O. Berger, *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[10] G. Taguchi, *Introduction to quality engineering: Designing quality into products and processes*. 1986.

[11] T. K.-L. Chen W Allen J and F Mistree, "A procedure for robust design: Minimizing variations caused by noise factors and control factors," *ASME Journal of Mechanical Design*, vol. 118, pp. 478–485, 1996.

[12] L. Cook and J. Jarrett, "Horsetail matching: A flexible approach to optimization under uncertainty," *Engineering Optimization*, pp. 1–19, 2017.

[13] T. Gerstner and M. Griebel, "Numerical integration using sparse grids," *Numerical algorithms*, vol. 18, no. 3, pp. 209–232, 1998.

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

[15] H. J. Kushner, "A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise," *Journal of Basic Engineering*, vol. 86, no. 1, pp. 97–106, 1964.

[16] J. Mockus, V. Tiesis, and A. Zilinskas, "The application of Bayesian methods for seeking the extremum," *Towards Global Optimization*, vol. 2, pp. 117–129, 1978.

[17] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," *arXiv preprint arXiv:0912.3995*, 2010.

[18] P. Hennig and C. J. Schuler, "Entropy search for information-efficient global optimization," *Journal of Machine Learning Research*, vol. 13, no. Jun, pp. 1809–1837, 2012.

[19] J. M. Hernández-Lobato, M. Gelbart, M. Hoffman, R. Adams, and Z. Ghahramani, "Predictive entropy search for Bayesian optimization with unknown constraints," in *International Conference on Machine Learning*, 2015, pp. 1699–1707.

[20] Z. Wang and S. Jegelka, "Max-value entropy search for efficient Bayesian optimization," *arXiv preprint arXiv:1703.01968*, 2017.

[21] J. Bect, D. Ginsbourger, L. Li, V. Picheny, and E. Vazquez, "Sequential design of computer experiments for the estimation of a probability of failure," *Statistics and Computing*, vol. 22, no. 3, pp. 773–793, 2012.

[22] R Schöbi, B. Sudret, and S. Marelli, "Rare event estimation using polynomial-chaos kriging," *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, vol. 3, no. 2, p. D4016002, 2016.

[23] A. Torn and A. Zilinskas, *Global optimization*. Springer-Verlag New York, Inc., 1989.

[24] E. Brochu, "Interactive Bayesian optimization: Learning user preferences for graphics and animation," PhD thesis, University of British Columbia, 2010.