BAYESIAN OPTIMAL DESIGN OF EXPERIMENTS FOR EXPENSIVE BLACK-BOX FUNCTIONS UNDER UNCERTAINTY

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Piyush Pandita

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Ilias Bilionis, Co-Chair

School of Mechanical Engineering

Dr. Jitesh H. Panchal, Co-Chair

School of Mechanical Engineering

Dr. Marcial Gonzalez

School of Mechanical Engineering

Dr. Alejandro H. Strachan

School of Materials Engineering

Approved by:

Dr. Jay P. Gore

Head of the School Graduate Program

ACKNOWLEDGMENTS

I start by thanking my advisors, Prof. Ilias Bilionis and Prof. Jitesh Panchal, without whom this work would not have been possible. They supported me and nurtured my interest in doctoral studies, while overlooking my numerous deficiencies, during my initial days at Purdue University. Next, I thank my committee members, Prof. Marcial Gonzalez and Prof. Alejandro Strachan, for their consistent support and insight into my research and its applications. I thank Dr. Gautham and his team at TRDDC, Pune, India for providing me with a practical problem for testing the methods. My deepest gratitude for my first mentors in the industry, Dr. Liping Wang and Dr. Jesper Kristensen at GE Research, Niskayuna, who exposed me to various challenges in probabilistic design and allowed me to learn at my own pace. The financial support provided by the National Science Foundation and the School of Mechanical Engineering at Purdue has been critical to the timely completion of this thesis. I thank all of my teachers, co-workers, friends and family members for their encouragement and patience.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure Page

Figure Page

ABSTRACT

Pandita, Piyush Ph.D., Purdue University, May 2019. Bayesian Optimal Design of Experiments for Expensive Black-Box Functions under Uncertainty. Major Professor: Ilias Bilionis, School of Mechanical Engineering.

Researchers and scientists across various areas face the perennial challenge of selecting experimental conditions or inputs for computer simulations in order to achieve promising results. The aim of conducting these experiments could be to study the production of a material that has great applicability. One might also be interested in accurately modeling and analyzing a simulation of a physical process through a high-fidelity computer code. The presence of noise in the experimental observations or simulator outputs, called aleatory uncertainty, is usually accompanied by limited amount of data due to budget constraints. This gives rise to what is known as epistemic uncertainty. This problem of designing of experiments with limited number of allowable experiments or simulations under aleatory and epistemic uncertainty needs to be treated in a Bayesian way. The aim of this thesis is to extend and uncover the state-of-the-art in Bayesian optimal design of experiments where one can optimize and infer statistics of the expensive experimental observation(s) or simulation output(s) under uncertainty.

1. INTRODUCTION

Research groups in academia and industry devote human resources in developing mathematical methods and experimental treatments and techniques to tackle challenging problems in their domain. An output of this expertise being used to solve challenging problems are high-fidelity computational simulators and expensive experimental rigs. Since, conducting a single experiment or a single simulation can cost thousands of dollars or take multiple days one has no option but to optimally allocate their resources. Design of experiments is an area of research with stakeholders across experimental and computational science. Examples of interest in identifying optimal budget allocation policies can be found in areas of computational physics [1], experimental chemistry [2], financial planning [3], human experiment design [4–6], etc.

The experimental observations are usually contaminated with noise, that may or may not be heteroscedastic. Similarly, computer simulations are often deterministic but in some cases the presence of hidden parameters in the model gives rise to stochastic code output(s). This type of uncertainty is called aleatory uncertainty. Secondly, a major challenge is posed by the computationally expensive nature of the computer codes or the experimental setup. This means that the number of code evaluations are finite. Hence, restricting the use of state-of-the-art conventional optimization algorithms. Uncertainty in one's state-of-knowledge due to limited data points is called epistemic uncertainty.

The researcher might be interested in acquiring information about a part of the experiment's observations, optimizing the expensive computer simulator, inferring the optimal values for the designs/parameters, etc. In the language of uncertainty quantification, one might be interested in solving an inverse problem, a stochastic optimization problem, inferring statistics in an uncertainty propagation task, etc. State-of-the-art methods in optimal design can be broadly divided based on the

process of selecting experiments. One can choose to select all the experiments at once using the alphabetical-optimal designs [7] and their Bayesian counterparts [8]. The other way to select the experiments is called sequential design of experiments (SDOE). This way of conducting simulations or experiments has a Bayesian foundation [9–13] and is consistent with utility theory.

The essence of SDOE methods based on expected improvement [12, 14–16], mutual information [17], maximum entropy [18], etc., is contained in their ability to augment their state of knowledge during the design of experiments process.

This thesis aims to do the following: a) propose SDOE methods for optimization of expensive experiments or simulations with single and multiple competing objectives and b) propose SDOE methods for estimating the statistics of the experiment or simulation, in scenarios of limited data and under uncertainty. The following sections highlight some aspects of the contents of this thesis.

1.1 Bayesian global optimization

Design optimization of engineering systems with multiple competing objectives is a painstakingly tedious process especially when the objective functions are 'expensive-to-evaluate' computer codes with parametric uncertainties. The effectiveness of the state-of-the-art techniques, like goal programming, goal attainment approach, weighted-sum method and heuristic methods like genetic algorithms, is greatly diminished mainly due to the following lacuna: 1) they generate solutions that are not 'optimal' and; 2) they require large number of objective evaluations, which makes them impractical for realistic problems. Bayesian global optimization (BGO), has been relatively successful in dealing with the above challenges in solving single-objective optimization problems and has recently been extended to multi-objective optimization (MOO) [19–21]. BGO models the objectives via probabilistic surrogates and uses the epistemic uncertainty to define an information acquisition function (IAF) that quantifies the merit of evaluating the objective at untried designs. The expensive objective is evaluated at the design

corresponding to the maximum value of the IAF, and the latest observation is used to update the surrogate. This iterative process continues until a stopping criterion is met. The most commonly used IAF is the Expected improvement (EI), which extends to the Expected improvement over the dominated Hyper volume (EIHV) when solving MOO problems. Unfortunately, the current versions of EI and EIHV are unable to deal with parametric uncertainties or uncertainties in measuring the objective(s). This thesis provides systematic reformulations of EI and EIHV, to deal with the problem of stochastic BGO in singles-objective and and multiple-objective scenarios, called extended EI and extended EIHV.

We demonstrate our approach for MOO problems on a real engineering problem of die pass design for a multi-pass steel wire drawing where the physical process is represented by an expensive Finite element solver (developed at TRDDC, Pune, India). The competing objectives in this problem are the ultimate tensile strength and the strain non-uniform factor of the drawn wire. The reduction ratios and the die angles at each pass are the design/process variables which have associated uncertainties due to unavoidable manufacturing tolerances as well as die wear during the process. The methodology provides flexibility to the designer to design the die parameters while quantifying the associated uncertainties.

On the selection of multiple experiments

After starting with an initial set of noisy measurements the methodology selects the experimental condition that maximizes the information acquisition function (IAF). However, it is infeasible in this problem to conduct just one experiment in one batch, so the methodology suggests multiple experiments in one batch which is an addendum of the work done in Chapter 2. This is done by adding to the data set/initial measurements, a sample from the modeled response surface (GPs) of the outputs corresponding to the experiment condition selected. This augmented set is used to run the BGO algorithm to suggest the next condition and the process iterates till

the required number of conditions for a batch of experiments to be conducted are obtained. This methodology shows promising results in a chemical vapor deposition experiment problem [22] where the quality of Graphene deposited on either side of copper foil were the objectives of the stochastic MOO problem.

1.2 Bayesian optimal design for inferring statistics

In some problems, dealing with expensive black-box codes, the goal is to sample regions of design space to estimate the statistics of the code output(s). These statistics are functions of the code output(s) or experimental observation(s) like the statistical expectation, variance of the output, probability of taking values lower than a fixed threshold, etc. These statistics can also include optimization scenarios. This question is answered via a combination of data-driven modeling and quantification of plausible gain in information possessed by a hypothetical design or experimental condition.

Towards this goal, we derive an optimal acquisition strategy, named EKLD, for obtaining the most informative simulations or experiments if one aims to estimate the value of a function of the expensive objective. This framework guides the designer towards evaluating the objective function sequentially to acquire information about any arbitrary quantity of interest to the engineer or scientist. We verify and validate the proposed methodology by applying it on synthetic test problems of different dimensionality and multiple number of modes. Comparisons with two classic state-of-the-art methods show promising results for the EKLD. We then demonstrate our approach on a real-world wire-manufacturing problem to estimate the statistics of the total frictional work done in the process.

2. OPTIMIZING SINGLE-OBJECTIVE BLACK-BOX FUNCTIONS UNDER UNCERTAINTY

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.

The following text is taken from the publication titled: *Extending Expected Improvement for High-Dimensional Stochastic Optimization of Expensive Black-Box Functions.*

2.1 Introduction

The majority of stochastic optimization techniques are based on Monte Carlo sampling, e.g., stochastic gradient descent [23], sample average approximation [24], and random search [25]. Unfortunately, the advantages offered by these techniques can be best leveraged [26] 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. [10, 27–32]. 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) [9], resulting in a version of BGO known as efficient global optimization (EGO), and the probability of improvement (PoI) [10]. The operations research literature has developed the concept of knowledge gradient (KG) [33–36], which is essentially a generalization of the EI, and the machine learning community has been experimenting with the expected information gain (EIG) [18, 37, 38].

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. 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 [39]. 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 chapter is as follows. We start Sec. 2.2 by providing the mathematical definition of the stochastic optimization problem that is being studied. In Sec. 2.2.1, 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. 2.2.2, we show how the epistemic uncertainty on the location of the optimum and the optimal value can be quantified. In Sec. 2.2.3, we derive our extension to EI suitable for stochastic optimization. The numerical results are presented in Sec. 2.3. In particular, in Sec. 2.3.1 and 2.3.2, 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. 2.3.3, the methodology is used to solve the oil-well placement problem with uncertainties in soil permeability and the oil price timeseries. The conclusions are presented in Sec. 2.4.

2.2 Methodology

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

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \mathbb{E}_{\boldsymbol{\xi}} \left[V(\mathbf{x}; \boldsymbol{\xi}) \right], \tag{2.1}$$

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.,

$$\mathbb{E}_{\boldsymbol{\xi}} \left[V(\mathbf{x}; \boldsymbol{\xi}) \right] = \int V(\mathbf{x}; \boldsymbol{\xi}) p(\boldsymbol{\xi}) d\boldsymbol{\xi}, \tag{2.2}$$

where $p(\boldsymbol{\xi})$ is the probability density function (PDF) of $\boldsymbol{\xi}$. We will develop a methodology for the solution of Eq. (2.1) 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,

$$\mathbf{x}_{1:n} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}, \tag{2.3}$$

and the corresponding observed noisy objective evaluations,

$$\mathbf{y}_{1:n} = \{y_1, \cdots, y_n\}. \tag{2.4}$$

What can be said about the solution of Eq. (2.1) using only the observed data $\mathbf{x}_{1:n}$ and $\mathbf{y}_{1:n}$? In the language of probability theory [40], 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. 2.2.1 and Sec. 2.2.2.

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 [41, 42]. As in standard BGO, we will rely on a sub-optimal 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. 2.2.3.

2.2.1 Gaussian process regression

Gaussian process regression [43] is the Bayesian interpretation of classical Kriging [44, 45]. 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 $\mathbf{y}_{1:n}$.

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:

$$p(f(\cdot)|\boldsymbol{\psi}) = \mathrm{GP}(f(\cdot)|m(\cdot;\boldsymbol{\psi}), k(\cdot,\cdot;\boldsymbol{\psi})), \tag{2.5}$$

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:

$$k(\mathbf{x},\mathbf{x}';\boldsymbol{\psi}) = s^2 \exp\left\{-\frac{1}{2}\sum_{i=1}^{d}\frac{(x_i - x_i{}')^2}{\ell_i^2}\right\}, \tag{2.6}$$

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:

$$p(\boldsymbol{\psi}) = p(s)\prod_{i=1}^{d} p(\ell_i), \tag{2.7}$$

where

$$p(s) \propto \frac{1}{s} \tag{2.8}$$

is the Jeffreys' prior [46], and

$$p(\ell_i) \propto \frac{1}{1+\ell_i^2} \tag{2.9}$$

is a log-logistic prior [47].

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:

$$p(\mathbf{y}_{1:n}|\mathbf{x}_{1:n},\boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{y}_{1:n}\middle|0,\mathbf{K}_n(\boldsymbol{\psi})+\sigma^2\mathbf{I}_n\right), \tag{2.10}$$

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,

$$\mathbf{K}_n(\boldsymbol{\psi}) = \begin{pmatrix} k(\mathbf{x}_1,\mathbf{x}_1;\boldsymbol{\psi}) & \dots & k(\mathbf{x}_1,\mathbf{x}_n;\boldsymbol{\psi}) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_n,\mathbf{x}_1;\boldsymbol{\psi}) & \dots & k(\mathbf{x}_n,\mathbf{x}_n;\boldsymbol{\psi}) \end{pmatrix}, \tag{2.11}$$

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:

$$p(\sigma) \propto \frac{1}{\sigma}. \tag{2.12}$$

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,

$$p(f(\cdot)|\mathbf{x}_{1:n},\mathbf{y}_{1:n},\boldsymbol{\theta}) = \mathrm{GP}\left(f(\cdot)|m_n(\mathbf{x};\boldsymbol{\theta}),k_n(\mathbf{x},\mathbf{x}';\boldsymbol{\theta})\right), \tag{2.13}$$

albeit with posterior mean and covariance functions,

$$m_n(\mathbf{x};\boldsymbol{\theta}) = \left(\mathbf{k}_n(\mathbf{x};\boldsymbol{\psi})\right)^T\left(\mathbf{K}_n(\boldsymbol{\psi})+\sigma^2\mathbf{I}_n\right)^{-1}\mathbf{y}_{1:n}, \tag{2.14}$$

and

$$\begin{aligned} k_n(\mathbf{x},\mathbf{x}';\boldsymbol{\theta}) &= k(\mathbf{x},\mathbf{x}';\boldsymbol{\psi}) \\ &\quad - \left(\mathbf{k}_n(\mathbf{x};\boldsymbol{\psi})\right)^T \left(\mathbf{K}_n(\boldsymbol{\psi}) + \sigma^2\mathbf{I}_N\right)^{-1} \mathbf{k}_n(\mathbf{x}';\boldsymbol{\psi}) \end{aligned} \tag{2.15}$$

respectively, where $\mathbf{k}_n(\mathbf{x};\boldsymbol{\psi}) = \left(k(\mathbf{x},\mathbf{x}_1;\boldsymbol{\psi}),\ldots,k(\mathbf{x},\mathbf{x}_n;\boldsymbol{\psi})\right)^T$, 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. (3.13) the *point-predictive probability density* conditioned on the hyperparameters $\boldsymbol{\theta}$:

$$p(f(\mathbf{x})|\mathbf{x}_{1:n},\mathbf{y}_{1:n},\boldsymbol{\theta}) = \mathcal{N}\left(f(\mathbf{x})\middle|m_n(\mathbf{x};\boldsymbol{\theta}),\sigma_n^2(\mathbf{x};\boldsymbol{\theta})\right), \tag{2.16}$$

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:

$$p(\boldsymbol{\theta}|\mathbf{x}_{1:n},\mathbf{y}_{1:n}) \propto p(\mathbf{y}_{1:n}|\mathbf{x}_{1:n},\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{2.17}$$

where $p(\boldsymbol{\theta}) = p(\boldsymbol{\psi})p(\sigma)$. Unfortunately, Eq. (2.17) 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) [48]. Formally, we write:

$$p(\boldsymbol{\theta}|\mathbf{x}_{1:n},\mathbf{y}_{1:n}) \approx \frac{1}{N}\sum_{i=1}^{N}\delta(\boldsymbol{\theta}-\boldsymbol{\theta}_i), \tag{2.18}$$

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. (2.17); 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.

2.2.2 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. (2.1) 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

$$p(Q|\mathbf{x}_{1:n}, \mathbf{y}_{1:n}) = (\int (\int \delta\left(Q - Q[f(\cdot)]\right) p(f(\cdot)|\mathbf{x}_{1:n}, \mathbf{y}_{1:n}, \boldsymbol{\theta}) \\ df(\cdot)) p(\boldsymbol{\theta}|\mathbf{x}_{1:n}, \mathbf{y}_{1:n}) d\boldsymbol{\theta}), \tag{2.19}$$

By sampling M functions, $f_1(\cdot), \ldots, f_M(\cdot)$ from Eq. (3.13) and using Eq. (2.18), we get the particle approximation:

$$p(Q|\mathbf{x}_{1:n}, \mathbf{y}_{1:n}) \approx \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \delta\left(Q - Q[f_i(\cdot)]\right). \tag{2.20}$$

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, [49–53]. In our numerical examples we use $M = 100$.

2.2.3 Extended expected improvement function

The classic definition of expected improvement, see [9], relies on the observed minimum $\tilde{y}_n = \min_{1 \leqslant i \leqslant n} y_i$. 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 $\boldsymbol{\theta}$:

$$\tilde{m}_n(\boldsymbol{\theta}) = \min_{1 \leqslant i \leqslant n} m_n(\mathbf{x}_i; \boldsymbol{\theta}), \tag{2.21}$$

where $m_n(\mathbf{x}; \boldsymbol{\theta})$ is the posterior mean of Eq. (5.12). Using $\tilde{m}_n(\boldsymbol{\theta})$, the improvement we would get if we observed $f(\mathbf{x})$ at design point $\mathbf{x}$ is:

$$I(\mathbf{x}, f(\mathbf{x}); \boldsymbol{\theta}) = \max\{0, \tilde{m}_n(\boldsymbol{\theta}) - f(\mathbf{x})\}. \tag{2.22}$$

This is identical to the improvement function formulated in Sequential kriging optimization (SKO) [54]. 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. 3.2.1,

$$\begin{aligned} \mathrm{EEI}_n(\mathbf{x}) = (\int \int I(\mathbf{x}, f(\mathbf{x}); \boldsymbol{\theta}) p(f(\mathbf{x}) | \mathbf{x}_{1:n}, \mathbf{y}_{1:n}, \boldsymbol{\theta}) df(\mathbf{x}) \\ p(\boldsymbol{\theta} | \mathbf{x}_{1:n}, \mathbf{y}_{1:n}) d\boldsymbol{\theta}), \end{aligned} \tag{2.23}$$

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. (5.15) and Eq. (2.17), 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. (2.18). The end result is:

$$\begin{aligned} \mathrm{EEI}_n(\mathbf{x}) \approx \frac{1}{N} \sum_{i=1}^{N} [\sigma_n(\mathbf{x}; \boldsymbol{\theta}_i) \phi \left(\frac{\tilde{m}_n(\boldsymbol{\theta}_i) - m_n(\mathbf{x}; \boldsymbol{\theta}_i)}{\sigma_n(\mathbf{x}; \boldsymbol{\theta}_i)} \right) \\ + (\tilde{m}_n(\boldsymbol{\theta}_i) - m_n(\mathbf{x}; \boldsymbol{\theta}_i)) \Phi \left(\frac{\tilde{m}_n(\boldsymbol{\theta}_i) - m_n(\mathbf{x}; \boldsymbol{\theta}_i)}{\sigma_n(\mathbf{x}; \boldsymbol{\theta}_i)} \right)]. \end{aligned} \tag{2.24}$$

Algorithm 3 demonstrates how the derived information acquisition criterion can be used in a modified version of BGO to obtain an approximation to Eq. (2.1). Note that

instead of attempting to maximize $\text{EEI}_n(\mathbf{x})$ 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 [55].

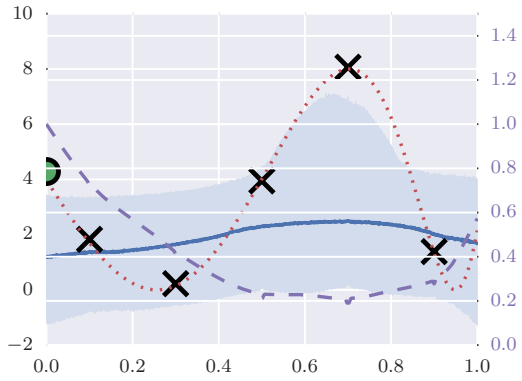
Algorithm 1 The Bayesian global optimization algorithm with the Extended expected improvement function

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.

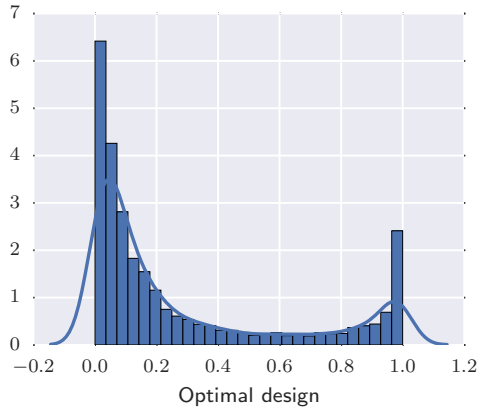
$s \leftarrow 0$.
while $s < S$ **do**
  Construct the particle approximation to the posterior of $\boldsymbol{\theta}$, Eq. (2.18).
  Generate a set of candidate test points $\hat{\mathbf{x}}_{1:n_d}$, e.g., via a latin hypercube design [55].
  Compute EEI on all of the candidate points $\hat{\mathbf{x}}_{1:n_d}$ using Eq. (2.24).
  Find the candidate point $\hat{\mathbf{x}}_j$ that exhibits the maximum EEI.
  if $\text{EEI}_{n+s}(\mathbf{x}_j) < \epsilon$ **then**
    Break.
  end if
  Evaluate the objective at $\hat{\mathbf{x}}_j$ measuring $\hat{y}$.
  $\mathbf{x}_{1:n+s+1} \leftarrow \mathbf{x}_{1:n+s} \cup \{\hat{\mathbf{x}}_j\}$.
  $\mathbf{y}_{1:n+s+1} \leftarrow \mathbf{y}_{1:n+s} \cup \{\hat{y}\}$.
  $s \leftarrow s + 1$.
end while

2.3 Numerical Results

We validate our approach, see Sec. 2.3.1 and 2.3.2, 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. 2.3.3, 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



(a)



(b)

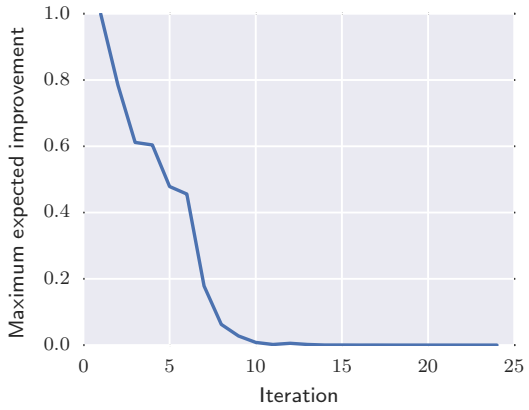
Figure 2.1. One-dimensional synthetic example ($s(x) = 0.1, n = 5$). Subfigure (a) depicts our initial state of knowledge about the true expected objective (dotted red line) conditioned on $n = 5$ noisy observations (black crosses). Subfigure (b), shows a histogram of the predictive distribution of the optimal design x^*.

paragraphs and they are the same for all examples. The only thing that we vary is the initial number of observations n.

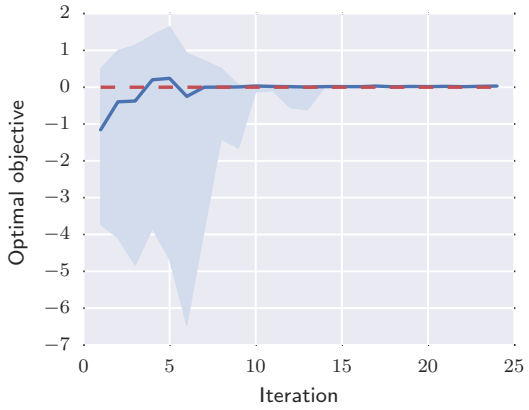
2.3.1 One-dimensional synthetic example

Consider the one-dimensional synthetic objective:

$$V(x, \xi) = 4\left(1 - \sin\left(6x + 8e^{6x-7}\right)\right) + s(x)\xi, \tag{2.25}$$



(a)

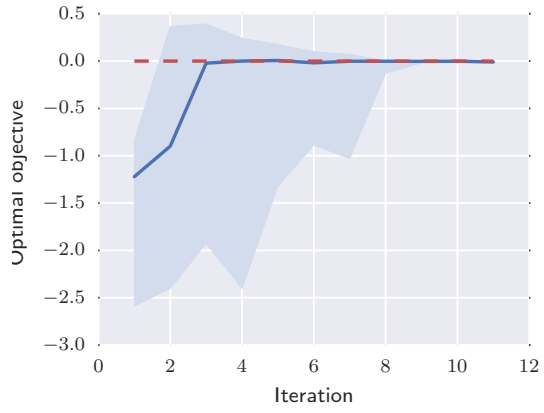


(b)

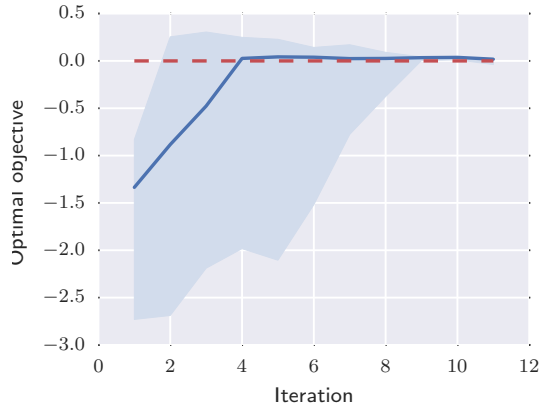
Figure 2.2. One-dimensional synthetic example ($s(x) = 0.1, n = 5$). The dashed red line in Subfigure (b) marks the real optimal value.

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. 2.1 (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. 2.1 (a), the solid blue line is the median of the predictive distribution of the GP and the



(a)



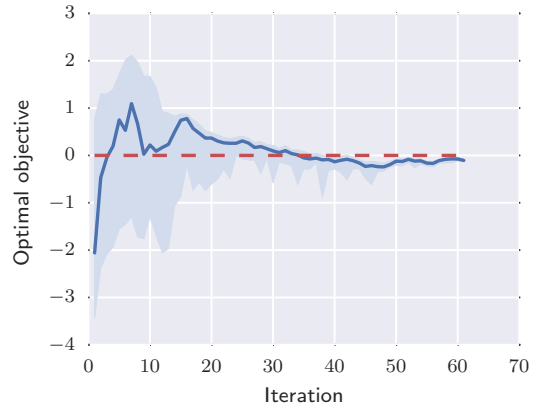
(b)

Figure 2.3. One-dimensional synthetic example ($n = 10$).

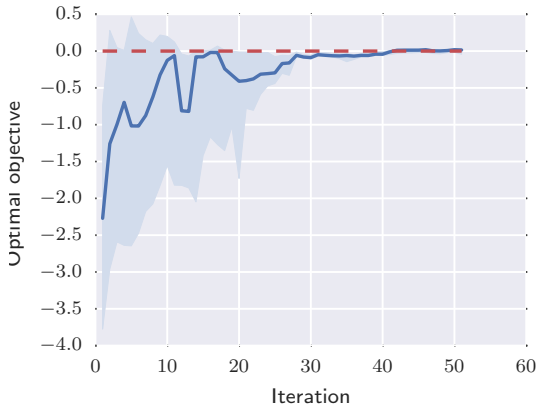
shaded blue area corresponds to a 95% prediction interval. Fig. 2.2 (a) and (b) depict the maximum EEI and the evolution of the 95% predictive bounds for the optimal objective value (PBOO), respectively, as a function of the iteration number. Fig. 2.3 (a) and (b) show the evolution of the PBOO for ($s(x) = 0.01$) and ($s(x) = 0.1$) respectively and Fig. 2.3 (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)

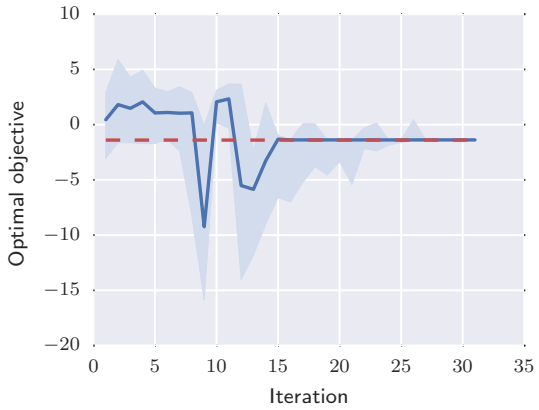
Figure 2.4. One-dimensional synthetic example ($n = 10$).

2.3.2 Two-dimensional synthetic example

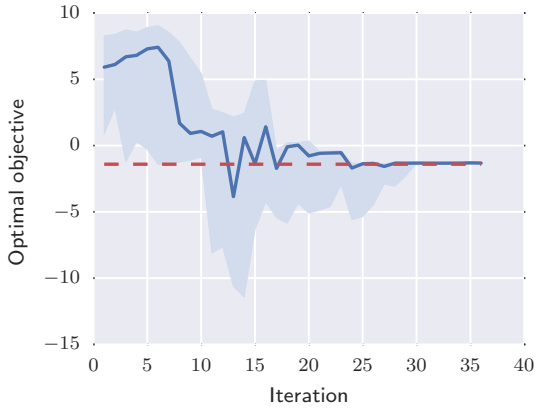
Consider the two-dimensional function [56]:

$$V(\mathbf{x};\xi) = 2 + \frac{(x_2 - x_1^2)^2}{100} + (1 - x_1)^2 + 2(2 - x_2)^2 \\ +7\sin(0.5x_2)\sin(0.7x_1x_2) + s(\mathbf{x})\xi, \tag{2.26}$$

for $\mathbf{x} \in [0,5]^2$, ξ a standard normal, and $s(\mathbf{x}) = 0.01, 0.1, 1$, or the heteroscedastic $s(\mathbf{x}) = (\frac{x_2 - x1}{3})^2$. 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.



(a)

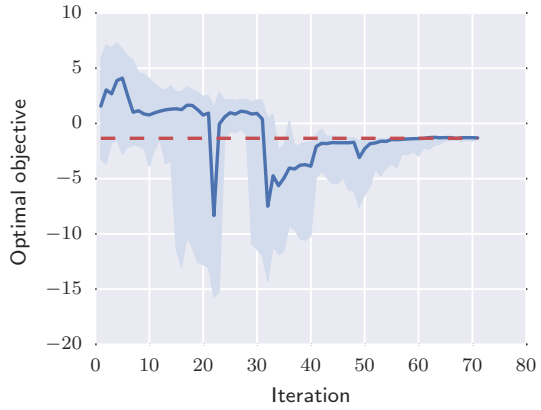


(b)

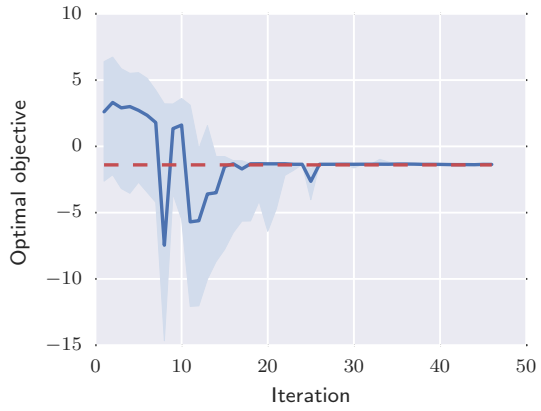
Figure 2.5. Two-dimensional synthetic example ($n = 20$).

Fig. 3.3 (a) and (b) show the PBOO for ($s(\mathbf{x}) = 0.01$) and ($s(\mathbf{x}) = 0.1$) and Fig. 2.6 (a) and (b) show the PBOO for ($s(\mathbf{x}) = 1$) and ($s(\mathbf{x}) = \left(\frac{x_2 - x1}{3}\right)^2$), respectively, as a function 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 [41, Ch. 5.2]. The existence of this effect means, however, that one needs to be very careful in choosing the stopping criterion.



(a)



(b)

Figure 2.6. Two-dimensional synthetic example ($n = 20$).

2.3.3 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 location 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,

$$\mathbf{C}(\mathbf{x}_s; \boldsymbol{\xi}_c) = e^{g(\mathbf{x}_s; \boldsymbol{\xi}_c)} c(\mathbf{x}_s) \mathbf{I}_3, \tag{2.27}$$

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 [57]), $g(\mathbf{x}_s; \boldsymbol{\xi}_c)$ 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 [58], 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. 2.7.

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 [59]. 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 $s_{\mathrm{WC}} = 0.2$, the irreducible oil saturation is $s_{\mathrm{OR}} = 0.2$, the water viscosity is set

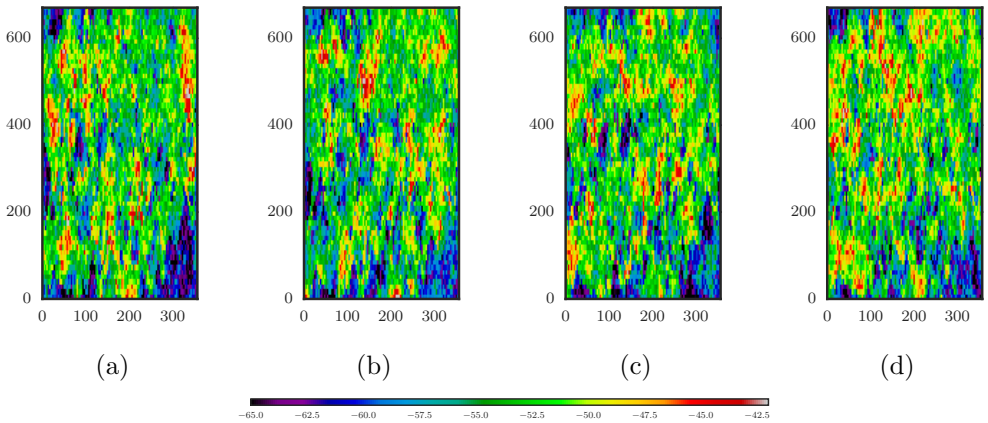


Figure 2.7. OWPP: Samples from the stochastic permeability model (in logarithmic scale) defined in Eq. (2.27).

to $\mu_w = 3 \times 10^{-4}$ Pa $\cdot$ s, the oil viscosity to $\mu_o = 3 \times 10^{-3}$ Pa $\cdot$ 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.

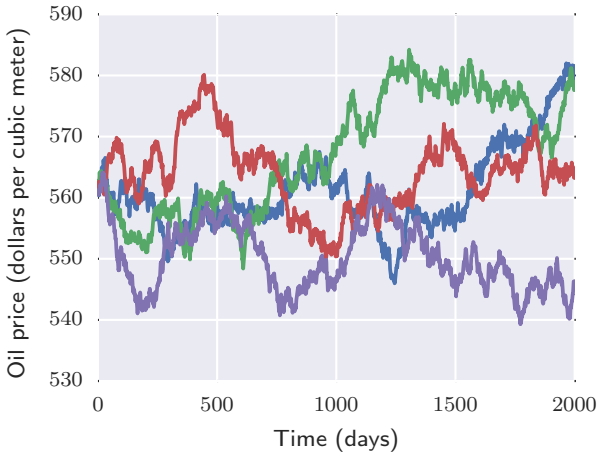


Figure 2.8. OWPP: Samples from the stochastic oil price model.

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:

$$W_{t+1} = W_t + \mu + \alpha\xi_{o,t}, \tag{2.28}$$

where the $\mu = 10^{-8}$, $\alpha = 10^{-3}$, and $\xi_{o,t}$ are independent standard normal random variables. Fig. 2.8 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

$S^{-}_{w,t} = \$0.30/\mathrm{m}^3$. Assuming a discount rate $r = 10\%$ and risk neutrality, our objective is to maximize the NPV of the investment. Equivalently, we wish to minimize:

$$V(\mathbf{x};\boldsymbol{\xi}) = 10^{-6} \sum_{t=1}^{2{,}000\,\text{days}} \left[S_{w,t} q_(t;\mathbf{x},\boldsymbol{\xi}_c) - S_{o,t}(\boldsymbol{\xi}_o) q_o(t;\mathbf{x},\boldsymbol{\xi}_c)\right] (1+r)^{-t/365\,\text{days}}, \tag{2.29}$$

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

Fig. 2.9 (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.

2.4 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,

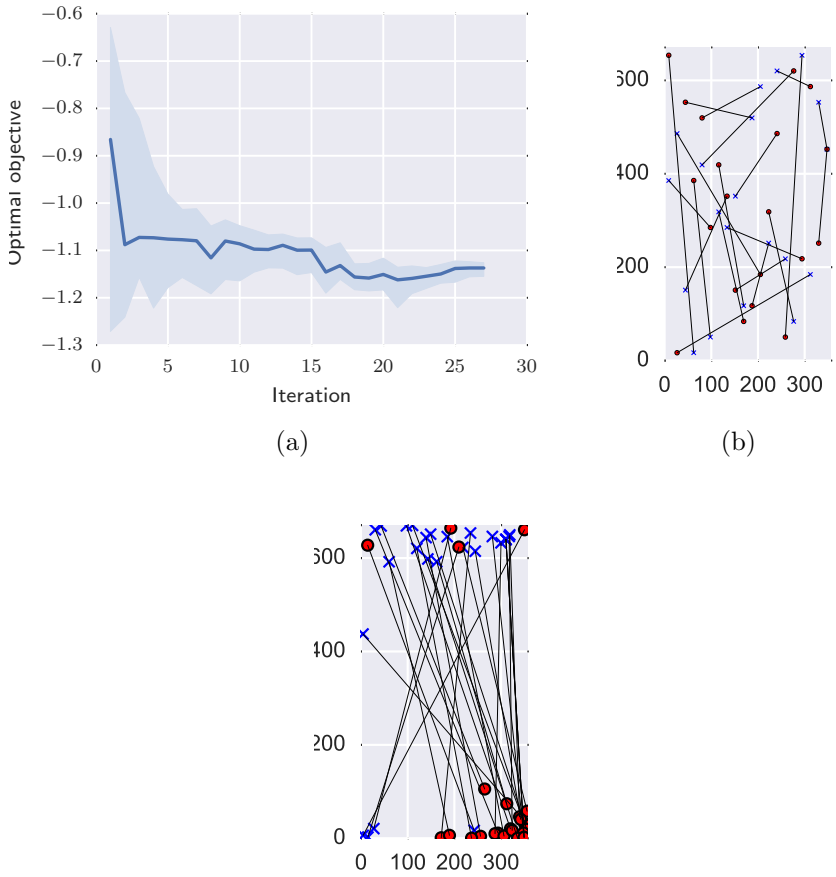


Figure 2.9. OWPP ($n = 20$).

the most important direction would be to construct surrogates that explicitly model heteroscedasticity and use them to extend the present methodology to robust stochastic optimization and, subsequently, to multi-objective stochastic optimization.

3. OPTIMIZING MULTI-OBJECTIVE BLACK-BOX FUNCTIONS UNDER UNCERTAINTY

Design optimization of engineering systems with multiple competing objectives is a painstakingly tedious process especially when the objective functions are expensive-to-evaluate computer codes with parametric uncertainties. The effectiveness of the state-of-the-art techniques is greatly diminished because they require a large number of objective evaluations, which makes them impractical for problems of the above kind. Bayesian global optimization (BGO), has managed to deal with these challenges in solving single-objective optimization problems and has recently been extended to multi-objective optimization (MOO). BGO models the objectives via probabilistic surrogates and uses the epistemic uncertainty to define an information acquisition function (IAF) that quantifies the merit of evaluating the objective at new designs. This iterative data acquisition process continues until a stopping criterion is met. The most commonly used IAF for MOO is the expected improvement over the dominated hypervolume (EIHV) which in its original form is unable to deal with parametric uncertainties or measurement noise. In this chapter, we provide a systematic reformulation of EIHV to deal with stochastic MOO problems. The primary contribution of this chapter lies in being able to filter out the noise and reformulate the EIHV without having to observe or estimate the stochastic parameters. An addendum of the probabilistic nature of our methodology is that it enables us to characterize our confidence about the predicted Pareto front. We verify and validate the proposed methodology by applying it to synthetic test problems with known solutions. We demonstrate our approach on an industrial problem of die pass design for a steel wire drawing process.

The following text is taken from the publication titled: *Stochastic multi-objective optimization on a budget: Application to multi-pass wire drawing with quantified uncertainties.*

3.1 Introduction

The goal of this paper is to derive a sequential information acquisition methodology that aims at efficiently discovering the Pareto set of a stochastic MOO problem. Stochastic MOOs are characterized by uncertain objective measurements, i.e., for a fixed design, repeated measurements of the objectives may vary. When the objectives are the outcomes of an experiment, this randomness may be due to manufacturing imperfections, operational uncertainties, wear and tear of the specimen, sensor malfunction, etc. When the objectives depend on a simulation model, then this randomness may be induced by uncertainty in the model parameters, e.g., boundary/initial conditions, parameters of constitutive relations, or artifact geometries. In the latter case, the designer chooses probability distributions for all uncertain parameters in an effort to accurately describe their state of knowledge about the artifact.

MOO techniques based on evolutionary algorithms [60], e.g., the strength Pareto evolutionary algorithm [61], the non-dominated sorting genetic algorithm II (NSGA-II) [62], require a significant number of objective evaluations, especially when coupled with a sample average approximation [25] to estimate the stochastic objectives. Other popular techniques like *goal programming* [63, 64] that involve a slight modification of the original MOO objectives face shortcomings [65] like selecting the relative importance of the objectives, or requiring the designer to have prior information about discontinuities in the objective space.

Bayesian global optimization (BGO) [9, 66] is a class of black-box optimization algorithms that can operate under a limited objective evaluation budget. BGO models the objectives using probabilistic surrogates, e.g., Gaussian process regression, and exploits the epistemic uncertainty to select which experiments/simulations to perform.

The latter is typically done by maximizing an information acquisition function (IAF) which quantifies the value of evaluating the objective at a specific design. The choice of the IAF depends on the details of the underlying optimization task. One of the most popular IAFs is the expected improvement (EI) [9, 54, 67–69]. The EI balances the exploration-exploitation trade-off better than other popular IAFs such as the probability of improvement (PI) or the upper confidence bound (UCB) [10]. Keane [70] extended the original version of EI to MOO by deriving the expected improvement over the dominated hypervolume (EIHV). The EIHV evaluates the expected improvement in the volume of the attained set induced by a hypothetical observation at an untried design. [21] derived a closed form representation which made the evaluation of EIHV computationally efficient. Research in EIHV has been gaining momentum over the past few years [71–73], but it has not yet been extended to cover the case of stochastic multi-objective optimization.

In this chapter, we propose an extension to the EIHV suitable for stochastic MOO, which is the main contribution of this paper. We will be referring to the proposed methodology as the *extended EIHV* (EEIHV). Our proposal is a generalization of the extended expected improvement (EEI), which we developed in Chapter 1., to deal with stochastic single-objective optimization. The methodology relies on building probabilistic surrogates of the objectives and uses the EEIHV IAF to quantify the merit of evaluating the expensive stochastic computer code at a new design. Leveraging the work done in [19] allows quantification of the uncertainty about the estimated PF at each stage/iteration.

The above methodology is applied to solve a multi-pass steel wire manufacturing problem under uncertainty. The competing objectives in this problem are the ultimate tensile strength (UTS) and the strain non-uniformity factor (SNUF) of the drawn wire. A finite element (FE) solver (developed at Tata Consulatancy Services (TCS), Pune, India) generates these objectives. The reduction ratios and the die angles at each pass are the design/process variables which have associated uncertainties due to unavoidable manufacturing tolerances as well as die wear during the process.

The outline of the chapter is as follows. At the very beginning, Sec. 3.2 by provides the mathematical definition of the stochastic MOO optimization problem that we are studying. In Sec. 3.2.1, we introduce Gaussian process regression (GPR) which is used to construct the probabilistic surrogates of the map between the design variables and the objectives. In Sec. 3.2.3, we derive our extension to EIHV suitable for stochastic multi-objective optimization. The numerical results are presented in Sec. 3.3. In particular, in Sec. 3.3.2 and 3.3.3, we validate our approach using two synthetic stochastic MOO problems with known analytical expressions, and we experiment with varying levels of stochasticity (to represent noisy measurements). In Sec. 3.4, the methodology is used to solve the wire drawing problem. The conclusions are presented in Sec. 3.5.

3.2 Methodology

Let X denote the set of feasible designs and $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We assume that X is a closed and bounded set of a Euclidean space. We have m stochastic quantities of interest (QoIs) which we represent as Borel-measurable functions $o_i : X \times \Omega \to \mathbb{R}, i = 1, \ldots, m$. Our goal is to find designs $\mathbf{x} \in X$ that maximize the expectations of these QoIs over $\omega \in \Omega$, i.e., we wish to maximize $O_i(\mathbf{x}) := \mathbb{E}[o_i(\mathbf{x}, \omega)] := \int o_i(\mathbf{x}, \omega) d\mathbb{P}(\omega)$. We say that $\mathbf{x} \in X$ *dominates* $\mathbf{x}' \in X$, and write $\mathbf{x} \succcurlyeq \mathbf{x}'$, if and only if

$$O_i(\mathbf{x}) \geqslant O_i(\mathbf{x}'), \forall i = 1, \ldots, m. \tag{3.1}$$

We say that $\mathbf{x}$ *strictly dominates* $\mathbf{x}'$, and write $\mathbf{x} \succ \mathbf{x}'$ if and only if $\mathbf{x} \succcurlyeq \mathbf{x}'$ and there exists $i \in \{1, \ldots, m\}$ such that $\mathbb{E}[o_i(\mathbf{x}, \omega)] > \mathbb{E}[o_i(\mathbf{x}', \omega)]$.

We wish to characterize the *set of optimal designs*, otherwise known as the *Pareto-efficient frontier*, induced by the preference relation '$\succcurlyeq$'. In words, the Pareto-efficient frontier, P_O, is the set of achievable objectives that are not dominated. Since P_O has Lebesgue measure zero, working with it directly is problematic. Instead, we will work

with the *attained set*, A_O, which is defined as the set of achievable objectives that are strictly dominated. P_O is simply part of the boundary of A_O.

We now proceed to the exact mathematical definition of A_O and, subsequently, P_O. At first glance, our definitions may seem unnecessarily complex. The benefit of such a rigorous approach is that it highlights the dependence of these quantities on the objectives $\mathbf{O}$. Explicitly denoting this dependence will help us appreciate the nature of our approximation to the Pareto frontier when $\mathbf{O}$ is replaced by a Gaussian process surrogate.

Select a point $\mathbf{r} = (r_1, \ldots, r_m) \in \mathbb{R}^m$ for which we have $\min_{\mathbf{x}\in X} O_i(\mathbf{x}) \geqslant r_i$. Since X is compact, such a point exists if $O_i(\mathbf{x})$ is continuous. $\mathbf{r}$ is known as the *reference point*. Consider the vector valued function $\mathbf{O} : X \to \mathbb{R}^m$ defined by $\mathbf{O} = (O_1, \ldots, O_m)$. $\mathbf{O}$ just joins all the expected objectives in a vector. The image $\mathbf{O}[X]$ of X under $\mathbf{O}$, defined by

$$\mathbf{O}[X] = \{\mathbf{y} \in \mathbb{R}^m : \exists \mathbf{x} \in X, \mathbf{y} = \mathbf{O}(\mathbf{x})\},$$

is the set of all achievable objectives. We do not know exactly how $\mathbf{O}[X]$ looks like. However, exploiting the definition of the reference point, we see that $\mathbf{O}[X]$ is fully contained in the m-dimensional cone $[\mathbf{r}, \infty] := \times_{i=1}^{m}[r_i, \infty)$, i.e.,

$$\mathbf{O}[X] \subset [\mathbf{r}, \infty).$$

Consider any subset B of $[\mathbf{r}, \infty)$. We define the attained set of B, denoted by $A[B]$, to be the set of points in $[\mathbf{r}, \infty)$ that are dominated by B, i.e.,

$$A[B] := \{\mathbf{y} \in [\mathbf{r}, \infty) : \exists \mathbf{y}' \in B, \mathbf{y}' \geqslant \mathbf{y}\}, \tag{3.2}$$

where $\mathbf{y}' \geqslant \mathbf{y}$ corresponds to element-wise comparison. The attained set of our multi-objective problem is just:

$$A_O := A[\mathbf{O}[X]]. \tag{3.3}$$

Finally, we define the Pareto frontier of B, denoted by $P[B]$, to be the set of points in B that are not dominated by any other point in B, i.e.,

$$P[B] := \{\mathbf{y} \in B : \{\mathbf{y}' \in B : \mathbf{y}' \succcurlyeq \mathbf{y}\} = \varnothing\}. \tag{3.4}$$

But we can get the Pareto frontier of B directly from the boundary of its attained set. Specifically, it is easy to prove that $P[B]$ is the top right boundary of $A[B]$, i.e.,

$$P[B] = \partial A[B] \backslash \cup_{i=1}^{m} \{\mathbf{r} + t(\max_{\mathbf{y} \in B} y_i)\mathbf{e}_i\}, \tag{3.5}$$

where $\mathbf{e}_i$ is the standard basis function of $\mathbb{R}^m$ pertaining to the i-th dimension. The Pareto front of our multi-objective problem is just:

$$P_O := P[\mathbf{O}[X]]. \tag{3.6}$$

Assume that we can choose to measure the QoIs at any design point $\mathbf{x} \in X$ we wish, albeit only a limited number of times n. Such measurements take place as follows. When we request information about $\mathbf{x}$, a latent process samples an *unobserved* $\omega \in \Omega$ according to the probability measure $\mathbb{P}$, and we observe a noisy version of the QoIs $\mathbf{y} = (o_1(\mathbf{x}, \omega), \ldots, o_m(\mathbf{x}, \omega))$. This setup is general enough to account for both simulation-based and experiment-based QoIs.

Assume that we have queried the information source at n design points.

$$\mathbf{x}_{1:n} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in X^n, \tag{3.7}$$

and that we have made the following noisy observations:

$$\mathbf{y}_{1:n} = (\mathbf{y}_1, \ldots, \mathbf{y}_n). \tag{3.8}$$

We address two problems:

1. What is our *state of knowledge* about the true Pareto-efficient frontier P_O given the observations $(\mathbf{x}_{1:n}, \mathbf{y}_{1:n})$?

2. How should we select $\mathbf{x}_{1:n}$ so that we come as close as possible to discovering the true Pareto-efficient frontier P_O?

In the language of probability theory [40], the former problem seeks to characterize the probability (a state of belief) of a design being optimal conditional on the observations. The uncertainty encoded in this probability is epistemic and it is induced by the fact that inference is based on just a small number of observations. We address this problem by leveraging the Bayesian nature of Gaussian process surrogates, see Sec. 3.2.1. Looking for an optimal information acquisition policy that solves the latter problem is a mathematically intractable task since the problem is equivalent to a non-linear stochastic dynamic program [41, 42]. We rely on a myopic/greedy one-step-look-ahead strategy (which is sub-optimal) by extending the definition of the standard EIHV, see Sec. 3.2.3, so that it can cope robustly with noise.

3.2.1 Gaussian process regression

Gaussian process (GP) regression [43] is the Bayesian interpretation of classical Kriging [44, 45]. It is a powerful non-linear and non-parametric regression technique that is able to quantify the epistemic uncertainty induced by limited data. We use GP regression to model our state of knowledge about the objectives, i.e., $O_i(\mathbf{x}) = \mathbb{E}[o_i(\mathbf{x}, \omega)], i = 1, \ldots, m$, as induced by a set of n observations $(\mathbf{x}_{1:n}, \mathbf{y}_{1:n})$. The methodology applies to each $i = 1, \ldots, m$ independently. For simplicity, we will write $f(\mathbf{x})$ for $O_i(\mathbf{x})$ and $y_{1:n}$ for $y_{i,1:n} = (y_{i1}, \ldots, y_{in})$.

Expressing prior beliefs

Let $(\Omega^e, \mathcal{F}^e, \mathbb{P}^e)$ be the probability space corresponding to our epistemic uncertainty. Note that this is different from $(\Omega, \mathcal{F}, \mathbb{P})$ which is associated with the problem

uncertainty. A GP $f^e(x, \omega^e)$ is a $(\Omega^e, \mathcal{F}^e, \mathbb{P}^e)$-random field indexed by $\mathbf{x} \in X$ with Gaussian finite dimensional distributions. That is, for any $\mathbf{x}_{1:n} \in X^n$ the random vector $f^e_{1:n} := (f^e(\mathbf{x}_1, \omega^e), \ldots, f^e(\mathbf{x}_n, \omega^e))$ follows a multivariate Gaussian. The interpretation is as follows. Nature has chosen a reality $\omega^e \in \Omega^e$, i.e., $f(\cdot) \equiv f^e(\cdot, \omega^e)$, that we cannot directly observe. $(\Omega^e, \mathcal{F}^e, \mathbb{P}^e)$ models our prior state of knowledge about this reality, in the sense that for all $B \in \mathcal{F}^e$ the probability that we give to $\omega^e \in B$ is $\mathbb{P}^e[B] = \int_B d\mathbb{P}^e(\omega^e)$.

A GP is characterized by a mean and a covariance function. Without loss of generality, we may assume that the mean function is zero, since the covariance can always be modified to include a non-zero mean trend. Mathematically, we write:

$$f^e|\theta^e \sim \mathrm{GP}(0, k), \tag{3.9}$$

where $k : X \times X \times \Theta^e \to \mathbb{R}$ is a covariance function parameterized by the epistemic random variable $\theta^e : \Omega^e \to \Theta^e$. According to the definition of the GP, our a priori beliefs about the values $f^e_{1:n}$ are captured by:

$$f^e_{1:n}|\mathbf{x}_{1:n}, \theta^e \sim \mathcal{N}(0, k(\mathbf{x}_{1:n}, \theta^e)), \tag{3.10}$$

where $\mathcal{N}(\lambda, \Sigma)$ denotes the multivariate Gaussian distribution with mean λ and covariance matrix Σ, for all $\mathbf{x}'_{1:n'} \in X^{n'}$ we define $k(\mathbf{x}_{1:n}, \mathbf{x}'_{1:n'}, \theta^e)$ to be the $n \times n'$ matrix with (i, j) element $k(\mathbf{x}_i, \mathbf{x}_j, \theta^e)$, and $k(\mathbf{x}_{1:n}, \theta^e) := k(\mathbf{x}_{1:n}, \mathbf{x}_{1:n}, \theta^e)$ is the covariance matrix. In our numerical examples, we use the Matern($\nu = \frac{3}{2}$) [43] covariance:

$$k(\mathbf{x}, \mathbf{x}', \theta^e) = s^2 \left(\exp\left\{ -\sqrt{3 \sum_{j=1}^{d} \frac{(x_j - x_j{}')^2}{\ell_j^2}} \right\} \right) \left(1 + \sqrt{3 \sum_{j=1}^{d} \frac{(x_j - x_j{}')^2}{\ell_j^2}} \right), \tag{3.11}$$

where d is the dimensionality of the design space, $s > 0$ and $\ell_j > 0$ can be interpreted as the signal strength of the response and the lengthscale along input dimension j, respectively, and $\theta^e = (s, \ell_1, \ldots, \ell_d) \in \mathbb{R}^d_+$.

Modeling the measurement process

In general, the noise that contaminates the measurement y is heteroscedastic, i.e., input-dependent. However, we approximate this noise as Gaussian with a fixed, but unknown, variance ν^2. Despite this fact, we observe numerically that the GP can still estimate the optimization objectives, i.e., expectation of y, when the noise to signal ratio is not too big. The *likelihood* of the model is:

$$p(y_{1:n}|\mathbf{x}_{1:n}, \theta^e) = \mathcal{N}\left(y_{1:n}|0, k(\mathbf{x}_{1:n}, \theta^e) + \nu^2 I_n\right), \tag{3.12}$$

where $I_n \in \mathbb{R}^{n\times n}$ is the identity matrix, $k(\mathbf{x}_{1:n}, \theta^e)$ is as in Eq. (3.10), and, for notational convenience, we have re-defined $\theta^e \leftarrow (\theta^e, \nu)$.

Posterior state of knowledge about the objectives

Bayes rule combines our prior beliefs with the data and yields a posterior probability measure on the space of meta-models. Conditioned on the hyperparameters θ^e, this measure is also a GP,

$$f^e|\mathbf{x}_{1:n}, y_{1:n}, \theta^e \sim \mathrm{GP}\left(\mu_n, k_n\right), \tag{3.13}$$

where the posterior mean and covariance functions are

$$\mu_n(\mathbf{x}; \theta^e) = k_n(\mathbf{x}, \mathbf{x}_{1:n}, \theta^e)\left[k(\mathbf{x}_{1:n}, \theta^e) + \nu^2 I_n\right]^{-1} y_{1:n}, \tag{3.14}$$

and

$$\begin{aligned} k_n(\mathbf{x}, \mathbf{x}', \theta^e) \quad = \quad & k(\mathbf{x}, \mathbf{x}', \theta^e) \\ & -k_n(\mathbf{x}, \mathbf{x}_{1:n}, \theta^e)\left[k(\mathbf{x}_{1:n}, \theta^e) + \sigma^2 I_n\right]^{-1} k_n(\mathbf{x}_{1:n}, \mathbf{x}, \theta^e) \end{aligned} \tag{3.15}$$

respectively. Restricting our attention to a specific design point $\mathbf{x}$, we can derive from Eq. (3.13) the *point-predictive* PDF conditioned on the hyperparameters θ^e:

$$f^e(\mathbf{x})|\mathbf{x}_{1:n}, y_{1:n}, \theta^e \sim \mathcal{N}\left(\mu_n(\mathbf{x};\theta^e), \sigma_n^2(\mathbf{x};\theta^e)\right), \tag{3.16}$$

where *predictive variance* is $\sigma_n^2(\mathbf{x};\theta^e) = k_n(\mathbf{x},\mathbf{x};\theta^e)$.

The hyper-parameters of the covariance function are estimated by maximizing the likelihood $p(y_{1:n}|\mathbf{x}_{1:n},\theta^e)$ with respect to θ^e. To avoid numerical instabilities, one typically works with the logarithm of the likelihood:

$$\begin{aligned}\mathcal{L}(\theta^e) = &-\frac{1}{2}{y_{1:n}}^T\left[k(\mathbf{x}_{1:n},\theta^e)+\nu^2 I_n\right]^{-1} y_{1:n}\\ &-\frac{1}{2}\log\det\left[k(\mathbf{x}_{1:n},\theta^e)+\nu^2 I_n\right]-\frac{n}{2}\log 2\pi.\end{aligned} \tag{3.17}$$

This maximization problem is solved using the BFGS algorithm [74]. To account for the positivity constraints we simply optimize with respect to the logarithms of the hyperparameters. The solution of this optimization problem, denoted by $\hat{\theta}^e$, is known as the maximum likelihood estimate (MLE) of θ^e. For notational convenience, in what follows we are not going to be explicitly indicating the dependence of μ_n and k_n on θ^e. Instead it will be understood that $\mu_n(\mathbf{x}) \equiv \mu_n(\mathbf{x},\hat{\theta}^e)$, $k_n(\mathbf{x},\mathbf{x}') \equiv k_n(\mathbf{x},\mathbf{x}',\hat{\theta}^e)$, and $\sigma_n(\mathbf{x}) \equiv \sigma_n(\mathbf{x},\hat{\theta}^e)$.

3.2.2 Characterization of the Pareto-efficient frontier using limited data

What is our state of knowledge about the true Pareto-efficient frontier P_O given $n \leqslant N$ observations $(\mathbf{x}_{1:n}, \mathbf{y}_{1:n})$? Let $\mathbf{f}^e = (f_1^e, \ldots, f_m^e)$ be the GPs representing our state of knowledge about each one of the m objectives. Our state of knowledge about the relation '$\succcurlyeq$' is now captured by the *random relation* '$\succcurlyeq^e$', namely $\mathbf{x} \succcurlyeq^e \mathbf{x}'$ if and only if $\mathbf{f}^e(\mathbf{x}) \geqslant \mathbf{f}^e(\mathbf{x}')$. Our state of knowledge about the attained set A_O of Eq. (3.3) is given by the random set $A[\mathbf{f}^e[X]]$. Similarly, our state of knowledge about the Pareto front P_O of Eq. (3.6) is represented by the random set $P[\mathbf{f}^e[X]]$.

The first step is to derive summary statistics of $A[\mathbf{f}^e[X]]$ that can be used to visualize our epistemic uncertainty about it. Following [19, 75], we achieve this by estimating the Vorob'ev expectation and deviation of the random set $A[\mathbf{f}^e[X]]$. Towards this end, we introduce the *attainment function* and its upper level sets. The attainment function $a_n^e : [\mathbf{r}, \infty) \to [0, 1]$ is defined to be the conditional probability, given $(\mathbf{x}_{1:n}, \mathbf{y}_{1:n})$, that a vector of objectives $\mathbf{y} \in [\mathbf{r}, \infty)$ can be attained, i.e., we define

$$a_n^e(\mathbf{y}) := \mathbb{P}^e \left[\{\omega^e \in \Omega^e : \mathbf{y} \in A[\mathbf{f}_{\omega^e}^e[X]]\} \,|\mathbf{x}_{1:n}, \mathbf{y}_{1:n}\right], \tag{3.18}$$

where $\mathbf{f}_{\omega^e}^e(\cdot) = (f_1^e(\cdot, \omega^e), \ldots, f_m^e(\cdot, \omega^e))$. For $\beta \in [0, 1]$, the upper level sets of the attainment function,

$$Q_{n,\beta}^e := \{\mathbf{y} \in [\mathbf{r}, \infty) : a_n^e(\mathbf{y}) \geqslant \beta\}, \tag{3.19}$$

are known as the β-quantiles of $A[\mathbf{f}^e[X]]$. Intuitively, Q_{n,β^*}^e can be seen as the set of objectives that are considered achievable with probability greater than or equal to β. The conditional *Vorob'ev expectation* [76] of $A[\mathbf{f}^e[X]]$ is defined to be the β^*-quantile Q_{n,β^*}^e for which:

$$\lambda(Q_{n,\beta}^e) \leqslant \mathbb{E}^e \left[\lambda \left(A[\mathbf{f}^e[X]]\right) |\mathbf{x}_{1:n}, \mathbf{y}_{1:n}\right] \leqslant \lambda(Q_{n,\beta^*}^e),\ \forall \beta \in [\beta^*, 1], \tag{3.20}$$

where λ is the Lebesgue measure on $\mathbb{R}^m$. In words, Q_{n,β^*}^e is the β-quantile that has the same Lebesgue measure as the conditional expectation of the Lebesgue measure of the attained set. Intuitively, Q_{n,β^*}^e and its top right boundary are our expectations about the attained set A_O and P_O, respectively, after observing $(\mathbf{x}_{1:n}, \mathbf{y}_{1:n})$.

Now, we are in a position to quantify our uncertainty about P_O. Consider the symmetric difference $Q_{n,\beta^*}^e \triangle A[\mathbf{f}^e[X]]$ between the set Q_{n,β^*}^e and $A[\mathbf{f}^e[X]]$ defined by

$$Q_{n,\beta^*}^e \triangle A[\mathbf{f}^e[X]] := \left(Q_{n,\beta^*}^e \cup A[\mathbf{f}^e[X]]\right) \setminus \left(Q_{n,\beta^*}^e \cap A[\mathbf{f}^e[X]]\right). \tag{3.21}$$

That is, a point $\mathbf{y}$ belongs in $Q^e_{n,\beta^*} \triangle A[\mathbf{f}^e[X]]$ only if it belongs to exactly one of these sets. Such points appear in the top right corner of $[\mathbf{r}, \infty)$ and are candidate points for the Pareto front. Therefore, we quantify our uncertainty about P_O through the *symmetric deviation function* $d^e_n : [\mathbf{r}, \infty) \to [0, 1]$ defined as the conditional probability that a vector of objectives $\mathbf{y} \in [\mathbf{r}, \infty)$ belongs to the symmetric difference $Q^e_{n,\beta^*} \triangle A[\mathbf{f}^e[X]]$, i.e.,

$$d^e_n(\mathbf{y}) = \mathbb{P}^e \left[\mathbf{y} \in Q^e_{n,\beta^*} \triangle A[\mathbf{f}^e[X]] | \mathbf{x}_{1:n}, \mathbf{y}_{1:n} \right]. \tag{3.22}$$

Unfortunately, it is not possible to characterize $a^e_n(\mathbf{y})$, Q^e_{n,β^*}, and $d^e_n(\mathbf{y})$ exactly. The difficulty arises from the fact that X may be infinite dimensional. To overcome this obstacle, we use a Monte Carlo (MC) approach. Let $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ be a new probability space associated with the MC approximation uncertainty. Let $\tilde{X}_s : \tilde{\Omega} \to X^{\tilde{n}}$, collectively denoted by $\tilde{X}_{1:S} = (\tilde{X}_1, \ldots, \tilde{X}_S)$, be independent identically distributed (iid) random variables in $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ with values in $X^{\tilde{n}}$. For each one, we have $\tilde{X}_s := \tilde{\mathbf{x}}_{1,1:\tilde{n}} := (\tilde{\mathbf{x}}_{s,1}, \ldots, \tilde{\mathbf{x}}_{s,\tilde{n}})$ The specific distribution of these variables is not important as soon they cover X. For convergence, it suffices to make all the $\tilde{\mathbf{x}}_{s,i}, s = 1, \ldots, S, i = 1, \ldots, \tilde{n}$ iid with a support that covers X. In our numerical examples, we take all these random variables to be independently uniform. Conditional on each $\tilde{X}_s$, define the epistemic random variable $\tilde{F}^e_s \in \mathbb{R}^{m\tilde{n}}$ associated with the values of the objectives on $\tilde{X}_s$. That is, $\tilde{F}^e_s := \tilde{\mathbf{f}}^e_{s,1:m,1:\tilde{n}} := (\tilde{f}^e_{s,1,1:\tilde{n}}, \ldots, \tilde{f}^e_{s,m,1:\tilde{n}})$, with $\tilde{f}_{s,i,1:\tilde{n}} := (f^e_i(\tilde{\mathbf{x}}_{s,1}), \ldots, f^e_i(\tilde{\mathbf{x}}_{s,\tilde{n}})) \in \mathbb{R}^{\tilde{n}}$. Note that, since we constructed each one of the GPs representing the objectives independently, we have that $\tilde{f}^e_{s,i,1:\tilde{n}}, s = 1, \ldots, S, i = 1, \ldots, m$ are independent. Making use of the posterior GP representing our state of knowledge about $f^e_i(\mathbf{x})$, see Eq. (3.13), we get that, conditional on $\tilde{\mathbf{x}}_{1:\tilde{n}}$ and $(\mathbf{x}_{1:n}, y_{i,1:n})$, $\tilde{f}^e_{s,i,1:\tilde{n}}$ is normally distributed:

$$\tilde{f}^e_{s,i,1:\tilde{n}} | \tilde{\mathbf{x}}_{s,1:\tilde{n}}, \mathbf{x}_{1:n}, y_{i,1:n} \sim \mathcal{N}\left(\mu_{i,n}(\tilde{\mathbf{x}}_{s,1:\tilde{n}}), k_{i,n}(\tilde{\mathbf{x}}_{s,1:\tilde{n}})\right), \tag{3.23}$$

where $\mu_{i,n}(\mathbf{x})$ and $k_{i,n}(\mathbf{x}, \mathbf{x}')$ are the posterior mean and posterior covariance functions ($\mu_n(\mathbf{x})$ and $k_n(\mathbf{x}, \mathbf{x}')$) of Sec. 3.2.1, respectively, if we make the substitution $y_{1:n} \leftarrow y_{i,1:n}$.

Using $\tilde{F}_s^e$, and the definition in Eq. (3.2) we denote the sampled attained set by $A[\tilde{F}_s^e]$ and the corresponding *sampled Pareto front* by $P[\tilde{F}_s^e]$. Now we can compute the *empirical attainment function* $\tilde{a}^e_{S,\tilde{n},n} : [\mathbf{r}, \infty) \to [0, 1]$:

$$\tilde{a}^e_{S,\tilde{n},n}(\mathbf{y}) = \frac{1}{S} \sum_{s=1}^{S} 1_{A[\tilde{F}_s^e]}(\mathbf{y}), \tag{3.24}$$

where $1_B(\mathbf{y})$ is the characteristic function of the set B. Using $\tilde{a}^e_{S,\tilde{n},n}(\mathbf{y})$ we can obtain estimates of the β-quantiles, say $\tilde{Q}^e_{S,\tilde{n},n,\beta}$. Just like [19], estimates of the β-quantiles can be used within a bisection algorithm to estimate the Vorob'ev expectation $\tilde{Q}^e_{S,\tilde{n},n,\beta*}$. Finally, we compute the *empirical symmetric deviation function*:

$$\tilde{d}^e_{S,\tilde{n},n}(\mathbf{y}) = \frac{1}{S} \sum_{s=1}^{S} 1_{\tilde{Q}_{S,\tilde{n},n,\beta*} \triangle A[\tilde{F}_s^e]}(\mathbf{y}), \tag{3.25}$$

which is an estimate of $d^e_n(\mathbf{y})$. In our numerical examples (in which $m = 2$) we represent $\tilde{a}^e_{S,\tilde{n},n}(\mathbf{y})$ and $\tilde{d}^e_{S,\tilde{n},n}(\mathbf{y})$ on a 64×64 grid defined on $\times_{i=1}^m [r_i, u_i]$, where $\mathbf{u} = (u_1, \ldots, u_m) \in \mathbb{R}^m$ is a point of the design space with $u_i \geqslant \max_{\mathbf{x} \in X} O_i(\mathbf{x}), i = 1, \ldots, m$. For larger number of objectives $m > 3$, more sophisticated techniques must be developed in order to overcome the curse of dimensionality. From the law of large numbers, we have that

$$\lim_{S \to \infty} \lim_{\tilde{n} \to \infty} \tilde{a}^e_{S,\tilde{n},n} = a^e_n, \tag{3.26}$$

$$\lim_{S \to \infty} \lim_{\tilde{n} \to \infty} \tilde{d}^e_{S,\tilde{n},n} = d^e_n. \tag{3.27}$$

We also expect that the attainment function a^e_n will converge to the characteristic function of the attained set A_O as $n \to \infty$ on a set of design points that becomes dense. The exact nature of the latter convergence is beyond the scope of the present work.

3.2.3 Extended expected improvement over dominated hypervolume

Given our current state of knowledge about P_O, how should we select the next observation $\mathbf{x}$? We derive a myopic one-step-look-ahead strategy that attempts to sequentially maximize the expected improvement in the volume of the attained set. Specifically, we define the *extended expected improvement over the dominated hypervolume* (EEIHV) as the expectation of the change in the Lebesgue measure of the attained set conditional on a hypothetical observation. Mathematically, we define for $\mathbf{x} \in X$:

$$\begin{aligned} \mathrm{EEIHV}(\mathbf{x}) \;=\; \mathbb{E}^e\Big[\;& \mathbb{E}^e\big[\lambda\left(A[\mathbf{f}^e[X]]\right)\big|\mathbf{x},\mathbf{y},\mathbf{x}_{1:n},\mathbf{y}_{1:n}\big] \\ & -\mathbb{E}^e\big[\lambda\left(A[\mathbf{f}^e[X]\right)|\mathbf{x}_{1:n},\mathbf{y}_{1:n}\big]\Big|\mathbf{x},\mathbf{x}_{1:n},\mathbf{y}_{1:n}\Big], \end{aligned} \tag{3.28}$$

where the outer expectation is over our state of knowledge about the hypothetical measurement $\mathbf{y}$ induced by the GPs of Sec. 2.2.1:

$$p(\mathbf{y}|\mathbf{x},\mathbf{x}_{1:n},\mathbf{y}_{1:n}) = \prod_{i=1}^{m} \mathcal{N}(y_i|\mu_{i,n}(\mathbf{x}),\sigma^2_{i,n}(\mathbf{x};\theta^e)+\nu^2), \tag{3.29}$$

where $\mu_{i,n}(\cdot) = \mu_{i,n}(\cdot;\theta^e_i)$ and $\sigma^2_{i,n}(\cdot) = \sigma^2_{i,n}(\cdot;\theta^e_i)$ are the posterior predictive mean and variance of the GP f^e_i pertaining to objective $i = 1,\ldots,m$, see Eq. (5.15). Our myopic strategy is outlined in Algorithm 3.

Eq. (3.28) is analytically intractable and must be approximated using the sampling methods of Sec. 3.2.2. This is computationally inefficient because it does not allow the use of gradient-based optimization algorithms such as BFGS. To overcome this difficulty, we derive an approximation that will allow us to make use of the analytical formulas derived by [21]. We have:

$$\begin{aligned} \mathbb{E}^e\big[\lambda\left(A[\mathbf{f}^e[X]]\right)\big|\mathbf{x}_{1:n},\mathbf{y}_{1:n}\big] \;&\geqslant\; \mathbb{E}^e\big[\lambda\left(A[\mathbf{f}^e[\mathbf{x}_{1:n}]]\right)|\mathbf{x}_{1:n},\mathbf{y}_{1:n}\big] \\ &\approx\; \lambda\left(A[\boldsymbol{\mu}_n[\mathbf{x}_{1:n}]]\right). \end{aligned}$$

The first row inequality comes from $\mathbf{x}_{1:n} \subset X$ implying $\mathbf{f}^e[\mathbf{x}_{1:n}] \subset \mathbf{f}^e[X]$ which, in turn, yields $A[\mathbf{f}^e[\mathbf{x}_{1:n}]] \subset A[\mathbf{f}^e[X]]$. For the approximation in the second row, start by noticing that $\mathbf{z} = \mathbf{f}^e[\mathbf{x}_{1:n}]$ conditioned on $\mathbf{x}_{1:n}$ and that $\mathbf{y}_{1:n}$ follows a multivariate Gaussian, see Eq. (3.13). Then, take the Taylor expansion of $\lambda(A[\mathbf{z}])$ about $\mathbf{z} = \mathbf{z}_0 = \boldsymbol{\mu}_n(\mathbf{x}_{1:n}) := (\mu_{1,n}(\mathbf{x}_{1:n}), \ldots, \mu_{m,n}(\mathbf{x}_{1:n}))$. The zero order term is the constant you see above, i.e., $\lambda\left(A[\boldsymbol{\mu}_n[\mathbf{x}_{1:n}]]\right)$. The expectation of the first order term vanishes and we ignore second and higher order terms. Thinking in the same way, we can get:

$$\begin{aligned}\mathbb{E}^e\left[\lambda\left(A[\mathbf{f}^e[X]]\right)\middle|\mathbf{x}, \mathbf{y}, \mathbf{x}_{1:n}, \mathbf{y}_{1:n}\right] &\geqslant \mathbb{E}^e\left[\lambda\left(A[\mathbf{f}^e[\mathbf{x}_{1:n} \cup \{\mathbf{x}\}]]\right)\middle|\mathbf{x}, \mathbf{y}, \mathbf{x}_{1:n}, \mathbf{y}_{1:n}\right] \\ &\approx \lambda\left(A[\boldsymbol{\mu}_{n,(\mathbf{x},\mathbf{y})}[\mathbf{x}_{1:n} \cup \{\mathbf{x}\}]]\right),\end{aligned}$$

where $\boldsymbol{\mu}_{n,(\mathbf{x},\mathbf{y})}$ is the posterior mean after seeing the hypothetical observation $(\mathbf{x}, \mathbf{y})$. Finally, we approximate the expectation over the hypothetical measurement as:

$$\begin{aligned}&\mathbb{E}^e\left[\lambda\left(A[\boldsymbol{\mu}_{n,(\mathbf{x},\mathbf{y})}[\mathbf{x}_{1:n} \cup \{\mathbf{x}\}]]\right)\middle|\mathbf{x}, \mathbf{x}_{1:n}, \mathbf{y}_{1:n}\right] \approx \\ &\mathbb{E}^e\left[\lambda\left(A[\boldsymbol{\mu}_{n,(\mathbf{x},\boldsymbol{f}^e(\mathbf{x}))}[\mathbf{x}_{1:n} \cup \{\mathbf{x}\}]]\right)\middle|\mathbf{x}, \mathbf{x}_{1:n}, \mathbf{y}_{1:n}\right].\end{aligned}$$

To see why this is possible, note that $\mathbf{y} = \mathbf{f}^e(\mathbf{x}) + \nu^2\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is Gaussian with zero mean and unit covariance, take the Taylor expansion of the integrand in the first line about $\boldsymbol{\epsilon} = \mathbf{0}$, and keep only the zero order term (the expectation of the first order term vanishes). Putting everything together, we get the (approximate) inequality:

$$\begin{aligned}\mathrm{EEIHV}(\mathbf{x}) &\tilde{\geqslant} \quad \overline{\mathrm{EEIHV}}(\mathbf{x}) \\ &:= \mathbb{E}^e\left[\lambda\left(A[\boldsymbol{\mu}_{n,(\mathbf{x},\boldsymbol{f}^e(\mathbf{x}))}[\mathbf{x}_{1:n} \cup \{\mathbf{x}\}]]\right)\middle|\mathbf{x}, \mathbf{x}_{1:n}, \mathbf{y}_{1:n}\right] \\ &\qquad -\lambda\left(A[\boldsymbol{\mu}_n[\mathbf{x}_{1:n}]]\right).\end{aligned} \tag{3.30}$$

The inequality is approximate because the first term on the right hand side is approximately greater than the second one. The accuracy is again second order and proving it requires taking the Taylor expansion of the integrand of the first term with respect to $\mathbf{z} \equiv \mathbf{f}^e(\mathbf{x})$ about $\mathbf{z} = \mathbf{z}_0 \equiv \boldsymbol{\mu}_n(\mathbf{x})$.

The important observation here is that the lower bound to EEIHV, i.e., $\overline{\text{EEIHV}}$ on right hand side of Eq. (3.30), is similar to the original EIHV of [21] with a few key differences. Specifically, $\overline{\text{EEIHV}}$ has the same analytical form as EIHV if in EIHV (i) we replace the observed targets with their projections to the posterior mean, i.e., if we work with the denoised measurements instead of the noisy ones; and (ii) we remove the noise variance from the predictive distribution of the GP. Therefore, the analytical formula for the calculation of EIHV found in [21] applies to $\overline{\text{EEIHV}}$ subject to the aforementioned substitutions. In all our numerical examples, we use $\overline{\text{EEIHV}}$. We maximize the lower bound over $\mathbf{x}$ using BFGS with multiple random restarts.

Algorithm 2 Information acquisition strategy for discovering the Pareto-frontier.

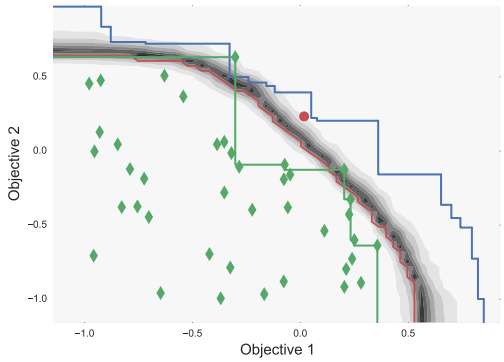
Require: Initially observed designs $\mathbf{x}_{1:n}$; Initial objective measurements $\mathbf{y}_{1:n}$; number of restarts of EEIHV optimization n_d; maximum number of allowed information source queries $N_{\max}$; EEIHV tolerance $\delta > 0$.

while $n < N_{\max}$ **do**
  Train the GP for each objective as described in Sec. 2.2.1.
  Find $\mathbf{x}_{n+1} = \arg\max_{\mathbf{x}\in X} \overline{\text{EEIHV}}(\mathbf{x})$ using n_d random restarts of BFGS.
  if $\overline{\text{EEIHV}}(\mathbf{x}_{n+1}) < \delta$ **then**
    Break.
  end if
  Evaluate the objectives at $\mathbf{x}_{n+1}$ measuring $\mathbf{y}_{n+1}$.
  $\mathbf{x}_{1:n+1} \leftarrow (\mathbf{x}_{1:n}, \mathbf{x}_{n+1})$.
  $\mathbf{y}_{1:n+1} \leftarrow (\mathbf{y}_{1:n}, \mathbf{y}_{n+1})$.
  $n \leftarrow n + 1$.
end while

3.3 Numerical Results

In Sec. 3.3.1 we use a synthetic example to visualize some of the concepts used through out this section. In Sections 3.3.2 and 3.3.3, we validate our approach using two synthetic stochastic optimization problems with known optimal solutions. To assess the robustness of the methodology, we experiment with various levels of stochasticity which causes the resultant noise in the outputs. In Sec. 3.4, we solve the steel wire drawing problem with uncertainties in the incoming wire diameters and the

die angles at each pass. In all the problems the objectives are scaled by subtracting and dividing by the emprical mean and standard deviation, respectively.



(a)

Figure 3.1. A synthetic example of the template followed throughout the paper depicting the Pareto front and the representation of the uncertainty around the Pareto front.

3.3.1 Correspondence between nomenclature and visualizations

Fig. 3.1 uses an $m = 2$ synthetic example to help us visualize and name some of the concepts used throughout this section. The dark blue staircase is an approximation of the true P_O, generated by taking the empirical Pareto frontier of sample averaged objective measurements at a large number of designs. The figure also shows a scatter plot of the denoised measurements $\boldsymbol{\mu}_n(\mathbf{x}_{1:n})$ (green dots), and as well as the corresponding empirical Pareto frontier $P[\boldsymbol{\mu}_n[\mathbf{x}_{1:n}]]$ (green line). The red dot marks the denoised measurement made at the design $\mathbf{x}_{n+1}$ that maximizes $\overline{\text{EEIHV}}(\mathbf{x})$. The red line is the top right boundary of the Vorob'ev expectation of $A[\mathbf{f}^e[X]]$ conditioned on the observed data $(\mathbf{x}_{1:n}, \mathbf{y}_{1:n})$, i.e., it is our expectation about $P[\mathbf{f}^e[X]]$ conditioned on our current state of knowledge. The gray contours show to the symmetric deviation $d_n^e(\mathbf{y})$ of $A[\mathbf{f}^e[X]]$ which corresponds to our uncertainty about $P[\mathbf{f}^e[X]]$.

3.3.2 Two-dimensional synthetic example

Consider the two-dimensional synthetic multi-objective problem taken from [77] which has been slightly modified for our use here:

$$o_1(\mathbf{x},\omega) = -\left(b_2 - \frac{5.1}{4\pi^2}{b_1}^2 + \frac{5}{\pi}b_1 - 6\right)^2 - 10\left[\left(1 - \frac{1}{8\pi}\right)\cos(b_1) + 1\right], \tag{3.31}$$

$$o_2(\mathbf{x},\omega) = \sqrt{|(10.5 - b_1)||(b_1 + 5.5)||(b_2 + 0.5)|} \frac{1}{30}\left(b_2 - \frac{5.1}{4\pi^2}{b_1}^2 - 6\right)^2 \frac{1}{3}\left[\left(1 - \frac{1}{8\pi}\right)\cos(b_1) + 1\right], \tag{3.32}$$

$$b_1(\mathbf{x},\omega) = 15(x_1 + s\xi(\omega)) - 5, \tag{3.33}$$

$$b_2(\mathbf{x},\omega) = 15(x_2 + s\xi(\omega)), \tag{3.34}$$

for $\mathbf{x} = (x_1, x_2) \in X = [0,1]^2$. The $(\Omega, \mathbb{P}, \mathcal{F})$ random variable ξ is standard normal, i.e., $\xi \sim \mathcal{N}(0,1)$. The parameter s controls the standard deviation of the noise infused by ξ. Notice that even though ξ is normal, the measured objectives $o_i(\mathbf{x},\omega)$ are not normally distributed due to the non-linearities. That is the statistics of the measurement process do not match our assumptions in Sec. 2.2.1. We do this on purpose. In real applications the statistics of the measurements process are not known and we would like to investigate to what extent the normality assumption produces robust results.

To validate our methodology, we must first estimate accurately the true P_O. We achieve this by finding the empirical Pareto frontier of a large number of designs (10000) while approximating $O_i(\mathbf{x}) = \mathbb{E}[o_i(\mathbf{x},\omega)]$ with 100 Monte Carlo samples. In this example, we aim to maximize the two objectives.

We start with $n = 20$ random initial observations and we add an additional 100 measurements selected according to Algorithm 3. Fig. 3.2 depicts our final state of

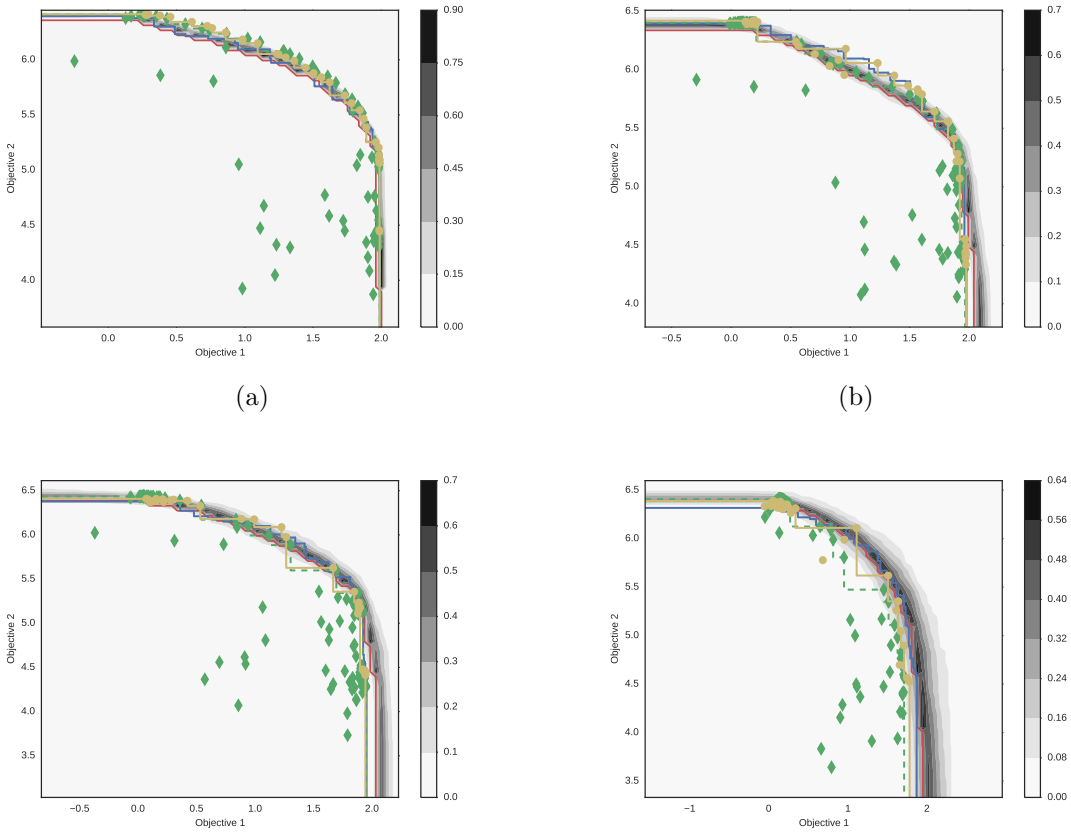


Figure 3.2. Two-dimensional synthetic example for starting from $n = 20$ initial measurements. Subfigures (a) ($s = 0.01$), (b) ($s = 0.03$), (c) ($s = 0.05$), and (d) ($s = 0.1$), depict our state of knowledge about the final $P[\mathbf{f}^e[\mathbf{x}_{1:n}]]$ after 100 measurements selected using Algorithm 3.

knowledge about $P[\mathbf{f}^e[\mathbf{x}_{1:n}]]$ for increasing noise levels $s = 0.01, 0.03, 0.05$, and 0.1. Another graphic that appears on this figure is the line joining the large yellow dots. These points represent the Pareto frontier obtained by a sampling average of the objectives at the Pareto optimal designs found by the methodology after the fixed number of iterations, i.e., an estimation of $P[\mathbf{O}[\mathbf{x}_{1:N_{\max}}]]$ which is to be contrasted to $P[\boldsymbol{\mu}_{N_{\max}}[\mathbf{x}_{1:N_{\max}}]]$. This Pareto frontier is a representation of the quality of the solution obtained by the methodology. With low levels of stochasticity the methodology neatly approximates the noise in the outputs as Gaussian, shown in Fig. 3.2 (a) and (b).

With an increase in the value of the stochasticity parameter, s, the final Pareto frontier obtained starts diverging from P_O, shown in Fig. 3.2 (c) and (d). In Fig. 3.2 (c) and (d), the methodology ends up exploiting the area near the two ends of the observed $P[\mathbf{O}[\mathbf{x}_{1:N_{\max}}]]$ only, and not the whole P_O which is possibly a manifestation of the methodology not being able to estimate and filter out the excessive non-Gaussian noise in these cases. The contours of the symmetric deviation (which can be understood as the probability of a particular set of objective values being achievable conditional on the observations made thus far) do reinforce greater knowledge about the plausibility of the achievable values even in regions which tend to dominate the approximated Pareto frontier. This means that with more simulations the methodology should eventually discover more Pareto efficient solutions across the complete boundary of the approximated Pareto frontier. So, the symmetric deviation allows the decision maker to realize the potential value that lies in doing further simulations.

3.3.3 Six-dimensional synthetic example

Consider the following test objective functions from [78]:

$$o_1(\mathbf{x},\omega) = \frac{1}{2}(x_1 + s\xi_1(\omega))(1+g), \tag{3.35}$$

$$o_2(\mathbf{x},\omega) = \frac{1}{2}(1 - (x_1 + s\xi_1(\omega)))(1+g), \tag{3.36}$$

$$g = 100\Bigg[5 + \sum_{i\in\{2,\cdots,6\}} \left((x_i + s\xi_i(\omega)) - 0.5\right)^2 - \cos(2\pi((x_i + s\xi_i(\omega)) - 0.5))\Bigg], \tag{3.37}$$

for $\mathbf{x} \in X = [0,1]^6$, where $\xi_i \sim \mathcal{N}(0,1), i = 1,\ldots,6$ are independent. As before, the expected objectives are not analytically available. We use the same approximation technique as in the previous example to estimate the ground truth of P_O for this test problem. Fig. 3.3 depicts our final state of knowledge about $P[\mathbf{f}^e[\mathbf{x}_{1:n}]]$ for increasing noise levels $s = 0.01, 0.03, 0.05$, and 0.1. As before, the larger the noise the harder it

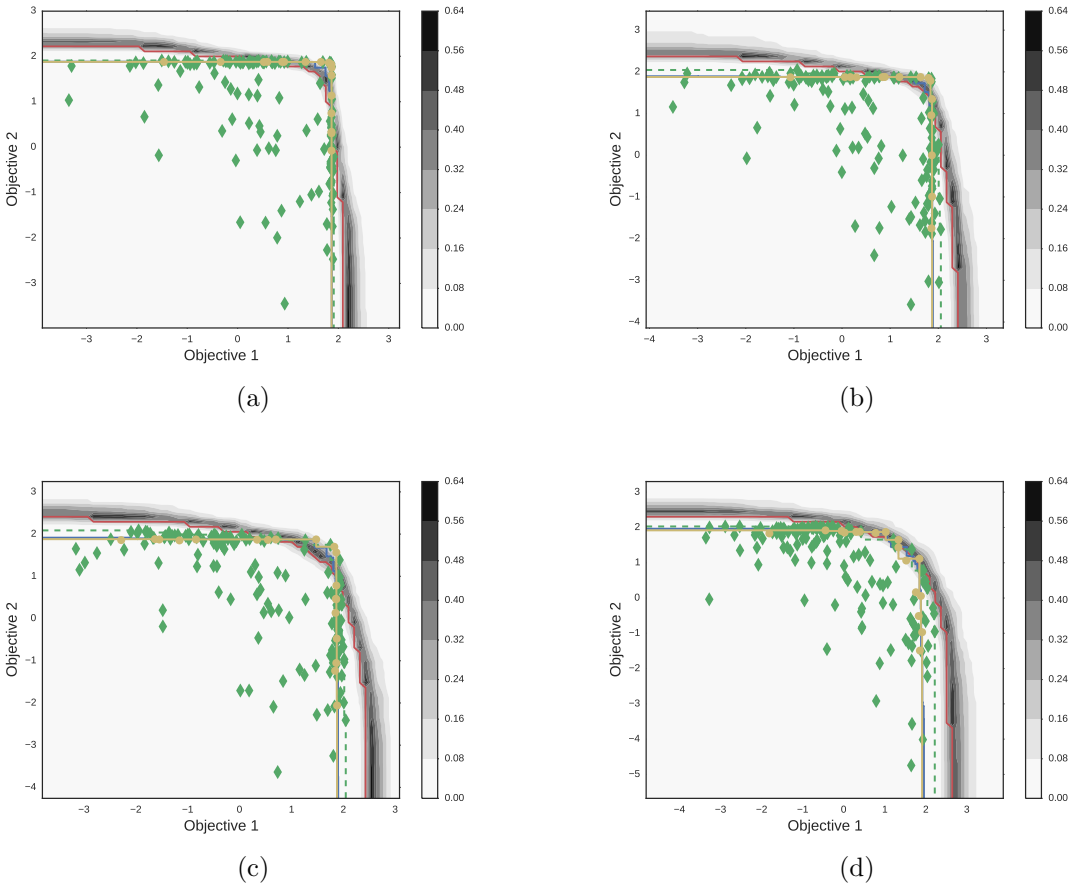


Figure 3.3. Six-dimensional synthetic example starting from ($n = 40$) initial measurements. Subfigures (a) ($s = 0.01$), (b) ($s = 0.03$), (c) ($s = 0.05$), and (d) ($s = 0.1$), depict our state of knowledge about the final $P[\mathbf{f}^e[\mathbf{x}_{1:n}]]$ after 100 measurements selected using Algorithm 3.

is for the methodology to discover P_O, the true Pareto frontier. In general, as can be seen in Fig. 3.3 the method is robust to noise as long as the noise is reasonably low for the given number of initial measurements. The powerfulness of the methodology can be observed through Fig. 3.3 (a) and (b) , where the final $P[\mathbf{f}^e[\mathbf{x}_{1:n}]]$ contains points that dominate the $P[\mathbf{O}[\mathbf{x}_{1:N_{\max}}]]$, when the noise parameter has relatively low values. The method, as expected, discovers very few points on P_O as the noise increases to $s = 0.1$ as can be seen in (d) of Fig. 3.3.

3.4 Wire drawing problem

The wire drawing process is designed to achieve the desired final diameter and mechanical properties such as ultimate tensile strength (UTS) and ductility through cold reduction of a larger diameter wire. The desired wire properties depend on applications – for example, high torsional ductility is required for application in tires, high strength wires used in machine tools for metal cutting. A typical reduction of the cross section the wire, based on the final properties required would be in the range of 70-90 percent and this is achieved by reducing the wire diameter in a number of passes. Each pass involves drawing through a conical die and the sequence of reductions and corresponding die angles at each pass would play an important role on the final properties as well as performance of operations. Here we consider a wire drawing process having a fixed number of passes (8 passes). An finite element analysis (FEA) based simulator, developed for an industrial operation was used to simulate this process. This wire drawing simulator includes wire deformation, heat generation and dissipation in the wire as well as dies, cooling of wire on the cooling drum and in the atmosphere and is based on large deformation theory. The model considers the process to be axisymmetric. The multi-pass drawing effect is modeled by considering carryover effect of previous pass such as residual stress, plastic strain and temperature. The FEA is done using four noded isoparametric elements. A penalty parameter approach is used for modeling the contact between the wire and the dies.

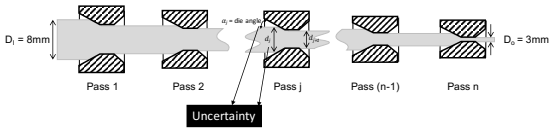
The simulator takes the input as wire material properties, input wire diameter, die pass schedule (reduction and die angle at each pass), wire drawing speed, cooling conditions, friction, etc.; and predicts the internal stress and strains in the wire and the die, load on each die and the drum, temperature of the wire and the die, properties indicative of final wire mechanical properties – UTS representing strength and strain non-uniformity factor (SNUF) representing relative ductility.

The plastic deformation across the cross section of the final wire should be as uniform as possible for enhanced ductility. The UTS is primarily governed by the total reduction but the non-uniform deformation has a significant secondary role on the final UTS. To understand this uniformity, the plastic strain distribution is modeled and is represented as SNUF. SNUF is a ratio of difference between the peak and average strain to average strain. Besides the properties of the drawn wire, process defects such as wire burst during drawing process is an important aspect to consider as central burst is highly undesired since it leads to wire breakage during drawing process and this effect is modeled through the measurement of triaxiality by a factor called the hydraulic failure factor (HFF). The coefficient of friction is assumed to be constant throughout the process. Here, we have the UTS and the SNUF as the two competing objectives for the process.

The design variables for this problem are the die angles (one at each pass) and the incoming wire diameter (implicit in the reduction ratio) at each pass. The outgoing wire diameter at a pass is same as the incoming wire diameter for the next pass. The incoming wire diameter d_j and the reduction ratio (rr_j) for a pass j are related by the formula given in (3.38).

$$rr_j = 1 - \frac{d_{j+1}^2}{d_j^2} \tag{3.38}$$

For this problem we take the case of drawing an 8mm wire into a 3mm wire Fig. 3.4. So, with the overall reduction ratio (and the incoming wire diameter for the first pass) fixed, the problem becomes that of two objectives with 15 design parameters (8 die angles and 7 incoming wire diameters). We apply our methodology to the wire drawing problem and demonstrate its ability to deal with the problem of stochasticity in the



(a)

Figure 3.4. WMP: The wire manufacturing process with the depiction of the sources of uncertainty, ie. the incoming wire diameter d_j and the die angle α_j, at an individual pass j.

objectives induced by our inability to fully control the design parameters, to obtain a set of Pareto optimal solutions. This uncertainty can be understood as the ubiquitous effect of the continuous wear and tear on the die which would cause the process to deviate from delivering ideal (no noise) outputs. Also, in any manufacturing process the tolerances need to be accounted for as the procured dies themselves would not have exact dimensions as required. The design space has been bounded by choosing a suitable range for design variables as follows:

1. For $i = 1, \ldots, 7$, $x_i \in [0, 1]$ represent the *incoming wire diameters*.

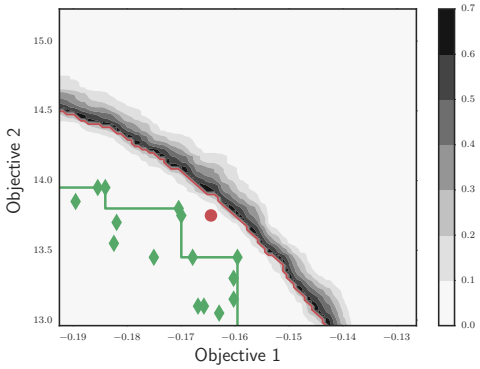
2. For $i = 1, \ldots, 8$, $x_{i+7} \in [0, 1]$ represent the *die angles*.

Specifically, we assume that when we try to implement a process with design $\mathbf{x}$, what we actually get is a process with design $\mathbf{x} + \mathbf{S}\boldsymbol{\xi}$, where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_{15}, \mathbf{I}_{15})$ and $S = \text{diag}(s_1, \ldots, s_{15})$ where $s_i = 0.05, \forall i \in [1, 7]$ and $s_i = 0.1, \forall i \in [8, 15]$. The above space $X = [0, 1]^{15}$ is a scaled representation of the real space for simplification purposes. The random vector from the real space $X = [7.2, 7.5] \times [6.6, 6.9] \times [5.8, 6.1] \times [5.1, 5.4] \times [4.4, 4.7] \times [3.9, 4.2] \times [3.3, 3.6] \times [8, 14]^8$, can be obtained by rescaling the random vector from the scaled space by using a simple linear transformation. The noisy objectives considered here are:

$$o_1(\mathbf{x}, \omega) = -\text{SNUF}\left(\mathbf{x} + \mathbf{S}\boldsymbol{\xi}(\omega)\right), \tag{3.39}$$

$$o_2(\mathbf{x}, \omega) = \text{UTF}\left(\mathbf{x} + \mathbf{S}\boldsymbol{\xi}\right). \tag{3.40}$$

The optimization problem involves maximizing the UTS and minimizing the SNUF. For simplifying the problem to the requirements of our code and software we convert it to an equivalent maximization problem where we maximize the UTS and maximize the negative of the SNUF. We consider a scenario with 15 initial observations of the MOO problem and limit our computational budget to allow for 50 additional simulations to be carried out sequentially.

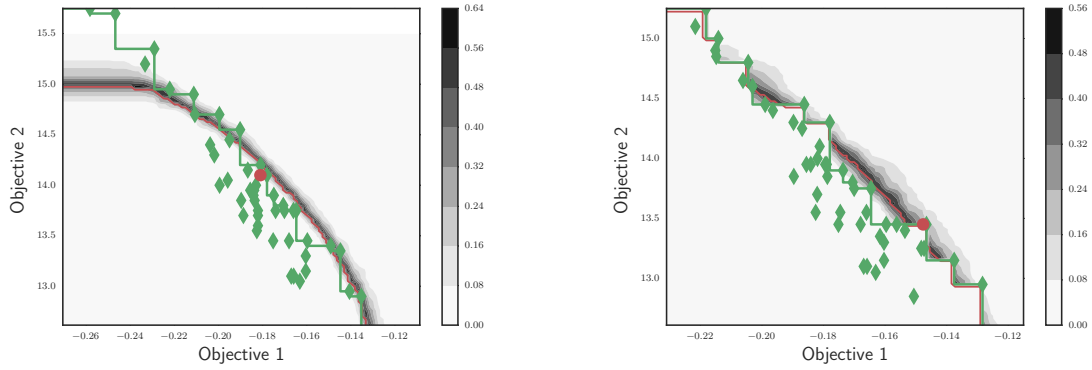


(a)

Figure 3.5. WMP: The $P[\mathbf{f}^e[\mathbf{x}_{1:n}]]$ for the inital observations using Eq. (5.12). Objective 1 is the -SNUF and Objective 2 is the UTS.

Fig. 3.5 shows the projected initial observations for the problem. We scale the measurements obtained by subtracting and dividing by the empirical mean and standard deviation just as in the case of the test function discussed above. This is done to maintain consistency with the assumption 3.2.1 of a zero mean (standard normal) GP for computational flexibility.

A key aspect of quantifying our knowledge about the state of the objectives is the Vorob'ev expectation which is computed by obtaining by sampling the design space X. However, it must be noted that in this case with 15 dimensions, it becomes very difficult to cover the whole design space as a result of which certain designs picked by the algorithm, end up outside the sampled designs. The overarching effect of this can be seen in Fig. 3.6 (a), where the Vorob'ev expectation can be seen lying below the points in the top left corner picked by the methodology. To circumvent this issue, we augment the set of sampled designs with the designs at which we have made observations. This provides a clearer picture, Fig. 3.6 (b), of the state as it reinforces the information obtained thus far while quantifying our beliefs about the state of the Pareto-efficient frontier.

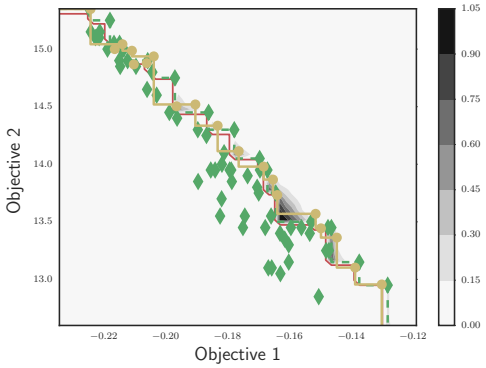


(a) (b)

Figure 3.6. WMP: The $P[\mathbf{f}^e[\mathbf{x}_{1:n}]]$. Subfigures (a) with the random sample design space, (b) after adding the observed designs to the sampled design space.

Fig. 3.7 depicts the state of the problem after the fiftieth iteration. Since, we do not have the computational resources to obtain $P[\mathbf{O}[\mathbf{x}_{1:N_{\max}}]]$ for comparison, we

sample average the value of the objectives, 100 times, corresponding to the final Pareto designs as shown in Fig. 3.6. This averaging gives us an estimate of the approximate true state of the Pareto-efficient frontier after the computational budget has been exhausted.



(a)

Figure 3.7. WMP: The $P[\mathbf{f}^e[\mathbf{x}_{1:n}]]$ after 50 additional measurements along with the *sampled averaged approximation of* $P[\mathbf{O}[\mathbf{x}_{1:N_{\max}}]]$ represented by the yellow line. Objective 1 is the -SNUF and Objective 2 is the UTS.

3.5 Conclusions

We constructed an extension to the EIHV information acquisition function which makes possible the application of BGO to stochastic multi-objective black-box optimization problems. In addition to the above, we have shown how the epistemic uncertainty induced by the limited number of simulations can be quantified and used, to represent the uncertainty around the PF at each stage. We have validated our approach by applying it on two, slightly modified to include stochastic parameters, synthetic test functions with known Pareto frontiers. Furthermore, we applied our method on the challenging steel wire drawing problem under parametric uncertainty in a scenario of simulation based design. The method offers a viable alternative to the state-of-the-art evolutionary optimization algorithms which rely heavily on sample averaging and are unaffordable under a limited budget scenario. Moreover, the

proposed extension to EIHV gives acceptable results under cases of moderate levels of noise with limited number of initial observations. There remain several open research questions. The most pressing direction to look in would be the efficient treatment of stochastic multi-objective problems under unknown and expensive constraints under a scenario of constrained computational resources.

4. BAYESIAN OPTIMAL DESIGN OF EXPERIMENTS FOR INFERRING THE STATISTICAL EXPECTATION OF A BLACK-BOX FUNCTION

Bayesian optimal design of experiments (BODE) has been successful in acquiring information about a quantity of interest (QoI) which depends on a black-box function. BODE is characterized by sequentially querying the function at specific designs selected by an infill-sampling criterion. However, most current BODE methods operate in specific contexts like optimization, or learning a universal representation of the black-box function. The objective of this chapter is to design a BODE for estimating the statistical expectation of a physical response surface. This QoI is omnipresent in uncertainty propagation and design under uncertainty problems. Our hypothesis is that an optimal BODE should be maximizing the expected information gain in the QoI. We represent the information gain from a hypothetical experiment as the Kullback-Liebler (KL) divergence between the prior and the posterior probability distributions of the QoI. The prior distribution of the QoI is conditioned on the observed data and the posterior distribution of the QoI is conditioned on the observed data and a hypothetical experiment. The main contribution of this chapter is the derivation of a semi-analytic mathematical formula for the expected information gain about the statistical expectation of a physical response. The developed BODE is validated on synthetic functions with varying number of input-dimensions. We demonstrate the performance of the methodology on a steel wire manufacturing problem.

The following text is taken from the publication titled: *Bayesian Optimal Design of Experiments For Inferring The Statistical Expectation Of A Black-Box Function.*

4.1 Introduction

Engineering problems require either computationally intensive computer codes [11] or expensive physical experiments [79]. With insufficient information about the analytic dependence of the physical response on the design parameters or experimental conditions, the engineer needs scores of physical response evaluations to make decisions with confidence. To overcome this issue, researchers have developed design of experiments (DOE) techniques that attempt to select the maximally informative physical response evaluations within a given budget [80–82]. Classical DOE techniques generate a single batch design [83] and, thus, they face several shortcoming in case of functions with local features, e.g., discontinuities, or sharp non-linearities [8]. Sometimes the DOE obtained can be equally spaced when the context requires more samples from certain regions of the domain. Such scenarios require a sequential DOE (SDOE) approach.

SDOE uses past observations to decide the next evaluation point [84, 85]. Over the past two decades, SDOE has been used in several applications spanning both physical experiments [79, 86–89] and computer simulations [12, 90–92]. One of the most theoretically sound SDOEs is Bayesian optimal design of experiments (BODE). Under BODE, one models the physical response using a statistical surrogate and selects the next evaluation point by attempting to maximize the expected value of information. The newly acquired information is used to condition one's belief about the physical response using Bayes' rule. The process is repeated until the marginal value of information is negative. The exact definition of the value of information depends on one's goals. For example, one could be interested in optimizing an objective [9, 14, 15, 30, 66, 93–100], learning an accurate representation of the physical response [13, 101–105] or estimating the probability of a rare event [106, 107].

Instead of the value of information, several BODE approaches attempt to maximize the information gain about a quantity of interest (QoI). The information gain can be quantified through the Kullback-Leibler divergence (KLD) [55, 108] (also known

as relative entropy). Over the years, KLD has been used to quantify information gain [109] about the objective function, from a hypothetical experiment (an untried design). The efficacy of the KLD has been extended and demonstrated on various applications including the sensor placement problem [96,110], surrogate modeling [111–113], learning missing parameters [114], optimizing an expensive physical response [18], calibrating a physical model [115,116], reliability design [117,118], efficient design space exploration [119], probabilistic sensitivity analysis [120], portfolio optimization [3], neural-network hyperparameter tuning [121].

Despite the significant progress, deriving BODE methods for new objectives remains a non-trivial task. In particular, there are no BODE methods for efficiently propagating input uncertainties through a physical response surface, e.g., estimating the statistical expectation, the variance, or higher order statistics of a physical quantity of interest. Uncertainty propagation is particularly important for characterizing the robustness of a simulation/experiment and, thus, being able to do it efficiently is essential for robust design. To address this need, the *objective* of this chapter is to develop a BODE methodology for estimating the statistical expectation of the physical response. The technical details of our approach are as follows. Much like the majority of the work in BODE, we use Gaussian process (GP) surrogates to emulate the physical response [122]. The expected information gain from a hypothetical experiment is defined to be the KLD between one's *prior* and *posterior* probability densities on the statistical expectation of the physical response. To derive analytical expressions of the prior and the posterior of this quantity of interest, we use the standard expressions for the mean and covariance of a GP conditioned on data. The EKLD of the statistical expectation of the physical response comes out to be an analytically tractable function which alleviates the need for sample averaging.

The outcomes of this chapter can be enumerated as follows: (a) The derivation of semi-analytical expressions for the expected information gain in one's state of knowledge about the statistical expectation of an expensive-to-evaluate physical response; (b) The numerical investigation of the performance of the resulting BODE

using synthetic examples; (c) Numerical comparisons to uncertainty sampling; (d) The application of the new scheme to solve an uncertainty propagation problem involving a steel wire manufacturing process simulated using finite elements; and (e) A freely available PYTHON implementation of our methodology[1].

The chapter is organized as follows: Sec. 4.2 describes in detail the methodology used, including GP regression Sec. 4.2.1 and the EKLD Sec. 4.2.2. The results obtained for four synthetic examples have been presented in Sec. 4.3. We compare the above proposed BODE methodology with uncertainty sampling which is a common design of experiments method used in practical engineering scenarios in Sec. **??**. The steel wire manufacturing problem is briefly explained and treated with the proposed methodology in Sec. 4.3.5. We summarize the nuances of the methodology including its weaknesses and comment on future research directions in Sec. 4.4.

4.2 Methodology

Throughout the paper we represent the various elements of our state of knowledge and objective as follows:

1. $\mathbf{X}_n$ are the n designs at which the simulation/experiment has been conducted, i.e., $\mathbf{X}_n = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$.

2. $\mathbf{Y}_n$ are the values of the physical response at the corresponding n designs, i.e., $\mathbf{Y}_n = \{y_1, \cdots, y_n\}$.

3. Collectively, we represent all observed data by $\mathbf{D}_n = \{\mathbf{X}_n, \mathbf{Y}_n\}$.

4. A hypothetical untried design is denoted by $\tilde{\mathbf{x}}$.

5. A hypothetical observation at $\tilde{\mathbf{x}}$ is denoted by $\tilde{y}$.

Let $\mathbf{x}$ be a random variable with probability density function (PDF) $p(\mathbf{x})$. Without loss of generality, we will assume that $p(\mathbf{x})$ is the uniform PDF supported on the

[1] `https://github.com/piyushpandita92/bode`

d-hypercube $\mathcal{X} = \times_{k=1}^{d}[0,1]$. The true physical response f is assumed to be a squared integrable function of $\mathbf{x} \in \mathcal{X}$, i.e., $f \in \mathcal{L}^2(\mathcal{X})$, where

$$\mathcal{L}^2(\mathcal{X}) = \left\{ f : \mathcal{X} \to \mathbb{R} \middle| \int_{\mathcal{X}} f^2(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} < \infty \right\}. \tag{4.1}$$

The QoI that we want to discover through the sequential design of experiments is the statistical expectation of the physical response. Mathematically,

$$\boldsymbol{Q}[f] = \int_{\mathcal{X}} f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{4.2}$$

This QoI is a bounded linear functional, an observation that leads to analytical progress. At each stage of the SDOE, we will update our beliefs about $\boldsymbol{Q}[f]$ in a Bayesian way, quantifying the epistemic uncertainty induced by limited data at the same time. The above QoI can also be approached using Quadrature methods [2, 123], however we restrict the focus of this work to sequential experiment design. We will select the new experiment by maximizing the expected information gain for $\boldsymbol{Q}[f]$.

4.2.1 Surrogate modeling

GP regression is a very popular non-parametric Bayesian regression technique. It allows one to express their prior beliefs about the underlying response surface, but it also quantifies epistemic uncertainty induced by limited observations. Here, we describe the GP regression very briefly. More details can be found in [43].

Prior Gaussian process

We model our prior beliefs about the physical response using a zero mean GP. The covariance function is defined by a radial basis function (RBF), also known as squared exponential. Mathematically,

$$f \sim \mathrm{GP}(0, k), \tag{4.3}$$

where

$$k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}'; \boldsymbol{\psi}) = s^2 \exp\left\{ -\frac{1}{2} \sum_{j=1}^{d} \frac{(x_j - x_j')^2}{\ell_j^2} \right\}. \tag{4.4}$$

The covariance function defined in Eq. (4.4) encodes our prior beliefs about the smoothness and magnitude of the response. The symbol $\ell_j > 0$ in Eq. (4.4) is the lengthscale of the j-dimension of the input space. This parameter quantifies the correlation between the function values at two different inputs. The s^2 in Eq. (4.4) is the signal strength of the GP. It incorporates the scale of the response. These parameters are the hyper-parameters of the covariance function and we will denote them by $\boldsymbol{\psi}$, i.e., $\boldsymbol{\psi} = \{s^2, \ell_1, \cdots, \ell_d\}$. A nonzero mean function can always be included with only minor modifications in what follows.

The data likelihood

The likelihood of the data $\mathbf{Y}_n$ a multivariate Gaussian. The mean vector of this Gaussian distribution is the vector of function output values $\mathbf{f}_n = \{f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)\}$ at observed designs. The covariance matrix can be computed using the structure defined in Eq. (4.4). The observations are assumed to be contaminated with Gaussian noise with variance σ^2. This noise variance is could will be very small relative to the signal strength in the case of computer simulation design. We augment the vector of hyper-parameters to include this additional parameter to get $\boldsymbol{\theta} = \{\boldsymbol{\psi}, \sigma^2\}$. Mathematically, the likelihood of the observed data is:

$$p(\mathbf{Y}_n | \mathbf{X}_n, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{Y}_n | 0, \mathbf{K}_n + \sigma^2 \mathbf{I}_n), \tag{4.5}$$

where $\mathbf{K}_n$ is a $n \times n$ covariance matrix defined according to Eq. (4.4), i.e., $K_{nij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

Training the hyper-parameters

Typically, the hyper-parameter values are fitted to the observed data by maximizing the likelihood of Eq. (4.5). However, this process may result in overfitting which is particularly problematic in the context of SDOE. In this work, we opt for a fully Bayesian treatment [124] which is more robust. We assume that the hyperparameters are a priori independent following an exponential prior distribution on the lengthscales and Gamma prior distribution on the signal strength. Since we do not treat noisy problems in this work, we fix the variance of the likelihood probability to 1e-6 which is a reasonably small value. Bayes' rule allows yields the hyperparameter posterior:

$$p(\boldsymbol{\theta}|\mathbf{D}_n) \propto p(\mathbf{Y}_n|\mathbf{X}_n, \boldsymbol{\psi}) p(\boldsymbol{\psi}). \tag{4.6}$$

Here, we employ a *parallel-chain* Markov chain Monte Carlo (MCMC) algorithm with an *affine invariance* sampler to sample from the posterior. More details on the inner workings of the MCMC algorithm can be found in [125]. The code for this MCMC algorithm is available online.[2]

Making predictions

Conditioned on the hyperparameters, our state of knowledge about f is also characterized by a GP:

$$f|\mathbf{D}_n, \boldsymbol{\theta} \sim \mathrm{GP}(f|m_n, k_n), \tag{4.7}$$

where

$$m_n(\mathbf{x}) = (\mathbf{k}_n(\mathbf{x}))^T \left(\mathbf{K}_n + \sigma^2 \mathbf{I}_n\right)^{-1} \mathbf{Y}_n, \tag{4.8}$$

with

$$\boldsymbol{\alpha}_n = (\mathbf{K}_n + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{Y}_n, \tag{4.9}$$

[2] `https://github.com/dfm/emcee`

is the *posterior mean* function, and

$$k_n(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - (\mathbf{k}_n(\mathbf{x}))^T \left(\mathbf{K}_n + \sigma^2 \mathbf{I}_n\right)^{-1} \mathbf{k}_n(\mathbf{x}'), \tag{4.10}$$

with $\mathbf{k}_n(\mathbf{x}) = (k(\mathbf{x}, \mathbf{x}_1), \ldots, k(\mathbf{x}, \mathbf{x}_n))^T$, is the *posterior covariance* function. In particular, at an untried design point $\tilde{\mathbf{x}}$ the point-predictive posterior probability density of $\tilde{y} = f(\tilde{\mathbf{x}})$ conditioned on the hyperparameters is:

$$p(\tilde{y}|\tilde{\mathbf{x}}, \mathbf{D}_n, \boldsymbol{\theta}) = \mathcal{N}\left(\tilde{y}\middle|m_n(\tilde{\mathbf{x}}; \boldsymbol{\theta}), \sigma_n^2(\tilde{\mathbf{x}}; \boldsymbol{\theta})\right), \tag{4.11}$$

where $\sigma_n^2(\tilde{\mathbf{x}}; \boldsymbol{\theta}) = k_n(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}; \boldsymbol{\theta})$. Finally, the *point-predictive posterior* PDF of $\tilde{y} = f(\tilde{\mathbf{x}})$ is:

$$p(\tilde{y}|\tilde{\mathbf{x}}, \mathbf{D}_n) = \int p(\tilde{y}|\tilde{\mathbf{x}}, \mathbf{D}_n, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{D}_n) d\boldsymbol{\theta}. \tag{4.12}$$

The latter is, of course, not analytically available, but one can derive sampling average approximations using the MCMC samples from $p(\boldsymbol{\theta}|\mathbf{D}_n)$.

4.2.2 Sequential design of experiments using the expected information gain

Given $\mathbf{D}_n$ observations, our state of knowledge about the QoI $\boldsymbol{Q}[f]$ is given by:

$$p(q|\boldsymbol{\theta}, \mathbf{D}_n) = \mathbb{E}\left[\delta\left(q - \boldsymbol{Q}[f]\right)|\boldsymbol{\theta}, \mathbf{D}_n\right], \tag{4.13}$$

where $\delta(\cdot)$ is Dirac's delta function and the expectation is over the function space measure defined by the posterior GP, see Eq. (4.7). The uncertainty in $p(q|\mathbf{D}_n)$ represents our epistemic uncertainty induced by the limited number of observations in $\mathbf{D}_n$. Now suppose that we did an experiment at $\tilde{\mathbf{x}}$ and observed the output $\tilde{y}$. The

posterior GP measure would become $p(q|\mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y})$ and, thus, our state of knowledge about $\boldsymbol{Q}[f]$ would be:

$$p(q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}) = \mathbb{E}\left[\delta\left(q - \boldsymbol{Q}[f]\right)|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}\right]. \tag{4.14}$$

According to information theory, the information gained through the hypothetical experiment $(\tilde{\mathbf{x}}, \tilde{y})$ conditioned on the hyperparameters, say $G(\tilde{\mathbf{x}}, \tilde{y}; \boldsymbol{\theta})$ is given by the KLD between $p(q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y})$ and $p(q|\boldsymbol{\theta}, \mathbf{D}_n)$. Mathematically, it is:

$$G(\tilde{\mathbf{x}}, \tilde{y}; \boldsymbol{\theta}) = \int_{-\infty}^{\infty} p(q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}) \log \frac{p(q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y})}{p(q|\boldsymbol{\theta}, \mathbf{D}_n)} dq. \tag{4.15}$$

The expected information gain of the hypothetical experiment, say $G(\tilde{\mathbf{x}})$, is obtained by taking the expectation of $G(\tilde{\mathbf{x}}, \tilde{y})$ over our current state of knowledge. Specifically,

$$G(\tilde{\mathbf{x}}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G(\tilde{\mathbf{x}}, \tilde{y}; \boldsymbol{\theta}) p(\tilde{y}|\boldsymbol{\theta}, \tilde{\mathbf{x}}, \mathbf{D}_n) p(\boldsymbol{\theta}|\mathbf{D}_n) d\tilde{y} d\boldsymbol{\theta}. \tag{4.16}$$

We pick the next experiment by solving:

$$\mathbf{x}_{n+1} = \arg\max_{\tilde{\mathbf{x}}} G(\tilde{\mathbf{x}}). \tag{4.17}$$

In the rest of this section, we derive analytical approximations of $p(q|\boldsymbol{\theta}, \mathbf{D}_n)$ (Sec. 5.2.3), $p(q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y})$ (Sec. 5.2.3), $G(\tilde{\mathbf{x}}, \tilde{y}; \boldsymbol{\theta})$ (Sec. 5.2.3), and a sampling average approximation for $G(\tilde{\mathbf{x}})$ (Sec. 5.2.3).

Quantification of the current state of knowledge about QoI

We now derive an analytical approximation of our current state of knowledge about the QoI, i.e., $p(q|\boldsymbol{\theta}, \mathbf{D}_n)$. Since the QoI $\boldsymbol{Q}[f]$, Eq. (5.3), is linear and the

point predictive PDF of $y = f(x)$ is Gaussian, Eq. (5.15), $p(q|\boldsymbol{\theta}, \mathbf{D}_n)$ is Gaussian. In particular, it is easy to show that:

$$p(q|\boldsymbol{\theta}, \mathbf{D}_n) = \mathcal{N}\left(q\middle|\mu_1, \sigma_1^2\right). \tag{4.18}$$

The mean μ_1 is given by:

$$\begin{aligned} \mu_1 &:= \mathbb{E}[\boldsymbol{Q}[f]|\boldsymbol{\theta}, D_n] \\ &= \mathbb{E}\left[\textstyle\int_{\mathcal{X}} f(\mathbf{x})p(x)d\mathbf{x}\middle|\boldsymbol{\theta}, D_n\right] \\ &= \textstyle\int_{\mathcal{X}} \mathbb{E}\left[f(\mathbf{x})|\boldsymbol{\theta}, D_n\right] p(\mathbf{x})d\mathbf{x} \\ &= \textstyle\int_{\mathcal{X}} m_n(\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &= \boldsymbol{\epsilon}_n^T \boldsymbol{\alpha}_n, \end{aligned} \tag{4.19}$$

where $\boldsymbol{\alpha}_n$ is defined in Eq. (5.13) and each component of $\boldsymbol{\epsilon}_n \in \mathbb{R}^n$ is given by:

$$\begin{aligned} \epsilon_{ni} &= \epsilon(\mathbf{x}_i) \\ &:= \textstyle\int_{\mathcal{X}} k(\mathbf{x}_i, \mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &= s^2 \left(\tfrac{\pi}{2}\right)^{\frac{d}{2}} \textstyle\prod_{k=1}^{d} \left\{\ell_k \left[\operatorname{erf}\left(\tfrac{1-x_{ik}}{\sqrt{2}\ell_k}\right) - \operatorname{erf}\left(-\tfrac{x_{ik}}{\sqrt{2}\ell_k}\right)\right]\right\}, \end{aligned} \tag{4.20}$$

with erf being the error function, and x_{ik} the k-th component of the observed input $\mathbf{x}_i$. The variance σ_1^2 is given by:

$$\begin{aligned} \sigma_1^2 &:= \mathbb{E}[\boldsymbol{Q}^2[f]|\boldsymbol{\theta}, D_n] - (\mathbb{E}[\boldsymbol{Q}[f]|\boldsymbol{\theta}, D_n])^2 \\ &= \mathbb{E}[(\textstyle\int_{\mathcal{X}} f(\mathbf{x})p(\mathbf{x})d\mathbf{x})^2||\boldsymbol{\theta}, D_n] - {\mu_1}^2 \\ &= \mathbb{E}[\textstyle\int_{\mathcal{X}} f(\mathbf{x})p(\mathbf{x})d\mathbf{x} \textstyle\int_{\mathcal{X}} f(\mathbf{x}')p(\mathbf{x}')d\mathbf{x}'||\boldsymbol{\theta}, D_n] - {\mu_1}^2 \\ &= \textstyle\int_{\mathcal{X}} \textstyle\int_{\mathcal{X}} \mathbb{E}[f(\mathbf{x})f(\mathbf{x}')|\boldsymbol{\theta}, D_n]p(\mathbf{x})p(\mathbf{x}')d\mathbf{x}d\mathbf{x}' - {\mu_1}^2 \\ &= \textstyle\int_{\mathcal{X}} \textstyle\int_{\mathcal{X}} [k_n(\mathbf{x}, \mathbf{x}') + m_n(\mathbf{x})m_n(\mathbf{x}')]p(\mathbf{x})p(\mathbf{x}')d\mathbf{x}d\mathbf{x}' - {\mu_1}^2 \\ &= \textstyle\int_{\mathcal{X}} \textstyle\int_{\mathcal{X}} k_n(\mathbf{x}, \mathbf{x}')p(\mathbf{x})p(\mathbf{x}')d\mathbf{x}d\mathbf{x}' \\ &= \sigma_0^2 - \boldsymbol{\epsilon}_n^T \left(\mathbf{K}_n + \sigma^2\right)^{-1} \boldsymbol{\epsilon}_n, \end{aligned} \tag{4.21}$$

where

$$\begin{aligned}\sigma_0^2 &= \int_{\mathcal{X}}\int_{\mathcal{X}} k(\mathbf{x},\mathbf{x}')p(\mathbf{x})p(\mathbf{x}')d\mathbf{x}d\mathbf{x}' \\ &= s^2\prod_{k=1}^{d}(2\ell_k^2\sqrt{\pi})\Bigg\{\tfrac{-1}{\sqrt{\pi}}+\tfrac{1}{\sqrt{\pi}}\exp\left(\tfrac{-1}{2\ell_k^2}\right)+ \\ &\qquad \tfrac{1}{\sqrt{2}\ell_k}\operatorname{erf}\left(\tfrac{1}{\sqrt{2}\ell_k}\right)\Bigg\}.\end{aligned} \tag{4.22}$$

Quantification of the hypothetical state of knowledge about QoI

To derive an analytical approximation of our hypothetical state of knowledge about the QoI, i.e., $p(q|\boldsymbol{\theta},\mathbf{D}_n,\tilde{\mathbf{x}},\tilde{y})$, we proceed as in Sec. 5.2.3, but with the remark that the posterior GP after adding the hypothetical observation will have mean function:

$$\tilde{\mu}_{n+1}(\mathbf{x}) = \mu_n(\mathbf{x}) + k_n(\mathbf{x},\tilde{\mathbf{x}})\frac{\tilde{y}-\mu_n(\tilde{\mathbf{x}})}{k_n(\tilde{\mathbf{x}},\tilde{\mathbf{x}})+\sigma^2}, \tag{4.23}$$

and covariance function:

$$\tilde{k}_{n+1}(\mathbf{x},\mathbf{x}') = k_n(\mathbf{x},\mathbf{x}') - \frac{k_n(\mathbf{x},\tilde{\mathbf{x}})k_n(\tilde{\mathbf{x}},\mathbf{x}')}{k_n(\tilde{\mathbf{x}},\tilde{\mathbf{x}})+\sigma^2}. \tag{4.24}$$

We get,

$$p(q|\boldsymbol{\theta},\mathbf{D}_n,\tilde{\mathbf{x}},\tilde{y}) = \mathcal{N}(q|\mu_2(\tilde{\mathbf{x}},\tilde{y}),\sigma_2^2(\tilde{\mathbf{x}})). \tag{4.25}$$

The mean $\mu_2(\tilde{\mathbf{x}},\tilde{y})$ is:

$$\begin{aligned}\mu_2(\tilde{\mathbf{x}},\tilde{y}) &:= \mathbb{E}[\boldsymbol{Q}[f]|\boldsymbol{\theta},D_n,\tilde{\mathbf{x}},\tilde{y}] \\ &= \int_{\mathcal{X}}\tilde{\mu}_{n+1}(\mathbf{x})d\mathbf{x} \\ &= \mu_1 + \tfrac{\nu(\tilde{\mathbf{x}})}{k_n(\tilde{\mathbf{x}},\tilde{\mathbf{x}})+\sigma^2}\left(\tilde{y}-\mu_n(\tilde{\mathbf{x}})\right)\end{aligned} \tag{4.26}$$

with

$$\nu(\tilde{\mathbf{x}}) := \epsilon(\tilde{\mathbf{x}}) - \boldsymbol{\epsilon}_n^T\left(\mathbf{K}_n+\sigma^2\right)^{-1}\mathbf{k}_n(\tilde{\mathbf{x}}), \tag{4.27}$$

where $\epsilon(\tilde{\mathbf{x}})$ as in Eq. (4.20) but with $\mathbf{x}_i$ replaced by $\tilde{\mathbf{x}}$. Using the expression for the posterior covariance from Eq. (4.24) one can simplify $\sigma_2^2(\tilde{\mathbf{x}})$ similar to the derivation in Eq. (5.36) to get:

$$\begin{aligned} \sigma_2^2(\tilde{\mathbf{x}}) &:= \mathbb{E}[\boldsymbol{Q}^2[f]|\boldsymbol{\theta}, D_n, \tilde{\mathbf{x}}, \tilde{y}] - (\mathbb{E}[\boldsymbol{Q}[f]|\boldsymbol{\theta}, D_n, \tilde{\mathbf{x}}, \tilde{y})^2 \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} \tilde{k}_{n+1}(\mathbf{x}, \mathbf{x}') p(\mathbf{x}) p(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \\ &= \sigma_1^2 - \tfrac{\nu^2(\tilde{\mathbf{x}})}{k_n(\tilde{\mathbf{x}},\tilde{\mathbf{x}})+\sigma^2}. \end{aligned} \tag{4.28}$$

Quantification of the expected information gain about the QoI

Since both Eq. (5.34) and Eq. (4.25) are Gaussian, the KL divergence between the hypothetical and the current state of knowledge about the QoI conditional on the hyper-parameters, $G(\mathbf{x}, \tilde{y}; \boldsymbol{\theta})$ of Eq. (5.31), is analytically tractable [126], i.e.,

$$G(\mathbf{x}, \tilde{y}; \boldsymbol{\theta}) = \log\left(\tfrac{\sigma_1}{\sigma_2(\tilde{\mathbf{x}})}\right) + \tfrac{\sigma_2^2(\tilde{\mathbf{x}})}{2\sigma_1^2} + \tfrac{(\mu_2(\tilde{\mathbf{x}},\tilde{y})-\mu_1)^2}{2\sigma_1^2} - \tfrac{1}{2}. \tag{4.29}$$

Furthermore, $G(\mathbf{x}, \tilde{y}; \boldsymbol{\theta})$ is a quadratic function of $\tilde{y}$, and $p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\theta}, \mathbf{D}_n)$ is Gaussian, see Eq. (5.15). Thus, we can analytically integrate $\tilde{y}$ out to obtain:

$$\begin{aligned} G(\tilde{\mathbf{x}}; \boldsymbol{\theta}) &= \int_{-\infty}^{\infty} G(\tilde{\mathbf{x}}, \tilde{y}; \boldsymbol{\theta}) p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\theta}, \mathbf{D}_n) d\tilde{y} \\ &= \log\left(\tfrac{\sigma_1}{\sigma_2(\tilde{\mathbf{x}})}\right) + \tfrac{1}{2}\tfrac{{\sigma_2}^2(\tilde{\mathbf{x}})}{{\sigma_1}^2} - \tfrac{1}{2} \\ &\quad + \tfrac{1}{2}\tfrac{v(\tilde{\mathbf{x}})^2}{\sigma_1^2(\sigma_n^2(\tilde{\mathbf{x}})+\sigma^2)}, \end{aligned} \tag{4.30}$$

Finally, we take the expectation of $G(\tilde{\mathbf{x}}; \boldsymbol{\theta})$ over the posterior of the hyperparameters, $p(\boldsymbol{\theta}|\mathbf{D}_n)$ of Eq. (4.6), using the MCMC samples $\left\{\boldsymbol{\theta}^{(s)}\right\}_{s=1}^{S}$ collected with the procedure described in [125, 127].

This yields:

$$\begin{aligned} G(\tilde{\mathbf{x}}) &= \int G(\tilde{\mathbf{x}}; \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{D}_n) d\boldsymbol{\theta} \\ &\approx \tfrac{1}{S} \sum_{s=1}^{S} G\left(\tilde{\mathbf{x}}; \boldsymbol{\theta}^{(s)}\right). \end{aligned} \tag{4.31}$$

Maximizing the expected information gain about the QoI

At each stage of our BODE algorithm, we optimize the EKLD $G(\tilde{\mathbf{x}})$ using Bayesian global optimization (BGO) based on the augmented expected improvement (AEI) [15]. This choice takes into account the noisy nature of the approximation of Eq. (5.41), and it reduces the computational time compared to a brute force or a multistart-and-gradient-based-optimization approach. See Algorithm 3 for pseudocode. In all our experiments, irrespective of the dimensionality, we use $T_n = 20$ BGO iterations to optimize the EKLD.

Algorithm 3 Optimize the EKLD using BGO with AEI.

Require: Initial number of EKLD evaluations T_i; maximum number of EKLD evaluations T_n; number of candidate designs n_d for BGO; MCMC samples from the posterior of the hyperparameters $\left\{\boldsymbol{\theta}^{(s)}\right\}_{s=1}^{S}$; stopping tolerance $\gamma_i > 0$.

Evaluate $G(\tilde{\mathbf{x}})$ using Eq. (5.41) at T_i random points to generate training data, $\tilde{\mathbf{X}}_{T_i} = \{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_{T_i}\}$ and $\mathbf{G}_{T_i} = \left\{\tilde{G}_1 = G(\mathbf{x}_1), \ldots, \tilde{G}_{T_i} = G(\mathbf{x}_{T_i})\right\}$, for BGO.
$t \leftarrow t_i$.
while $t < T_n$ **do**
    Fit a standard GP on the input-output pairs $\tilde{\mathbf{X}}_t$-$\tilde{\mathbf{G}}_t$ using maximum likelihood to approximate $G(\tilde{\mathbf{x}})$.
    Generate a set of candidate test points $\hat{\mathbf{X}}_{n_d} = \{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_{n_d}\}$ using Latin Hypercube Sampling (LHS) [55].
    Compute the AEI of all of the candidate points in $\hat{\mathbf{X}}_{n_d}$.
    Find the candidate point $\hat{\mathbf{x}}_j$ that exhibits the maximum AEI.
    if If the maximum AEI is smaller than γ_i **then**
        Break.
    end if
    Use Eq. (5.41) to evaluate $G(\tilde{\mathbf{x}})$ at $\hat{\mathbf{x}}_j$ measuring $\hat{G}_j = G(\hat{\mathbf{x}}_j)$.
    $\tilde{x}_{t+1} \leftarrow \hat{x}_j$.
    $\tilde{G}_{t+1} \leftarrow \hat{G}_j$.
    $\mathbf{X}_{t+1} \leftarrow \tilde{\mathbf{X}}_t \cup \{\tilde{\mathbf{x}}_{t+1}\}$.
    $\mathbf{G}_{t+1} \leftarrow \mathbf{G}_t \cup \{\tilde{G}_{t+1}\}$.
    $t \leftarrow t + 1$.
end while
return $\arg\max_{\tilde{\mathbf{x}}_{T_n}} \tilde{\mathbf{G}}_{T_n}$.

4.2.3 Selecting the Initial Set of Designs

In most literature, as a rule of thumb, $10d$ number of initial samples are used. We resort to using lesser number of initial data points to test the performance of the methodology when it starts from the low-sample regime. Readers interested in the problem of the optimal selection of initial data size can refer to the work of Sòbester et. al. [128] where the authors discuss the problem in the context of optimization. The problem of selecting an optimal number of initial points is beyond the scope of the work presented here.

4.2.4 Selecting the Covariance kernel

Selecting the form of the covariance kernel is the problem of optimal model selection which is a challenging problem in itself. However, optimal model selection is not the focus of this work. For consistency across the results for different problems, throughout this work, we use the squared exponential (RBF) covariance kernel for GP regression modeling.

4.2.5 Complete BODE framework

In Algorithm 4, we provide pseudocode implementation of the proposed BODE framework. The algorithm stops when a predetermined number of experiments have been performed. Alternatively, one could stop the algorithm when the expected information gain is below a threshold.

4.3 Numerical Results

We apply the methodology on two one-dimensional mathematical functions (synthetic examples), a three-dimensional problem, and a five-dimensional problem. For the first two synthetic examples the input domain simply becomes $[0, 1]$ whereas for the third synthetic example the input domain is $[-2, 6]^3$. The inputs for the five

Algorithm 4 Bayesian optimal design of experiments maximizing the expected information gain about the statistical expectation of a physical response.

Require: Initially observed inputs $\mathbf{X}_{n_i}$; initially observed outputs $\mathbf{Y}_{n_i}$; maximum number of allowed experiments N.

1: $n \leftarrow n_i$.
2: **while** $n < N$ **do**
3: Sample from the posterior of the hyperparameters, Eq. (4.6), to obtain $\left\{\boldsymbol{\theta}^{(s)}\right\}_{s=1}^{S}$.
4: Find the next experiment $\mathbf{x}_{n+1}$ using Algorithm 3 to solve Eq. (4.17).
5: Evaluate the objective at $\mathbf{x}_{n+1}$ measuring $y_{n+1} = f(\mathbf{x}_{n+1})$.
6: $\mathbf{X}_{n+1} \leftarrow \mathbf{X}_n \cup \{\mathbf{x}_{n+1}\}$.
7: $\mathbf{Y}_{n+1} \leftarrow \mathbf{Y}_n \cup \{y_{n+1}\}$.
8: $t \leftarrow t + 1$.
9: **end while**

dimensional numerical example lie in the hyper-cube $[0, 1]^5$. The number of initial data points is denoted by n_i. The number of initial data points is taken as low as possible for the numerical examples. In most literature, as a rule of thumb, $10d$ number of initial samples are used. We resort to using lesser number of initial data points to test the performance of the methodology when it starts from the low-sample regime. Readers interested in the problem of the optimal selection of initial data size can refer to the work of Sòbester et. al. [128] where the authors discuss the problem in the context of optimization. The problem of selecting an optimal number of initial points is beyond the scope of the work presented here.

4.3.1 Synthetic example no. 1

Consider the function

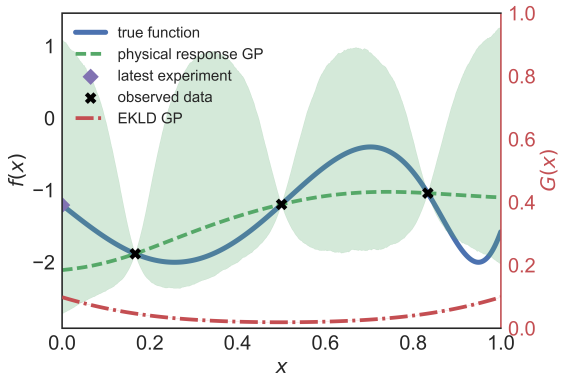
$$f(x) = 4\left(1 - \sin\left(6x + 8e^{6x-7}\right)\right), \tag{4.32}$$

defined on $[0, 1]$. This function is smooth throughout its domain, but it exhibits two local minima. We will apply our methodology to estimate the statistical expectation:

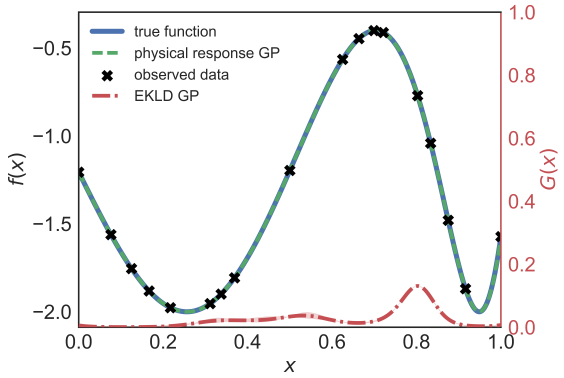
$$\boldsymbol{Q}[f] = \int_0^1 f(x)dx.$$

The true value of $\boldsymbol{Q}[f]$ is analytically available, $\boldsymbol{Q}[f] = -1.3599$. We apply our methodology to this problem starting from $n_i = 3$ and sample a total of $N = 28$ points. The number of MCMC chains for the results shown below is six, and the number of steps per chain is 500. For further details on the MCMC part of training the GP, we refer the readers to [125, 127].

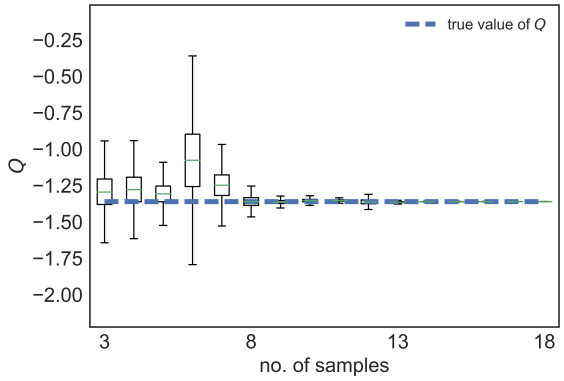
Figs. 4.1 (a) and (b) show the initial and final state of Algorithm 4. The thick blue line represents the true function f, Eq. (4.32). The black crosses are the observed data at the given stage. In subfigure (a), the next experiment selected by maximizing the EKLD, see Algorithm 3, is corresponds to the purple diamond. The mean of the GP fit to the expected information gain $G(\tilde{\mathbf{x}})$ constructed by BGO in Algorithm 3. The predictive mean of the EKLD is shown by the dotted light blue line. This dotted line represents the response surface of the EKLD after the BGO has ended and the red shaded area around it represents the uncertainty (2.5 percentile and 97.5 percentile) around it. As expected, the mean of the EKLD is very small or close to zero at points where experiments have been performed. Thus, the point selected by the methodology (purple diamond) is located in the input space where the EKLD has high mean. The posterior mean of the GP of the black-box function is represented by the dashed bottle-green line. The bottle-green shaded area represents the uncertainty (2.5 percentile and 97.5 percentile) around it. The final set of inputs, space-filling, selected by the methodology can be seen in Fig. 4.1 (b). Fig. 4.1 (c) shows the $p(q|\mathbf{D}_n)$ plotted against the number of data samples while showing convergence towards the true value of $\boldsymbol{Q}[f]$. The gradual reduction of predictive uncertainty of $\boldsymbol{Q}[f]$ from the initial to the final stage of the algorithm is seen in Fig. 4.1 (c).



(a)



(b)



(c)

Figure 4.1. One-dimensional synthetic example ($n_i = 3$). Subfigures (a) and (b) show the state of the function (1st iteration) at the start and the end (15th iteration) of the algorithm. Subfigure (c) represents the convergence to the true expectation of the function and the reduction in uncertainty about the QoI after the end of the algorithm.

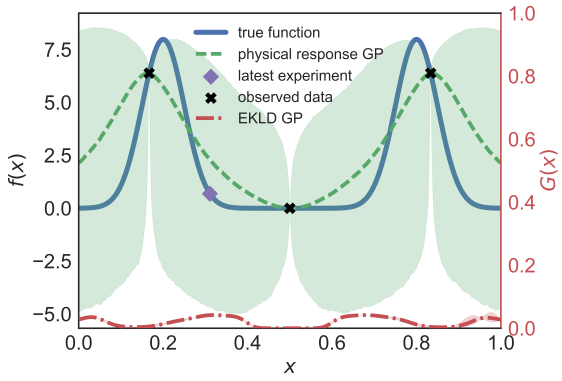
4.3.2 Synthetic example no. 2

We consider the following Gaussian mixture function to test and validate our methodology further.

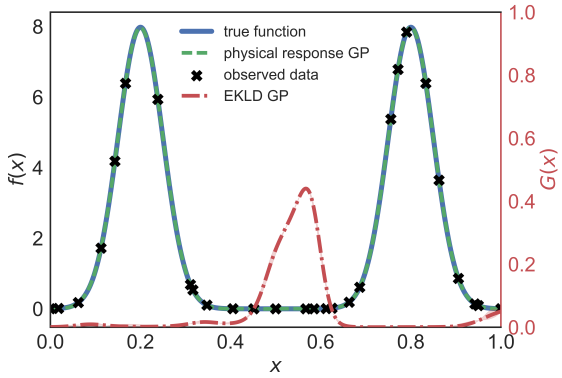
$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}s_1}\exp\left\{-\frac{(x-m_1)^2}{2{s_1}^2}\right\} \\ &\quad + \frac{1}{\sqrt{2\pi}s_2}\exp\left\{-\frac{(x-m_2)^2}{2{s_2}^2}\right\}, \end{aligned} \tag{4.33}$$

where $m_1 = 0.2$ and $s_1 = 0.05$, $m_2 = 0.8$ and $s_2 = 0.05$. As can be seen from Eq. (4.33), the function is a sum of probability densities of two Gaussian distributions. The notoriety of the function lies in two relatively sharp but smaller areas of high magnitude. The true value of $\boldsymbol{Q}[f]$ is analytically available, $\boldsymbol{Q}[f] = 2.0$. We apply our methodology to this problem starting from $n_i = 3$ and sample another 25 points. The final state of sampling can be seen in Fig. 4.2 (b), which shows a fairly equally spaced spread of designs. It is important to note that Fig. 4.2 (b) can mislead the reader into perceiving the sampling to be less dense in the areas where the function is sharply peaked. This is an illusion due to the starkly varying ordinates of the sampled points near the peaks of the function. The convergence of the estimated mean to the true value of $\boldsymbol{Q}[f]$ and the reduction in uncertainty around the $\boldsymbol{Q}[f]$ can be seen in Fig. 4.2 (c).

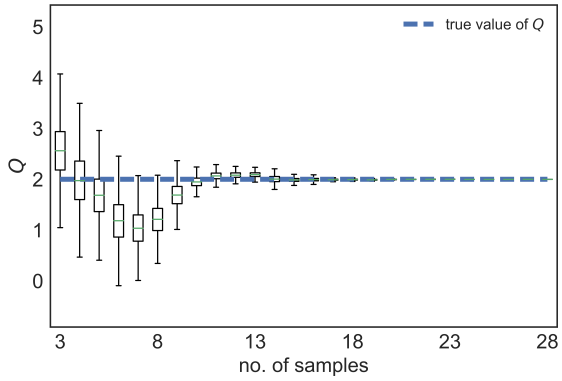
In Fig. 4.2 (a) and (b), the EKLD is shown by the dotted line and the true function is shown by the dashed red line. The solid line represents the mean of the GP model and the orange shaded areas around it represent the 2.5th and the 97.5th percentile of the GP. We plot the relative maximum mean EKLD as a function of the number of samples in Fig. 4.3 for both the synthetic functions. This relative maximum EKLD is the ratio of the maximum predictive mean of the EKLD for the current iteration and the overall maximum predictive mean of the EKLD obtained across all iterations. The plots in Fig. 4.3 show a characteristic typical of BODE functions i.e. of increasing in magnitude for the first few iterations and then falling sharply. This predicted mean value of the EKLD asymptotically goes to zero for both the synthetic functions here.



(a)



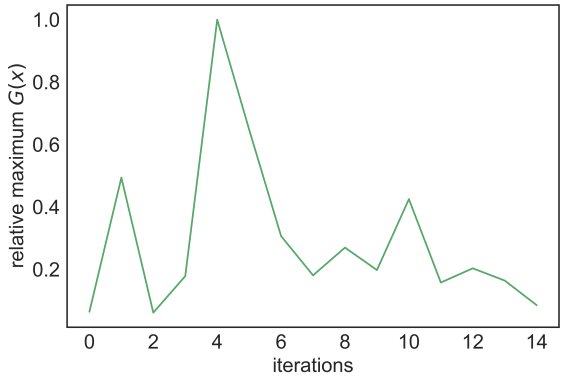
(b)



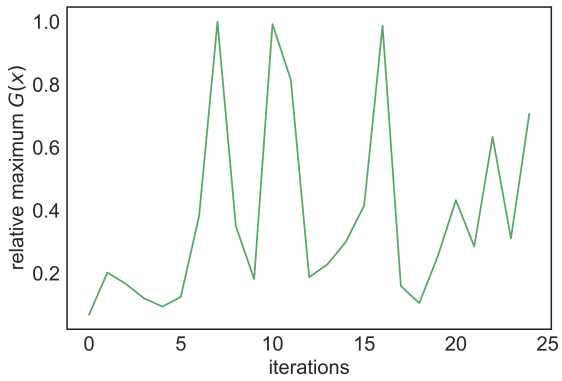
(c)

Figure 4.2. One-dimensional synthetic example ($n_i = 3$). Subfigures (a) and (b) show the state of the function at the start (1st iteration) and the end (25th iteration) of the algorithm. Subfigures (c) represents the convergence to the true expectation of the function and the reduction in uncertainty about the QoI after the end of the algorithm.

The number of MCMC chains for the results shown below is six, and the number of steps per chain is 500.



(a)



(b)

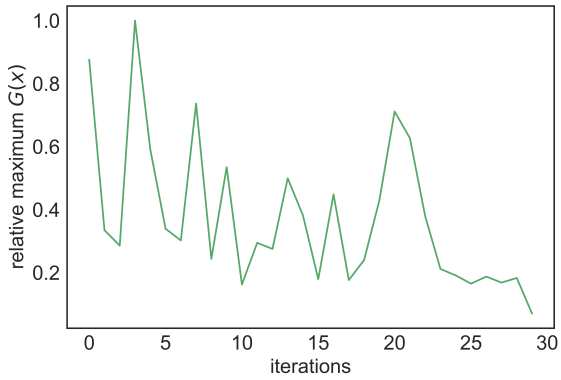
Figure 4.3. One-dimensional synthetic examples. Subfigures (a) and (b) show the predictive mean of the EKLD, for synthetic example no. 1 ($n_i = 3$) and synthetic example no. 2 ($n_i = 4$) respectively.

4.3.3 Synthetic example no. 3

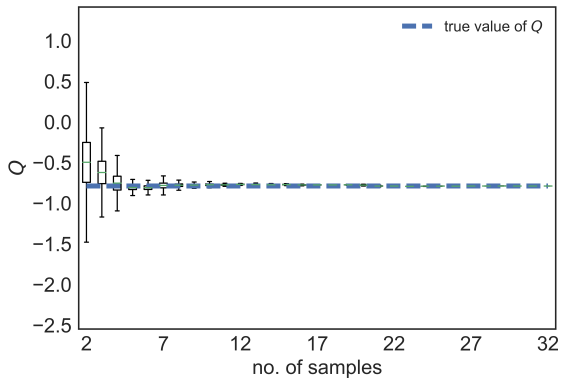
We consider the following three dimensional function from [129] to test and validate our methodology further.

$$\begin{aligned} f(\mathbf{x}) &= 4(x_1 + 8x_2 - 8x_2^2 - 2)^2 + (3 - 4x_2)^2 \\ &\quad +16\sqrt{x_3 + 1}(2x_3 - 1)^2. \end{aligned} \tag{4.34}$$

The major difference between this function Eq. (4.34) and the the first two synthetic examples is the dimensionality of the problem. The true value of $\boldsymbol{Q}[f]$ is analytically available, $\boldsymbol{Q}[f] = -0.7864$. We apply our methodology to this problem starting from $n_i = 2$ and sample another 30 points. Fig. 4.4 (b) shows that the methodology started with a highly uncertain estimate of the true value and eventually converged to a sharp peaked Gaussian distribution around the true value. The approximation to $\boldsymbol{Q}[f]$ at each stage of the algorithm is shown in Fig. 4.4 (b). The gradual reduction in uncertainty around $\boldsymbol{Q}[f]$ also can be seen in Fig. 4.4 (b). Fig. 4.4 (a) demonstrates how the relative EKLD fluctuates while seemingly approaching zero.



(a)



(b)

Figure 4.4. Three-dimensional synthetic example ($n_i = 2$). Subfigure (a) shows the decay of the EKLD from the 1st iteration to the end of the 30th iteration of the algorithm. Subfigures (b) show the convergence to the true value of the QoI respectively.

4.3.4 Synthetic example no. 4

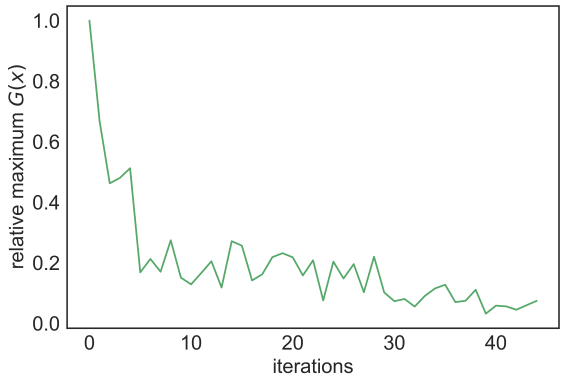
The following five dimensional function is taken from [130].

$$f(\mathbf{x}) \quad = \quad 10\sin(\pi x_1 x_2) + 20(x_3 - 5)^2 + 10x_4 + 5x_5. \tag{4.35}$$

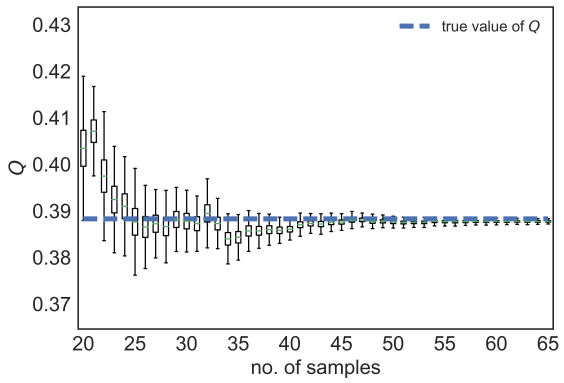
This function Eq. (4.35) is reasonably high-dimensional and challenging due to the non-linear input-output relation. The true value of $\boldsymbol{Q}[f]$ is analytically available, $\boldsymbol{Q}[f] = 0.3883$. We apply our methodology to this problem starting from $n_i = 20$ and sample another 45 points. Fig. 4.5 (a) demonstrates how the mean of the relative EKLD tends to approach zero by the end of the sampling process. The iteration-wise convergence of the $\boldsymbol{Q}[f]$ to its true value is shown in Fig. 4.5 (b). Fig. 4.5 (b) can present an illusion to the reader as it shows that the mean of the QoI is very close to the true value at the start of sampling. This is misleading because of the relatively large variance around the mean which means that the methodology is not confident of being close to the true value. As a result of this it can be seen, in the subsequent iterations, that the mean of the QoI goes to either side of the true value with a gradual decrease in variance. This might happen due to the methodology discovering different modes of the underlying function. As more data are accumulated, the uncertainty around the estimate decreases.

4.3.5 Wire drawing problem

The wire drawing process aims to achieve a required reduction in the cross section of the incoming wire, while aiming to monitor or optimize the mechanical properties of the outgoing wire. The incoming wire is passed through a series of dies (8 dies) to achieve an overall reduction in wire diameter. Each pass reduces the cross section of the incoming wire. The wire drawing process here is represented by an expensive computer code of which only a small number of evaluations are possible. The frictional work per Tonne (FWT) is one of the outputs of the expensive code. Large deformation



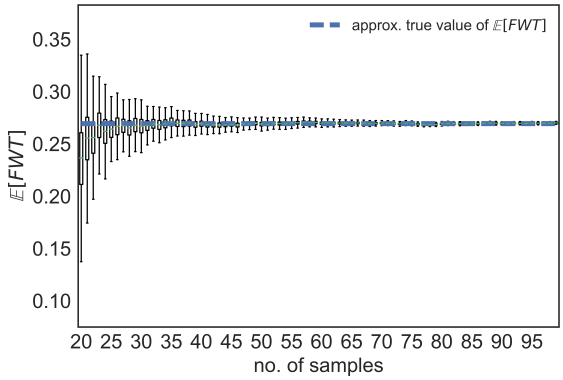
(a)



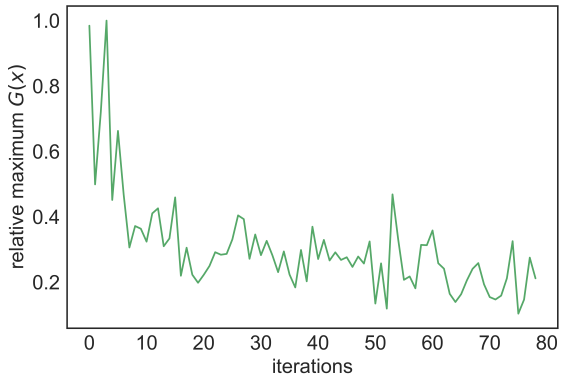
(b)

Figure 4.5. Five-dimensional synthetic example ($n_i = 20$). Subfigure (a) shows the decay of the EKLD from the 1st iteration to the end of the 45th iteration of the algorithm. Subfigure (b) shows convergence to the true value of the QoI.

theory is used to model the wire deformation, heat generation and dissipation in the wire and the dies at each pass, cooling of wire on the cooling drum and in the atmosphere. The model considers the process to be axisymmetric. The multi-pass drawing effect is modeled by considering carryover effect of previous pass such as residual stress, plastic strain and temperature. The FEA is done using four noded isoparametric elements. The contact between the wire and the dies is modeled using a penalty parameter approach. The statistical expectation value of the FWT is of importance for various stakeholders as the work done by the friction on the passing wire determines the power consumed, the wear on the final wire, etc. The FWT is the aggregate of the frictional work done at each pass. In our problem, we consider the die angle as design variables for each pass. The outgoing diameters at each pass are fixed to reasonable values. Thus, we deal with a total of 8 design variables. We start the methodology with 20 initial data points and add another 80 samples. We approximate the true value of the expectation of FWT, by averaging the outputs at 6,000 designs generated by Latin-hypercube sampling (LHS), as $\boldsymbol{Q}[FWT] \approx 0.2694$. The results in Fig. 4.6 show the gradual convergence of the methodology's mean estimate of the QoI towards the approximated true value. Fig. 4.6 (a) shows the mean and variance of the expectation of FWT as the mean approaches the approximate true value while the variance around it decreases gradually. The reduction in variance around the QoI from the start of the sampling to the end can be seen in Fig. 4.6 (b). This is intuitive as the number of collected samples increases, the variance around the QoI decreases. The comparison of the performance of the EKLD to that of the US is seen in Fig. 4.6 (c). The mean of the statistical expectation value of FWT for the EKLD converges to the approximate true value as more samples are added, while that for the US makes gradual drifts either side of the approximate true value. The US requires more samples to approach the approximate true value. This difference may be explained by the context specific functional form of the derived EKLD compared to the agnostic US which, although is a reduced form of the KLD in the design variables, seems to be slower in higher dimensions.



(a)



(b)

Figure 4.6. Wire drawing problem ($n_i = 20$) after 80 iterations.

4.3.6 Comparison with Uncertainty Sampling

As a demonstration of the performance of the methodology in contrast to a ubiquitous state-of-the-art sampling technique, namely uncertainty sampling (US), the methodology is tested on the synthetic examples given in Sec. 4.3.1, Sec. 4.3.2, Sec. 4.3.3 and Sec. 4.3.4. The uncertainty sampling technique works on the principle of reducing the uncertainty around the predictive response surface. Interestingly it has been shown that maximizing the information gain in the parameters reduces to uncertainty sampling under certain assumptions [101]. Moreover, US, in its functional form, as an IAF is agnostic to the context (QoI) in the problem. Hence, it serves as an ideal benchmark to compare with the EKLD. An explanation of the US methodology is as follows. The methodology selects a design with the maximum magnitude of predictive variance and follows this procedure until it sequentially acquires the required number of samples. The surrogate modeling process for the US works the same way as for the EKLD. The overall algorithm remains the same as Algorithm 4, but for the change in the sampling criterion.

The convergence to the QoI for the synthetic example in Sec. 4.3.1 and Sec. 4.3.2 is seen in Fig. 5.10 (a) and Fig. 5.10 (b) respectively. Overall, the two methodologies converge to the true value within reasonable time of one another. With the two peaked one-dimensional function of Sec. 4.3.2, the EKLD takes more iterations to converge as seen in Fig. 5.10 (b). The US can be seen as being quicker in reaching very close to the true value of the QoI compared to the EKLD for the synthetic example no. 2 whereas EKLD takes slightly fewer iterations to estimate the true statistical expectation value for synthetic example no. 1.

As the complexity of the problems increases, convergence for the EKLD becomes quicker compared to US as shown in Fig. 5.10 (a), (b) and (c). With the three-dimensional problem Fig. 5.10 (c), the mean estimate of the QoI for the EKLD converges after 20 samples have been collected. For the same problem, US takes almost 30 samples to converge. This saving of almost 10 samples could be useful in

engineering problems where each sample is collected at the expense of thousands of dollars of effort or a computational burden of multiple days.

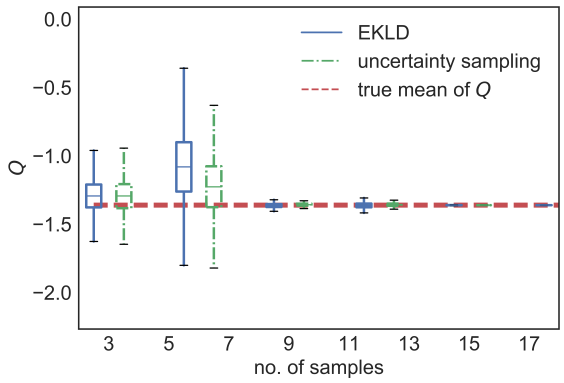
For the five-dimensional synthetic example, Fig. 5.11 (b) shows how the EKLD starts to approach the true value of the QoI as the number of iterations increases, whereas US tends to shows jaggedness in its patterns of convergence. After 65 samples have been collected US shows convergence, but convergence can be seen for the EKLD as early as the addition of the 45th sample. This observation is further strengthened by looking at the decay of the EKLD in Fig. 4.5 (b). The comparison in Fig. 5.11 (b) highlights the capability of the methodology to infer the QoI in a limited number of iterations. This is useful in the context of problems with expensive black-box functions where each evaluation of the expensive function has a very high cost. Moving on to the wire-problem in Fig. 5.11 (b), it can be seen that the convergence to the approximate true value is achieved by the EKLD and US albeit with more samples for US.

Another important feature of the comparisons in Fig. 5.10 and Fig. 5.11 is the faster reduction in the uncertainty for EKLD compared to US. This observation hints at the faster convergence of the EKLD across all numerical examples. For expensive problems, with very high-dimensional parameter, space reduced-order model based techniques [131] need to be used for the context of inferring the statistical expectation of the black-box function. Approaching such problems is beyond the scope of this work.

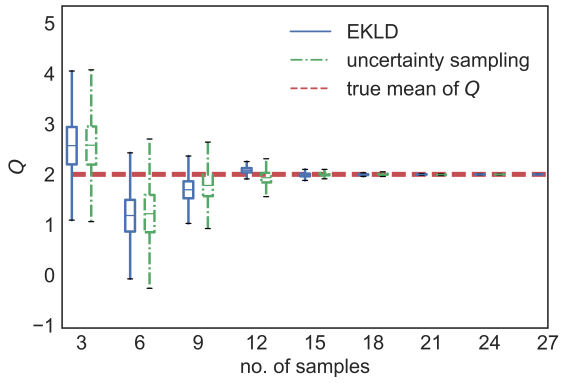
4.3.7 Insight into EKLD

We summarize our thoughts and observations, based on the above experiments, as follows:

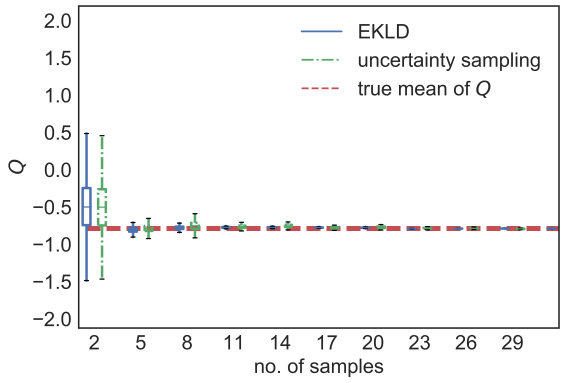
1. We observe that EKLD and US exhibit similar behavior in low-dimensions, but that EKLD is clearly better in higher-dimensions both in terms of point-wise estimation error and reduction in epistemic uncertainty. For one-dimensional



(a)

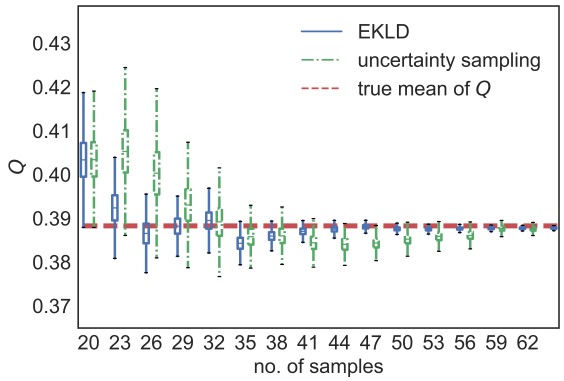


(b)

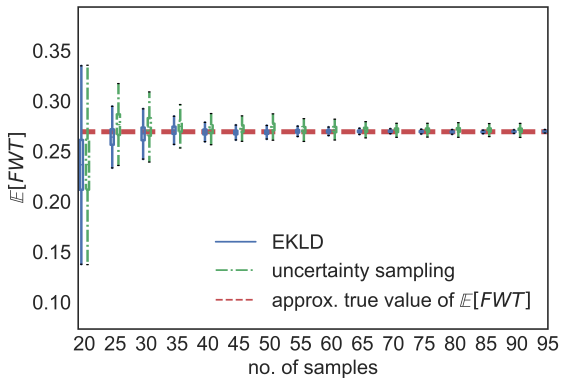


(c)

Figure 4.7. Subfigures (a), (b), and (c) show the comparison of the EKLD to uncertainty sampling, for synthetic example nos. 1, 2, and 3 respectively.



(a)



(b)

Figure 4.8. Subfigures (a) and (b) show the comparison of the EKLD to uncertainty sampling, for synthetic example no.4 and the wire-drawing problem respectively.

problems, US took fewer samples to converge to the true value in one of the numerical examples.

2. The EKLD quantifies the information gain in the statistical expectation whereas US quantifies the information gain in the parameters (design variables in this work) while selecting the most informative experiment. More work needs to be done to truly analyze and point out the difference between the two methodologies. The use of non-stationary GPs is a natural way to fully test the merits and pitfalls of the two methodologies, as it would results in locally adapted designs.

3. The number of initial data points differs for each of the above toy problems. This is done on purpose to test the limits of the methodology for examples of varying dimensionality. Thus far the experiments do not reveal a concrete rule for choosing the number of initial data points. However, starting the methodology with too few points can lead to delayed convergence. As a rule of thumb $5d$ number of initial points would be considered enough to start the methodology.

4. It is also observed that the MCMC samples needed to approximate the EKLD using sample averaging can cause numerical issues. If the MCMC samples are selected from a very short ensemble of chains, the EKLD will be noisy. This would need a more rigorous treatment, than the AEI-based BGO, to optimize the EKLD. To circumvent this issue we do not start the MCMC from scratch at iteration. Instead we use the last particle of the trace from the previous iteration to initialize the MCMC for a given iteration. This results in shorter thermalization times for the MCMC.

5. The MCMC details for each problem are in similar vein. The results presented above mention the number of chains and the number of steps per chain for each problem. We observe that the *emcee* [127] MCMC sampler performs well consistently with a reasonable number of chains and number of steps per chain. One of the requirements of the *emcee* sampler is that the number of chains

should be greater than or equal to twice the number of hyper-parameters of the GP model. Thus, the number of chains grows as the dimensionality of the problem increases leading to increased computational cost.

4.4 Conclusions

We presented a methodology for designing experiments to infer the value of a particular QoI, the statistical expectation of a physical response. The methodology leverages the expected KL divergence to compute the information gain in the QoI, from a hypothetical design. The work presented in this chapter is different from previous work done in sequential design of experiments using KL divergence as it quantifies the information gain in the QoI, instead of the information gain in the model parameters. The analytical tractability of the final expressions derived for the expected KL divergence, for learning the statistical expectation of a physical response, obviates computational hurdles induced by sample averaging.

One weakness of our methodology is the assumption that the covariance function of the GP model is stationary. The modeling of the hyperparameters of the GP should instead be based on a non-stationary covariance function for more locally adapted designs. However, the problem of implementing a non-stationary GP is not trivial. Another area of limited research is the selection of number of initial data points, i.e., before starting sequential design of experiments. A vast majority of literature on BODE uses *ad hoc* criteria for selection of this initial DOE. We accept that this is an open problem and more work is needed in this direction to ensure optimal allocation of budget. In similar vein, the methodology can be well extended to design experiments to infer generic statistics or quantities of interest which depend on a noisy black-box function. Some of these challenges are addressed in the next chapter.

5. BAYESIAN OPTIMAL DESIGN OF EXPERIMENTS TO INFER STATISTICS OF BLACK-BOX FUNCTIONS

Estimating statistical quantities of interest (QoIs), that are non-linear functions of an expensive black-box computer code or a laboratory experiment, is a challenging problem. Traditional methods are either context specific or require hundreds of thousands of evaluations of the black-box code. Bayesian optimal design of experiments (BODE) is a family of methods that define an optimal design of experiments under different contexts such as optimizing the black-box objective, estimating the statistical expectation of the black-box objective, inferring the response surface of the black-box objective, etc. in a limited number of function evaluations or laboratory experiments. Under BODE methods, sequential design of experiments (SDOE) accomplishes this task by selecting an optimal sequence of experiments. SDOE methods use data-driven probabilistic surrogate models that emulate the expensive black-box function and quantify one's current state-of-knowledge. Probabilistic predictions from the surrogate model are used to compute an information-criterion that quantifies the plausible information gain (IG) in a hypothetical experiment. The next experiment is selected by maximizing this IG. In this chapter, we extend a Kullback-Liebler (KL) divergence based BODE heuristic, which has been previously applied to the case of inferring the statistical expectation, to estimate generic QoIs. The computation of the information gain in a hypothetical experiment is done using numerical approximation via sample averaging. This is done by averaging over samples of the QoI and a hypothetical value of the physical response at the hypothetical design, both of which are obtained using the probabilistic surrogate model. Surrogate models, commonly vanilla Gaussian process (GP) based, often fall short in capturing inconspicuous characteristics of the underlying physical process such as spatially-varying smoothness, discontinuities and hetereoscedastic noise. We model the black-box physical response as fully-Bayesian

non-stationary GP (FBNSGP) probabilistic models. This FBNSGP model does not require one to have strong assumptions on the smoothness and scale of the underlying function as it infers local estimates of these properties as functions of the input. This NSGP model does not require one to have strong assumptions on the smoothness and scale of the underlying function. We demonstrate the performance of the BODE methodology on four numerical examples and a practical engineering problem of steel wire manufacturing. The final BODE algorithm is tested for convergence and analyzed on different criteria like the final set of designs obtained, the QoI, etc. We make comparisons to two traditional methods used in SDOE namely, expected improvement (EI), and uncertainty sampling (US).

5.1 Introduction

Researchers and scientists often simulate real-world physical phenomena as computer codes [11] or laboratory experiments [79]. The use of sophisticated mathematical models or advanced laboratory equipment makes the simulators near precise. This accuracy comes at either a huge computational cost attached to the computer code or painful logistic overheads with the laboratory experiment. As a result, the number of simulations or experiments that can be queried is finite. Another important facet of such expensive codes is that they are almost always handled in a non-intrusive manner. This is because most of these codes and test-rigs, in research groups and laboratories across the globe, are outcomes of years of expertise and insight. So, the task at hand is to optimally allocate one's available budget while acquiring information about a quantity of interest (QoI) resulting from the simulation or experiment. The researcher might be interested in augmenting their state-of-knowledge about statistics of the physical response being simulated, like the maximum value of the physical response [11, 17, 18], the statistical expectation or other percentiles of the code output [132], etc.

When applied to black-box expensive simulators, traditional design of experiments (DOE) methods usually face two major hurdles: a) they require hundreds of thousands of expensive simulations, and b)they require gradients of the simulator. Modern Bayesian methods, in the paradigm of Bayesian Optimal Design of Experiments (BODE), used in the design of computer experiments require few evaluations of the simulator and are gradient-free. Among BODE methods, sequential design of experiments (SDOE) methods [9, 14, 15, 30, 93, 94, 133] like expected improvement (EI), probability of improvement (PI), uncertainty sampling (US), random sampling (RS), entropy search, are well-suited to task of designing experiments under a limited budget. However, there are very few methods that are agnostic to the context of the SDOE problem. For example, the expected improvement (EI) [32, 134, 135] is used when the designer is interested in the optimal value of the parameter or design. The uncertainty sampling (US) [101] is a BODE heuristic derived from maximizing the information gain in the parameters. In practice US sequentially reduces the epistemic uncertainty in the physical response and results in learning the response surface.

The common SDOE methods use probabilistic surrogate models [136, 137], called emulators, to model the expensive simulator. Probabilistic predictions of the physical response from this surrogate model are used to compute an information gain criterion arbitrary design or experiment possesses. The solution of this optimization problem is the design at which the experiment or simulation is to be performed. This sequence continues until one exhausts the available budget unless a predefined stopping criterion is met. A common assumption made by surrogates, like commonly used Gaussian Process Regression (GPR) [122], about the physical response is that it has constant smoothness and constant signal strength across the input domain. This assumption can have ill-effects on the modeling and the subsequent sequential design process when the underlying function has discontinuities or sharp changes in smoothness as shown in [138, 139].

In recent years, the use of the KLD has been extended and demonstrated on various applications including the sensor placement problem [96, 110, 140], surrogate

modeling [111–113], learning missing parameters [114], optimizing an expensive physical response [18], calibrating a physical model [115, 116], reliability design [117], efficient design space exploration [119], probabilistic sensitivity analysis [120], portfolio optimization [3], neural-network hyperparameter tuning [121] and human experiment design [141].

In this chapter, we extend the Kullback-Leibler divergence (KLD) [55, 108] to quantify the information gain in the QoI to select the next experiment or design. The information gain is the KLD from the posterior probability distribution to the prior probability distribution on the statistical QoI. The expected information gain is made numerically tractable using Monte Carlo (MC) sample averaging. This MC estimate of the information gain is called the Expected KLD (EKLD) throughout this paper. In addition to this, we define a fully Bayesian non-stationary Gaussian process (FBNSGP) surrogate, building on the work done in [138, 139, 142, 143], to model the physical response. This surrogate model allows the scientist's expertise to be incorporated, as prior knowledge, in the surrogate model at the highest hierarchical level of the parameterization. Previous work done in developing non-stationary emulators for such problems include the Treed GP model of Grammacy et. al. [144], the GP-experts based model of Rasmussen et.al. [145], point estimates of local smoothness based non-stationary GP modeling by [146] and [147].

The aim of this work is two-fold: a) to examine the performance of the EKLD in SDOE for arbitrary statistics of the black-box code output, b) to compare and contrast the convergence of the EKLD to state-of-the-art methods. We comment on the pros and cons of the derived estimator of the information gain qualitatively and quantitatively in the sections that follow. A free version of the implementation of the methodology in the PYTHON programming language is under preparation.

We perform experiments on synthetic examples with varying characteristics such as the no. of input dimensions, no. of initial samples, and levels of smoothness. Comparison to state-of-the-art BODE methods namely, US and EI, are presented

in different contexts. We apply the proposed information gain based BODE on a problem of steel wire manufacturing.

The chapter is organized as follows: The details of our methodology can be found in 5.2, our main results on synthetic examples are in 5.3, comparison studies are in 5.4, and the conclusions follow in 5.6.

5.2 Methodology

Consider the following situation faced by a scientist working on a laboratory experiment or an expensive black-box computer simulation:

1. Set of n conditions or inputs, i.e., $\mathbf{X}_n$.
2. Set of n outputs, i.e., $\mathbf{Y}_n$.
3. Data set $\mathbf{D}_n = \{\mathbf{X}_n, \mathbf{Y}_n\}$.
4. A plausible next experiment or simulation, i.e., $\tilde{\mathbf{x}}$.
5. A plausible observation at $\tilde{\mathbf{x}}$, i.e., $\tilde{y}$.

Let $\mathbf{x}$ be a random variable with probability density function (PDF) $p(\mathbf{x})$. Without loss of generality, we will assume that $p(\mathbf{x})$ is the uniform PDF supported on the d-hypercube $\mathcal{X} = \times_{k=1}^{d}[0,1]$. The true physical response f is assumed to be a squared integrable function of $\mathbf{x} \in \mathcal{X}$, i.e., $f \in \mathcal{L}^2(\mathcal{X})$, where

$$\mathcal{L}^2(\mathcal{X}) = \left\{ f : \mathcal{X} \to \mathbb{R} \middle| \int_{\mathcal{X}} f^2(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} < \infty \right\}. \tag{5.1}$$

Mathematically, we define a QoI for our purposes as:

$$\boldsymbol{Q} : \mathcal{L}^2 \to \mathbb{R} \tag{5.2}$$

for example,

$$\boldsymbol{Q}[f] = \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x}, \tag{5.3}$$

the statistical expectation which happens to be a bounded linear functional. In this work, we aim to estimate $\boldsymbol{Q}[f]$ where $\boldsymbol{Q}[\cdot]$ is any operator on our function $f(\cdot)$. We define the restricted input space as, $\mathcal{X} = \times_{k=1}^{d}[\mathbf{x}_{u,k}, \mathbf{x}_{l,k}]$. Without loss of generality, the same becomes $\mathcal{X} = \times_{k=1}^{d}[0, 1]$ in this paper.

At each stage of SDOE, we will update our beliefs about $\boldsymbol{Q}[f]$ in a Bayesian way [148], quantifying the epistemic uncertainty induced by limited data at the same time. We will select the new experiment by maximizing the expected information gain for $\boldsymbol{Q}[f]$.

5.2.1 Surrogate Modeling

Numerous surrogate modeling techniques are used in problems where a black-box expensive function is used to model a physical process. For a comprehensive perspective, on such techniques, we refer the readers to [43, 136, 149]. Gaussian process (GP) regression [43, 150] is a commonly used non-parametric probabilistic surrogate modeling technique. Recent advances in GP regression have seen methods to tackle input-dependent noise [16], incorporating local properties while modeling smoothness [146, 147, 151] of the underlying physical response.

GPs that incorporate effects of local features are broadly known as non-stationary GPs (NSGP). This is because a GP, with a zero mean, can be fully specified by a covariance function which defines a functional relationship between the responses across the input domain. Vanilla implementations of GPs often use a stationary kernel to define the covariance function. Commonly used kernel functions in GPs are Radial Basis Function (RBF) kernels, Matern class of kernels, Exponential, etc. For a detailed discussion of covariance kernels we refer the reader to Chapters 4 and 5 of [43].

In NSGP surrogate models, the major difference is the choice of the so-called non-stationary covariance kernel. A stationary kernel is one whose functional formulation has only the Eucledian distance between two inputs conditioned on a single set of constant hyperparameters. However, in a non-stationary kernel, the hyperparameters themselves become functions of the input values. Thus, with a non-stationary kernel which has all the properties required to simulate a GP, see Chapter 4 of [43], one can recover the same prior and posterior beliefs over the black-box function as in the case of stationary GPs. In this work, we model the black-box function as a fully-Bayesian NSGP.

We perform fully-Bayesian inference using Hamiltonian Monte Carlo (HMC) [152–154] to generate samples from the posterior distribution of the hyperparameters of the NSGP model. The application of HMC has been demonstrated in the works of Kramer et. al. [155] and Girolami et. al. [156]. Recent advancements in HMC methods include the works of [157, 158]. The following sections have the necessary details of our NSGP modeling process.

Modeling prior beliefs

We represent our prior beliefs about the black-box function as a zero mean GP which, mathematically, corresponds to the following:

$$f \sim GP(0, k) \tag{5.4}$$

Since a GP can be fully specified by a covariance kernel we resort to incorporating the local properties of the input domain in the functional form of the stationary covariance kernel used in the vanilla GP. Building on the work done in [146,147,159] we model this

spatial covariance with the functional form used by Heinonen et. al. [138]. This results in a covariance function that has the following form for a d-dimensional input-space:

$$k(\mathbf{x}, \mathbf{x}^{'}) = k(\mathbf{x}, \mathbf{x}^{'}) = \prod_{i=1}^{d} s_i(x_i) s_i(x_i^{'}) \sqrt{\frac{l_i(x_i) l_i(x_i^{'})}{l_i^2(x_i) + l_i^2(x_i^{'})}} \exp\left(- \frac{(x_i - x_i^{'})^2}{l_i^2(x_i) + l_i^2(x_i^{'})} \right). \quad (5.5)$$

The covariance kernel in Eq. (5.5) has point estimates of the lengthscale and the signal-strength in each dimension. Thus, having the ability to model input-dependent smoothness and variance. In our work we model the $s_i(\cdot)$s and $l_i(\cdot)$s in Eq. (5.5) as functions of the inputs by modeling the logarithms of $s_i(\cdot)$s and $l_i(\cdot)$s as stationary GPs. The hyperparameters of these GPs have prior probability distributions, usually non-informative, that allow the scientist to encode specific information about the function at the root of the model. Mathematically, this means:

$$\log s_i \sim GP(\hat{m}_{s,i}(\cdot), \hat{k}_{s,i}(\cdot, \cdot)) \quad (5.6)$$

and,

$$\log l_i \sim GP(\hat{m}_{l,i}(\cdot), \hat{k}_{l,i}(\cdot, \cdot)) \quad (5.7)$$

The covariance kernel of these GPs has the following form:

$$\hat{k}(x, x^{'}) = \hat{k}(x, x^{'}; \hat{\boldsymbol{\psi}}) = \hat{s}^2 \exp - \frac{(x - x')^2}{2\hat{l}^2}. \quad (5.8)$$

The kernel in Eq. (5.8) has two hyperparameters, i.e. $\hat{s}$ and $\hat{l}$. These hyperparameters can either be estimated at the beginning of the algorithm or can be optimized over a range of values via grid search. Differentiating our work from that of Heinonen et. al. [138], we employ a fully-Bayesian Hamiltonian Monte Carlo (FBHMC) [154] based scheme to obtain samples from the posterior of these hyperparameters. Thus, the hyperparameters at the root of the surrogate model can be chosen according to the prior beliefs of the scientist. The hyperparameters of each kernel and the mean function modeling the signal-strength GP and the lengthscale GP will be

denoted by the vector $\hat{\boldsymbol{\psi}}_{s,i}$ and $\hat{\boldsymbol{\psi}}_{l,i}$ respectively for the ith input dimension. The FBHMC also includes inferring the latent values of training points for these latent GPs modeling the input dependent signal-strength and lengthscale in each dimension. Thus, we denote all the hyperparameters of our NSGP model at any stage by $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = \{(\hat{\boldsymbol{\psi}}_{s,1}, \hat{\boldsymbol{\psi}}_{l,i}), \cdots, (\hat{\boldsymbol{\psi}}_{s,d}, \hat{\boldsymbol{\psi}}_{l,d})\}$. More details of the HMC sampling can be found in [154]. We shall describe in detail how prior beliefs, called hyperpriors, on all the hyperparameters in $\boldsymbol{\theta}$ are chosen in later sections.

Modeling observed data

The likelihood probability of the observed data $\mathbf{Y}_n$ is modeled as a multivariate Gaussian, the mean of which is the vector of function values $\mathbf{f}_n = f(\mathbf{x}_1, \cdots, f(\mathbf{x}_n))$. The covariance can be computed using the formulation in Eq. (5.5). The observations are presumed to be contaminated with noise denoted by σ^2. In this work, we assume the computer simulations to be very accurate and fix the value of σ^2 to 1e-6. The likelihood can be represented as follows:

$$p(\mathbf{Y}_n|\mathbf{X}_n;\boldsymbol{\theta}) = \mathcal{N}(\mathbf{Y}_n|0, \mathbf{K}_n + \sigma^2 I_n), \tag{5.9}$$

where an element of $\mathbf{K}_n$ is obtained as $K_{nij} = k(\mathbf{x}_i, \mathbf{x}_j)$ based on Eq. (5.5).

Inferring hyperparameters

The structure of NSGP modeling described in Sec. 5.2.1 requires one to simulate values of $\boldsymbol{\theta}$ from the posterior state of knowledge on $\boldsymbol{\theta}$. Unfortunately, this is not possible directly because the posterior of $\boldsymbol{\theta}$ is known only upto a proportionality constant as:

$$p(\boldsymbol{\theta}|\mathbf{Y}_n, \mathbf{X}_n) \propto p(\mathbf{Y}_n|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{5.10}$$

This situation is frequently encountered in surrogate modeling and is approached through different techniques depending on the scenario. Since we approach the SDOE

problem in the low-sample regime, it makes complete sense to infer $\boldsymbol{\theta}$ using a Bayesian approach. This often corresponds to obtaining samples from the stationary distribution of $\boldsymbol{\theta}$ using Markov chain Monte Carlo methods. One such method that has gained popularity owing to its fast convergence rates is Hamiltonian Monte Carlo (originally Hybrid Monte Carlo). We use the methodology implemented by [154], available as open source software, to model the NSGP for this work. More development on HMC methods is an ongoing process with more intelligent versions being proposed in recent times.

Hyperpriors

Prior probability distributions $p(\boldsymbol{\theta})$ on the hyperparameters $\boldsymbol{\theta}$ are chosen to be largely uninformative in our work. For all the latent GPs we place a $\mathcal{G}(1,1)$, Gamma distribution, prior on their lengthscales and the signal-strengths.

The process of selecting a combination of hyperpriors includes choosing between a set of hyperpriors for each mean constant. The subtle differences arise mainly due to the dimensionality of the problem. This aspect is elucidated further in Sec. 5.3.

Making predictions

Conditioned on the hyperparameters, our state of knowledge about f is also characterized by a GP:

$$f|\mathbf{D}_n, \boldsymbol{\theta} \sim \mathrm{GP}(f|m_n, k_n), \tag{5.11}$$

where

$$m_n(\mathbf{x}) = (\mathbf{k}_n(\mathbf{x}))^T \left(\mathbf{K}_n + \sigma^2 \mathbf{I}_n\right)^{-1} \mathbf{Y}_n, \tag{5.12}$$

with

$$\boldsymbol{\alpha}_n = (\mathbf{K}_n + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{Y}_n, \tag{5.13}$$

is the *posterior mean* function, and

$$k_n(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - (\mathbf{k}_n(\mathbf{x}))^T \left(\mathbf{K}_n + \sigma^2 \mathbf{I}_n\right)^{-1} \mathbf{k}_n(\mathbf{x}'), \tag{5.14}$$

with $\mathbf{k}_n(\mathbf{x}) = (k(\mathbf{x}, \mathbf{x}_1), \ldots, k(\mathbf{x}, \mathbf{x}_n))^T$, is the *posterior covariance* function. In particular, at an untried design point $\tilde{\mathbf{x}}$ the point-predictive posterior probability density of $\tilde{y} = f(\tilde{\mathbf{x}})$ conditioned on the hyperparameters is:

$$p(\tilde{y}|\tilde{\mathbf{x}}, \mathbf{D}_n, \boldsymbol{\theta}) = \mathcal{N}\left(\tilde{y}\middle|m_n(\tilde{\mathbf{x}}; \boldsymbol{\theta}), \sigma_n^2(\tilde{\mathbf{x}}; \boldsymbol{\theta})\right), \tag{5.15}$$

where $\sigma_n^2(\tilde{\mathbf{x}}; \boldsymbol{\theta}) = k_n(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}; \boldsymbol{\theta})$. Finally, the *point-predictive posterior* PDF of $\tilde{y} = f(\tilde{\mathbf{x}})$ is:

$$p(\tilde{y}|\tilde{\mathbf{x}}, \mathbf{D}_n) = \int p(\tilde{y}|\tilde{\mathbf{x}}, \mathbf{D}_n, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{D}_n) d\boldsymbol{\theta}. \tag{5.16}$$

5.2.2 Karhunen-Loève expansion of a NSGP

We seek a Karhunen-Loève expansion (KLE) of the posterior NSGP to obtain samples of f. Eventually each of these samples of f will contribute to the quantification of our state of knowledge about the QoI, $\boldsymbol{Q}[\cdot]$, that we seek to infer. The KLE [160] of the posterior NSGP is:

$$f(\mathbf{x}; \boldsymbol{\xi}, \boldsymbol{\theta}) = m_n(\mathbf{x}; \boldsymbol{\theta}) + \sum_{i=1}^{\infty} \xi_i \sqrt{\lambda_{n,i}} \phi_{n,i}(\mathbf{x}; \boldsymbol{\theta}), \tag{5.17}$$

where $m_n(\mathbf{x}; \boldsymbol{\theta})$ is simply the posterior predictive mean of the function, Eq. (5.15). The random variables $\boldsymbol{\xi}$ are independent identically distributed (iid) standard normal. We truncate Eq. (5.17) at order W,

$$f(\mathbf{x}; \boldsymbol{\xi}, \boldsymbol{\theta}) \approx m_n(\mathbf{x}; \boldsymbol{\theta}) + \sum_{i=1}^{W} \xi_i \sqrt{\lambda_{n,i}} \phi_{n,i}(\mathbf{x}; \boldsymbol{\theta}). \tag{5.18}$$

The scalars $\lambda_{n,i}$'s and functions $\phi_{n,i}(\mathbf{x};\boldsymbol{\theta})$s are the eigenvalues and eigenfunctions of the posterior covariance function constructed using the quadrature given in Section 2.8.1 of [161]. More details about the same can be found in [161]. The number of terms, W, is determined by specifying the percentage β of the total sum of the eigenvalues to be retained as follows:

$$\sum_{i=1}^{W}\lambda_{n,i} = \beta\sum_{i=1}^{\infty}\lambda_{n,i}. \tag{5.19}$$

The use of KLE is motivated by its optimality in the mean squared sense [162, 163]. This means that the truncated KLE converges in $\mathcal{L}^2(\mathcal{X})$ as $W \to \infty$. Intuitively, the more correlated the physical response, the fewer the number of non-zero eigenvalues, $\lambda_{n,i}$s in Eq. (5.18), required.

For our experiments we take the value of β equal to 0.95, i.e., the truncated KLE explains 95% of the total variance of the posterior GP.

Conditioning the state-of-knowledge on a hypothetical observation

Now, consider an untried design $\tilde{\mathbf{x}}$ and a hypothetical observation at $\tilde{\mathbf{x}}$ denoted by $\tilde{y}$. The point distribution of $\tilde{y}$ conditioned on $\boldsymbol{\xi}$ is a Gaussian distribution with the noise variance σ^2 of the NSGP.

$$p(\tilde{y}|\tilde{\mathbf{x}},\boldsymbol{\xi},\mathbf{D}_n;\boldsymbol{\theta}) = \mathcal{N}(\tilde{y}|m_n(\tilde{\mathbf{x}};\boldsymbol{\theta}) + \sum_{i=i}^{W}\xi_i\sqrt{\lambda_{n,i}}\phi_{n,i}(\tilde{\mathbf{x}};\boldsymbol{\theta}),\sigma^2). \tag{5.20}$$

Deriving the posterior of $\boldsymbol{\xi}$ conditional on the $\mathbf{D}_n$, $\tilde{\mathbf{x}}$ and $\tilde{y}$ by completing the squares, results in the following:

$$\begin{aligned} p(\boldsymbol{\xi}|\tilde{\mathbf{x}},\tilde{y},\mathbf{D}_n;\boldsymbol{\theta}) &\propto p(\tilde{y}|\boldsymbol{\xi},\tilde{\mathbf{x}},\mathbf{D}_n;\boldsymbol{\theta})p(\boldsymbol{\xi}) \\ \Rightarrow p(\boldsymbol{\xi}|\tilde{\mathbf{x}},\tilde{y},\mathbf{D}_n;\boldsymbol{\theta}) &= \mathcal{N}(\boldsymbol{\xi}|\tilde{\boldsymbol{\mu}}_p,\tilde{\boldsymbol{\Sigma}}_p), \end{aligned} \tag{5.21}$$

where,

$$\tilde{\boldsymbol{\mu}}_p = \tilde{\boldsymbol{\Sigma}}_p \mathbf{a}^T \left(\frac{\tilde{y} - m_n(\tilde{\mathbf{x}})}{\sigma^2} \right), \tag{5.22}$$

$$\tilde{\boldsymbol{\Sigma}}_p = \left(\mathbf{I}_W + \mathbf{a}^T \mathbf{a} \frac{1}{\sigma^2} \right)^{-1}, \tag{5.23}$$

with,

$$\mathbf{a} = \left[\sqrt{\lambda_{n,1}} \phi_{n,1}(\mathbf{x}; \boldsymbol{\theta}), \cdots, \sqrt{\lambda_{n,W}} \phi_{n,W}(\mathbf{x}; \boldsymbol{\theta}) \right], \tag{5.24}$$

The Sherman-Morrison formula [164] allows us to express the posterior covariance of $\boldsymbol{\xi}$ from Eq. (5.23) as:

$$\tilde{\boldsymbol{\Sigma}}_p = \mathbf{I}_W - \frac{\mathbf{a}^T \mathbf{a}}{\sigma^2 + \mathbf{a}\mathbf{a}^T}, \tag{5.25}$$

where $\mathbf{a}$ is a row vector as defined in Eq. (5.24). An element of the posterior covariance matrix of $\boldsymbol{\xi}$ can be expressed as:

$$\tilde{\boldsymbol{\Sigma}}_{p,ij} = \delta_{ij} - \frac{\sqrt{\lambda_{n,i}} \sqrt{\lambda_{n,j}} \phi_{n,i}(\mathbf{x}) \phi_{n,j}(\mathbf{x})}{\sigma^2 + \sum_{i=1}^{W} (\sqrt{\lambda_{n,i}} \phi_{n,i}(\mathbf{x}))^2}, \tag{5.26}$$

where δ_{ij} is the Kronecker delta. Simplifying the notation, we define:

$$\tilde{\boldsymbol{\mu}}_p = \tilde{\boldsymbol{\mu}}_{cp} \left(\frac{\tilde{y} - m_n(\tilde{\mathbf{x}}; \boldsymbol{\theta})}{\sigma^2} \right), \tag{5.27}$$

where,

$$\tilde{\boldsymbol{\mu}}_{cp} = \tilde{\boldsymbol{\Sigma}}_p \mathbf{a}^T. \tag{5.28}$$

Obtaining samples of the QoI

A sample of $\boldsymbol{\xi}$ from the independent multivariate normal distribution allows one to sample $f|\mathbf{D}_n; \boldsymbol{\theta}$ using the truncated expansion in Eq. (5.18). This is all we need to obtain a priori samples of the $\boldsymbol{Q}[f]$.

With a sample of $\boldsymbol{\xi}$ from $p(\boldsymbol{\xi}|\tilde{\mathbf{x}}, \tilde{y}, \mathbf{D}_n; \boldsymbol{\theta})$ one can sample $f|\mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}; \boldsymbol{\theta}$ using the same construction as in Eq. (5.18) albeit with posterior $\boldsymbol{\xi}$. This allows one to obtain a posteriori samples of the $\boldsymbol{Q}[f]$. These a priori and a posteriori samples of the $\boldsymbol{Q}[f]$ will be used in the following section while estimating the information gain in the $\boldsymbol{Q}[f]$.

5.2.3 Sequential design of experiments using the expected information gain

Given the observed data, $\mathbf{D}_n$, our state of knowledge about the QoI $\boldsymbol{Q}[f]$ can be written as follows:

$$p(Q|\boldsymbol{\theta}, \mathbf{D}_n) = \mathbb{E}\left[\delta\left(Q - \boldsymbol{Q}[f]\right)|\boldsymbol{\theta}, \mathbf{D}_n\right], \tag{5.29}$$

where δ is Dirac's delta function and the expectation is over the function space measure defined by the posterior GP, see Eq. (5.11). The uncertainty in $p(Q|\mathbf{D}_n)$ represents our epistemic uncertainty induced by the limited number of observations in $\mathbf{D}_n$. Now consider the hypothetical output $\tilde{y}$ at a hypothetical experiment $\tilde{\mathbf{x}}$. With the posterior GP measure denoted by $p(Q|\mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y})$, our state of knowledge becomes:

$$p(Q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}) = \mathbb{E}\left[\delta\left(Q - \boldsymbol{Q}[f]\right)|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}\right]. \tag{5.30}$$

The information gained through the hypothetical experiment $(\tilde{\mathbf{x}}, \tilde{y})$ conditioned on the hyperparameters, say $G(\tilde{\mathbf{x}}, \tilde{y}; \boldsymbol{\theta})$ is given by the KLD between $p(Q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y})$ and $p(Q|\boldsymbol{\theta}, \mathbf{D}_n)$. Mathematically, it is:

$$G(\tilde{\mathbf{x}}, \tilde{y}; \boldsymbol{\theta}) = \int_{-\infty}^{\infty} p(Q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}) \log \frac{p(Q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y})}{p(Q|\boldsymbol{\theta}, \mathbf{D}_n)} dQ. \tag{5.31}$$

The expected information gain of the hypothetical experiment, say $G(\tilde{\mathbf{x}})$, is obtained by taking the expectation of $G(\tilde{\mathbf{x}}, \tilde{y})$ over our current state of knowledge. Specifically,

$$G(\tilde{\mathbf{x}}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G(\tilde{\mathbf{x}}, \tilde{y}; \boldsymbol{\theta}) p(\tilde{y}|\boldsymbol{\theta}, \tilde{\mathbf{x}}, \mathbf{D}_n) p(\boldsymbol{\theta}|\mathbf{D}_n) d\tilde{y} d\boldsymbol{\theta}. \tag{5.32}$$

The next experiment or simulation is selected by solving:

$$\mathbf{x}_{n+1} = \arg\max_{\tilde{\mathbf{x}}} G(\tilde{\mathbf{x}}). \tag{5.33}$$

Quantification of the current state of knowledge about QoI

We now derive a sample averaged (SA) estimator of our current state of knowledge about the QoI, i.e., $p(Q|\boldsymbol{\theta}, \mathbf{D}_n)$. Since the QoI, Eq. (5.3), is not always linear we resort to a Gaussian approximation to our state of knowledge about Q. This means that going forward we will approximate $p(Q|\boldsymbol{\theta}, \mathbf{D}_n)$ and $p(Q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y})$ as Gaussian distributions. If $p(Q|\boldsymbol{\theta}, \mathbf{D}_n)$ is approximated by a Gaussian distribution $\mathcal{N}(\mu_1, \sigma_1^2)$ then we can write:

$$p(Q|\boldsymbol{\theta}, \mathbf{D}_n) \approx \mathcal{N}\left(Q|\mu_1, \sigma_1^2\right). \tag{5.34}$$

Samples of Q, denoted by q, can be easily obtained by substituting samples of $\boldsymbol{\xi}$ in Eq. (5.18). Using these samples, the unbiased estimators for the mean and variance of the Gaussian approximation can be written as:

$$\mu_1 = \frac{1}{M} \sum_{i=1}^{M} q_i, \tag{5.35}$$

and,

$$\sigma_1^2 = \frac{1}{M-1} \sum_{i=1}^{M} (q_i - \mu_1)^2, \tag{5.36}$$

where q_is are samples of Q from $p(Q|\boldsymbol{\theta}, \mathbf{D}_n)$.

Quantification of the hypothetical state of knowledge about QoI

Proceeding with the same spirit as above we derive an estimator of our hypothetical state of knowledge about the QoI, i.e., $p(Q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y})$, we get:

$$\mu_2(\tilde{\mathbf{x}}, \tilde{y}) \;=\; \tfrac{1}{M}\sum_{i=1}^{M} q_i, \tag{5.37}$$

and the variance $\sigma_2^2(\tilde{\mathbf{x}}, \tilde{y})$ is given by:

$$\sigma_2^2(\tilde{\mathbf{x}}, \tilde{y}) \;=\; \tfrac{1}{M-1}\sum_{i=1}^{M}(q_i - \mu_2(\tilde{\mathbf{x}}, \tilde{y}))^2, \tag{5.38}$$

where q_is are M samples of Q from $p(Q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y})$. These samples of Q can be easily obtained by substituting samples of $\boldsymbol{\xi}$ from its posterior distribution i.e. Eq. (5.21) in Eq. (5.20).

Quantification of the expected information gain about the QoI

Since both $p(Q|\boldsymbol{\theta}, \mathbf{D}_n)$ and $p(Q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y})$ are Gaussian, the KL divergence between the hypothetical and the current state of knowledge about the QoI conditional on the hyper-parameters, $G(\mathbf{x}, \tilde{y}; \boldsymbol{\theta})$ of Eq. (5.31), has a functional form [126], i.e.,

$$G(\tilde{\mathbf{x}}, \tilde{y}; \boldsymbol{\theta}) \;=\; \log\left(\tfrac{\sigma_1}{\sigma_2(\tilde{\mathbf{x}}, \tilde{y})}\right) + \tfrac{\sigma_2^2(\tilde{\mathbf{x}}, \tilde{y})}{2\sigma_1^2} + \tfrac{(\mu_2(\tilde{\mathbf{x}}, \tilde{y}) - \mu_1)^2}{2\sigma_1^2} - \tfrac{1}{2}. \tag{5.39}$$

Since $p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\theta}, \mathbf{D}_n)$ is Gaussian, see Eq. (5.15) we can sample average out $\tilde{y}$ to obtain:

$$\begin{aligned} G(\tilde{\mathbf{x}}; \boldsymbol{\theta}) \;&=\; \textstyle\int_{-\infty}^{\infty} G(\tilde{\mathbf{x}}, \tilde{y}; \boldsymbol{\theta}) p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\theta}, \mathbf{D}_n) d\tilde{y} \\ &\approx\; \tfrac{1}{B}\textstyle\sum_{b=1}^{B} G(\tilde{\mathbf{x}}, \tilde{y}^b; \boldsymbol{\theta}), \end{aligned} \tag{5.40}$$

Finally, we take the expectation of $G(\tilde{\mathbf{x}};\boldsymbol{\theta})$ over the posterior of the hyperparameters, $p(\boldsymbol{\theta}|\mathbf{D}_n)$ of Eq. (5.10), using the MCMC samples $\left\{\boldsymbol{\theta}^{(s)}\right\}_{s=1}^{S}$ collected with the procedure described in [154]. This yields:

$$\begin{aligned} G(\tilde{\mathbf{x}}) &\approx \int G(\tilde{\mathbf{x}};\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{D}_n) d\boldsymbol{\theta} \\ &\approx \frac{1}{S}\sum_{s=1}^{S} G\left(\tilde{\mathbf{x}};\boldsymbol{\theta}^{(s)}\right). \end{aligned} \tag{5.41}$$

For most of our numerical experiments we fix the values of M, B and S to 50 each.

Maximizing the EKLD defined in Eq. (5.33) might ideally need a multi-start-optimization algorithm, but we resort to a Bayesian global optimization algorithm (see Algorithm 5), same as in Chapter 4 to maximize the EKLD. In our experiments, we use $T_n = 30$ BGO iterations to optimize the EKLD for one-dimensional functions. For multi-dimensional functions $T_n = 40$ BGO iterations are used to optimize the EKLD.

5.2.4 Quantities of interest

The framework described in the previous sections is applied to infer the following statistics, $\boldsymbol{Q}[f]$, of the function:

1. statistical expectation denoted by $\mathbb{E}[f]$
2. statistical variance denoted by $\mathbb{V}[f]$
3. kth percentile denoted by $\boldsymbol{P}_k[f]$
4. maximum scalar value denoted by $\max[f]$
5. minimum scalar value denoted by $\min[f]$

The methodology is compared to the uncertainty sampling (US) for the first three $\boldsymbol{Q}[f]$s and compared to the classic expected improvement (EI) for the last two $\boldsymbol{Q}[f]$s.

Algorithm 5 Optimize the EKLD using BGO with AEI.

Require: Initial number of EKLD evaluations T_i; maximum number of EKLD evaluations T_n; number of candidate designs n_d for BGO; MCMC samples from the posterior of the hyperparameters $\left\{\boldsymbol{\theta}^{(s)}\right\}_{s=1}^{S}$; stopping tolerance $\gamma_i > 0$.

Evaluate $G(\tilde{\mathbf{x}})$ using Eq. (5.41) at T_i random points to generate training data, $\tilde{\mathbf{X}}_{T_i} = \{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_{T_i}\}$ and $\mathbf{G}_{T_i} = \left\{\tilde{G}_1 = G(\mathbf{x}_1), \ldots, \tilde{G}_{T_i} = G(\mathbf{x}_{T_i})\right\}$, for BGO.
$t \leftarrow t_i$.
while $t < T_n$ **do**
    Fit a standard GP on the input-output pairs $\tilde{\mathbf{X}}_t$-$\tilde{\mathbf{G}}_t$ using maximum likelihood to approximate $G(\tilde{\mathbf{x}})$.
    Generate a set of candidate test points $\hat{\mathbf{X}}_{n_d} = \{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_{n_d}\}$ using Latin Hypercube Sampling (LHS) [55].
    Compute the AEI of all of the candidate points in $\hat{\mathbf{X}}_{n_d}$.
    Find the candidate point $\hat{\mathbf{x}}_j$ that exhibits the maximum AEI.
    if If the maximum AEI is smaller than γ_i **then**
        Break.
    end if
    Use Eq. (5.41) to evaluate $G(\tilde{\mathbf{x}})$ at $\hat{\mathbf{x}}_j$ measuring $\hat{G}_j = G(\hat{\mathbf{x}}_j)$.
    $\tilde{x}_{t+1} \leftarrow \hat{x}_j$.
    $\tilde{G}_{t+1} \leftarrow \hat{G}_j$.
    $\mathbf{X}_{t+1} \leftarrow \tilde{\mathbf{X}}_t \cup \{\tilde{\mathbf{x}}_{t+1}\}$.
    $\mathbf{G}_{t+1} \leftarrow \mathbf{G}_t \cup \{\tilde{G}_{t+1}\}$.
    $t \leftarrow t + 1$.
end while
return $\arg\max_{\tilde{\mathbf{x}}_{T_n}} \tilde{\mathbf{G}}_{T_n}$.

5.3 Numerical Results

We present results for the methodology's performance on two one-dimensional mathematical functions, a three-dimensional problem, and a five-dimensional problem. The input domain for the first two synthetic examples is $[0, 1]$ whereas for the third synthetic example the input domain is $[-2, 6]^3$. The input space for the five dimensional synthetic example no. becomes the hyper-cube $[0, 1]^5$. The number of initial data points is denoted by n_i.

The prior distributions for the one-dimensional functions are chosen as follows: a) the mean function on the log-lengthscale GP is fixed to a negative integer constant.

Algorithm 6 Bayesian optimal design of experiments maximizing the expected information gain a statistical QoI of a physical response.

Require: Initially observed inputs $\mathbf{X}_{n_i}$; initially observed outputs $\mathbf{Y}_{n_i}$; maximum number of allowed experiments N.

1: $n \leftarrow n_i$.
2: **while** $n < N$ **do**
3: Sample from the posterior of the hyperparameters, Eq. (5.10), to obtain $\{\boldsymbol{\theta}^{(s)}\}_{s=1}^{S}$.
4: Find the next experiment $\mathbf{x}_{n+1}$ using Algorithm 5 to solve Eq. (5.33).
5: Evaluate the objective at $\mathbf{x}_{n+1}$ measuring $y_{n+1} = f(\mathbf{x}_{n+1})$.
6: $\mathbf{X}_{n+1} \leftarrow \mathbf{X}_n \cup \{\mathbf{x}_{n+1}\}$.
7: $\mathbf{Y}_{n+1} \leftarrow \mathbf{Y}_n \cup \{y_{n+1}\}$.
8: $t \leftarrow t + 1$.
9: **end while**

The reason behind it is that we wish to encode prior information about the lengthscale taking values as low as of the order of 1e-1. Thus, defining a lower bound or threshold on the point estimates of the lengthscale values. In this work, this constant mean function is fixed at -2 for one-dimensional problems. For higher dimensional problems, this constant is fixed to 0. b) for the signal-strength we choose a Gaussian prior, $\mathcal{N}(0, 4)$, on the mean function of the log-signal-strength GP for the one-dimensional problems. c) for the multi-dimensional problems we fixed the mean function to a value of 0.

Another technique to choose these prior distributions could be the Bayesian information criterion (BIC) [165]. Under the BIC combinations of prior distributions on the hyperparameters are compared against each other based on value of the data likelihood and a penalty criterion which is a function of the number of data and parameters. Since, the number of data and parameters would be the same, the BIC would boil down to maximizing the likelihood. We look to choose the priors based on some basic intuition about GPs and some prior knowledge about the function. The same prior distributions are used for all one-dimensional functions. We do this because we wish to test a set of non-informative priors that can be used across different problems without the need for any user intervention. The number of HMC samples

for each problem is fixed at 11,500, from which the first 1500 samples are discarded. For further details on the HMC part of training the NSGP, we refer the readers to [153, 166].

We mentioned in the previous section that the values of the number of posterior samples of $\boldsymbol{\theta}$, denoted by S, number of samples of $\tilde{y}$ denoted by B and the number of samples of the Q denoted by M are fixed at 50. This is done to ensure a default setting for the different controls of the algorithm irrespective of the function or the QoI being inferred. In some cases, for some QoIs like the estimating the 2.5th percentile, or inferring the minimum or maximum values of a multi-modal function the default settings are changed to obtain smooth convergence results which can be explained better. However, we do ensure the settings remain same for the ELKD and other methods in comparison studies in all cases.

5.3.1 Synthetic example no. 1

Consider the following function:

$$f(x) = 4\left(1 - \sin\left(6x + 8e^{6x-7}\right)\right), \tag{5.42}$$

defined on $[0, 1]$. This function is smooth throughout its domain, but it exhibits two local minima. We will apply our methodology to estimate the QoIs in Sec. 5.2.4 including the case of inferring the 2.5th percentile of the function. The true values of the five $\boldsymbol{Q}[f]$s enumerated in Sec. 5.2.4 are:

1. $\mathbb{E}[f] = -1.36$
2. $\mathbb{V}[f] = 0.30$
3. $\max[f] = -0.40$
4. $\min[f] = -2.00$
5. $\boldsymbol{P}_{97.5}[f] = -1.99$

We apply our methodology to this problem starting from $n_i = 3$ and sample a total of $N = 18$ points.

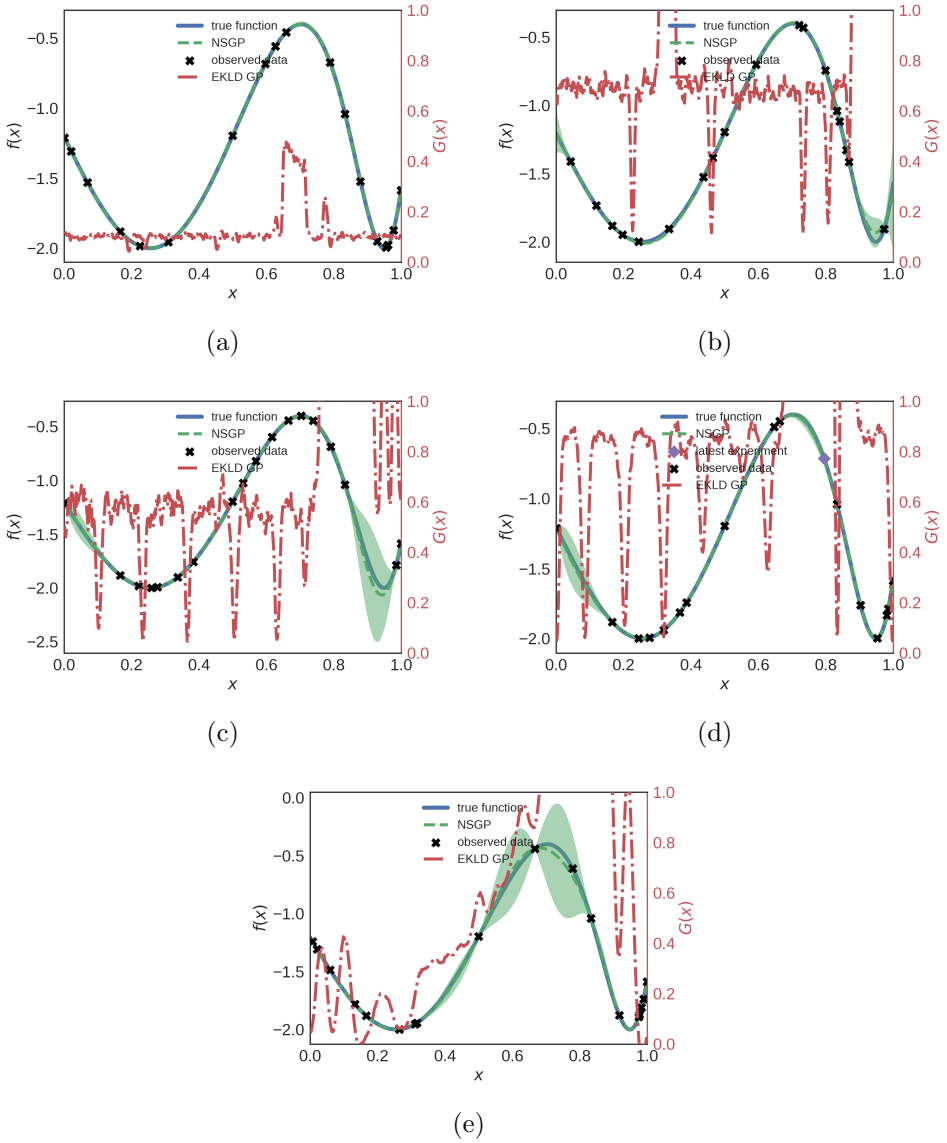


Figure 5.1. One-dimensional synthetic example ($n_i = 3$) shows the state of the function at the end (15th iteration) of the algorithm for inferring: Subfigures (a) $\mathbb{E}[f]$, (b) $\mathbb{V}[f]$, (c) $\max[f]$, (d) $\min[f]$, and (e) $\boldsymbol{P}_{2.5}[f]$.

Fig. 5.1 shows the final state of designs for the different $\boldsymbol{Q}[f]$s. The thick blue line represents the true function f, Eq. (5.42). The black crosses are the observed data at the final stage. In subfigure (a), the next experiment selected by maximizing the

EKLD, see Algorithm 3, corresponds to the purple diamond. The mean of the GP fit to the expected information gain $G(\tilde{\mathbf{x}})$ constructed by BGO in Algorithm 3. The predictive mean of the EKLD is shown by the dotted light blue line. This dotted line represents the response surface of the EKLD after the BGO has ended and the red shaded area around it represents the uncertainty (2.5 percentile and 97.5 percentile) around the mean. As expected, the mean of the EKLD is very small or close to zero at points where experiments have been performed. Thus, the point selected by the methodology (purple diamond) is located in the input space where the EKLD has high mean. The posterior mean of the GP of the black-box function is represented by the dashed bottle-green line. The bottle-green shaded area represents the uncertainty (2.5 percentile and 97.5 percentile) around it.

The inferred lengthscale and signal-strength GPs are shown in 5.3 (b) and (c) respectively for the case of inferring the statistical expectation. These plots simply show the posterior mean using each of the S posterior samples. The lengthscale is larger in the region $[0, 0.6]$ compared to $[0.6, 1]$. This can be understood by comparing the waviness of the function in these regions. Similarly, the inferred signal-strength has higher absolute value corresponding to those taken by f. With limited data, fully-Bayesian HMC allows one to obtain such estimates of uncertainty around the inferred value of lengthscale and signal-strength. Other approaches which come at a lower computational cost, like maximum a posteriori (MAP) or maximum likelihood estimate (MLE) would need a significantly larger number of training points to infer meaningful point estimates for the lengthscale and the signal-strength. In the low-sample regime the MAP estimate for the hyperparameters are prone to mislead the methodology either by selecting a single sample i.e. a local optima of the posterior of the hyperparameters. Multiple optimization restarts for MAP and MLE approaches might be beneficial but only slightly unless the number of restarts is of the order of 100. This usually increases the computational burden without making the solution significantly better. The fully-Bayesian approach used here makes a compelling case

to infer the lengthscale and signal-strength GPs under epistemic uncertainty albeit at a higher computational cost.

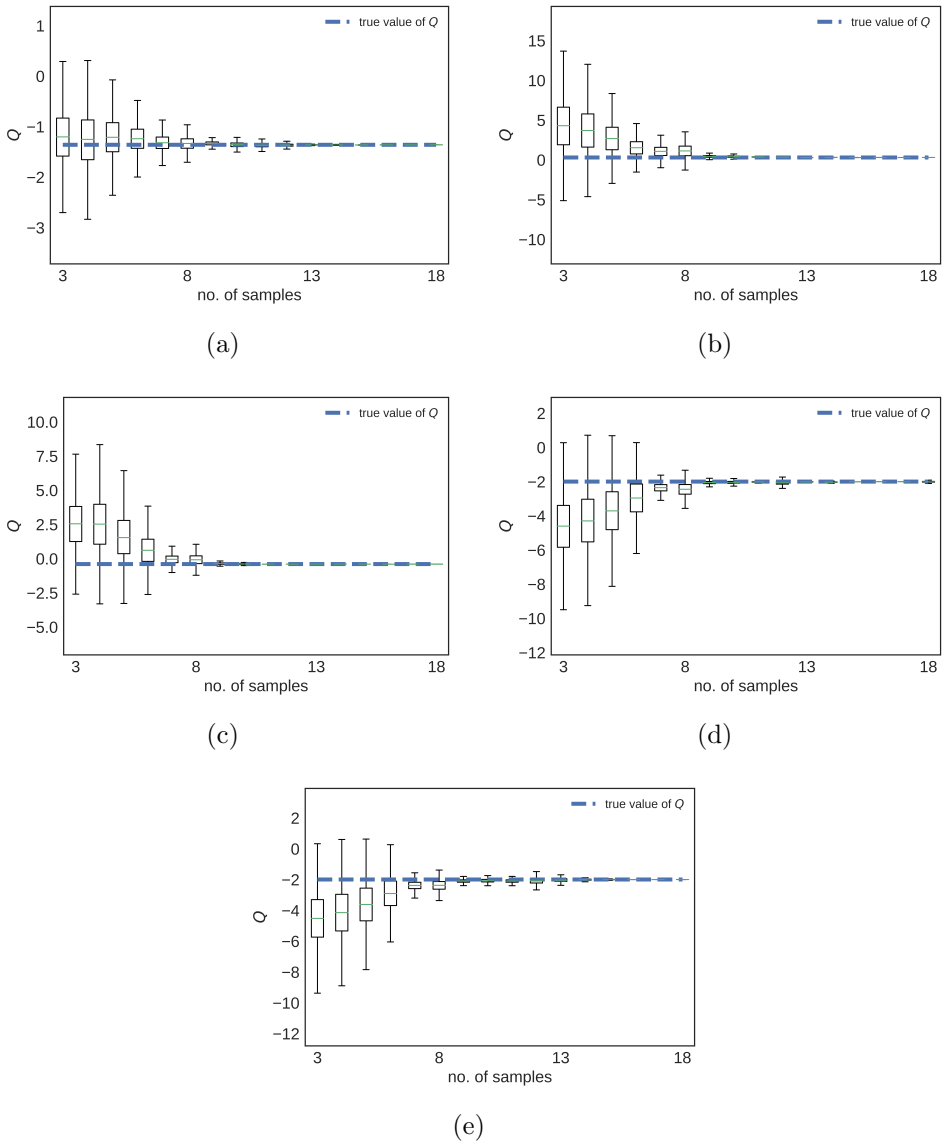


Figure 5.2. One-dimensional synthetic example ($n_i = 3$) shows the convergence of EKLD to the true value of Q for inferring: Subfigures (a) $\mathbb{E}[f]$, (b) $\mathbb{V}[f]$, (c) $\max[f]$, (d) $\min[f]$, and (e) $\boldsymbol{P}_{2.5}[f]$.

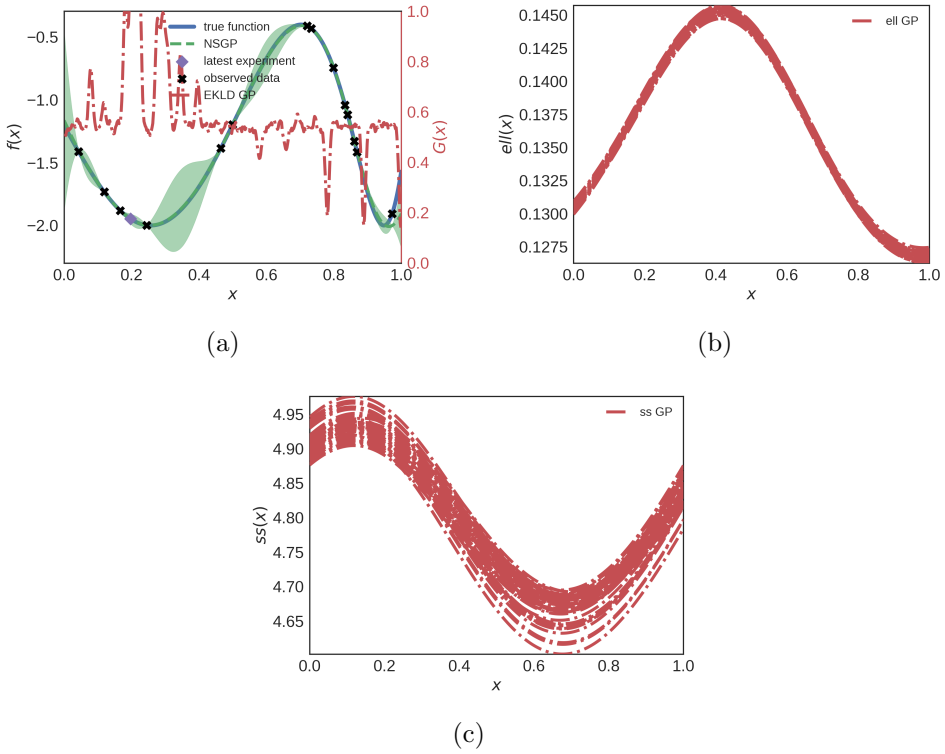


Figure 5.3. One-dimensional synthetic example ($n_i = 3$) shows the statistics of the FBNSGP at the the 12th iteration of the sampling where: Subfigure (a) shows the state of the sampling, (b) shows the inferred point estimates of the lengthscale and (c) shows the inferred point estimates of the signal-strength.

5.3.2 Synthetic example no. 2

We consider the following Gaussian mixture function to test and validate our methodology further.

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}s_1}\exp\left\{-\frac{(x-m_1)^2}{2{s_1}^2}\right\} \\ &\quad + \frac{1}{\sqrt{2\pi}s_2}\exp\left\{-\frac{(x-m_2)^2}{2{s_2}^2}\right\}, \end{aligned} \tag{5.43}$$

where $m_1 = 0.2$ and $s_1 = 0.05$, $m_2 = 0.8$ and $s_2 = 0.05$. As can be seen from Eq. (5.43), the function is a sum of probability densities of two Gaussian distributions. The function has two narrow areas of high magnitude. The true value of the $\boldsymbol{Q}[f]$s are analytically available, and take the following values:

1. $\mathbb{E}[f] = 2.00$
2. $\mathbb{V}[f] = 7.28$
3. $\max[f] = 8.00$
4. $\min[f] = 0.00$
5. $\boldsymbol{P}_{97.5}[f] = 7.91$

We use the same hyperpriors for this problem as in example no. 1. The inferred signal-strength and lengthscale can be seen in Fig. 5.6 for iteration no. 22 of sampling. The lengthscale values in Fig. 5.6 (b) show high values in the middle region of the input space and lower values in the areas where the input value is 0.2 and 0.8 respectively. This behavior of the inferred lengthscale GP is in concurrence with the true function being inferred in Fig. 5.6 (a) where the methodology has almost learned the true function. The lengthscale values should be small as the waviness is high in areas of the two sharp peaks in the function. Similarly, the lengthscale values should be high where the function is very smooth or in other words flat. The inferred signal-strength, shown in Fig. 5.6 (c), also corresponds to the scalar value of the true function in the corresponding regions of the input space.

For this problem, the methodology starts with $n_i = 5$ and samples another 25 points. The final state of sampling for each $\boldsymbol{Q}[f]$ can be seen in Fig. 5.4. The final states show the different sets of designs obtained for different QoIs. The convergence of the estimated mean to the true value for each $\boldsymbol{Q}[f]$ and the reduction in uncertainty around the $\boldsymbol{Q}[f]$ can be seen in Fig. 5.5.

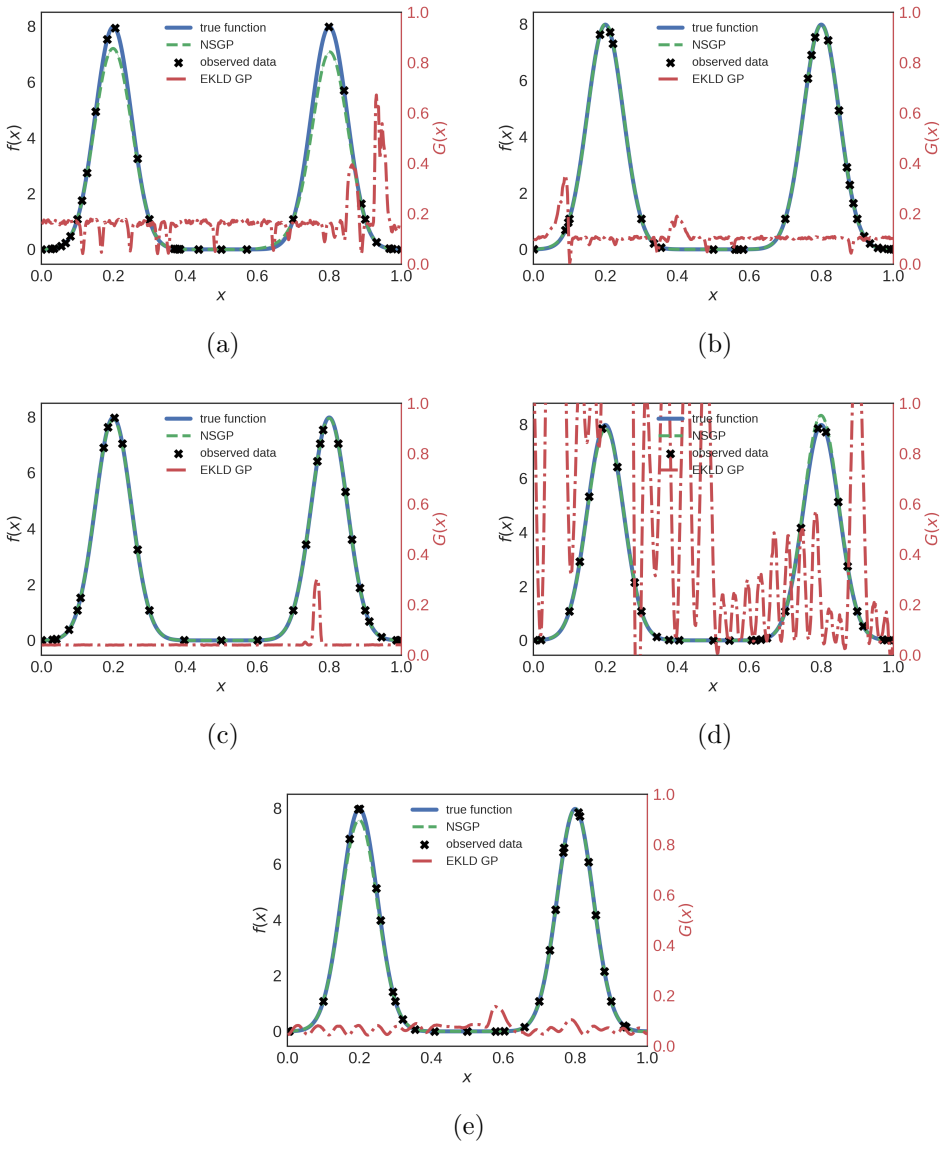


Figure 5.4. One-dimensional synthetic example ($n_i = 5$) shows the state of the function at the end (25th iteration) of the algorithm for inferring: Subfigures (a) $\mathbb{E}[f]$, (b) $\mathbb{V}[f]$, (c) $\max[f]$, (d) $\min[f]$, and (e) $\boldsymbol{P}_{97.5}[f]$.

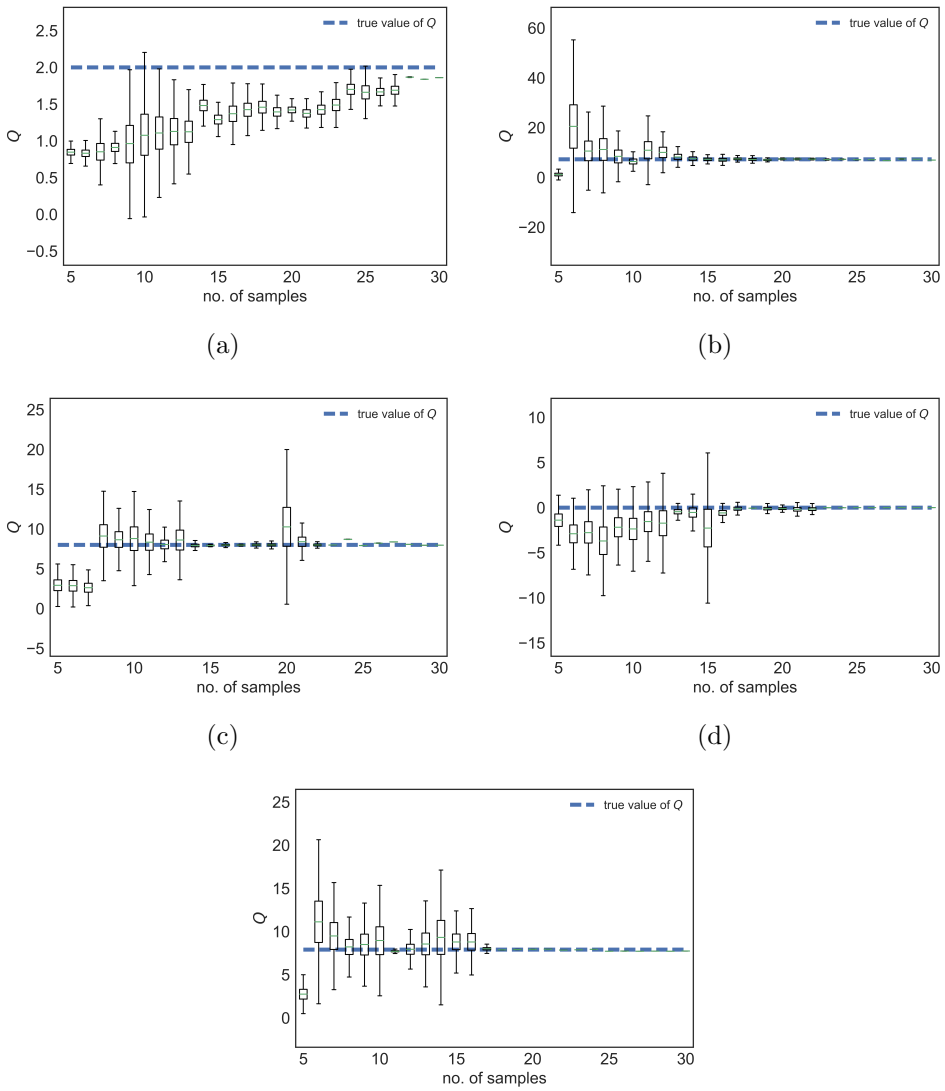


Figure 5.5. One-dimensional synthetic example ($n_i = 5$) shows the convergence of EKLD to the true value of Q for inferring: Subfigures (a) $\mathbb{E}[f]$, (b) $\mathbb{V}[f]$, (c) $\max[f]$, (d) $\min[f]$, and (e) $\boldsymbol{P}_{97.5}[f]$.

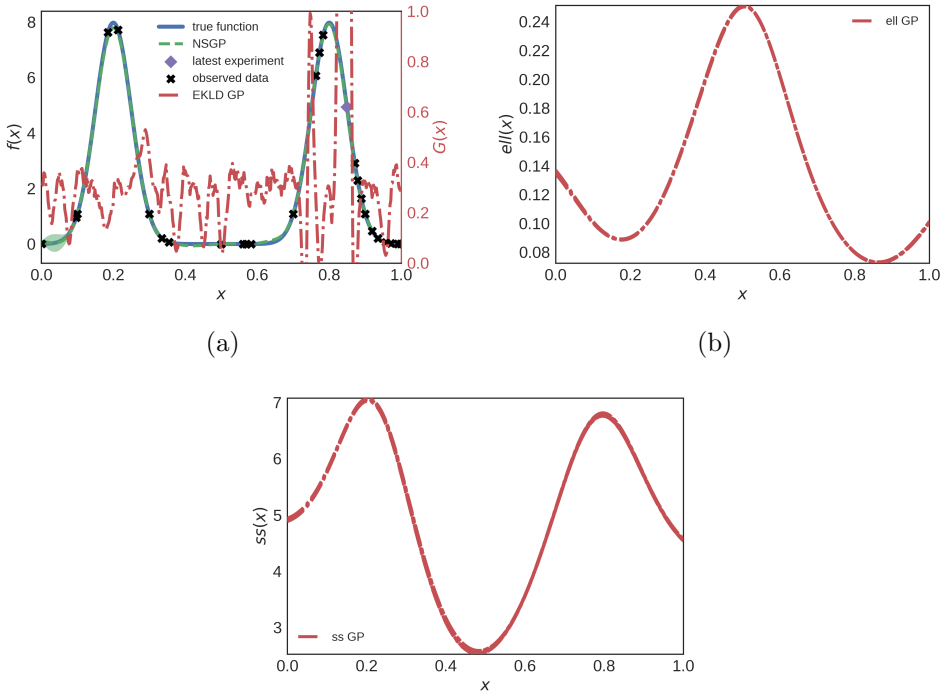


Figure 5.6. One-dimensional synthetic example ($n_i = 5$) shows the statistics of the FBNSGP at the the 22nd iteration of the sampling where: Subfigure (a) shows the state of the sampling, (b) shows the inferred point estimates of the lengthscale and (c) shows the inferred point estimates of the signal-strength.

5.3.3 Synthetic example no. 3

Consider the following three-dimensional function from [129] to test our methodology further.

$$\begin{aligned} f(\mathbf{x}) &= 4(x_1 + 8x_2 - 8x_2^2 - 2)^2 + (3 - 4x_2)^2 \\ &\quad +16\sqrt{x_3 + 1}(2x_3 - 1)^2. \end{aligned} \tag{5.44}$$

One difference between this function Eq. (5.44) and the the first two synthetic examples is the dimensionality of the problem. This is crucial because the NSGP modeling framework is expected to behave in a slightly different manner in multiple input dimensions. Unlike the one-dimensional synthetic examples discussed above, we proceed forward with a constant zero mean function of the GPs that model the logarithms of the lengthscale and the signal-strength. We also find this to be consistent with the BIC model selection at the beginning of the SDOE. An intuitive explanation about this different behaviour of lengthscale values in higher dimensions is given in [167]. The true values of the $\boldsymbol{Q}[f]$s, analytically available, are:

1. $\mathbb{E}[f] = -0.7864$
2. $\mathbb{V}[f] = 0.0209$
3. $\max[f] = -0.0575$
4. $\min[f] = -0.9999$
5. $\boldsymbol{P}_{2.5}[f] = -0.9899$

We apply our methodology to this problem starting from $n_i = 10$ and sample another 40 points. Fig. 5.7 (b) shows that the methodology started with a highly uncertain estimate of the true value and eventually converged to a sharp peaked Gaussian distribution around the true value. The approximation to each $\boldsymbol{Q}[f]$ at each stage of the algorithm is shown in Fig. 5.7. The gradual reduction in uncertainty around each $\boldsymbol{Q}[f]$ also can be seen in Fig. 5.7.

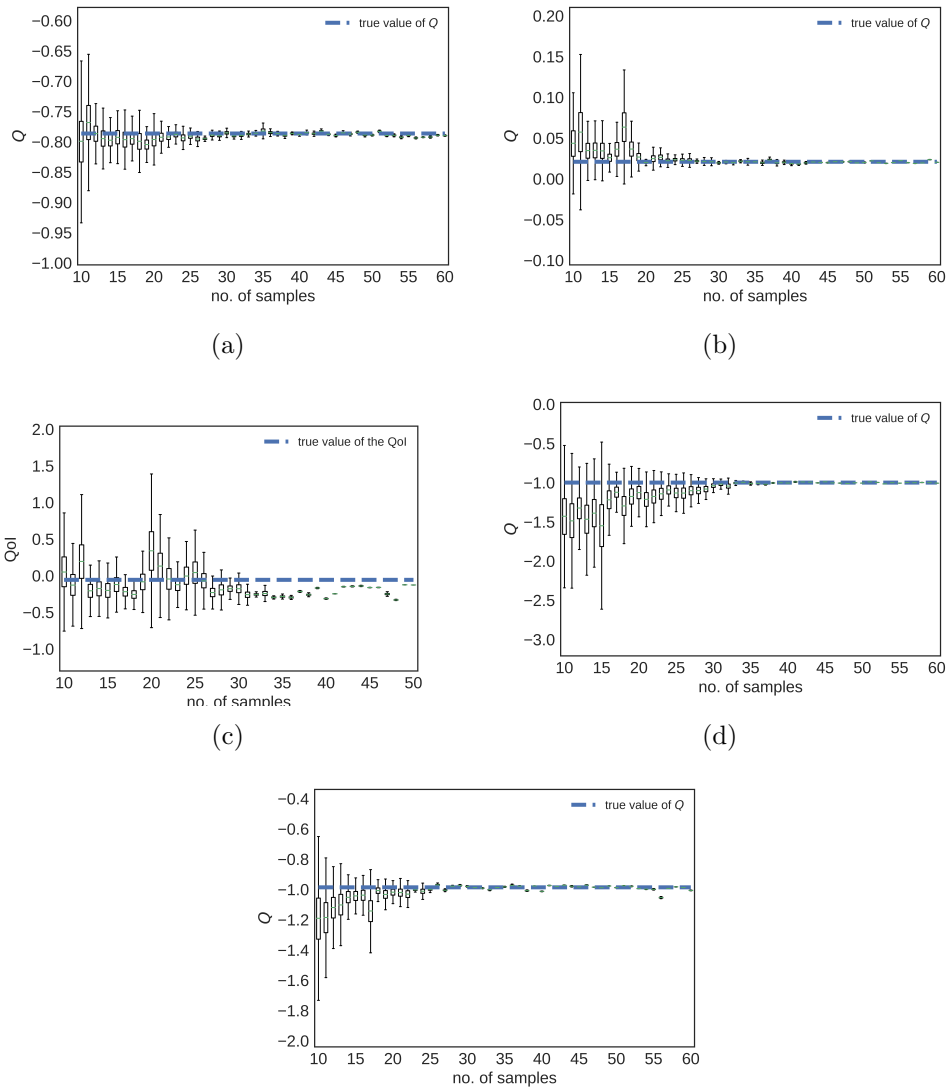


Figure 5.7. Three-dimensional synthetic example ($n_i = 10$) shows the convergence of EKLD to the true value of Q for inferring: Subfigures (a) $\mathbb{E}[f]$, (b) $\mathbb{V}[f]$, (c) $\max[f]$, (d) $\min[f]$, and (e) $\boldsymbol{P}_{2.5}[f]$.

5.3.4 Synthetic example no. 4

The following five-dimensional function is taken from [130].

$$f(\mathbf{x}) \quad = \quad 10\sin(\pi x_1 x_2) + 20(x_3 - 5)^2 + 10x_4 + 5x_5. \tag{5.45}$$

This function Eq. (5.45) is reasonably high-dimensional and challenging due to the non-linear input-output relation. The true values of the $\boldsymbol{Q}[f]$s, analytically available, are:

1. $\mathbb{E}[f] = 0.3882$
2. $\mathbb{V}[f] = 1.0896$
3. $\max[f] = 2.4941$
4. $\min[f] = -1.5906$
5. $\boldsymbol{P}_{2.5}[f] = -1.2782$

We apply our methodology to this problem starting from $n_i = 10$ and sample another 60 points for inferring $\mathbb{E}[f]$ and $\mathbb{V}[f]$. For inferring the 2.5th percentile Fig. 5.8(e) of f we start with 10 initial points and collect another 60 samples. In the cases shown in Fig. 5.8 (c) and (d) for inferring $\max[f]$ and $\min[f]$ respectively, the methodology starts with 20 initial points and samples another 50 points using the EKLD. The iteration-wise convergence of the $\boldsymbol{Q}[f]$s to the respective true value is shown in Fig. 5.8. Along expected lines, as more samples are collected by the EKLD, the uncertainty around the mean of $\boldsymbol{Q}[f]$ reduces. This uncertainty becomes negligible around the 50th sample mark for each of the five $\boldsymbol{Q}[f]$s in Fig. 5.8.

5.3.5 Wire drawing problem

This problem is a special case of the wire problem discussed in Sec. 4.3.5 as the number of passes in this case are five instead of eight. Rest of the technical

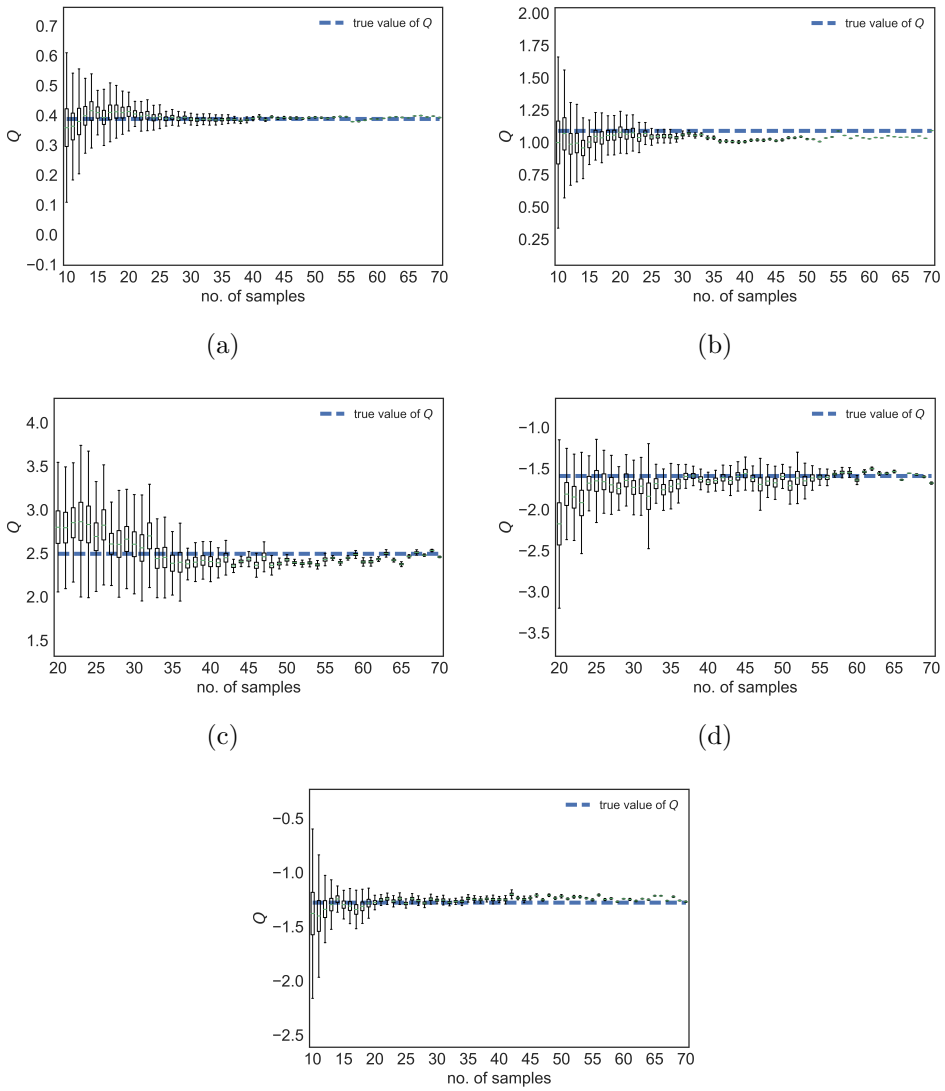


Figure 5.8. Five-dimensional synthetic example ($n_i = 10$) shows the convergence of EKLD to the true value of Q for inferring: Subfigures (a) $\mathbb{E}[f]$, (b) $\mathbb{V}[f]$, (c) $\max[f]$, (d) $\min[f]$, and (e) $\boldsymbol{P}_{2.5}[f]$.

regarding the micro-structure modeling using FEM remain the same as in Sec. 4.3.5. The iteration-wise convergence of the $\boldsymbol{Q}[f]$s to the respective true value is shown in Fig. 5.9. It is interesting to note the noise in the convergence for QoIs no.3 and no.4 in Fig. 5.9 (c) and (d). This is because the number of samples M needed to approximate to QoIs at each iteration for cases where a global minima is located in a small region becomes very high. One way around this could be to take more M samples albeit at a very high computational cost. For the QoIs no.1, no.2 and no.5 we have convergence as the number of samples reaches 30.

The true values of the $\boldsymbol{Q}[f]$s, analytically available, are:

1. $\mathbb{E}[f] = -2.2402$
2. $\mathbb{V}[f] = 0.1805$
3. $\max[f] = -0.8136$
4. $\min[f] = -3.5724$
5. $\boldsymbol{P}_{2.5}[f] = -3.027$

We apply our methodology to this problem starting from $n_i = 10$ and sample another 60 points for inferring the different QoIs. The iteration-wise convergence of the $\boldsymbol{Q}[f]$s to the respective true value is shown in Fig. 5.9. Along expected lines, as more samples are collected by the EKLD, the uncertainty around the mean of $\boldsymbol{Q}[f]$ reduces. This uncertainty becomes negligible around the 30th sample mark for each of the five $\boldsymbol{Q}[f]$s in Fig. 5.9.

5.4 Comparison studies

We compare the EKLD to two classic SDOE methods, namely uncertainty sampling (US) and expected improvement (EI). This is done in order to ascertain to some extent the convergence pattern of the EKLD. A comparison with US when the QoI is the mean or the variance or an extreme percentile of the function is done. This is because

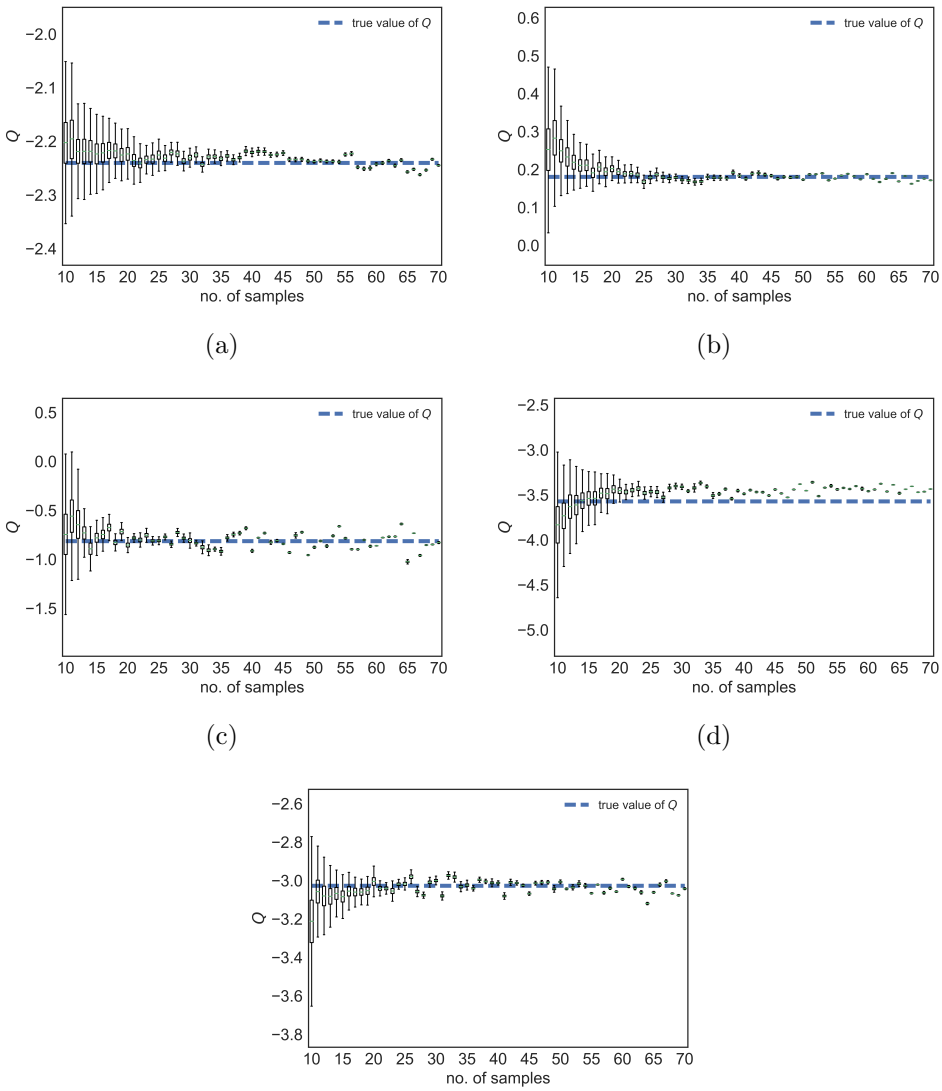


Figure 5.9. Wire-drawing problem ($n_i = 10$) shows the convergence of EKLD to the true value of Q for inferring: Subfigures (a) $\mathbb{E}[f]$, (b) $\mathbb{V}[f]$, (c) $\max[f]$, (d) $\min[f]$, and (e) $\boldsymbol{P}_{2.5}[f]$.

US is agnostic to the QoI unlike EI which is used for comparison when the QoI is the minimum or the maximum of the function.

The comparison studies show mixed results. For the one-dimensional function in Fig. 5.1 (a) and (b), the EKLD and US appear to converge to the true value at almost the same number of samples for inferring the statistical expectationFig. 5.10 (a) and varianceFig. 5.10 (b) of f. However, the EKLD converges sooner for inferring the 2.5th percentile of f. The EI and the EKLD show similar trends on converging to the truth for synthetic example no.1 when the QoI is the maximum Fig. 5.10 (c) and the minimum Fig. 5.10 (d) of f.

For the one-dimensional function in Sec. 5.3.2 the US and EKLD converge at around the same stage of sampling which is clearly seen in Fig. 5.11.

The three-dimensional function in Sec. 5.3.3 is a problem with multiple dimensions. The EKLD converges sooner compared to the US for inferring the expectation, variance and the 2.5th percentile when a total of 40 additional samples are collected. In the optimization scenarios the EKLD comes close to convergence for estimating the maximum value of f, whereas the EI is unable to estimate this value even after 40 iterations of the SDOE algorithm. Estimating the minimum of f, throws up results that put the EKLD and EI at the same level of performance.

The five-dimensional problem in Sec. 5.3.4 shows similar results as for the three-dimensional problem in the comparison study of the EKLD with US while inferring the statistical expectation, variance, and the 2.5th percentile of f. EKLD converges sooner, near the 35 sample mark compared to the US which takes almost 50 samples to converge, for the three QoIs. The optimization cases in Fig. 5.8 (c) and (d) and Fig. 5.13 (c) and (d) provide similar convergence results for the EKLD and EI, with both methods converging at almost the same number of samples.

Comparison studies for the wire-drawing problem show a slightly mixed pattern of convergence with the EKLD and US taking almost same number of samples for inferring the three QoIs on which they are compared. Whereas for the optimization cases both EKLD and EI seem to be slow in identifying the true minimum of f as can

be seen in Fig. 5.9 (d) and Fig. 5.14 (d) respectively. Estimating the maximum value takes fewer samples for the EI compared to the EKLD shown in Fig. 5.14 (c) and Fig. 5.9 (c) respectively. Thus, results for the wire problem are not sufficient to draw a conclusion about the performance of the EKLD when compared to EI for inferring $\min[f]$ and $\max[f]$. The next step will be to run the methodologies for more number of iterations in order to establish convergence.

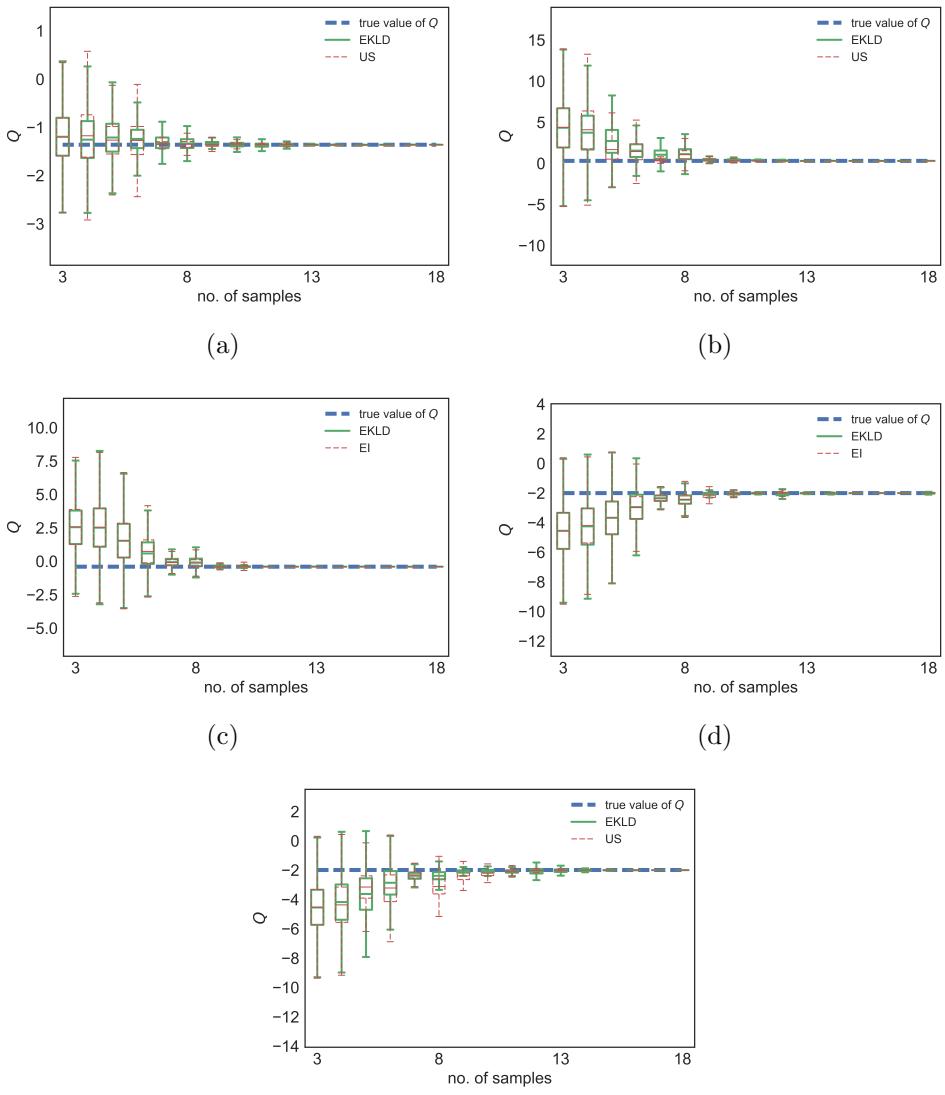


Figure 5.10. Comparison studies for example no.1. Subfigures (a), and (b), convergence of US for $\mathbb{E}[f]$ and $\mathbb{V}[f]$. Subfigures (c), and (d), convergence of EI for $\max[f]$ and $\min[f]$. Subfigure (e), convergence of US for inferring $\boldsymbol{P}_{2.5}[f]$.

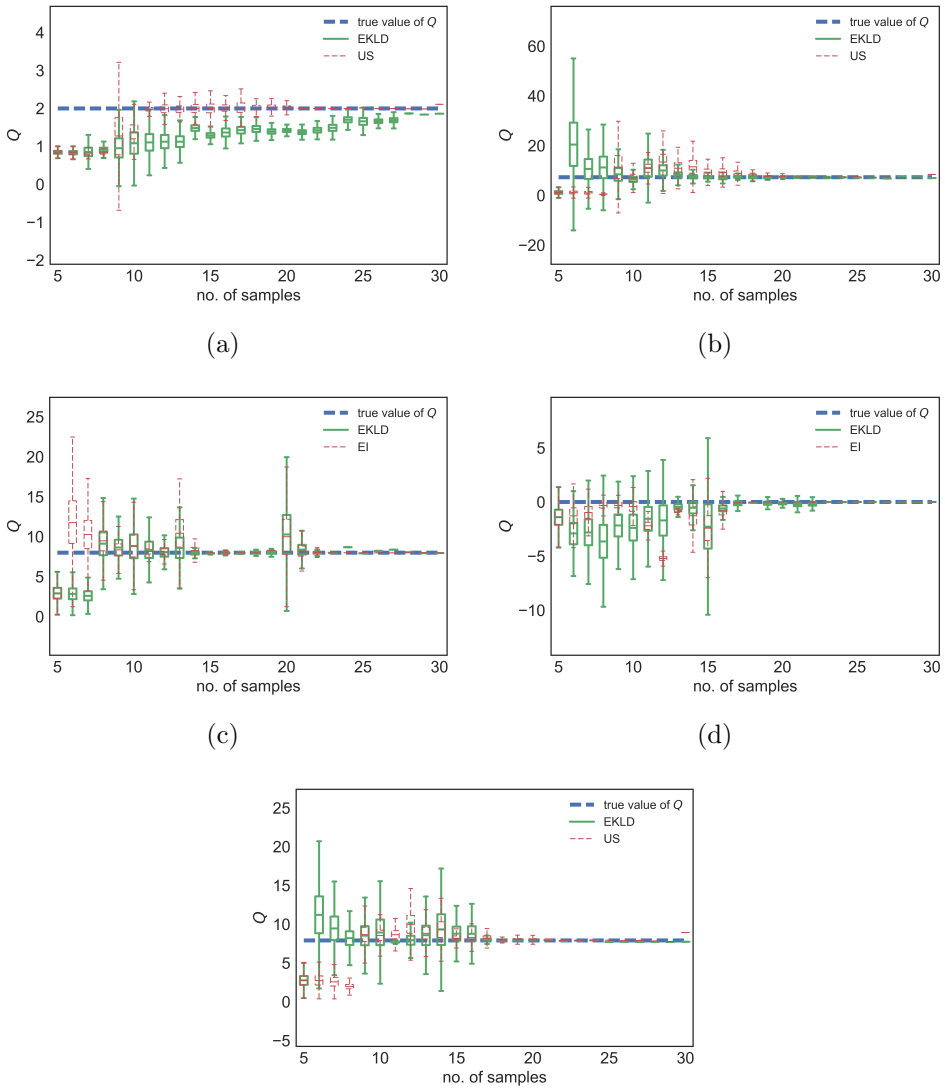


Figure 5.11. Comparison studies for example no.2. Subfigures (a), and (b), convergence of US for $\mathbb{E}[f]$ and $\mathbb{V}[f]$. Subfigures (c), and (d), convergence of EI for $\max[f]$ and $\min[f]$. Subfigure (e), convergence of US for inferring $\boldsymbol{P}_{97.5}[f]$.

5.5 Useful findings and insights

We highlight some salient features of EKLD and its comparison studies with US and EI below.

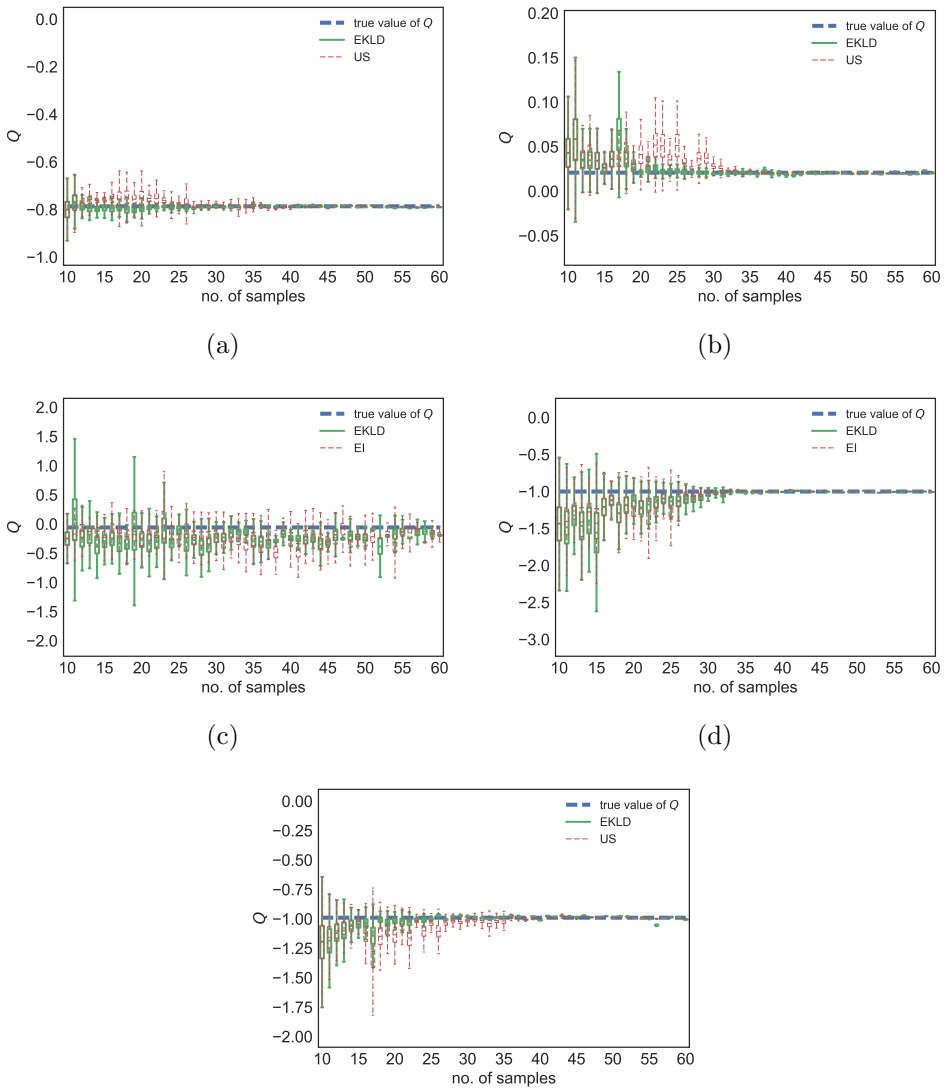


Figure 5.12. Comparison studies for example no.3. Subfigures (a), and (b), convergence of US for $\mathbb{E}[f]$ and $\mathbb{V}[f]$. Subfigures (c), and (d), convergence of EI for $\max[f]$ and $\min[f]$. Subfigure (e), convergence of US for inferring $\boldsymbol{P}_{2.5}[f]$.

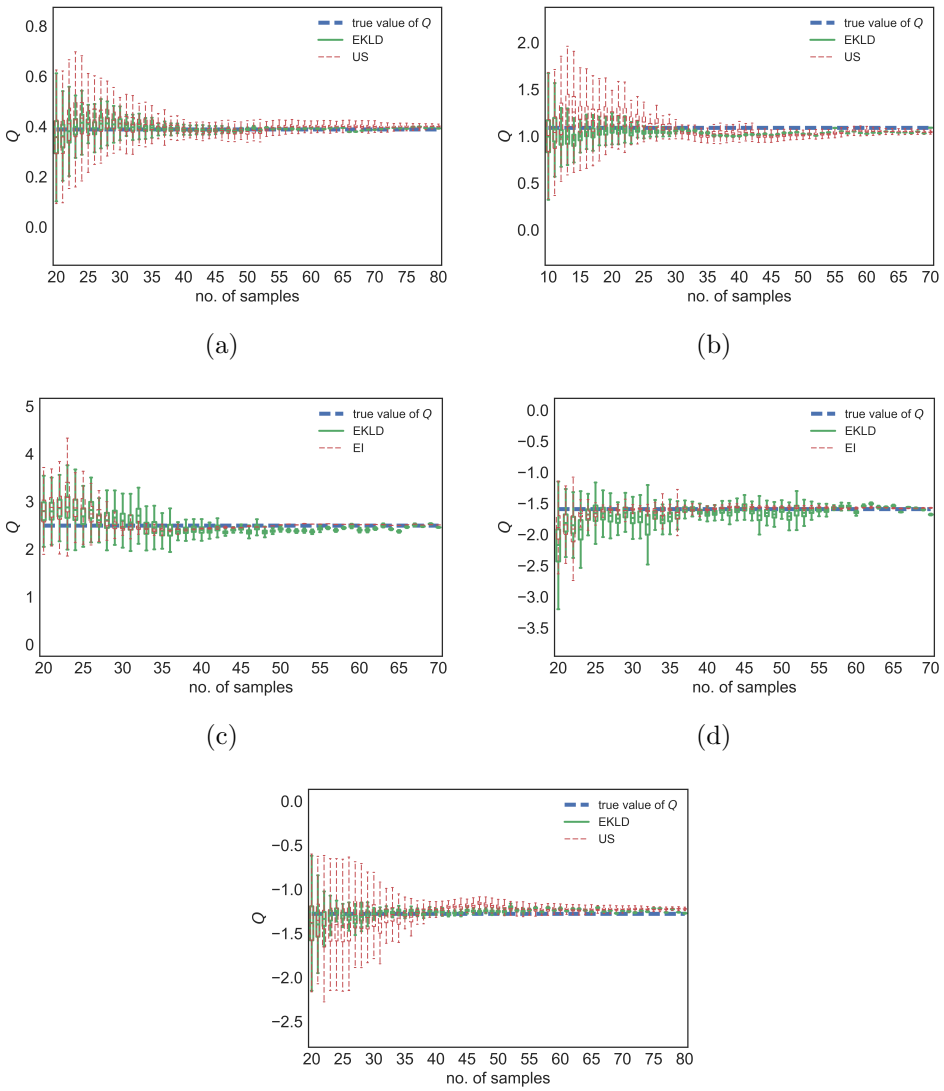


Figure 5.13. Comparison studies for example no.4. Subfigures (a), and (b), convergence of US for $\mathbb{E}[f]$ and $\mathbb{V}[f]$. Subfigures (c), and (d), convergence of EI for $\max[f]$ and $\min[f]$. Subfigure (e), convergence of US for inferring $\boldsymbol{P}_{2.5}[f]$.

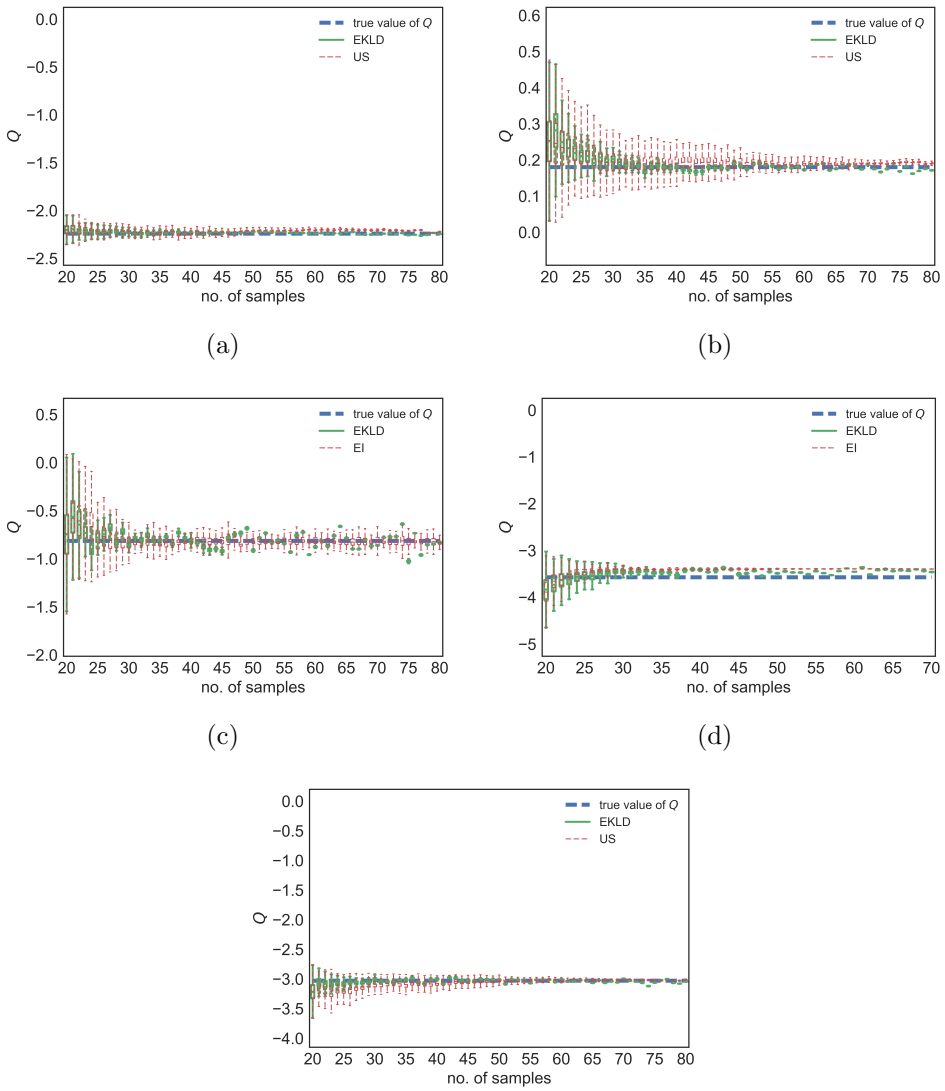


Figure 5.14. Comparison studies for the wire-drawing problem. Subfigures (a), and (b), convergence of US for $\mathbb{E}[f]$ and $\mathbb{V}[f]$. Subfigures (c), and (d), convergence of EI for $\max[f]$ and $\min[f]$. Subfigure (e), convergence of US for inferring $\boldsymbol{P}_{2.5}[f]$.

1. The derived estimator for EKLD samples the next point in a region of high uncertainty and/or high posterior mean of $\boldsymbol{Q}[f]$.

2. The point mentioned above also means that the EKLD, similar to the EI, balances the exploration-exploitation trade-off.

3. The EKLD performs better or equally good when compared to US and EI for the respective QoIs.

4. Some initial calibration needs to be done to select hyperpriors especially in the one-dimensional problems, where some functions can be explained better by a certain combination of hyperpriors. In the synthetic examples presented here we use uninformative priors for all hyperparameters. A default setting for the hyperpriors has been chosen which remains the same for all problems with one input. Similarly, a default setting for the hyperpriors for problems in multiple inputs is demonstrated with promising results.

5. The FBNSGP framework enables incorporation of point estimates of local smoothness and signal-strength even in low-sample regime. For the one-dimensional synthetic examples the inferred input-dependent lengthscales and signal-strengths have been shown. The inferred values of the lengthscale and signal-strength across the input space have been sampled for each of the S posterior samples of $\boldsymbol{\theta}$. This allows us to quantify the epistemic uncertainty around the point estimates of the lengthscale and the signal-strength across the input space.

6. High input-dimensionality will pose certain challenges for the EKLD. Since, training the NSGP model involves inferring 3 parameters each for the lengthscale and signal-strength GPs per input dimension. This means that at every stage of model training, $6d$ number of hyperparameters need to be inferred. This task becomes computationally cumbersome when one is faced with problems greater than *single-digit* input dimensions.

7. This research shows how some critical statistics can be inferred with fully-Bayesian quantification of epistemic uncertainty for problems of different dimensions and input-dependent lengthscales and signal-strengths. This nuance of the proposed methodology is highly useful in designing simulations and experiments that take multiple days to finish.

8. An interesting point that we have not covered is the application of the EKLD framework mentioned above to suggesting multiple simulations or experiments at each iteration. This scheme, if extended from the current EKLD, holds great promise because this would enable practical use of computational or laboratory resources. Secondly, it might also be cheaper to suggest multiple experiments in one iteration for problems in dimensions greater than five.

5.6 Conclusions

We derive an estimator to quantify the information gain in a hypothetical experiment when a scientist wishes to estimate a QoI which depends on some output of the experiment. The information gain is the Kullback-Leibler divergence between a prior state of knowledge about the QoI and a posterior state of knowledge about the QoI. This methodology is augmented by a robust and flexible response surface modeling approach. The fully Bayesian non-stationary Gaussian process surrogate model allows the user to incorporate prior knowledge about the input-dependent smoothness and variance of the underlying physical response. The performance of the SDOE heuristic is demonstrated on four numerical examples and an engineering problem of eight input dimensions. The convergence tests for different numerical examples and the engineering problem have been compared to state-of-the-art methods namely uncertainty sampling, expected improvement and probability of improvement. These state-of-the-art SDOE methods are commonly suited for certain QoIs which is further highlighted by the comparison tests. The derived SDOE heuristic converges at the same level or better as the other methods for problems which differ on accounts of

dimensionality and context. More work can be done on the presented methodology to suggest multiple experiments or designs at a single iteration, thereby allowing parallel use of laboratory or computational resources. This direction of research rhymes well with the spirit of batch optimization [168] and parallel data acquisition.

6. SUMMARY

We are at a stage where designing experiments to optimize expensive black-box functions can be treated for single (SOO in Chapter 2.) and multi objective (MOO in Chapter 2.) scenarios. The methodologies proposed in Chapters 2. and 3. also enable the quantification of uncertainty about the optimal designs in a SOO problem and around the Pareto Frontier in a MOO problem. An extension to the methodology in Chapter 3., suggesting multiple experiments or batch design has been demonstrated on a problem of chemical vapor deposition for Graphene manufacturing in a collaborative work [22, 89, 169].

In Chapters 4. and 5., we derive an estimator for quantifying the information gain (IG) in a hypothetical experiment to design experiments for estimating QoIs that are statistics of the expensive experiment or simulation. This IG is the KL divergence from the posterior probability distribution to the prior probability distribution of the statistical QoI.

The derived estimator of the EKLD IAF has been used for SDOE in Chapters 4. and 5. on synthetic functions and the wire-drawing problem. The performance of this estimator demonstrates convergence to the *ground truth* values of the statistical QoI being inferred using only a finite number of simulations. Comparisons with US and EI for different QoIs yield mixed results. However, a general takeaway is that the EKLD performs better than US and EI with increasing dimensionality of the input space.

An important lacuna in SDOE is the scalabilty of the method to very high input dimensions and large data. Our formulation suffers tremendously as the dimensionality increases beyond single digits. More work in the area of extending the use of EKLD to higher dimensions in necessary. Further work on deriving theoretical guarantees for the derived EKLD estimator holds high potential. Suggesting multiple points for

batch design, especially needed for experimental problems, is another direction in which the derived EKLD estimator can be extended.

REFERENCES

REFERENCES

[1] Jesper Kristensen, Ilias Bilionis, and Nicholas Zabaras. Relative entropy as model selection tool in cluster expansions. *Physical Review B*, 87(17):174112, 2013.

[2] Chris J Oates, Mark Girolami, and Nicolas Chopin. Control functionals for monte carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718, 2017.

[3] Joan Gonzalvez, Edmond Lezmi, Thierry Roncalli, and Jiali Xu. Financial applications of gaussian processes and bayesian optimization. *arXiv preprint arXiv:1903.04841*, 2019.

[4] Ashish M Chaudhari. *Crowdsourcing for engineering design: theoretical and experimental studies*. PhD thesis, Purdue University, 2017.

[5] Ashish M Chaudhari and Jitesh H Panchal. An experimental study of human decisions in sequential information acquisition in design: Impact of cost and task complexity. In *Research into Design for a Connected World*, pages 321–332. Springer, 2019.

[6] Murtuza Shergadwala, Ilias Bilionis, Karthik N Kannan, and Jitesh H Panchal. Quantifying the impact of domain knowledge and problem framing on sequential decisions in engineering design. *Journal of Mechanical Design*, 140(10):101402, 2018.

[7] Anthony Atkinson, Alexander Donev, and Randall Tobias. *Optimum experimental designs, with SAS*, volume 34. Oxford University Press, 2007.

[8] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

[9] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

[10] Donald R Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, 21(4):345–383, 2001.

[11] Jerome Sacks, William J Welch, Toby J Mitchell, and Henry P Wynn. Design and analysis of computer experiments. *Statistical science*, pages 409–423, 1989.

[12] Robert B Gramacy and Michael Ludkovski. Sequential design for optimal stopping problems. *SIAM Journal on Financial Mathematics*, 6(1):748–775, 2015.

[13] Robert B Gramacy and Herbert KH Lee. Adaptive design and analysis of supercomputer experiments. *Technometrics*, 51(2):130–145, 2009.

[14] Nicholas John Gaul. *Modified Bayesian Kriging for noisy response problems and Bayesian confidence-based reliability-based design optimization.* The University of Iowa, 2014.

[15] Deng Huang, Theodore T Allen, William I Notz, and Ning Zeng. Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of global optimization*, 34(3):441–466, 2006.

[16] Mickael Binois, Robert B Gramacy, and Mike Ludkovski. Practical heteroscedastic gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 27(4):808–821, 2018.

[17] Emile Contal, Vianney Perchet, and Nicolas Vayatis. Gaussian process optimization with mutual information. In *International Conference on Machine Learning*, pages 253–261, 2014.

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

[19] M Binois, D Ginsbourger, and O Roustant. Quantifying uncertainty on Pareto fronts with Gaussian process conditional simulations. *European Journal of Operational Research*, 243(2):386–394, June 2015.

[20] Michael Emmerich, André H Deutz, and Jan Willem Klinkenberg. Hypervolume-based expected improvement: Monotonicity properties and exact computation. In *Evolutionary Computation (CEC), 2011 IEEE Congress on*, pages 2147–2154. IEEE, 2011.

[21] Michael Emmerich and Jan-willem Klinkenberg. The computation of the expected improvement in dominated hypervolume of pareto front approximations. *Rapport technique, Leiden University*, 2008.

[22] Majed A Alrefae, Anurag Kumar, Piyush Pandita, Aaditya Candadai, Ilias Bilionis, and Timothy S Fisher. Process optimization of graphene growth in a roll-to-roll plasma cvd system. *AIP Advances*, 7(11):115102, 2017.

[23] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[24] Anton J Kleywegt, Alexander Shapiro, and Tito Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

[25] Daniel P Heyman and Matthew J Sobel. *Stochastic Models in Operations Research: Stochastic Optimization*, volume 2. Courier Corporation, 2003.

[26] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pages 2595–2603, 2010.

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

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

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

[30] D. Lizotte. *Practical Bayesian Optization*. Thesis, 2008.

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

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

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

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

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

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

[37] Julien Villemonteix, Emmanuel Vazquez, and Eric Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44(4):509–534, 2009.

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

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

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

[41] Warren B Powell and Ilya O Ryzhov. *Optimal learning*, volume 841. John Wiley & Sons, 2012.

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

[43] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, MA, 2006.

[44] Noel Cressie. The origins of kriging. *Mathematical geology*, 22(3):239–252, 1990.

[45] Tony E Smith and Jacob Dearmon. Gaussian process regression and bayesian model averaging: An alternative approach to modeling spatial phenomena. 2014.

[46] H. Jeffreys. 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):453–461, 1946.

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

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

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

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

[51] Ilias Bilionis and Nicholas Zabaras. Multidimensional adaptive relevance vector machines for uncertainty quantification. *SIAM Journal on Scientific Computing*, 34(6):B881–B908, 2012.

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

[53] Peng Chen, Nicholas Zabaras, and Ilias Bilionis. Uncertainty propagation using infinite mixture of gaussian processes and variational bayesian inference. *Journal of Computational Physics*, 284:291–333, 2015.

[54] Deng Huang, Theodore T Allen, William I Notz, and Ning Zeng. Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of global optimization*, 34(3):441–466, 2006.

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

[56] Michael James Sasena. *Flexibility and efficiency enhancements for constrained global design optimization with kriging approximations*. PhD thesis, General Motors, 2002.

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

[58] Roger G Ghanem and Pol D Spanos. *Stochastic finite elements: a spectral approach*. Courier Corporation, 2003.

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

[60] Kalyanmoy Deb. Introduction to evolutionary multiobjective optimization. In *Multiobjective Optimization*, pages 59–96. Springer, 2008.

[61] Eckart Zitzler, Marco Laumanns, Lothar Thiele, Eckart Zitzler, Eckart Zitzler, Lothar Thiele, and Lothar Thiele. Spea2: Improving the strength pareto evolutionary algorithm, 2001.

[62] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *Evolutionary Computation, IEEE Transactions on*, 6(2):182–197, 2002.

[63] Abraham Charnes and William Wager Cooper. Goal programming and multiple objective optimizations: Part 1. *European Journal of Operational Research*, 1(1):39–54, 1977.

[64] R Timothy Marler and Jasbir S Arora. Survey of multi-objective optimization methods for engineering. *Structural and multidisciplinary optimization*, 26(6):369–395, 2004.

[65] Milan Zeleny. The pros and cons of goal programming. *Computers & Operations Research*, 8(4):357–359, 1981.

[66] Jonas Mockus. *Bayesian approach to global optimization: theory and applications*, volume 37. Springer Science & Business Media, 2012.

[67] J Močkus. On bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference*, pages 400–404. Springer, 1975.

[68] David Ginsbourger, Rodolphe Le Riche, and Laurent Carraro. A multi-points criterion for deterministic parallel global optimization based on kriging. In *NCP07*, 2007.

[69] Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

[70] Andy J Keane. Statistical improvement criteria for use in multiobjective design optimization. *AIAA journal*, 44(4):879–891, 2006.

[71] Marc Tesch, Jurgen Schneider, and Howie Choset. Expensive multiobjective optimization for robotics. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 973–980. IEEE, 2013.

[72] Koji Shimoyama, Koma Sato, Shinkyu Jeong, and Shigeru Obayashi. Updating kriging surrogate models based on the hypervolume indicator in multi-objective optimization. *Journal of Mechanical Design*, 135(9):094503, 2013.

[73] Paul Feliot, Julien Bect, and Emmanuel Vazquez. A bayesian approach to constrained multi-objective optimization. In *Learning and Intelligent Optimization*, pages 256–261. Springer, 2015.

[74] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[75] Clément Chevalier, David Ginsbourger, Julien Bect, and Ilya Molchanov. Estimating and quantifying uncertainties on level sets using the vorob'ev expectation and deviation with gaussian process models. In *mODa 10–Advances in Model-Oriented Design and Analysis*, pages 35–43. Springer, 2013.

[76] I Molchanov. Theory of Random Sets. Springer-Verlag, London, 2005.

[77] James Parr. *Improvement criteria for constraint handling and multiobjective optimization*. PhD thesis, University of Southampton, 2013.

[78] Joshua Knowles. Parego: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *Evolutionary Computation, IEEE Transactions on*, 10(1):50–66, 2006.

[79] Nancy Flournoy. A clinical experiment in bone marrow transplantation: Estimating a percentage point of a quantal response curve. In *case studies in Bayesian Statistics*, pages 324–336. Springer, 1993.

[80] L Eriksson, E Johansson, N Kettaneh-Wold, C Wikström, and S Wold. Design of experiments. *Principles and Applications, Learn ways AB, Stockholm*, 2000.

[81] Mark J Anderson and Patrick J Whitcomb. *Design of experiments*. Wiley Online Library, 2000.

[82] Alen Alexanderian, Noemi Petra, Georg Stadler, and Omar Ghattas. A-optimal design of experiments for infinite-dimensional bayesian linear inverse problems with regularized \ell_0-sparsification. *SIAM Journal on Scientific Computing*, 36(5):A2122–A2148, 2014.

[83] Douglas C Montgomery. *Design and analysis of experiments*. John wiley & sons, 2017.

[84] Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959.

[85] Herbert Robbins. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, pages 169–177. Springer, 1985.

[86] Jos Havinga, Gerrit Klaseboer, and AH Van den Boogaard. Sequential optimization of strip bending process using multiquadric radial basis function surrogate models. In *Key engineering materials*, volume 554, pages 911–918. Trans Tech Publ, 2013.

[87] J Havinga, Antonius H van den Boogaard, and G Klaseboer. Sequential improvement for robust optimization using an uncertainty measure for radial basis functions. *Structural and multidisciplinary optimization*, 55(4):1345–1363, 2017.

[88] Majed A Alrefae. *Process Characterization and Optimization of Roll-to-roll Plasma Chemical Vapor Deposition for Graphene Growth*. PhD thesis, Purdue University, 2018.

[89] Kimberly R Saviers. *Scaled-Up Production and Transport Applications of Graphitic Carbon Nanomaterials*. PhD thesis, Purdue University, 2017.

[90] Matthias Schonlau. Computer experiments and global optimization. 1997.

[91] Timothy W Simpson, Dennis KJ Lin, and Wei Chen. Sampling strategies for computer experiments: design and analysis. *International Journal of Reliability and Applications*, 2(3):209–240, 2001.

[92] Xun Huan. *Accelerated Bayesian experimental design for chemical kinetic models*. PhD thesis, Massachusetts Institute of Technology, 2010.

[93] Marco Locatelli. Bayesian algorithms for one-dimensional global optimization. *Journal of Global Optimization*, 10(1):57–76, 1997.

[94] Peter I Frazier, Warren B Powell, and Savas Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.

[95] Paul D Arendt, Daniel W Apley, and Wei Chen. Objective-oriented sequential sampling for simulation based robust design considering multiple sources of uncertainty. *Journal of Mechanical Design*, 135(5):051005, 2013.

[96] Xun Huan and Youssef Marzouk. Gradient-based stochastic optimization methods in bayesian experimental design. *International Journal for Uncertainty Quantification*, 4(6), 2014.

[97] Remi Lam, Karen Willcox, and David H Wolpert. Bayesian optimization with a finite budget: An approximate dynamic programming approach. In *Advances in Neural Information Processing Systems*, pages 883–891, 2016.

[98] Alonso Marco, Philipp Hennig, Jeannette Bohg, Stefan Schaal, and Sebastian Trimpe. Automatic lqr tuning based on gaussian process global optimization. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 270–277. IEEE, 2016.

[99] Jesper Kristensen, Ilias Bilionis, and Nicholas Zabaras. Adaptive simulation selection for the discovery of the ground state line of binary alloys with a limited computational budget. In *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, pages 185–211. Springer, 2017.

[100] J Andrés Christen and Bruno Sansó. Advances in the sequential design of computer experiments based on active learning. *Communications in Statistics-Theory and Methods*, 40(24):4467–4483, 2011.

[101] David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

[102] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(Feb):235–284, 2008.

[103] Rémi Stroh, Séverine Demeyer, Nicolas Fischer, Julien Bect, and Emmanuel Vazquez. Sequential design of experiments to estimate a probability of exceeding a threshold in a multi-fidelity stochastic simulator. *arXiv preprint arXiv:1707.08384*, 2017.

[104] Joakim Beck and Serge Guillas. Sequential design with mutual information for computer experiments (mice): emulation of a tsunami model. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):739–766, 2016.

[105] Gabriel Terejanu, Rochan R Upadhyay, and Kenji Miki. Bayesian experimental design for the active nitridation of graphite by atomic nitrogen. *Experimental Thermal and Fluid Science*, 36:178–193, 2012.

[106] Mustafa A Mohamad. *Direct and adaptive quantification schemes for extreme event statistics in complex dynamical systems*. PhD thesis, Massachusetts Institute of Technology, 2017.

[107] Mustafa A Mohamad and Themistoklis P Sapsis. A sequential sampling strategy for extreme event statistics in nonlinear dynamical systems. *arXiv preprint arXiv:1804.07240*, 2018.

[108] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[109] Panagiotis Tsilifis, Roger G Ghanem, and Paris Hajali. Efficient bayesian experimentation using an expected information gain lower bound. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):30–62, 2017.

[110] Paromita Nath, Zhen Hu, and Sankaran Mahadevan. Sensor placement for calibration of spatially varying model parameters. *Journal of Computational Physics*, 343:150–169, 2017.

[111] Liang Yan, Xiaojun Duan, Bowen Liu, and Jin Xu. Gaussian processes and polynomial chaos expansion for regression problem: Linkage via the rkhs and comparison via the kl divergence. *Entropy*, 20(3):191, 2018.

[112] Seung-Kyum Choi, Ramana V Grandhi, Robert A Canfield, and Chris L Pettit. Polynomial chaos expansion with latin hypercube sampling for estimating response variability. *AIAA journal*, 42(6):1191–1198, 2004.

[113] Mohammad Hadigol and Alireza Doostan. Least squares polynomial chaos expansion: A review of sampling strategies. *Computer Methods in Applied Mechanics and Engineering*, 2017.

[114] G Terejanu, CM Bryant, and K Miki. Bayesian optimal experimental design for the shock-tube experiment. In *Journal of Physics: Conference Series*, volume 410, page 012040. IOP Publishing, 2013.

[115] Carlos Guestrin, Andreas Krause, and Ajit Paul Singh. Near-optimal sensor placements in gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 265–272. ACM, 2005.

[116] Xun Huan and Youssef M Marzouk. Simulation-based optimal bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, 232(1):288–317, 2013.

[117] Victor Picheny, David Ginsbourger, Olivier Roustant, Raphael T Haftka, and Nam-Ho Kim. Adaptive designs of experiments for accurate approximation of a target region. *Journal of Mechanical Design*, 132(7):071008, 2010.

[118] Ning-Cong Xiao, Ming J Zuo, Chengning Zhou, et al. A new adaptive sequential sampling method to construct surrogate models for efficient reliability analysis. *Reliability Engineering and System Safety*, 169(C):330–338, 2018.

[119] Haitao Liu, Shengli Xu, Ying Ma, Xudong Chen, and Xiaofang Wang. An adaptive bayesian sequential sampling approach for global metamodeling. *Journal of Mechanical Design*, 138(1):011404, 2016.

[120] Huibin Liu, Wei Chen, and Agus Sudjianto. Relative entropy based method for probabilistic sensitivity analysis in engineering design. *Journal of Mechanical Design*, 128(2):326–336, 2006.

[121] Jian Wu, Saul Toscano-Palmerin, Peter I Frazier, and Andrew Gordon Wilson. Practical multi-fidelity bayesian optimization for hyperparameter tuning. *arXiv preprint arXiv:1903.04703*, 2019.

[122] Anthony O'Hagan. Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering & System Safety*, 91(10-11):1290–1300, 2006.

[123] François-Xavier Briol, Chris Oates, Mark Girolami, and Michael A Osborne. Frank-wolfe bayesian quadrature: Probabilistic integration with theoretical guarantees. In *Advances in Neural Information Processing Systems*, pages 1162–1170, 2015.

[124] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.

[125] Jonathan Goodman and Jonathan Weare. Ensemble samplers with affine invariance. *Communications in applied mathematics and computational science*, 5(1):65–80, 2010.

[126] John Duchi. Derivations for linear algebra and optimization. *Berkeley, California*, 2007.

[127] Daniel Foreman-Mackey, David W Hogg, Dustin Lang, and Jonathan Goodman. emcee: the mcmc hammer. *Publications of the Astronomical Society of the Pacific*, 125(925):306, 2013.

[128] András Sóbester, Stephen J Leary, and Andy J Keane. On the design of optimization strategies based on global response surface approximation models. *Journal of Global Optimization*, 33(1):31–59, 2005.

[129] Holger Dette and Andrey Pepelyshev. Generalized latin hypercube design for computer experiments. *Technometrics*, 52(4):421–429, 2010.

[130] Joshua Knowles. Parego: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation*, 10(1):50–66, 2006.

[131] Tan Bui-Thanh, Karen Willcox, and Omar Ghattas. Model reduction for large-scale systems with high-dimensional parametric input space. *SIAM Journal on Scientific Computing*, 30(6):3270–3288, 2008.

[132] Jeremy Oakley. Estimating percentiles of uncertain computer code outputs. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1):83–93, 2004.

[133] Paul Feliot, Julien Bect, and Emmanuel Vazquez. User preferences in bayesian multi-objective optimization: the expected weighted hypervolume improvement criterion. *arXiv preprint arXiv:1809.05450*, 2018.

[134] Felipe AC Viana, Raphael T Haftka, and Layne T Watson. Efficient global optimization algorithm assisted by multiple surrogate techniques. *Journal of Global Optimization*, 56(2):669–689, 2013.

[135] Emmanuel Vazquez and Julien Bect. Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *Journal of Statistical Planning and inference*, 140(11):3088–3095, 2010.

[136] Russell R Barton and Martin Meckesheimer. Metamodel-based simulation optimization. *Handbooks in operations research and management science*, 13:535–574, 2006.

[137] Timothy W Simpson, Timothy M Mauery, John J Korte, and Farrokh Mistree. Kriging models for global approximation in simulation-based multidisciplinary design optimization. *AIAA journal*, 39(12):2233–2241, 2001.

[138] Markus Heinonen, Henrik Mannerström, Juho Rousu, Samuel Kaski, and Harri Lähdesmäki. Non-stationary gaussian process regression with hamiltonian monte carlo. In *Artificial Intelligence and Statistics*, pages 732–740, 2016.

[139] Hossein Mohammadi, Peter Challenor, Marc Goodfellow, and Daniel Williamson. Emulating computer models with step-discontinuous outputs using gaussian processes. *arXiv preprint arXiv:1903.02071*, 2019.

[140] Xiao Lin, Asif Chowdhury, Xiaofan Wang, and Gabriel Terejanu. Approximate computational approaches for bayesian sensor placement in high dimensions. *Information Fusion*, 46:193–205, 2019.

[141] Adam Foster, Martin Jankowiak, Eli Bingham, Paul Horsfall, Yee Whye Teh, Tom Rainforth, and Noah Goodman. Variational estimators for bayesian optimal experimental design. *arXiv preprint arXiv:1903.05480*, 2019.

[142] Mark N Gibbs. *Bayesian Gaussian processes for regression and classification*. PhD thesis, Citeseer, 1998.

[143] Christopher Joseph Paciorek. *Nonstationary Gaussian processes for regression and spatial modelling*. PhD thesis, Citeseer, 2003.

[144] Robert B Gramacy and Herbert K H Lee. Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130, 2008.

[145] Carl E Rasmussen and Zoubin Ghahramani. Infinite mixtures of gaussian process experts. In *Advances in neural information processing systems*, pages 881–888, 2002.

[146] Christopher J Paciorek and Mark J Schervish. Nonstationary covariance functions for gaussian process regression. In *Advances in neural information processing systems*, pages 273–280, 2004.

[147] Christian Plagemann, Kristian Kersting, and Wolfram Burgard. Nonstationary gaussian process regression using point estimates of local smoothness. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 204–219. Springer, 2008.

[148] Edwin T Jaynes. *Probability theory: the logic of science*. Cambridge university press, 2003.

[149] Raymond H Myers, Douglas C Montgomery, and Christine M Anderson-Cook. *Response surface methodology: process and product optimization using designed experiments*. John Wiley & Sons, 2016.

[150] David JC MacKay. Introduction to gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168:133–166, 1998.

[151] Kristian Kersting, Christian Plagemann, Patrick Pfaff, and Wolfram Burgard. Most likely heteroscedastic gaussian process regression. In *Proceedings of the 24th international conference on Machine learning*, pages 393–400. ACM, 2007.

[152] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[153] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

[154] De G Matthews, G Alexander, Mark Van Der Wilk, Tom Nickson, Keisuke Fujii, Alexis Boukouvalas, Pablo León-Villagrá, Zoubin Ghahramani, and James Hensman. Gpflow: A gaussian process library using tensorflow. *The Journal of Machine Learning Research*, 18(1):1299–1304, 2017.

[155] Andrei Kramer, Ben Calderhead, and Nicole Radde. Hamiltonian monte carlo methods for efficient parameter estimation in steady state dynamical systems. *BMC bioinformatics*, 15(1):253, 2014.

[156] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[157] Nilesh Tripuraneni, Mark Rowland, Zoubin Ghahramani, and Richard Turner. Magnetic hamiltonian monte carlo. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3453–3461. JMLR. org, 2017.

[158] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[159] Christopher J Paciorek and Mark J Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics: The official journal of the International Environmetrics Society*, 17(5):483–506, 2006.

[160] Roger G Ghanem and Pol D Spanos. Stochastic finite element method: Response statistics. In *Stochastic finite elements: a spectral approach*, pages 101–119. Springer, 1991.

[161] Ilias Bilionis and Nicholas Zabaras. Bayesian uncertainty propagation using gaussian processes. *Handbook of Uncertainty Quantification*, pages 1–45, 2016.

[162] Olivier Le Maître and Omar M Knio. *Spectral methods for uncertainty quantification: with applications to computational fluid dynamics*. Springer Science & Business Media, 2010.

[163] John L Lumley. *Stochastic tools in turbulence*. Courier Corporation, 2007.

[164] Jack Sherman and Winifred J Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1950.

[165] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[166] Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[167] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer, 2001.

[168] Pratibha Vellanki, Santu Rana, Sunil Gupta, David Rubin de Celis Leal, Alessandra Sutti, Murray Height, and Svetha Venkatesh. Bayesian functional optimisation with shape prior. *arXiv preprint arXiv:1809.07260*, 2018.

[169] Kimberly R Saviers, Majed A Alrefae, and Timothy S Fisher. Roll-to-roll production of graphitic petals on carbon fiber tow. *Advanced Engineering Materials*, 20(8):1800004, 2018.

VITA

VITA

Piyush Pandita completed his undergraduate studies in Mechanical Engineering at Punjab Engineering College, Chandigarh, India, in 2014. He subsequently joined the graduate school at Purdue University to pursue a Ph.D. in Mechanical Engineering. After the completion of his studies, at Purdue, he wishes to pursue his research interests further in industry and academia.