Bayesian Optimal Design of Experiments for Inferring the Statistical Expectation of Expensive Black-Box Functions

Piyush Pandita[1]
School of Mechanical Engineering,
Purdue University,
West Lafayette, IN 47907
e-mail: ppandit@purdue.edu

Ilias Bilionis
School of Mechanical Engineering,
Purdue University,
West Lafayette, IN 47907
e-mail: ibilion@purdue.edu

Jitesh Panchal
School of Mechanical Engineering,
Purdue University,
West Lafayette, IN 47907
e-mail: panchal@purdue.edu

Bayesian optimal design of experiments (BODEs) have 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 paper 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 paper 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. [DOI: 10.1115/1.4043930]

Keywords: optimal experimental design, Kullback–Leibler divergence, uncertainty quantification, information gain, mutual information, Gaussian processes, Bayesian inference

1 Introduction

Engineering problems require either computationally intensive computer codes [1] or expensive physical experiments [2]. 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 experiment (DOE) techniques that attempt to select the maximally informative physical response evaluations within a given budget [3–5]. Classical DOE techniques generate a single batch design, [6] and thus, they face several shortcomings in case of functions with local features, e.g., discontinuities, or sharp nonlinearities [7]. 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 [8,9]. Over the past two decades, SDOE has been used in several applications spanning both physical experiments [2,10–13] and computer simulations [14–17]. One of the most theoretically sound SDOEs is Bayesian optimal design of experiments (BODEs). 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 [18–30], learning an accurate representation of the physical response [31–36], or estimating the probability of a rare event [37,38].

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) [39,40] (also known as relative entropy). Over the years, KLD has been used to quantify information gain [41] 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 [26,42], surrogate modeling [43–45], learning missing parameters [46], optimizing an expensive physical response [47], calibrating a physical model [48,49], reliability design [50,51], efficient design space exploration [52], probabilistic sensitivity analysis [53], portfolio optimization [54], and neural-network hyperparameter tuning [55].

Despite the significant progress, deriving BODE methods for new objectives remains a nontrivial 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 the robust design. To address this need, the *objective* of this paper 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 [56]. The expected information gain from a

[1]Corresponding author.

Contributed by the Design Automation Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received July 26, 2018; final manuscript received May 28, 2019; published online July 10, 2019. Assoc. Editor: James T. Allison.

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 expected KLD (EKLD) of the statistical expectation of the physical response comes out to be an analytically tractable function which alleviates the need for sample averaging.

In summary, our main contributions are 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 (US); (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.[2]

The rest of the paper is organized as follows: Sec. 2 describes in detail the methodology used, including GP regression (Sec. 2.1) and the EKLD (Sec. 2.2). The results obtained for four synthetic test problems have been presented in Sec. 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. 3.6. The steel wire manufacturing problem is briefly explained and treated with the proposed methodology in Sec. 3.5. We summarize the nuances of the methodology including its weaknesses and comment on future research directions in Sec. 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., $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.
(2) $\mathbf{Y}_n$ are the values of the physical response at the corresponding n designs, i.e., $\mathbf{Y}_n = \{y_1, \ldots, y_n\}$.
(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 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

$$\mathcal{L}^2(\mathcal{X}) = \left\{ f : \mathcal{X} \to \mathbb{R} \middle| \int_{\mathcal{X}} f^2(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} < \infty \right\} \tag{1}$$

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

$$\boldsymbol{Q}[f] = \int_{\mathcal{X}} f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \tag{2}$$

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 [57,58]; however, we restrict the focus of this work to the sequential experiment design. We will select the new experiment by maximizing the expected information gain for $\boldsymbol{Q}[f]$.

2.1 Surrogate Modeling. GP regression is a very popular nonparametric 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 Ref. [59].

2.1.1 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,

$$f \sim \mathrm{GP}(0, k) \tag{3}$$

where

$$k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}'; \boldsymbol{\psi}) = s^2 \exp\left\{ -\frac{1}{2} \sum_{j=1}^{d} \frac{(x_j - x_j')^2}{\ell_j^2} \right\} \tag{4}$$

The covariance function defined in Eq. (4) encodes our prior beliefs about the smoothness and the magnitude of the response. The symbol $\ell_j > 0$ in Eq. (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) is the signal strength of the GP. It incorporates the scale of the response. These parameters are the hyperparameters of the covariance function, and we will denote them by $\boldsymbol{\psi}$, i.e., $\boldsymbol{\psi} = \{s^2, \ell_1, \ldots, \ell_d\}$. A nonzero mean function can always be included with only minor modifications in what follows.

2.1.2 The Data Likelihood. The likelihood of the data $\mathbf{Y}_n$ is 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). The observations are assumed to be contaminated with Gaussian noise with variance σ^2. This noise variance is very small relative to the signal strength in the case of computer simulation design. We augment the vector of hyperparameters to include this additional parameter to get $\boldsymbol{\theta} = \{\boldsymbol{\psi}, \sigma^2\}$. Mathematically, the likelihood of the observed data is

$$p(\mathbf{Y}_n | \mathbf{X}_n, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{Y}_n | \mathbf{0}, \mathbf{K}_n + \sigma^2 \mathbf{I}_n) \tag{5}$$

where $\mathbf{K}_n$ is a $n \times n$ covariance matrix defined according to Eq. (4), i.e., $K_{nij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

2.1.3 Training the Hyperparameters. Typically, the hyperparameter values are fitted to the observed data by maximizing the likelihood of Eq. (5). However, this process may result in overfitting which is particularly problematic in the context of SDOE. In this work, we opt for a full Bayesian treatment [60] which is more robust. We assume that the hyperparameters are a priori independent following an exponential prior distribution on the lengthscales and the 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 1×10^{-6}, which is a reasonably small value. Bayes' rule allows yielding the hyperparameter posterior

$$p(\boldsymbol{\theta} | \mathbf{D}_n) \propto p(\mathbf{Y}_n | \mathbf{X}_n, \boldsymbol{\psi}) p(\boldsymbol{\psi}) \tag{6}$$

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 Ref. [61]. The code for this MCMC algorithm is available online.[3]

[2] https://github.com/piyushpandita92/bode

[3] https://github.com/dfm/emcee

2.1.4 Making Predictions. Conditioned on the hyperparameters, our state of knowledge about f is also characterized by a GP

$$f|\mathbf{D}_n, \boldsymbol{\theta} \sim GP(f|m_n, k_n) \tag{7}$$

where

$$m_n(\mathbf{x}) = (\mathbf{k}_n(\mathbf{x}))^T \left(\mathbf{K}_n + \sigma^2 \mathbf{I}_n\right)^{-1} \mathbf{Y}_n \tag{8}$$

with

$$\boldsymbol{\alpha}_n = (\mathbf{K}_n + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{Y}_n \tag{9}$$

is the *posterior mean* function, and

$$k_n(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - (\mathbf{k}_n(\mathbf{x}))^T \left(\mathbf{K}_n + \sigma^2 \mathbf{I}_n\right)^{-1} \mathbf{k}_n(\mathbf{x}') \tag{10}$$

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

$$p(\tilde{y}|\tilde{\mathbf{x}}, \mathbf{D}_n, \boldsymbol{\theta}) = N\left(\tilde{y}|m_n(\tilde{\mathbf{x}}; \boldsymbol{\theta}), \sigma_n^2(\tilde{\mathbf{x}}; \boldsymbol{\theta})\right) \tag{11}$$

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

$$p(\tilde{y}|\tilde{\mathbf{x}}, \mathbf{D}_n) = \int p(\tilde{y}|\tilde{\mathbf{x}}, \mathbf{D}_n, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{D}_n) d\boldsymbol{\theta} \tag{12}$$

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

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 as

$$p(q|\boldsymbol{\theta}, \mathbf{D}_n) = \mathbb{E}\left[\delta\left(q - \boldsymbol{Q}[f]\right)|\boldsymbol{\theta}, \mathbf{D}_n\right] \tag{13}$$

where $\delta(\cdot)$ is Dirac's delta function and the expectation is over the function space measure defined by the posterior GP (see Eq. (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

$$p(q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}) = \mathbb{E}\left[\delta\left(q - \boldsymbol{Q}[f]\right)|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}\right] \tag{14}$$

According to the 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

$$G(\tilde{\mathbf{x}}, \tilde{y}; \boldsymbol{\theta}) = \int_{-\infty}^{\infty} p(q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}) \log \frac{p(q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y})}{p(q|\boldsymbol{\theta}, \mathbf{D}_n)} dq \tag{15}$$

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,

$$G(\tilde{\mathbf{x}}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G(\tilde{\mathbf{x}}, \tilde{y}; \boldsymbol{\theta}) p(\tilde{y}|\boldsymbol{\theta}, \tilde{\mathbf{x}}, \mathbf{D}_n) p(\boldsymbol{\theta}|\mathbf{D}_n) d\tilde{y} d\boldsymbol{\theta} \tag{16}$$

We pick the next experiment by solving

$$\mathbf{x}_{n+1} = \arg\max_{\tilde{\mathbf{x}}} G(\tilde{\mathbf{x}}) \tag{17}$$

In the rest of this section, we derive analytical approximations of $p(q|\boldsymbol{\theta}, \mathbf{D}_n)$ (Sec. 2.2.1), $p(q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y})$ (Sec. 2.2.2), $G(\tilde{\mathbf{x}}, \tilde{y}; \boldsymbol{\theta})$ (Sec. 2.2.3), and a sampling average approximation for $G(\tilde{\mathbf{x}})$ (Sec. 2.2.3).

2.2.1 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. (2)) is linear and the point predictive PDF of $y = f(x)$ is Gaussian (Eq. (11)), $p(q|\boldsymbol{\theta}, \mathbf{D}_n)$ is Gaussian. In particular, it is easy to show that

$$p(q|\boldsymbol{\theta}, \mathbf{D}_n) = \mathcal{N}\left(q|\mu_1, \sigma_1^2\right) \tag{18}$$

The mean μ_1 is given as

$$\begin{aligned} \mu_1 &:= \mathbb{E}[\boldsymbol{Q}[f]|\boldsymbol{\theta}, D_n] \\ &= \mathbb{E}\left[\int_{\mathcal{X}} f(\mathbf{x}) p(x) d\mathbf{x} \middle| \boldsymbol{\theta}, D_n\right] \\ &= \int_{\mathcal{X}} \mathbb{E}[f(\mathbf{x})|\boldsymbol{\theta}, D_n] p(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathcal{X}} m_n(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\ &= \boldsymbol{\epsilon}_n^T \boldsymbol{\alpha}_n \end{aligned} \tag{19}$$

where $\boldsymbol{\alpha}_n$ is defined in Eq. (9) and each component of $\boldsymbol{\epsilon}_n \in \mathbb{R}^n$ is given as

$$\begin{aligned} \epsilon_{\mathrm{ni}} &= \epsilon(\mathbf{x}_i) \\ &:= \int_{\mathcal{X}} k(\mathbf{x}_i, \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\ &= s^2 \left(\frac{\pi}{2}\right)^{d/2} \prod_{k=1}^{d} \left\{ \ell_k \left[\mathrm{erf}\left(\frac{1 - x_{ik}}{\sqrt{2}\ell_k}\right) - \mathrm{erf}\left(-\frac{x_{ik}}{\sqrt{2}\ell_k}\right) \right] \right\} \end{aligned} \tag{20}$$

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

$$\begin{aligned} \sigma_1^2 &:= \mathbb{E}[\boldsymbol{Q}^2[f]|\boldsymbol{\theta}, D_n] - (\mathbb{E}[\boldsymbol{Q}[f]|\boldsymbol{\theta}, D_n])^2 \\ &= \mathbb{E}\left[\left(\int_{\mathcal{X}} f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}\right)^2 \| \boldsymbol{\theta}, D_n\right] - {\mu_1}^2 \\ &= \mathbb{E}\left[\int_{\mathcal{X}} f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \int_{\mathcal{X}} f(\mathbf{x}') p(\mathbf{x}') d\mathbf{x}' \| \boldsymbol{\theta}, D_n\right] - {\mu_1}^2 \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} \mathbb{E}[f(\mathbf{x}) f(\mathbf{x}')|\boldsymbol{\theta}, D_n] p(\mathbf{x}) p(\mathbf{x}') d\mathbf{x} d\mathbf{x}' - {\mu_1}^2 \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} [k_n(\mathbf{x}, \mathbf{x}') + m_n(\mathbf{x}) m_n(\mathbf{x}')] p(\mathbf{x}) p(\mathbf{x}') d\mathbf{x} d\mathbf{x}' - {\mu_1}^2 \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} k_n(\mathbf{x}, \mathbf{x}') p(\mathbf{x}) p(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \\ &= \sigma_0^2 - \boldsymbol{\epsilon}_n^T \left(\mathbf{K}_n + \sigma^2\right)^{-1} \boldsymbol{\epsilon}_n \end{aligned} \tag{21}$$

where

$$\begin{aligned} \sigma_0^2 &= \int_{\mathcal{X}} \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}') p(\mathbf{x}) p(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \\ &= s^2 \prod_{k=1}^{d} (2\ell_k^2 \sqrt{\pi}) \left\{ \frac{-1}{\sqrt{\pi}} + \frac{1}{\sqrt{\pi}} \exp\left(\frac{-1}{2\ell_k^2}\right) \right. \\ &\quad \left. + \frac{1}{\sqrt{2}\ell_k} \mathrm{erf}\left(\frac{1}{\sqrt{2}\ell_k}\right) \right\} \end{aligned} \tag{22}$$

2.2.2 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. 2.2.1, but with the remark that the posterior GP after adding the hypothetical observation will have mean function

$$\tilde{\mu}_{n+1}(\mathbf{x}) = \mu_n(\mathbf{x}) + k_n(\mathbf{x}, \tilde{\mathbf{x}}) \frac{\tilde{y} - \mu_n(\tilde{\mathbf{x}})}{k_n(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + \sigma^2} \tag{23}$$

and covariance function

$$\tilde{k}_{n+1}(\mathbf{x}, \mathbf{x}') = k_n(\mathbf{x}, \mathbf{x}') - \frac{k_n(\mathbf{x}, \tilde{\mathbf{x}}) k_n(\tilde{\mathbf{x}}, \mathbf{x}')}{k_n(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + \sigma^2} \tag{24}$$

We get

$$p(q|\boldsymbol{\theta}, \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}) = \mathcal{N}(q|\mu_2(\tilde{\mathbf{x}}, \tilde{y}), \sigma_2^2(\tilde{\mathbf{x}})) \tag{25}$$

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

$$\begin{aligned} \mu_2(\tilde{\mathbf{x}}, \tilde{y}) &:= \mathbb{E}[\boldsymbol{Q}[f]|\boldsymbol{\theta}, D_n, \tilde{\mathbf{x}}, \tilde{y}] \\ &= \int_{\mathcal{X}} \tilde{\mu}_{n+1}(\mathbf{x}) d\mathbf{x} \\ &= \mu_1 + \frac{\nu(\tilde{\mathbf{x}})}{k_n(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + \sigma^2} \left(\tilde{y} - \mu_n(\tilde{\mathbf{x}})\right) \end{aligned} \tag{26}$$

with

$$\nu(\tilde{\mathbf{x}}) := \epsilon(\tilde{\mathbf{x}}) - \boldsymbol{\epsilon}_n^T \left(\mathbf{K}_n + \sigma^2\right)^{-1} \mathbf{k}_n(\tilde{\mathbf{x}}) \tag{27}$$

where $\epsilon(\tilde{\mathbf{x}})$ is as in Eq. (20) but with $\mathbf{x}_i$ replaced by $\tilde{\mathbf{x}}$. Using the expression for the posterior covariance from Eq. (24), one can simplify $\sigma_2^2(\tilde{\mathbf{x}})$ similar to the derivation in Eq. (21) to get

$$\begin{aligned} \sigma_2^2(\tilde{\mathbf{x}}) &:= \mathbb{E}[\boldsymbol{Q}^2[f]|\boldsymbol{\theta}, D_n, \tilde{\mathbf{x}}, \tilde{y}] - (\mathbb{E}[\boldsymbol{Q}[f]|\boldsymbol{\theta}, D_n, \tilde{\mathbf{x}}, \tilde{y}])^2 \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} \tilde{k}_{n+1}(\mathbf{x}, \mathbf{x}') p(\mathbf{x}) p(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \\ &= \sigma_1^2 - \frac{\nu^2(\tilde{\mathbf{x}})}{k_n(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + \sigma^2} \end{aligned} \tag{28}$$

2.2.3 Quantification of the Expected Information Gain About the QoI. Since both Eqs. (18) and (25) are Gaussian, the Kullback–Liebler (KL) divergence between the hypothetical and the current state of knowledge about the QoI conditional on the hyperparameters, $G(\mathbf{x}, \tilde{y}; \boldsymbol{\theta})$ of Eq. (15), is analytically tractable [62], i.e.,

$$G(\mathbf{x}, \tilde{y}; \boldsymbol{\theta}) = \log\left(\frac{\sigma_1}{\sigma_2(\tilde{\mathbf{x}})}\right) + \frac{\sigma_2^2(\tilde{\mathbf{x}})}{2\sigma_1^2} + \frac{(\mu_2(\tilde{\mathbf{x}}, \tilde{y}) - \mu_1)^2}{2\sigma_1^2} - \frac{1}{2} \tag{29}$$

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. (11)). Thus, we can analytically integrate $\tilde{y}$ out to obtain

$$\begin{aligned} G(\tilde{\mathbf{x}}; \boldsymbol{\theta}) &= \int_{-\infty}^{\infty} G(\tilde{\mathbf{x}}, \tilde{y}; \boldsymbol{\theta}) p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\theta}, \mathbf{D}_n) d\tilde{y} \\ &= \log\left(\frac{\sigma_1}{\sigma_2(\tilde{\mathbf{x}})}\right) + \frac{1}{2}\frac{{\sigma_2}^2(\tilde{\mathbf{x}})}{{\sigma_1}^2} - \frac{1}{2} \\ &\quad + \frac{1}{2}\frac{v(\tilde{\mathbf{x}})^2}{\sigma_1^2(\sigma_n^2(\tilde{\mathbf{x}}) + \sigma^2)} \end{aligned} \tag{30}$$

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. (6), using the MCMC samples $\{\boldsymbol{\theta}^{(s)}\}_{s=1}^S$ collected with the procedure described in Refs. [61,63]. This yields

$$G(\tilde{\mathbf{x}}) = \int G(\tilde{\mathbf{x}}; \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{D}_n) d\boldsymbol{\theta} \approx \frac{1}{S} \sum_{s=1}^{S} G\left(\tilde{\mathbf{x}}; \boldsymbol{\theta}^{(s)}\right) \tag{31}$$

2.2.4 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) [21]. This choice takes into account the noisy nature of the approximation of Eq. (31), and it reduces the computational time compared to a brute force or a multistart-and-gradient-based-optimization approach. See Algorithm 1 for pseudocode. In all our experiments, irrespective of the dimensionality, we use $T_n = 20$ BGO iterations to optimize the EKLD.

Algorithm 1 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 $\{\boldsymbol{\theta}^{(s)}\}_{s=1}^S$; stopping tolerance $\gamma_i > 0$.

1. Evaluate $G(\tilde{\mathbf{x}})$ using Eq. (31) 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 $\tilde{\mathbf{G}}_{T_i} = \{\tilde{G}_1 = G(\mathbf{x}_1), \ldots \tilde{G}_{T_i} = G(\mathbf{x}_{T_i})\}$, for BGO.
2. $t \leftarrow t_i$.
3. **while** $t < T_n$ **do**
4. Fit a standard GP on the input–output pairs $\tilde{\mathbf{X}}_t$–$\tilde{\mathbf{G}}_t$ using maximum likelihood to approximate $G(\tilde{\mathbf{x}})$.
5. 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) [40].
6. Compute the AEI of all of the candidate points in $\hat{\mathbf{X}}_{n_d}$.
7. Find the candidate point $\hat{\mathbf{x}}_j$ that exhibits the maximum AEI.
8. **if** If the maximum AEI is smaller **than** γ_i **then**
9. Break.
10. **end if**
11. Use Eq. (31) to evaluate $G(\tilde{\mathbf{x}})$ at $\hat{\mathbf{x}}_j$ measuring $\hat{G}_j = G(\hat{\mathbf{x}}_j)$.
12. $\tilde{x}_{t+1} \leftarrow \hat{x}_j$.
13. $\tilde{G}_{t+1} \leftarrow \hat{G}_j$.
14. $\mathbf{X}_{t+1} \leftarrow \tilde{\mathbf{X}}_t \cup \{\tilde{\mathbf{x}}_{t+1}\}$.
15. $\mathbf{G}_{t+1} \leftarrow \mathbf{G}_t \cup \{\tilde{G}_{t+1}\}$.
16. $t \leftarrow t + 1$.
17. **end while**
18. return $\arg\max_{\tilde{\mathbf{X}}_{T_n}} \tilde{\mathbf{G}}_{T_n}$.

2.2.5 Selecting the Initial Set of Designs. In most literature, as a rule of thumb, $10d$ number of initial samples are used. We resort 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. [64] 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.

2.2.6 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.

2.2.7 Complete BODE Framework. In Algorithm 2, 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.

3 Results

We apply the methodology on two one-dimensional mathematical functions (synthetic problems), a three-dimensional problem and a five-dimensional problem. For the first two synthetic problems, the input domain simply becomes [0,1], whereas for the third synthetic problem, the input domain is $[-2, 6]^3$. The inputs for the five-dimensional numerical example lie in the hypercube $[0, 1]^5$.

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 in order to test the performance of the methodology in the low-sample regime. Readers interested in the problem of selecting the optimal number of initial-data can refer to the work of Sòbester et al. [64] where the authors discuss the problem in the context of optimization. The problem of selecting an optimal number of initial-data is beyond the scope of the work presented here.

Algorithm 2 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. (6), to obtain $\{\boldsymbol{\theta}^{(s)}\}_{s=1}^{S}$.
4. Find the next experiment $\mathbf{x}_{n+1}$ using Algorithm 1 to solve Eq. (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**

3.1 Synthetic Problem No. 1.

Consider the function

$$f(x) = 4\left(1 - \sin\left(6x + 8e^{6x-7}\right)\right) \tag{32}$$

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 following 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 Refs. [61,63].

Figures 1(*a*) and 1(*b*) show the initial and final state of Algorithm 2. The solid line represents the true function f (Eq. (32)). The black crosses are the observed data at the given stage. In Fig. 1(*a*), the next experiment selected by maximizing the EKLD, see Algorithm 1, corresponds to the diamond. The mean of the GP fit to the expected information gain $G(\tilde{\mathbf{x}})$ constructed by BGO in Algorithm 1. The predictive mean of the EKLD is shown by the dotted line. This dotted line represents the response surface of the EKLD after the BGO has ended, and the 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 (diamond) is located in the input space where the EKLD has high mean. The posterior mean of the GP is represented by the dashed line. The shaded area around it represents the uncertainty (2.5 percentile and 97.5 percentile).

The final set of inputs, space-filling, selected by the methodology can be seen in Fig. 1(*b*). Figure 1(*c*) shows the $p(q|\mathbf{D}_n)$ plotted against the number of data samples while showing convergence toward 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. 1(*c*).

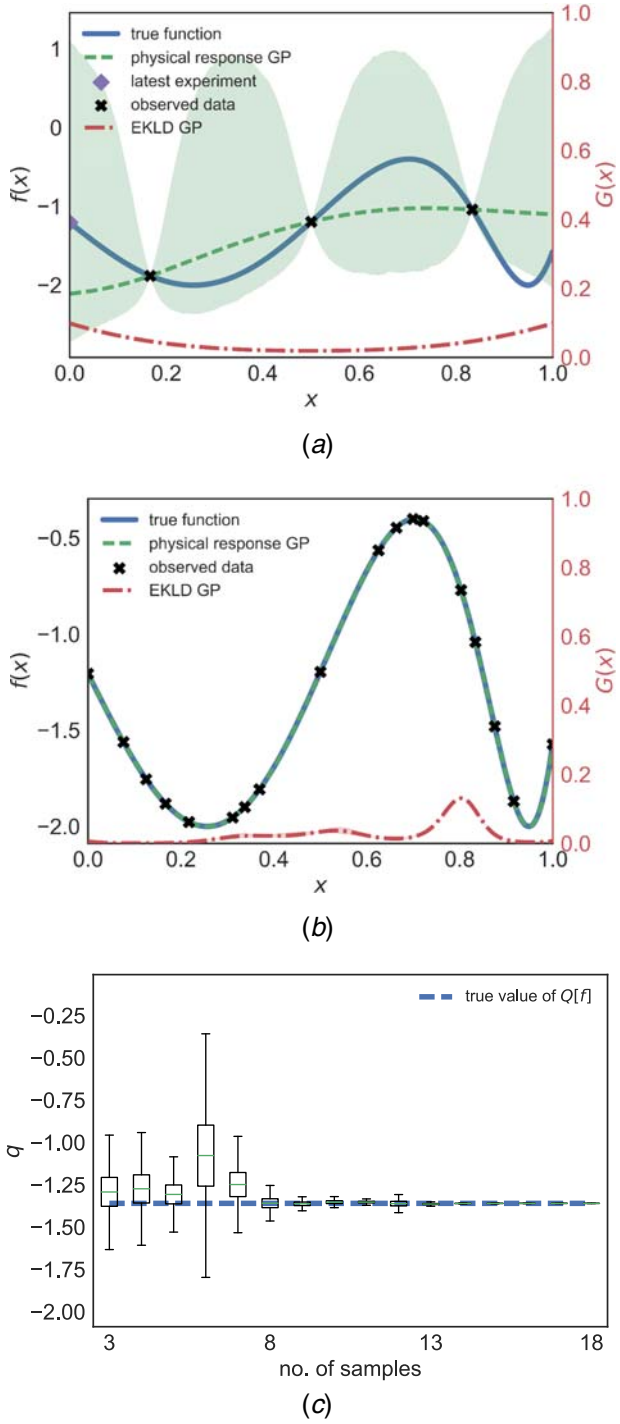

Fig. 1 One-dimensional synthetic problem ($n_i = 3$). (*a*) and (*b*) The state of the function (1st iteration) at the start and the end (15th iteration) of the algorithm. (*c*) The convergence to the true expectation of the function and the reduction in uncertainty about the QoI after the end of the algorithm.

3.2 Synthetic Problem No. 2.

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

$$f(x) = \frac{1}{\sqrt{2\pi}s_1}\exp\left\{-\frac{(x-m_1)^2}{2{s_1}^2}\right\} + \frac{1}{\sqrt{2\pi}s_2}\exp\left\{-\frac{(x-m_2)^2}{2{s_2}^2}\right\} \tag{33}$$

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. (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. 2(*b*), which shows a fairly equally spaced spread of designs. It is important to note that Fig. 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. 2(*c*).

In Figs. 2(*a*) and 2(*b*), the EKLD is shown by the dotted line and the true function is shown by the solid line. The dashed line

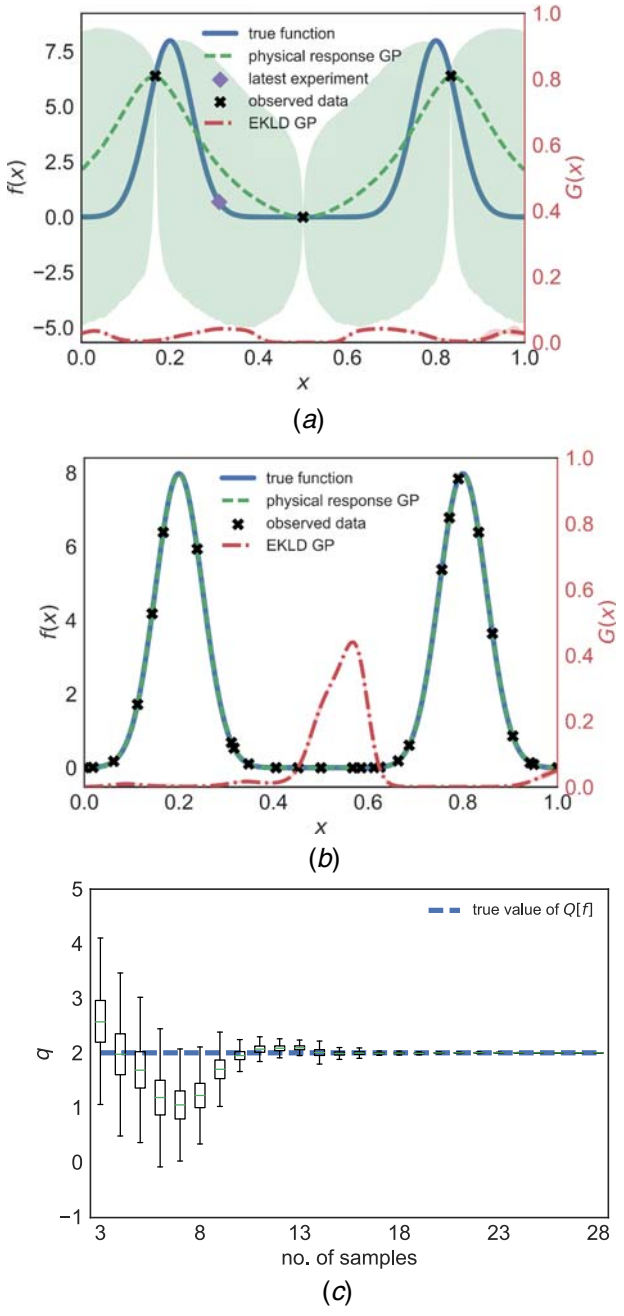

Fig. 2 One-dimensional synthetic example ($n_i = 3$). (*a*) and (*b*) The state of the function at the start (1st iteration) and the end (25th iteration) of the algorithm. (c) The convergence to the true expectation of the function and the reduction in uncertainty about the QoI after the end of the algorithm.

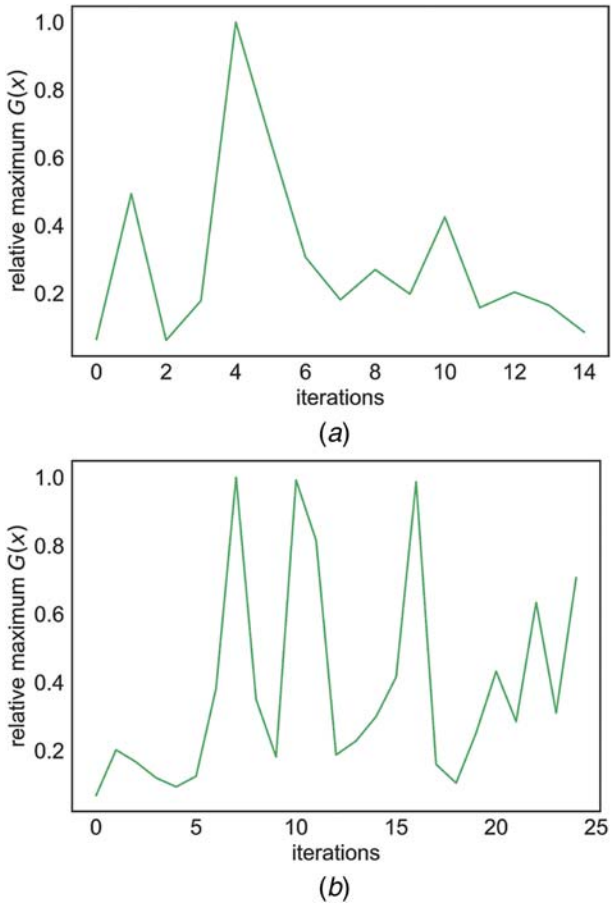

Fig. 3 One-dimensional synthetic examples. (*a*) and (*b*) The predictive mean of the EKLD for synthetic problem no. 1 ($n_i = 3$) and synthetic problem no. 2 ($n_i = 4$), respectively.

represents the mean of the GP model, and the 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 Figs. 3(*a*) and 3(*b*) for both the synthetic function 1 and 2 respectively. 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 plot in Fig. 3(*a*) show a characteristic that is typical of BODE functions, i.e., an increase in magnitude in the first few iterations followed by a sudden decline. This predicted mean value of the EKLD asymptotically goes to zero for both the synthetic functions here. The number of MCMC chains for the results shown below is six, and the number of steps per chain is 500.

3.3 Synthetic Problem No. 3. We consider the following three-dimensional function from [65] to test and validate our methodology further:

$$f(\mathbf{x}) = 4(x_1 + 8x_2 - 8x_2^2 - 2)^2 + (3 - 4x_2)^2 + 16\sqrt{x_3 + 1}(2x_3 - 1)^2 \tag{34}$$

The major difference between this function (Eq. (34)) and 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. Figure 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(*b*). The gradual reduction in uncertainty around

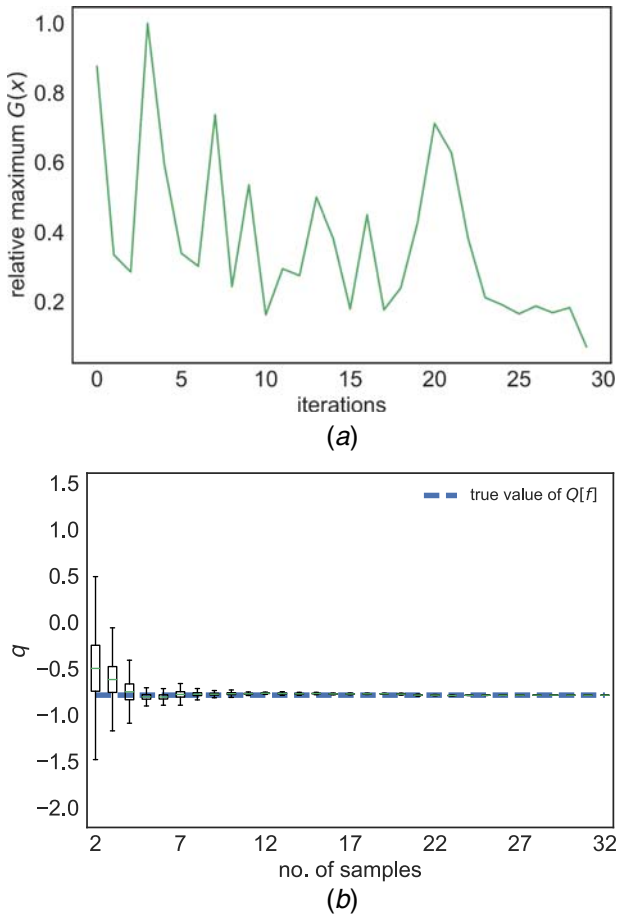

Fig. 4 Three-dimensional synthetic example ($n_i = 2$). (*a*) The decay of the EKLD from the 1st iteration to the end of the 30th iteration of the algorithm. (*b*) The convergence to the true value of the QoI.

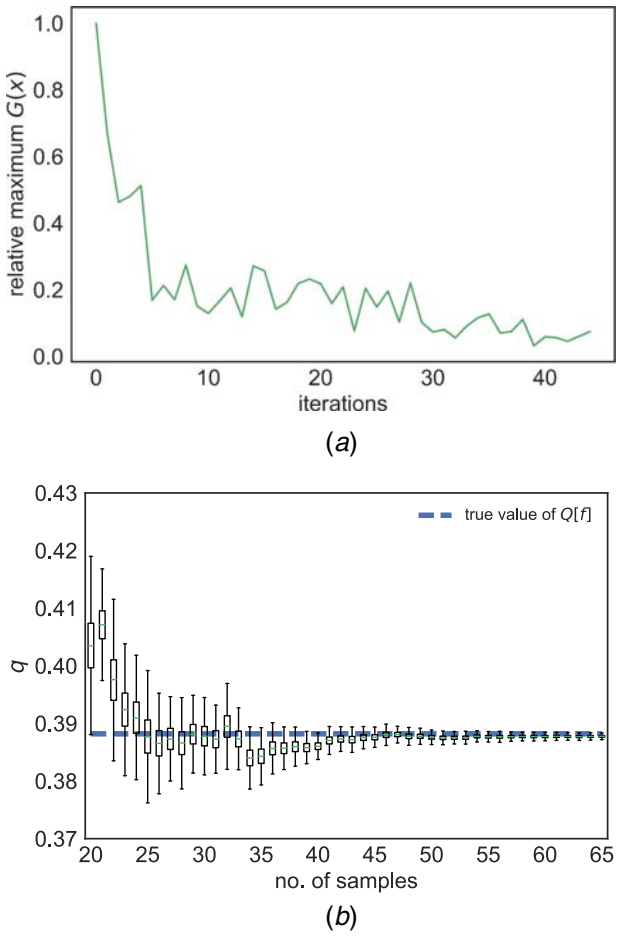

Fig. 5 Five-dimensional synthetic example ($n_i = 20$). (*a*) The decay of the EKLD from the 1st iteration to the end of the 45th iteration of the algorithm. (*b*) Convergence to the true value of the QoI.

$\boldsymbol{Q}[f]$ also can be seen in Fig. 4(*b*). Figure 4(*a*) demonstrates how the relative EKLD fluctuates while seemingly approaching zero.

3.4 Synthetic Problem No. 4. The following five-dimensional function is taken from Ref. [66]:

$$f(\mathbf{x}) = 10\sin(\pi x_1 x_2) + 20(x_3 - 5)^2 + 10x_4 + 5x_5 \tag{35}$$

This function (Eq. (35)) is reasonably high-dimensional and challenging due to the nonlinear 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. Figure 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. 5(*b*). Figure 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.

3.5 Steel 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 (eight 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 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 multipass drawing effect is modeled by considering a carryover effect of previous pass such as residual stress, plastic strain, and temperature. The finite element analysis 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 eight design variables. We start the methodology with 20 initial data points and add another 75 samples. We approximate the true value of the expectation of FWT, by averaging the outputs at 6000 designs generated by LHS, as $\boldsymbol{Q}[\text{FWT}] \approx 0.2694$. The results in Fig. 6 show the gradual convergence of the methodology's mean estimate of the QoI toward the approximated true value. Figure 6(*a*) shows the mean and variance of the expectation of FWT

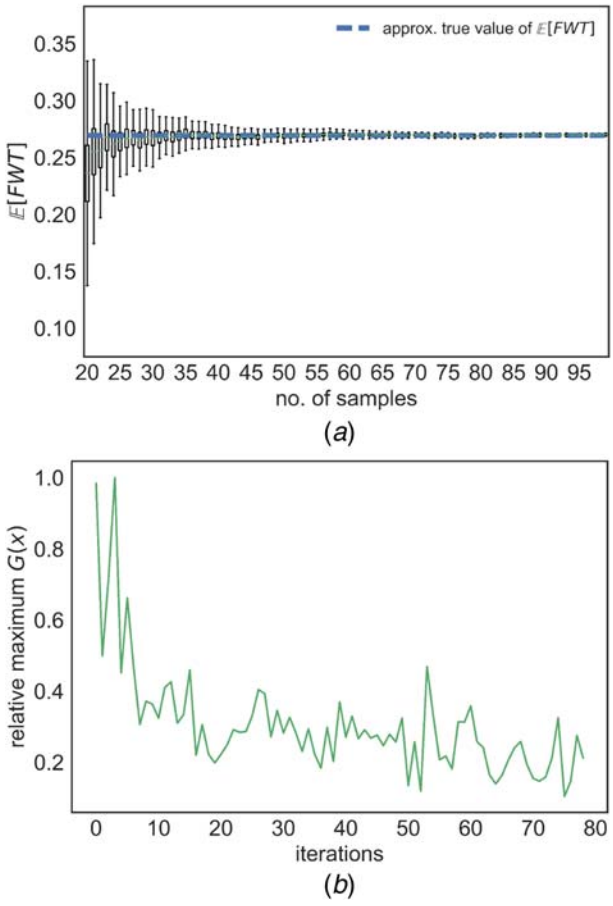

Fig. 6 Wire drawing problem ($n_i = 20$) after 75 iterations

as the mean approaches the approximate true value while the variance around it decreases gradually. The reduction in EKLD from the start of the sampling to the end can be seen in Fig. 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. 8(*b*). 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 on 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.

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, US, the methodology is tested on the synthetic examples given in Secs. 3.1–3.4. 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 [31]. Moreover, US, in its functional form, 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 2, but for the change in the sampling criterion.

The convergence to the QoI for the synthetic problem in Secs. 3.1 and 3.2 is seen in Figs. 7(*a*) and 7(*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. 3.2, the EKLD takes more iterations to converge as seen in Fig. 7(*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 problem no. 2, whereas EKLD takes slightly fewer iterations to estimate the true statistical expectation value for synthetic problem no. 1.

As the complexity of the problems increases, convergence for the EKLD becomes quicker compared to US as shown in

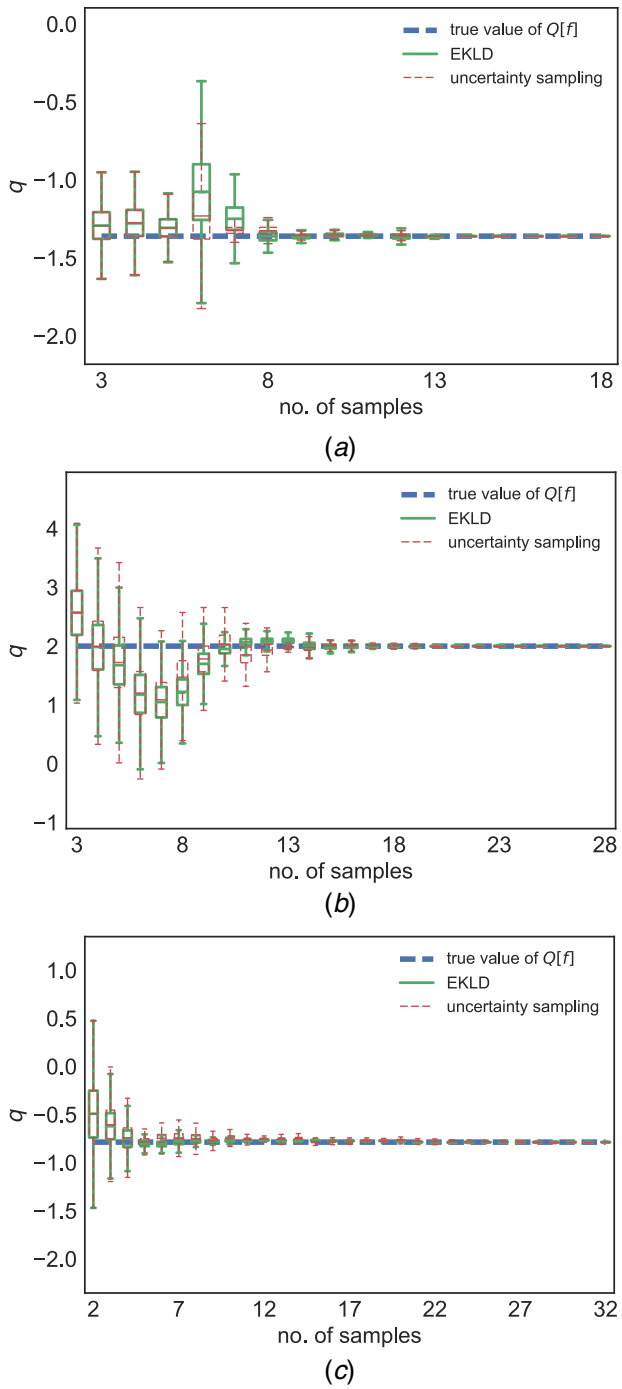

Fig. 7 (*a*)–(*c*) The comparison of the EKLD to uncertainty sampling for synthetic problem nos. 1, 2, and 3, respectively

Figs. 7(*a*)–7(*c*). With the three-dimensional problem (Fig. 7(*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 problem, Fig. 8(*b*) shows how the EKLD starts to approach the true value of the QoI as the number of iterations increases, whereas US tends to show 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. 5(*b*). The comparison in Fig. 8(*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. 8(*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 Figs. 7 and 8 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 [67] 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.

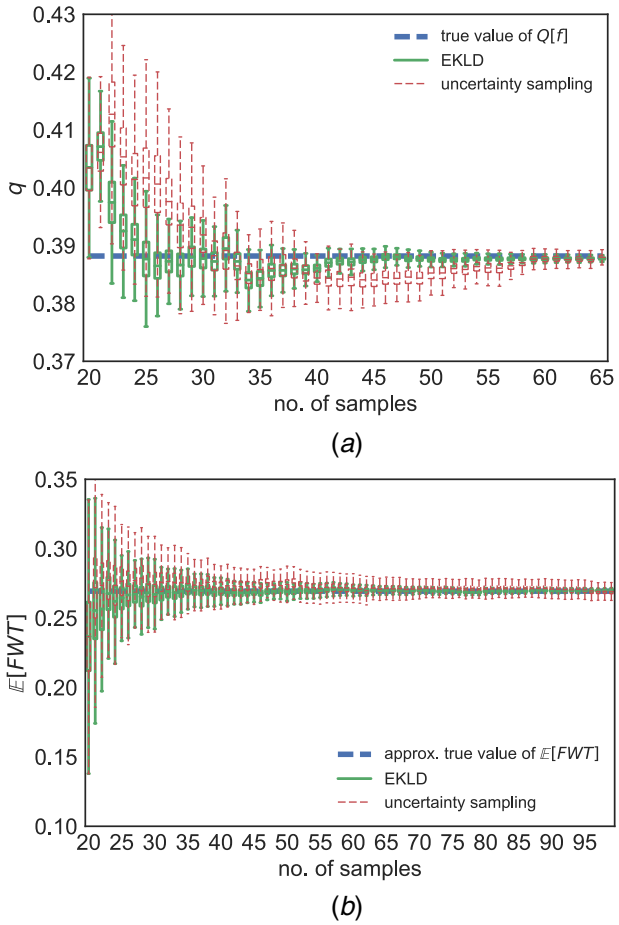

Fig. 8 (*a*) and (*b*) The comparison of the EKLD to uncertainty sampling for synthetic problem no. 4 and the wire-drawing problem, respectively

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 EKLD is clearly better in higher-dimensions both in terms of point-wise estimation error and reduction in epistemic uncertainty. For one-dimensional 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 need to be done to truly analyze and point out the difference between the two methodologies. The use of nonstationary GPs is a natural way to fully test the merits and pitfalls of the two methodologies, as it would result 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* [63] 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 hyperparameters of the GP model. Thus, the number of chains grows as the dimensionality of the problem increases leading to increased computational cost.

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. This work 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 nonstationary covariance function for more locally adapted designs. However, the problem of implementing a nonstationary GP is not trivial. Another area of limited research is the selection of a 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. We plan to address these challenges in our future work.

Acknowledgment

The authors thank their collaborators at TRDDC, Tata Consultancy Services, Pune, India, for providing the steel wire manufacturing problem.

Funding Data

- National Science Foundation (NSF) (Grant No. 1662230).

References

[1] Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P., 1989, "Design and Analysis of Computer Experiments," Stat. Sci., **4**(4), pp. 409–423.
[2] Flournoy, N., 1993, *Bayesian Statistics in Science and Technology: Case Studies*, C. Gatsonis, J. Hodges, R. E. Kass, and N. Singpurwalla, eds., Springer, New York, pp. 324–336.
[3] Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wikström, C., and Wold, S., 2000, *Principles and Applications*, Learn Ways AB, Stockholm.
[4] Anderson, M. J., and Whitcomb, P. J., 2000, *Design of Experiments*, Wiley Online Library, New York.
[5] Alexanderian, A., Petra, N., Stadler, G., and Ghattas, O., 2014, "A-Optimal Design of Experiments for Infinite-Dimensional Bayesian Linear Inverse Problems With Regularized L0 Sparsification," SIAM J. Sci. Comput., **36**(5), pp. A2122–A2148.
[6] Montgomery, D. C., 2017, *Design and Analysis of Experiments*, John Wiley & Sons, New York.
[7] Chaloner, K., and Verdinelli, I., 1995, "Bayesian Experimental Design: A Review," Stat. Sci., **10**(3), pp. 273–304.
[8] Chernoff, H., 1959, "Sequential Design of Experiments," Ann. Math. Stat., **30**(3), pp. 755–770.
[9] Robbins, H., 1952, "Some Aspects of the Sequential Design of Experiments," Bulletin of the American Mathematical Society, **58**(5), pp. 527–535.
[10] Havinga, J., Klaseboer, G., and Van den Boogaard, A., 2013, *Key Engineering Materials:The Current State-of-the-Art on Material Forming*, R. Alves de Sousa and R. Valente, eds., softcover, Vol. 554, Trans Tech Publications, Aveiro, Portugal, pp. 911–918.
[11] Havinga, J., van den Boogaard, A. H., and Klaseboer, G., 2017, "Sequential Improvement for Robust Optimization Using an Uncertainty Measure for Radial Basis Functions," Struct. Multidiscipl. Optim., **55**(4), pp. 1345–1363.
[12] Alrefae, M. A., 2018, "Process Characterization and Optimization of Roll-to-Roll Plasma Chemical Vapor Deposition for Graphene Growth," Ph.D. thesis, Purdue University, West Lafayette, IN.
[13] Saviers, K. R., 2017, "Scaled-Up Production and Transport Applications of Graphitic Carbon Nanomaterials," Ph.D. thesis, Purdue University, West Lafayette, IN.
[14] Schonlau, M., 1997, "Computer Experiments and Global Optimization," Ph.D. thesis, University of Waterloo, Ontario, Canada.
[15] Simpson, T. W., Lin, D. K., and Chen, W., 2001, "Sampling Strategies for Computer Experiments: Design and Analysis," Int. J. Reliab. Appl., **2**(3), pp. 209–240.
[16] Gramacy, R. B., and Ludkovski, M., 2015, "Sequential Design for Optimal Stopping Problems," SIAM J. Financ. Math., **6**(1), pp. 748–775.
[17] Huan, X., 2010, "Accelerated Bayesian Experimental Design for Chemical Kinetic Models," Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.
[18] Locatelli, M., 1997, "Bayesian Algorithms for One-Dimensional Global Optimization," J. Global Optim., **10**(1), pp. 57–76.
[19] Jones, D. R., Schonlau, M., and Welch, W. J., 1998, "Efficient Global Optimization of Expensive Black-Box Functions," J. Global Optim., **13**(4), pp. 455–492.
[20] Gaul, N. J., 2014, *Modified Bayesian Kriging for Noisy Response Problems and Bayesian Confidence-Based Reliability-Based Design Optimization*, The University of Iowa, Iowa City, IA.
[21] Huang, D., Allen, T. T., Notz, W. I., and Zeng, N., 2006, "Global Optimization of Stochastic Black-Box Systems Via Sequential Kriging Meta-Models," J. Global Optim., **34**(3), pp. 441–466.
[22] Lizotte, D.,2008, Ph.D. thesis, University of Alberta, Edmonton, Alberta, Canada.
[23] Frazier, P. I., Powell, W. B., and Dayanik, S., 2008, "A Knowledge-Gradient Policy for Sequential Information Collection," SIAM J. Control Optim., **47**(5), pp. 2410–2439.
[24] Mockus, J., 2012, *Bayesian Approach to Global Optimization: Theory and Applications*, Vol. 37, Springer Science & Business Media,New York.
[25] Arendt, P. D., Apley, D. W., and Chen, W., 2013, "Objective-Oriented Sequential Sampling for Simulation Based Robust Design Considering Multiple Sources of Uncertainty," ASME J. Mech. Des., **135**(5), p. 051005.
[26] Huan, X., and Marzouk, Y., 2014, "Gradient-Based Stochastic Optimization Methods in Bayesian Experimental Design," Int. J. Uncertain. Quantif., **4**(6), pp. 479–510.
[27] Lam, R., Willcox, K., and Wolpert, D. H., 2016, *Advances in Neural Information Processing Systems*, Curran Associates Inc., pp. 883–891.
[28] Marco, A., Hennig, P., Bohg, J., Schaal, S., and Trimpe, S., 2016, "Automatic lqr Tuning Based on Gaussian Process Global Optimization," 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, IEEE, pp. 270–277.
[29] Kristensen, J., Bilionis, I., and Zabaras, N., 2017, "Adaptive Simulation Selection for the Discovery of the Ground State Line of Binary Alloys With a Limited Computational Budget," *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, Springer, New York, pp. 185–211.
[30] Christen, J. A., and Sansó, B., 2011, "Advances in the Sequential Design of Computer Experiments Based on Active Learning," Commun. Stat. Theory Methods, **40**(24), pp. 4467–4483.
[31] MacKay, D. J., 1992, "Information-Based Objective Functions for Active Data Selection," Neural Comput., **4**(4), pp. 590–604.
[32] Krause, A., Singh, A., and Guestrin, C., 2008, "Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies," J. Mach. Learn. Res., **9**(Feb), pp. 235–284.
[33] Stroh, R., Demeyer, S., Fischer, N., Bect, J., and Vazquez, E., 2017, "Sequential Design of Experiments to Estimate a Probability of Exceeding a Threshold in a Multi-Fidelity Stochastic Simulator," 61th World Statistics Congress of the International Statistical Institute (ISI 2017), Marrakech, Morocco, July 16–21.
[34] Beck, J., and Guillas, S., 2016, "Sequential Design With Mutual Information for Computer Experiments (Mice): Emulation of a Tsunami Model," SIAM/ASA J. Uncertain. Quantif., **4**(1), pp. 739–766.
[35] Gramacy, R. B., and Lee, H. K., 2009, "Adaptive Design and Analysis of Supercomputer Experiments," Technometrics, **51**(2), pp. 130–145.
[36] Terejanu, G., Upadhyay, R. R., and Miki, K., 2012, "Bayesian Experimental Design for the Active Nitridation of Graphite by Atomic Nitrogen," Exp. Therm. Fluid. Sci., **36**, pp. 178–193.
[37] Mohamad, M. A., 2017, "Direct and Adaptive Quantification Schemes for Extreme Event Statistics in Complex Dynamical Systems," Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.
[38] Mohamad, M. A., and Sapsis, T. P., 2018, "A Sequential Sampling Strategy for Extreme Event Statistics in Nonlinear Dynamical Systems," preprint arXiv:1804.07240.
[39] Kullback, S., and Leibler, R. A., 1951, "On Information and Sufficiency," Ann. Math. Stat., **22**(1), pp. 79–86.
[40] Mckay, M. D., Beckman, R. J., and Conover, W. J., 2000, "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code," Technometrics, **42**(1), pp. 55–61.
[41] Tsilifis, P., Ghanem, R. G., and Hajali, P., 2017, "Efficient Bayesian Experimentation Using an Expected Information Gain Lower Bound," SIAM/ASA J. Uncertain. Quantif., **5**(1), pp. 30–62.
[42] Nath, P., Hu, Z., and Mahadevan, S., 2017, "Sensor Placement for Calibration of Spatially Varying Model Parameters," J. Comput. Phys., **343**, pp. 150–169.
[43] Yan, L., Duan, X., Liu, B., and Xu, J., 2018, "Gaussian Processes and Polynomial Chaos Expansion for Regression Problem: Linkage Via the Rkhs and Comparison Via the KL Divergence," Entropy, **20**(3), p. 191.
[44] Choi, S.-K., Grandhi, R. V., Canfield, R. A., and Pettit, C. L., 2004, "Polynomial Chaos Expansion With Latin Hypercube Sampling for Estimating Response Variability," AIAA J., **42**(6), pp. 1191–1198.
[45] Hadigol, M., and Doostan, A., 2017, "Least Squares Polynomial Chaos Expansion: A Review of Sampling Strategies," Comput. Methods Appl. Mech. Eng., **332**, pp. 382–407.
[46] Terejanu, G., Bryant, C., and Miki, K., 2013, "Bayesian Optimal Experimental Design for the Shock-Tube Experiment," J. Phys.: Conf. Ser., **410**(1), pp. 012040.
[47] Hennig, P., and Schuler, C. J., 2012, "Entropy Search for Information-Efficient Global Optimization," J. Mach. Learn. Res., **13**(Jun.), pp. 1809–1837.
[48] Guestrin, C., Krause, A., and Singh, A. P., 2005, "Near-Optimal Sensor Placements in Gaussian Processes," Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, ACM, pp. 265–272.
[49] Huan, X., and Marzouk, Y. M., 2013, "Simulation-Based Optimal Bayesian Experimental Design for Nonlinear Systems," J. Comput. Phys., **232**(1), pp. 288–317.
[50] Picheny, V., Ginsbourger, D., Roustant, O., Haftka, R. T., and Kim, N.-H., 2010, "Adaptive Designs of Experiments for Accurate Approximation of a Target Region," ASME J. Mech. Des., **132**(7), p. 071008.
[51] Xiao, N.-C., Zuo, M. J., and Zhou, C., 2018, "A New Adaptive Sequential Sampling Method to Construct Surrogate Models for Efficient Reliability Analysis," Reliab. Eng. Syst. Saf., **169**(C), pp. 330–338.
[52] Liu, H., Xu, S., Ma, Y., Chen, X., and Wang, X., 2016, "An Adaptive Bayesian Sequential Sampling Approach for Global Metamodeling," ASME J. Mech. Des., **138**(1), p. 011404.
[53] Liu, H., Chen, W., and Sudjianto, A., 2006, "Relative Entropy Based Method for Probabilistic Sensitivity Analysis in Engineering Design," ASME J. Mech. Des., **128**(2), pp. 326–336.
[54] Gonzalvez, J., Lezmi, E., Roncalli, T., and Xu, J., 2019, "Financial Applications of Gaussian Processes and Bayesian Optimization," preprint arXiv:1903.04841.
[55] Wu, J., Toscano-Palmerin, S., Frazier, P. I., and Wilson, A. G., 2019, "Practical Multi-Fidelity Bayesian Optimization for Hyperparameter Tuning," preprint arXiv:1903.04703.

[56] OHagan, A., 2006, "Bayesian Analysis of Computer Code Outputs: A Tutorial," Reliab. Eng. Syst. Saf., **91**(10–11), pp. 1290–1300.
[57] Briol, F.-X., Oates, C., Girolami, M., and Osborne, M. A., 2015, "Frank–Wolfe Bayesian Quadrature: Probabilistic Integration With Theoretical Guarantees," Advances in Neural Information Processing Systems, pp. 1162–1170.
[58] Oates, C. J., Girolami, M., and Chopin, N., 2017, "Control Functionals for Monte Carlo Integration," J. R. Stat. Soc.: Ser. B (Stat. Methodol.), **79**(3), pp. 695–718.
[59] Rasmussen, C. E., and Williams, C. K. I., 2006, *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning*, MIT Press, Cambridge, MA.
[60] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B., 2014, *Bayesian Data Analysis*, Vol. 2, CRC Press, Boca Raton, FL.
[61] Goodman, J., and Weare, J., 2010, "Ensemble Samplers With Affine Invariance," Commun. Appl. Math. Comput. Sci., **5**(1), pp. 65–80.
[62] Duchi, J., 2007, *Derivations for Linear Algebra and Optimization*, Stanford University, Berkeley, CA.
[63] Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J., 2013, "EMCEE: The MCMC Hammer," Publ. Astron. Soc. Pacific, **125**(925), p. 306.
[64] Sóbester, A., Leary, S. J., and Keane, A. J., 2005, "On the Design of Optimization Strategies Based on Global Response Surface Approximation Models," J. Global Optim., **33**(1), pp. 31–59.
[65] Dette, H., and Pepelyshev, A., 2010, "Generalized Latin Hypercube Design for Computer Experiments," Technometrics, **52**(4), pp. 421–429.
[66] Knowles, J., 2006, "Parego: A Hybrid Algorithm With On-Line Landscape Approximation for Expensive Multiobjective Optimization Problems," IEEE Trans. Evol. Comput., **10**(1), pp. 50–66.
[67] Bui-Thanh, T., Willcox, K., and Ghattas, O., 2008, "Model Reduction for Large-Scale Systems With High-Dimensional Parametric Input Space," SIAM J. Sci. Comput., **30**(6), pp. 3270–3288.