entropy

MDPI

Article

Gaussian Process Based Expected Information Gain Computation for Bayesian Optimal Design

Zhihang Xu [1,2,3] and Qifeng Liao [1,*]

[1] School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China; xuzhh@shanghaitech.edu.cn
[2] Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: liaoqf@shanghaitech.edu.cn

Received: 6 January 2020; Accepted: 19 February 2020; Published: 24 February 2020

Abstract: Optimal experimental design (OED) is of great significance in efficient Bayesian inversion. A popular choice of OED methods is based on maximizing the expected information gain (EIG), where expensive likelihood functions are typically involved. To reduce the computational cost, in this work, a novel double-loop Bayesian Monte Carlo (DLBMC) method is developed to efficiently compute the EIG, and a Bayesian optimization (BO) strategy is proposed to obtain its maximizer only using a small number of samples. For Bayesian Monte Carlo posed on uniform and normal distributions, our analysis provides explicit expressions for the mean estimates and the bounds of their variances. The accuracy and the efficiency of our DLBMC and BO based optimal design are validated and demonstrated with numerical experiments.

Keywords: Bayesian Monte Carlo; Bayesian optimal experimental design; Bayesian optimization

1. Introduction

As acquiring data in experiments is generally computationally demanding and time-consuming, maximizing the informativeness of experimental data is of crucial importance. For example, in the area of climate science, a complex system with various stochastic inputs is integrated to represent the real climate situation. Usually, the locations and the time of putting sensors to collect climate observations are optional. With limited resources, careful selections of sensor placements are required (see Reference [1] for a detailed discussion). Many works focus on finding experimental data carrying more information, and this topic is usually referred to as optimal experimental design (OED) [2]. In this paper, we consider the OED problem in the context of the Bayesian inverse problem. Given a forward problem, the inverse problem is to infer parameters inherent of the forward model through a set of design points and the corresponding responses. In the Bayesian setting, the parameters of interest are viewed as random variables, and hence the posterior distribution of the parameters can be obtained via the Bayes' rules [3–6]. In linear cases, the OED problem includes various criteria such as the D-optimal design criterion and the A-optimal design criterion. The D-optimal design criterion seeks to maximize the determinant of the information matrix of the design, whereas the A-optimal design criterion considers minimizing the trace of the inverse of the information matrix which results in minimizing the average variance [7–9].

We focus on the decision theoretic approach that considers maximizing the expectation of Kullback-Leibler (KL) divergence from the posterior distribution to the prior distribution [10]. This decision theoretic approach is a nonlinear generalization of the Bayesian D-optimal criterion [11]. The objective function we want to maximize, the expectation of KL divergence, is also referred to as the expected information gain (EIG) over parameters. Computing EIG is usually a challenging

problem, since it does not have an analytical form for nonlinear problems in general. In the literature, the following attempts have been made for this problem. Efficient surrogates for the forward models are introduced to make the problem tractable [1,12–15]. Recently, as the rise of novel computational methods, new approaches are actively developed to evaluate the EIG, including the quasi-Monte Carlo method [16] and the layered importance sampling [17] in the context of the focused optimal design. As the goal of this kind of optimal design is to find the maximizer of EIG, it is crucial to apply efficient optimization strategies. In Reference [12], a curve fitting surrogate of Monte Carlo experiments is proposed to result in an efficient optimization scheme. For the Bayesian optimal design problem focusing on the risk from an optimal terminal decision, the Bayesian optimization (BO) approach is proposed to find the minimizer of the risk [9]. We note that, while BO is proposed for the Bayesian optimal design problem for minimizing the risk in Reference [9], BO is considered for maximizing the EIG in this work. In addition, efficient approximate coordinate exchange strategies are proposed for Bayesian design in Reference [18]. For intractable likelihood models, Gaussian process (GP) models are built to emulate the likelihood function in Reference [19]. Gradient-based optimization methods are proposed to compute the maximizer of EIG in References [13,14]. Review of modern computational methods for decision theoretic optimal experimental design is provided in Reference [20].

The main purpose of this work is to propose an efficient Gaussian process (GP) based Bayesian optimal design strategy, where Bayesian Monte Carlo (BMC) and Bayesian optimization (BO) which are both based on GP are used [21–23]. As the Monte Carlo simulation for EIG involves an inner layer simulation and an outer layer simulation (see Reference [14]), we develop a novel efficient double-loop Bayesian Monte Carlo (DLBMC) method, which employs BMC [24–26] for both layers. However, the EIG is generally computationally expensive and its gradient information is typically not given explicitly. We propose a BO method [27–29] to compute the maximizer of EIG, where the gradient information of EIG is not required. To summarize, the contributions of this work are three-fold: first we develop a novel DLBMC to efficiently compute EIG; second we analyze the BMC for the normal and the uniform distributions; third we propose BO to obtain the maximizer of EIG.

This paper is organized as follows. In Section 2, we review the Bayesian optimal experimental design problem and formulate the expected information gain (EIG) criterion. In Section 3, we derive a double-loop Bayesian Monte Carlo estimator for the EIG and propose a Bayesian optimization approach to obtain the maximizer of the approximated EIG. Detailed analysis of BMC for the normal and the uniform distributions are conducted in this section. In Section 4, we demonstrate the efficiency of our GP based Bayesian optimal design with three test problems. Section 5 concludes the paper and provides discussions of the advantages and disadvantages of our algorithm.

2. Formulation of Experimental Design

In this section, we review the setting of the Bayesian optimal design problem following the presentation in [14]. In the Bayesian setting, the unknown parameters are viewed as random variables. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space, where Ω is a sample space, $\mathcal{F}$ is a σ-field, and $\mathcal{P}$ is the probability measure on $(\Omega, \mathcal{F})$. Let $\boldsymbol{\theta} : \Omega \to \mathbb{R}^{n_\theta}$ denote the parameters of interest, where n_θ is the dimension of the unknown parameters. Assume that $\boldsymbol{\theta}$ is associated with a prior measure μ on $\mathbb{R}^{n_\theta}$ satisfying $\mu(A) = \mathcal{P}(\boldsymbol{\theta}^{-1}(A))$ for $A \in \mathbb{R}^{n_\theta}$. Throughout this paper, we assume that all the random variables have densities with respect to the Lebesgue measure. Let $\boldsymbol{d} \in \mathcal{D} \subset \mathbb{R}^{n_d}$ denotes the design variable, where n_d is the number of design variables and $\mathcal{D}$ denotes the design space. Let $\boldsymbol{y} \in \mathbb{R}^{n_y}$ denote the response associated with $\boldsymbol{d}$ where n_y is the dimension of response. The inference of $\boldsymbol{\theta}$ can be obtained based on the prior distribution and observations via Bayes' rule,

$$\underbrace{p(\boldsymbol{\theta}|\boldsymbol{d},\boldsymbol{y})}_{\text{Posterior}} = \frac{\overbrace{p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d})}^{\text{Likelihood}}\ \overbrace{p(\boldsymbol{\theta}|\boldsymbol{d})}^{\text{Prior}}}{\underbrace{p(\boldsymbol{y}|\boldsymbol{d})}_{\text{Evidence}}}. \tag{1}$$

The likelihood function is often determined by a deterministic forward model and a statistical model for measurement of model noises. Here we model the relation of the design variable and the observation by a deterministic model $\boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{d})$ and additive Gaussian noises $\boldsymbol{\epsilon}$,

$$\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{d}) + \boldsymbol{\epsilon}\,, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})\,, \tag{2}$$

where $\boldsymbol{G}$ is the forward model. In many practical problems, the forward model is computationally expensive, and its explicit form is not given. So we can just view it as a black box whose internal structure is unknown, whereas we can generate noisy responses given fixed design variables and parameters.

Following the decision theoretic approach [10], we set the utility function as the KL divergence from the posterior distribution to the prior distribution,

$$u(\boldsymbol{d}, \boldsymbol{y}, \boldsymbol{\theta}) = D_{\mathrm{KL}}(p(\boldsymbol{\theta}|\boldsymbol{d}, \boldsymbol{y} \| p(\boldsymbol{\theta}))) = u(\boldsymbol{d}, \boldsymbol{y})\,. \tag{3}$$

This term is actually independent of $\boldsymbol{\theta}$. Noting that $u(\boldsymbol{d}, \boldsymbol{y})$ is a function of both $\boldsymbol{d}$ and $\boldsymbol{y}$, therefore we further take expectation of u over $\boldsymbol{y}$ to define the expected information gain (EIG):

$$U(\boldsymbol{d}) = \int_{\mathcal{Y}} u(\boldsymbol{d}, \boldsymbol{y}) p(\boldsymbol{y}|\boldsymbol{d})\, \mathrm{d}\boldsymbol{y} = \int_{\mathcal{Y}} \int_{\Theta} p(\boldsymbol{\theta}|\boldsymbol{d}, \boldsymbol{y}) \log \left[\frac{p(\boldsymbol{\theta}|\boldsymbol{d}, \boldsymbol{y})}{p(\boldsymbol{\theta})} \right] \mathrm{d}\boldsymbol{\theta} p(\boldsymbol{y}|\boldsymbol{d})\, \mathrm{d}\boldsymbol{y}\,.$$

Then the optimal experimental design is to find a design point which maximizes the expected utility, that is,

$$\boldsymbol{d}^\star = \arg\max_{\boldsymbol{d} \in \mathcal{D}} U(\boldsymbol{d})\,. \tag{4}$$

A double-loop Monte Carlo (DLMC) estimator of EIG is proposed in [30]. Rewrite $U(\boldsymbol{d})$ as

$$\begin{aligned} U(\boldsymbol{d}) &= \int_{\mathcal{Y}} \int_{\Theta} p(\boldsymbol{\theta}|\boldsymbol{d}, \boldsymbol{y}) \log \left[\frac{p(\boldsymbol{\theta}|\boldsymbol{d}, \boldsymbol{y})}{p(\boldsymbol{\theta})} \right] \mathrm{d}\boldsymbol{\theta} p(\boldsymbol{y}|\boldsymbol{d})\, \mathrm{d}\boldsymbol{y} \\ &= \int_{\mathcal{Y}} \int_{\Theta} \{\log[p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{d})] - \log[p(\boldsymbol{y}|\boldsymbol{d})]\} p(\boldsymbol{y}, \boldsymbol{\theta}|\boldsymbol{d})\, \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\boldsymbol{y}\,, \end{aligned}$$

and note that $p(\boldsymbol{\theta}|\boldsymbol{d}) = p(\boldsymbol{\theta})$, since the specification of $\boldsymbol{d}$ does not provide further information about inference of $\boldsymbol{\theta}$. Then, the DLMC method approximates $U(\boldsymbol{d})$ as

$$U(\boldsymbol{d}) \approx \frac{1}{n_{\mathrm{out}}} \sum_{i=1}^{n_{\mathrm{out}}} \left[\log(p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)}, \boldsymbol{d})) - \log(p(\boldsymbol{y}^{(i)}|\boldsymbol{d})) \right]\,, \tag{5}$$

where $\boldsymbol{\theta}^{(i)}$ are drawn from the prior $p(\boldsymbol{\theta})$, and $\boldsymbol{y}^{(i)}$ are drawn from the conditional distribution $p(\boldsymbol{y}|\boldsymbol{\theta} = \boldsymbol{\theta}^{(i)}, \boldsymbol{d})$ (i.e., the likelihood), and hence $p(\boldsymbol{y}^{(i)}|\boldsymbol{d})$ can be estimated via the importance sampling technique,

$$p(\boldsymbol{y}^{(i)}|\boldsymbol{d}) = \int_{\Theta} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}, \boldsymbol{d}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \approx \frac{1}{n_{\mathrm{in}}} \sum_{j=1}^{n_{\mathrm{in}}} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i,j)}, \boldsymbol{d})\,. \tag{6}$$

Combining (6) and (5) yields a biased estimator $\widetilde{U}(\boldsymbol{d})$ of $U(\boldsymbol{d})$. However, if we sample $\{\boldsymbol{\theta}^{(i,j)}\}_{j=1}^{n_{\mathrm{in}}}$ for every $y^{(i)}$ $(i = 1, \ldots, n_{\mathrm{out}})$, the complexity of this method is $\mathcal{O}(n_{\mathrm{out}} n_{\mathrm{in}})$. To reduce the computational cost, a sample reuse technique is employed in [14]. That is, for every $\boldsymbol{d}$, we drawn a fresh batch from the prior $\{\boldsymbol{\theta}^{(k)}\}_{k=1}^{n_{\mathrm{out}}}$ and use this set for both the outer Monte Carlo and the inner Monte Carlo (i.e., $\boldsymbol{\theta}^{(\cdot,j)} = \boldsymbol{\theta}^{(k)}$). Consequently, the complexity is reduced to $\mathcal{O}(n_{\mathrm{out}})$.

3. GP Based Framework for Bayesian Optimal Design

Our main framework for Bayesian optimal design is based on two powerful tools according to Gaussian processes: the Bayesian Monte Carlo (BMC) method and the Bayesian optimization (BO) method. In this section, we first review BMC and conduct the analysis of BMC for the normal and the uniform distributions, and then present our novel double loop Bayesian Monte Carlo (DLBMC) for EIG. After that, we propose BO to find the maximizer of the approximated EIG. Finally, we review the classical Markov chain Monte Carlo (MCMC) method for Bayesian parameter inference.

3.1. Bayesian Monte Carlo

Consider the integral problem $I := \int_{\mathcal{X}} f(x)p(x)\,\mathrm{d}x$, where $p(x)$ is the density of x, $f(x) : \mathbb{R}^n \to \mathbb{R}$ is the integrand, and $\mathcal{X}$ is the support of x, that is, the integration domain. The idea of BMC is to formulate an integral problem into a Bayesian inference problem by placing a prior over the integrand f and to obtain the posterior distribution of f conditioning on the data collected. A natural way of putting a prior over function is through a Gaussian process, which is completely characterized by its mean function $\mu(x)$ and kernel function $k(x, x')$, that is, $f \sim \mathcal{GP}(\mu(\cdot), k(\cdot,\cdot))$. A commonly used choice for the kernel function is the Gaussian kernel (or known as squared exponential kernel), $k(x, x') = \sigma_f^2 \exp(-\|x - x'\|_2^2/(2l^2))$, where both σ_f and l are hyperparameters of the kernel function. The choice of hyperparameters affects the result to a large extent. Therefore the hyperparameters need to be determined carefully. Having collected noisy observations $D = \{x^{(i)}, f^{(i)}\}_{i=1}^N$, where $f^{(i)} = f(x^{(i)}) + \epsilon^{(i)}$ with Gaussian noises $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$, a Gaussian process provides a posterior distribution for an arbitrary new point $x^\star$, $f(x^\star)|D, x^\star \sim \mathcal{N}(\mu_N(x^\star), \sigma_N(x^\star))$, with mean function $\mu_N(x^\star)$ and variance function $\sigma_N(x^\star)$ given by

$$\begin{aligned}\mu_N(x^\star) &:= k_N(x^\star)^T(K_N + \sigma^2 I)^{-1} f_N\,,\\ \sigma_N(x^\star) &:= k(x^\star, x^\star) - k_N(x^\star)^T(K_N + \sigma^2 I)^{-1} k_N(x^\star)\,,\end{aligned}$$

where $k_N(x^\star) = [k(x^\star, x^{(i)})]_{N\times 1}$, $K_N = [k(x^{(i)}, x^{(j)})]_{N\times N}$, and $f_N = [f^{(i)}]_{N\times 1}$. In some special cases (For example, when x is distributed with Gaussian and the kernel is a Gaussian kernel [25]), GP allows us to estimate the integration in a closed form, of which the posterior mean is given by,

$$\begin{aligned}\mathbb{E}_{f|D}[I] &= \iint_{\mathcal{X}} f(x)p(x)\,\mathrm{d}x p(f|D)\,\mathrm{d}f = \int_{\mathcal{X}} \left[\int f(x)p(f|D)\,\mathrm{d}f\right] p(x)\,\mathrm{d}x\\ &= \int_{\mathcal{X}} \mathbb{E}_{f|D}(f)p(x)\,\mathrm{d}x = \int_{\mathcal{X}} \mu_N(x)p(x)\,\mathrm{d}x = \int_{\mathcal{X}} k_N(x)^T(K_N + \sigma^2 I)^{-1} f_N p(x)\,\mathrm{d}x\\ &= \int_{\mathcal{X}} k_N(x)^T p(x)\,\mathrm{d}x (K_N + \sigma^2 I)^{-1} f_N = z^T(K_N + \sigma^2 I)^{-1} f_N\,,\end{aligned} \tag{7}$$

where $z := \int_{\mathcal{X}} k_N(x)^T p(x)\,\mathrm{d}x$, and the posterior variance is given by

$$\begin{aligned}\mathbb{V}_{f|D}[I] &= \mathbb{E}_{f|D}\{[I - \mathbb{E}_{f|D}(I)]^2\} = \int \left[\int_{\mathcal{X}} f(x)p(x)\,\mathrm{d}x - \int_{\mathcal{X}} \mu_N(x')p(x')\,\mathrm{d}x'\right]^2 p(f|D)\,\mathrm{d}f\\ &= \iint_{\mathcal{X}\times\mathcal{X}} \int [f(x) - \mu_N(x')]^2 p(f|D)\,\mathrm{d}f p(x)p(x')\,\mathrm{d}x\,\mathrm{d}x' = \iint_{\mathcal{X}\times\mathcal{X}} \mathbb{V}(f(x))p(x)p(x')\,\mathrm{d}x\,\mathrm{d}x'\\ &= \iint_{\mathcal{X}\times\mathcal{X}} k(x, x')p(x)p(x')\,\mathrm{d}x\,\mathrm{d}x' - z^T(K_N + \sigma^2 I)^{-1} z\,.\end{aligned} \tag{8}$$

Note that in the above analytical expression of posterior mean and variance, when the hyperparameters and the kernel function are given, z and K are determined by $\{x^{(i)}\}_{i=1}^N$ and they are independent of the observations $\{f^{(i)}\}_{i=1}^N$. Therefore, the computation procedure for the posterior mean and variance of the BMC estimator proceeds the following two steps: first, an input sample set $\{x^{(i)}\}_{i=1}^N$ is generated, and z and K are computed; second, the corresponding observation set $\{f^{(i)}\}_{i=1}^N$ is collected, and the

posterior mean and variance are computed through (7) and (8) respectively. Next, when x is an uniform or Gaussian random vector, we provide detailed derivations for the posterior mean and variance of BMC estimators.

Theorem 1 (Bayesian Monte Carlo for the standard Gaussian distribution)**.** *Consider the integral $I = \int_{\mathcal{X}} f(x)p(x)dx$, where x is the standard Gaussian random variable vector in $\mathcal{X} = \mathbb{R}^n$. The prior mean function is assumed to be a zero function, and the kernel is assumed to be the Gaussian kernel $k(x, x') = \sigma_f^2 \exp(-\|x - x'\|_2^2/(2l^2))$ with predetermined hyperparameters σ_f and l. Having collected noisy observations $D = \{x^{(i)}, f^{(i)}\}_{i=1}^N$ where $f^{(i)} = f(x^{(i)}) + \epsilon^{(i)}$ and $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$, the posterior mean and variance of BMC are given by*

$$\begin{aligned}\mathbb{E}_{f|D}[I] &= z^T(K_N + \sigma^2 I)^{-1} f_N\,,\\ \mathbb{V}_{f|D}[I] &= \sigma_f^2 \left(\frac{l^2}{l^2+2}\right)^{n/2} - z^T(K_N + \sigma^2 I)^{-1} z\,,\end{aligned}$$

where $K_N = [k(x^{(i)}, x^{(j)})]_{N\times N}$, $f_N = [f^{(1)}, \ldots, f^{(N)}]_{N\times 1}$, and the components of z are

$$z_i = \sigma_f^2 \left(\frac{l^2}{l^2+1}\right)^{n/2} \exp\left(-\frac{\|x^{(i)}\|^2}{2(l^2+1)}\right), \tag{9}$$

for $i = 1, \ldots, N$.

Proof. The components of z can be computed analytically, for $i = 1, \ldots, N$,

$$\begin{aligned} z_i &= \int_{\mathcal{X}} \sigma_f^2 \exp\left(-\frac{\|x - x^{(i)}\|_2^2}{2l^2}\right) \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{x^T x}{2}\right) \mathrm{d}x\\ &= \frac{\sigma_f^2}{\sqrt{(2\pi)^n}} \int_{\mathcal{X}} \exp\left[-\frac{1}{2} x^T \frac{(l^2+1)I}{l^2} x + \frac{x^T x^{(i)}}{l^2} - \frac{\|x^{(i)}\|^2}{2l^2}\right] \mathrm{d}x\\ &= \sigma_f^2 \left|\frac{l^2}{l^2+1} I\right|^{1/2} \exp\left[-\frac{\|x^{(i)}\|^2}{2l^2} - \frac{\|x^{(i)}\|_2^2}{2l^2(l^2+1)}\right]\\ &= \sigma_f^2 \left(\frac{l^2}{l^2+1}\right)^{n/2} \exp\left(-\frac{\|x^{(i)}\|_2^2}{2(l^2+1)}\right). \end{aligned}$$

Moreover, the variance of BMC is

$$\begin{aligned} \mathbb{V}_{f|D}[I] &= \iint_{\mathcal{X}\times\mathcal{X}} k(x, x')p(x)\,\mathrm{d}x p(x')\,\mathrm{d}x' - z^T K_N^{-1} z\\ &= \int_{\mathcal{X}} \frac{\sigma_f^2}{(2\pi)^{n/2}} (2\pi)^{n/2} \left|\frac{l^2}{l^2+1} I\right|^{1/2} \exp\left(-\frac{\|x'\|_2^2}{2l^2} + \frac{\|x'\|_2^2}{2l^2(l^2+1)}\right) p(x')\,\mathrm{d}x - z^T K_N^{-1} z\\ &= \sigma_f^2 \left(\frac{l^2}{l^2+1}\right)^{n/2} \int_{\mathcal{X}} \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{l^2+2}{2(l^2+1)}\|x'\|_2^2\right) \mathrm{d}x - z^T K_N^{-1} z\\ &= \sigma_f^2 \left(\frac{l^2}{l^2+1}\right)^{n/2} \frac{1}{(2\pi)^{n/2}} (2\pi)^{n/2} \left(\frac{l^2+1}{l^2+2}\right)^{n/2} - z^T K_N^{-1} z\\ &= \sigma_f^2 \left(\frac{l^2}{l^2+2}\right)^{n/2} - z^T K_N^{-1} z\,. \end{aligned}$$

□

We note that Theorem 1 is presented in References [22,25], but we give the above detailed proof for completeness.

Theorem 2 (Bayesian Monte Carlo for the uniform distribution)**.** *Consider the integral* $I = \int_{\mathcal{X}} f(x)p(x)dx$*, where* x *is a random vector uniformly distributed in the hypercube* $\mathcal{X} = [l_1, r_1] \times \cdots \times [l_n, r_n]$*. The prior mean function is assumed to be a zero function, and the kernel is assumed to be the Gaussian kernel* $k(x, x') = \sigma_f^2 \exp(-\|x - x'\|_2^2/(2l^2))$ *with predetermined hyperparameters* σ_f *and* l*. Having collected noisy observations* $D = \{x^{(i)}, f^{(i)}\}_{i=1}^N$ *where* $f^{(i)} = f(x^{(i)}) + \epsilon^{(i)}$ *and* $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$*, the posterior mean of BMC and the upper bound of variance of BMC are given by*

$$\begin{aligned}\mathbb{E}_{f|D}[I] &= z^T(K_N + \sigma^2 I)f\,,\\ \mathbb{V}_{f|D}[I] &< \frac{\sigma_f^2\sqrt{(2\pi l^2)^n}}{|\mathcal{X}|} - z^T(K_N + \sigma^2 I)z\,,\end{aligned}$$

where $K_N = [k(x^{(i)}, x^{(j)})]_{N\times N}$*,* $f_N = [f^{(1)}, \ldots, f^{(N)}]_{N\times 1}$*, and the components of* z *are given by*

$$z_i = \frac{\sigma_f^2}{|\mathcal{X}|}\prod_{j=1}^{n}(\Phi(x_j; x_j^{(i)}, l, r_j) - \Phi(x_j; x_j^{(i)}, l, l_j))\,, \tag{10}$$

for $i = 1, \ldots, N$*, with* $\Phi(x; \mu, \sigma, t)$ *being the cumulative distribution function (CDF) of the Gaussian distribution* $\mathcal{N}(\mu, \sigma^2)$*.*

Proof. The components of z can be computed analytically, for $i = 1, \ldots, N$,

$$\begin{aligned}z_i &= \int_{\mathcal{X}} k(x, x^{(i)})p(x)\,\mathrm{d}x = \frac{\sigma_f^2}{|\mathcal{X}|}\int_{\mathcal{X}} \exp\left(-\frac{\|x - x^{(i)}|\|^2}{2l^2}\right)\mathrm{d}x\\ &= \frac{\sigma_f^2}{|\mathcal{X}|}\int_{\mathcal{X}} \exp\left(-\frac{\sum_{j=1}^n (x_j - x_j^{(i)})^2}{2l^2}\right)\mathrm{d}x = \frac{\sigma_f^2}{|\mathcal{X}|}\prod_{j=1}^{n}\int_{l_j}^{r_j} \exp\left(-\frac{(x_j - x_j^{(i)})^2}{2l^2}\right)\mathrm{d}x_j\\ &= \frac{\sigma_f^2}{|\mathcal{X}|}\prod_{j=1}^{n}(\Phi(x_j; x_j^{(i)}, l, r_j) - \Phi(x_j; x_j^{(i)}, l, l_j))\,,\end{aligned}$$

where $\Phi(x; \mu, \sigma, t) = \int_{\infty}^{t} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x$ is the CDF of the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ and v_j denotes the j-th component of the vector v. However, since the double integral of the Gaussian density function has no analytical form, the variance of the estimator cannot be obtained, and we give an upper bound of the variance,

$$\begin{aligned}\mathbb{V}_{f|D}[I] &= \int_{\mathcal{X}}\int_{\mathcal{X}} k(x, x')p(x)\,\mathrm{d}x p(x')\,\mathrm{d}x' - z^T K_N^{-1} z < \int_{\mathcal{X}}\int_{\mathbb{R}^n} k(x, x')p(x)\,\mathrm{d}x p(x')\,\mathrm{d}x' - z^T K_N^{-1} z\\ &= \frac{\sigma_f^2\sqrt{(2\pi l^2)^n}}{|\mathcal{X}|}\int_{\mathcal{X}} p(x')\,\mathrm{d}x' - z^T K_N^{-1} z = \frac{\sigma_f^2\sqrt{(2\pi l^2)^n}}{|\mathcal{X}|} - z^T K_N^{-1} z\,.\end{aligned}$$

□

3.2. Estimating the Expected Information Gain Using Double-Loop BMC

In general, the value of the EIG has no closed form and has to be approximated via numerical methods. Based on the idea of BMC for efficiently evaluating integrals, we develop a double-loop BMC (DLBMC) scheme to approximate the EIG.

Letting $e(\boldsymbol{y},\boldsymbol{d}) = p(\boldsymbol{y}|\boldsymbol{d})$, $g(\boldsymbol{d},\boldsymbol{y},\boldsymbol{\theta}) = \{\ln[p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d})] - \ln[p(\boldsymbol{y}|\boldsymbol{d})]\}p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d}) = \{\ln[p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d})] - \ln[e(\boldsymbol{y},\boldsymbol{d})]\}p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d})$, and $h(\boldsymbol{d},\boldsymbol{y}) = \int_\Theta g(\boldsymbol{d},\boldsymbol{y},\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$, it is known that $U(\boldsymbol{d})$ can be rewritten as

$$\begin{aligned} U(\boldsymbol{d}) &= \int_{\mathcal{Y}}\int_\Theta \{\ln[p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d})] - \ln[e(\boldsymbol{y},\boldsymbol{d})]\}p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{y} \\ &= \int_{\mathcal{Y}}\int_\Theta g(\boldsymbol{d},\boldsymbol{y},\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{y} = \int_{\mathcal{Y}} h(\boldsymbol{d},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\,. \end{aligned}$$

First, we consider the straightforward detailed calculation of $U(\boldsymbol{d})$ for any fixed $\boldsymbol{d}$. To compute $U(\boldsymbol{d}) = \int_{\mathcal{Y}} h(\boldsymbol{d},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}$, we need samples $\{\boldsymbol{y}^{(i)}, h^{(i)} := h(\boldsymbol{d},\boldsymbol{y}^{(i)})\}_{i=1}^{n_{\text{out}}}$, where $\boldsymbol{y}^{(i)} \sim \mathcal{U}(\mathcal{Y})$ for $i = 1,\ldots,n_{\text{out}}$, n_{out} denotes the sample size of the outer layer. To compute $h^{(i)} := \int_\Theta g(\boldsymbol{d},\boldsymbol{y}^{(i)},\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$, samples $\{\boldsymbol{\theta}^{(i,j)}, g^{(i,j)} := g(\boldsymbol{d},\boldsymbol{y}^{(i)},\boldsymbol{\theta}^{(i,j)})\}_{j=1}^{n_{\text{in}}}$ are needed, where n_{in} denotes the sample size of the inner layer. Again, $g^{(i,j)} = \{\ln[p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i,j)},\boldsymbol{d})] - \ln[p(\boldsymbol{y}^{(i)}|\boldsymbol{d})]\}p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i,j)},\boldsymbol{d})$ also involves another integration $p(\boldsymbol{y}^{(i)}|\boldsymbol{d}) = \int_\Theta p(\boldsymbol{y}^{(i)}|\boldsymbol{d},\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$, and therefore samples $\{\boldsymbol{\theta}^{(i,j)}, p(\boldsymbol{y}^{(i)}|\boldsymbol{d},\boldsymbol{\theta}^{(i,j)})\}_{j=1}^{n_{\text{in}}}$ are needed and $\boldsymbol{\theta}^{(i,j)}$ are generated from the prior. We propose using the BMC method to evaluate integrals $\{e^{(i)}\}_{i=1}^{n_{\text{out}}}$, $\{h^{(i)}\}_{i=1}^{n_{\text{out}}}$, and U. So far, there are two problems. First, the computation complexity for the forward model is $\mathcal{O}(n_{\text{in}}n_{\text{out}})$, which grows fast with the increase of the problem dimension. Second, since we usually have no prior knowledge of the support $\mathcal{Y}$ of $\boldsymbol{y}$, we can not uniformly sample $\{\boldsymbol{y}^{(i)}\}_{i=1}^{n_{\text{out}}}$.

To overcome these obstacles, we employ the sample reuse technique [14] that sets $\boldsymbol{\theta}^{(\cdot,j)} = \boldsymbol{\theta}^{(j)}$, and the computational complexity is reduced to $\mathcal{O}(n_{\text{in}})$. Besides, it allows us to generate samples $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{n_{\text{in}}}$ in advance, and we can use the corresponding forward model outputs to estimate $\mathcal{Y}$—suppose the corresponding forward model values are given by $\{G^{(j)} = \boldsymbol{G}(\boldsymbol{d},\boldsymbol{\theta}^{(j)})\}_{j=1}^{n_{\text{in}}}$, and then $\mathcal{Y}$ can be approximated by $\widetilde{\mathcal{Y}} := [\min(G) - \sigma, \max(G) + \sigma]$ where $G = [G^{(1)},\ldots,G^{n_{\text{in}}}]^T$. In this way, we can sample $\{\boldsymbol{y}^{(i)}\}_{i=1}^{n_{\text{in}}} \sim \mathcal{U}(\widetilde{\mathcal{Y}})$. Intuitively speaking, the approximated $\widetilde{\mathcal{Y}}$ is slightly smaller than the actual field $\mathcal{Y}$, and consequently, bias is induced in the estimator. With increased sample size n_{in}, $\widetilde{\mathcal{Y}}$ can be captured more accurately and the bias can be reduced.

In the process of estimating U, we propose using the BMC method to compute e, h and U. Since two layers of integration are involved, let the hyperparameters of BMC for the inner layer and the outer layer be $\{l_{\text{in}}, (\sigma_f)_{\text{in}}\}$ and $\{l_{\text{out}}, (\sigma_f)_{\text{out}}\}$ respectively. In our previous discussion about BMC in (7)–(8), z and K can be computed, once the input sample set $\{x^{(i)}\}_{i=1}^{N}$ is given. Therefore, for computational simplicity, after $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{n_{\text{in}}}$ and $\{\boldsymbol{y}^{(i)}\}_{i=1}^{n_{\text{out}}}$ are generated, we can compute $\{z_{\text{in}}, K_{\text{in}}\}$ and $\{z_{\text{out}}, K_{\text{out}}\}$ ahead. Taking the prior being the standard normal distribution for example, z_{in} and K_{in} are given by,

$$\begin{aligned} (z_{\text{in}})_j &= (\sigma_f)_{\text{in}}^2 \left(\frac{l_{\text{in}}^2}{l_{\text{in}}^2+1}\right)^{n_\theta/2} \exp\left(-\frac{\|\boldsymbol{\theta}^{(j)}\|_2^2}{2(l_{\text{in}}^2+1)}\right), \quad \text{for } j = 1,\ldots,n_{\text{in}}\,, \\ K_{\text{in}} &= [(\sigma_f)_{\text{in}}^2 \exp(-\|\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}^{(j)}\|_2^2/(2l_{\text{in}}^2))]_{n_{\text{in}}\times n_{\text{in}}}\,. \end{aligned}$$

Details of our DLBMC method for estimating U are summarized in Algorithm 1. Note that we only use the mean estimates of the DLBMC estimator in the following. The variance of DLBMC is potentially useful, but as discussed in Section 3.1, the variance of BMC typically does not have a closed form, and we are not able to derive a closed form for the variance of DLBMC in this work. We will consider the variance of DLBMC in our future work. It is also possible to consider other numerical integration

methods to compute EIG, for example, the sparse grid quadrature rules [31,32] and their combination with physical model reduction techniques [33], but they are out of the scope of this paper.

Algorithm 1 Double-Loop Bayesian Monte Carlo (DLBMC) for estimating EIG

Input: Design points $\boldsymbol{d}$, prior $p(\boldsymbol{\theta})$, standard deviation of noise σ, hyperparameters $\{l_{\text{in}}, (\sigma_f)_{\text{in}}\}$ and $\{l_{\text{out}}, (\sigma_f)_{\text{out}}\}$.
Data preparation: Sample $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{n_{\text{in}}} \sim p(\boldsymbol{\theta})$ and compute $G^{(j)} = \boldsymbol{G}(\boldsymbol{d}, \boldsymbol{\theta}^{(j)})$ for $j = 1, \ldots, n_{\text{in}}$.
Sample $\{\boldsymbol{y}^{(i)}\}_{i=1}^{n_{\text{out}}} \sim \mathcal{U}(\min(G) - \sigma, \max(G) + \sigma)$.
Compute $\{z_{\text{in}}, K_{\text{in}}\}$ and $\{z_{\text{out}}, K_{\text{out}}\}$.
for $i = 1, \ldots, n_{\text{out}}$ **do**
    for $j = 1, \ldots, n_{\text{in}}$ **do**
        Compute the likelihood $f^{(i,j)} = p(\boldsymbol{y}^{(i)} | \boldsymbol{\theta}^{(j)}, \boldsymbol{d})$.
    end for
    Let $f^{(i)} = [f^{(i,1)}, \ldots, f^{(i,n_{\text{in}})}]^T$.
    Compute the evidence $e^{(i)} = z_{\text{in}}^T K_{\text{in}}^{-1} f^{(i)}$.
    for $j = 1, \ldots, n_{\text{in}}$ **do**
        $g^{(i,j)} = [\log(f^{(i,j)}) - \log(e^{(i)})] f^{(i,j)}$.
    end for
    Let $g^{(i)} = [g^{(i,1)}, \ldots, g^{(i,n_{\text{in}})}]^T$.
    Compute $h^{(i)} = z_{\text{in}}^T K_{\text{in}} g^{(i)}$.
end for
Let $h = [h^{(1)}, \ldots, h^{(n_{\text{out}})}]^T$.
Compute $\widehat{U}(\boldsymbol{d}) = z_{\text{out}}^T K_{\text{out}}^{-1} h$.
Output: the estimated EIG $\widehat{U}(\boldsymbol{d})$.

3.3. Bayesian Optimization

The ultimate goal of the optimal experimental design problem is to find the optimizer $\boldsymbol{d}^\star$ in (4). In this problem, since the computing of the function value $U(\boldsymbol{d})$ and the gradient ∇U is prohibitively expensive, it is challenging to apply function-value-based or gradient-based optimization methods. As the Bayesian optimization (BO) method [27,28,34,35] typically only requires a low objective function evaluation budget [36] and does not require any gradient information, it can be suitable for this problem. In this section, we give a brief review of BO and apply it to obtain the maximizer of EIG (4).

To compute the maximizer of EIG $U : \mathbb{R}^{n_d} \to \mathbb{R}$ (see (4)), for a given maximum number of iterations $t_{\max}$, that is, the evaluation budget, Bayesian optimization begins with putting a GP prior on $U \sim \mathcal{GP}(\mu_0(\boldsymbol{d}), k(\boldsymbol{d}, \boldsymbol{d}'))$, and then randomly chooses an initial point $\boldsymbol{d}_1$ and collects the corresponding response $U_1 = U(\boldsymbol{d}_1)$. Next, the posterior mean function $\mu_1(\boldsymbol{d})$ and variance $\sigma_1(\boldsymbol{d})$ are updated via collected data set $\mathcal{S}_1 = \{\boldsymbol{d}_1, U_1\}$. Usually $\boldsymbol{d}_1$ alone is inadequate to find the maximum and therefore we need a strategy to choose the next design point. Typically, the next point is determined through maximizing an acquisition function A, that is, at t-th iteration, $\boldsymbol{d}_{t+1} = \arg\max_{\boldsymbol{d}} A(\boldsymbol{d} | \mathcal{S}_{1:t})$. After the next point $\boldsymbol{d}_2$ is obtained, we sample the objective function U_2. The collected data set is then augmented as $\mathcal{S}_{1:2} = \mathcal{S}_1 \cup \{\boldsymbol{d}_2, U_2\} = \{\boldsymbol{d}_i, U_i\}_{i=1,2}$, and the posterior mean function $\mu_2(\boldsymbol{d})$ and the variance function $\sigma_t(\boldsymbol{d})$ are also updated. The above procedure repeats until the given maximum budget $t_{\max}$ is reached.

In the case that $\mathcal{D}$ is infinite, the process of finding the next design point $\boldsymbol{d}_{t+1} = \arg\max_{\boldsymbol{d} \in \mathcal{D}} A(\boldsymbol{d} | S_{1:t})$ is demanding. However, performing global search over the discretized space is usually effective [27,37], since we assume that evaluating U is more costly than computing the GP surrogate. Therefore, the design space $\mathcal{D}$ is discretized over equidistant grids and we denote the

discretized design space as $\overline{\mathcal{D}}$. Supposing we have collected data set $\mathcal{S}_{1:t}$, the posterior of U is a GP distribution with mean $\mu_t(\boldsymbol{d})$, kernel $k(\boldsymbol{d}, \boldsymbol{d}')$ and variance $\sigma_t^2(\boldsymbol{d})$,

$$\mu_t(\boldsymbol{d}) = k_t(\boldsymbol{d})^T (K_t + \sigma^2 I)^{-1} U_{1:t}\,, \tag{11}$$

$$\sigma_t(\boldsymbol{d}) = k(\boldsymbol{d}, \boldsymbol{d}) - K_t(\boldsymbol{d})^T (K_t + \sigma^2 I)^{-1} k_t(\boldsymbol{d})\,, \tag{12}$$

where $k_t(\boldsymbol{d}) = [k(\boldsymbol{d}, \boldsymbol{d}_i)]_{t\times 1}$, $K_t = [k(\boldsymbol{d}_i, \boldsymbol{d}_j)]_{t\times t}$ and $U_{1:t} = [U_i]_{t\times 1}$. We note the design space considered in this paper is assumed to be bounded, such that it can be directly discretized. For unbounded design spaces, an unbounded Bayesian optimization approach is developed through gradually extending regions with regularization in [38].

Choosing a proper acquisition function is crucial for the Bayesian optimization algorithm since it guides the search for the optimum. Popular choices of acquisition function include maximizing the probability of improvement (PI) [39,40], and maximizing the expected improvement (EI) in the efficient global optimization (EGO) algorithm [41,42]. A review for the selection of acquisition functions is in [27]. Suggested by [37], we apply the `GP-UCB` algorithm to choose the next point—the acquisition function is set to a linear combination of the posterior mean function and the posterior variance function,

$$\boldsymbol{d}_t = \arg\max_{\boldsymbol{d}\in\mathcal{D}} \mu_{t-1}(\boldsymbol{d}) + \sqrt{\beta_{t-1}}\sigma_{t-1}(\boldsymbol{d})\,,$$

where $\mu_{t-1}(\boldsymbol{d}) + \sqrt{\beta_{t-1}}\sigma_{t-1}(\boldsymbol{d})$ can be considered as the upper confidence bound of the current Gaussian process. It is clear that maximizing the acquisition function $\mu_{t-1}(\boldsymbol{d}) + \sqrt{\beta_{t-1}}\sigma_{t-1}(\boldsymbol{d})$ shows a trade-off between exploring the point with potential high function value and exploiting the point with high uncertainty. Here, β_{t-1} is the parameter balancing exploring and exploiting. Details of our BO strategy for optimal design are shown in Algorithm 2.

We set the prior mean function to $\mu_0(\boldsymbol{d}) = 0$, and set the kernel to be the Gaussian kernel given by $k(\boldsymbol{d}, \boldsymbol{d}') = \sigma_f^2 \exp(-\|\boldsymbol{d} - \boldsymbol{d}'\|^2/(2l^2))$. The design space $\mathcal{D}$ is discretized over equidistant grids and we denote the discretized design space as $\overline{\mathcal{D}}$. In the step of maximizing the acquisition function, $\boldsymbol{d}_t$ is located through a grid search over $\overline{\mathcal{D}}$.

A natural measure in performance of the Bayesian optimization method is defined through average cumulative regret. Supposing the maximum $U(\boldsymbol{d}^\star)$ is known, the instantaneous regret for iteration t is defined as $r_t = U(\boldsymbol{d}^\star) - U(\boldsymbol{d}_t)$ and the cumulative regret R_T after T iterations is defined as the sum of the instantaneous regrets $R_T = \sum_{t=1}^{T} r_t$. Then the average cumulative regret R_T/T is defined as $R_T/T = \sum_{t=1}^{T} r_t/T$. It should be noted that neither r_t nor R_T can be obtained directly from the Algorithm 2. In [37], it is proven that, for finite design space $\overline{\mathcal{D}}$, setting $\delta \in (0,1)$ and $\beta_t = 2\log(|\overline{\mathcal{D}}|t^2\pi^2/6\delta)$, the Bayesian optimization method is no-regret with high probability, that is, $\lim_{T\to\infty} R_T/T = 0$.

Algorithm 2 Bayesian optimization (BO) for optimal design

1: **Input:** Design space $\mathcal{D}$ and its discretized design space $\overline{\mathcal{D}}$, prior $\mu_0(\boldsymbol{d}) = 0$, hyperparameters l, σ_f of the Gaussian kernel, hyperparameter δ, and maximum number of iterations $t_{\max}$.
2: **for** $t = 1, \ldots, t_{\max}$ **do**
3: Find the maximizer of the acquisition function: $\boldsymbol{d}_t = \arg\max_{d\in\mathcal{D}} \mu_{t-1}(\boldsymbol{d}) + \sqrt{\beta_{t-1}}\sigma_{t-1}(\boldsymbol{d})$.
4: Sample the objective function $U_t = \widehat{U}(\boldsymbol{d}_t)$ using Algorithm 1.
5: Augment the data set $\mathcal{S}_{1:t} = \{\boldsymbol{d}_i, U_i\}_{i=1}^{t}$.
6: Perform Bayesian update to obtain μ_t and σ_t over $\overline{\mathcal{D}}$ using (11) and (12) respectively.
7: Update β_t .
8: **end for**
9: **Output:** Optimal design: $d^\star = \arg\max_{t=1,\ldots,t_{\max}} U_t$

3.4. Bayesian Parameter Inference

After the optimal design points are selected and the corresponding noisy observations $D = \{\boldsymbol{d}^{(i)}, \boldsymbol{y}^{(i)}\}_{i=1}^{N}$ are collected, we can conduct Bayesian inference for the system parameters, that is, to assess the posterior distribution $p(\boldsymbol{\theta}|D)$. The posterior $p(\boldsymbol{\theta}|D)$ can be calculated via Bayes' rule (1). However, as there is no closed-form for the evidence in (1) in many practical problems, the Markov chain Monte Carlo (MCMC) method is often used to generate samples of the posterior distribution. Next, we give a brief review of the MCMC algorithm.

The basic idea of MCMC is to construct a Markov chain over the state space until the chain has reached a stationary distribution, which is assumed to be a target distribution. Here our target distribution is set to be the posterior distribution $p(\boldsymbol{\theta}|D)$. Although there are many variants of MCMC, we focus on the Metropolis-Hastings MCMC (MH-MCMC) method [43–45]. The basic idea of constructing the Markov chain in MH-MCMC algorithm is that at each step, given current state $\boldsymbol{\theta}$, a candidate state $\boldsymbol{\theta}^{\text{cand}}$ is proposed with probability $q(\boldsymbol{\theta}^{\text{cand}}|\boldsymbol{\theta})$, where $q(\cdot|\cdot)$ is referred to as the *proposal distribution*. Note that proposal distribution can be arbitrary. A commonly used proposal is the symmetric Gaussian distribution, that is, $q(\boldsymbol{\theta}^{\text{cand}}|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}, \lambda I)$, where λ denotes the *stepsize*. Whether to accept the candidate state is determined by the acceptance probability α, given by the following formula,

$$\alpha = \min\left\{1, \frac{p(\boldsymbol{\theta}^{\text{cand}}|D)/q(\boldsymbol{\theta}^{\text{cand}}|\boldsymbol{\theta})}{p(\boldsymbol{\theta}|D)/q(\boldsymbol{\theta}|\boldsymbol{\theta}^{\text{cand}})} = \frac{p(\boldsymbol{\theta}^{\text{cand}}|D)}{p(\boldsymbol{\theta}|D)}\right\}.$$

Note that the equation holds only when the proposal distribution is symmetric. If $p(\boldsymbol{\theta}^{\text{cand}}|D)/p(\boldsymbol{\theta}|D) > 1$, it means that $\boldsymbol{\theta}^{\text{cand}}$ is more possible than the current state $\boldsymbol{\theta}$, we accept the proposal with probability $\alpha = 1$. Otherwise, we accept the proposal with probability α. The detailed MH-MCMC algorithm is summarized in Appendix B.

4. Numerical Experiments

In this section, we consider three numerical examples: a standard linear Gaussian model, a nonlinear simple model, and a partial differential equation (PDE) model. Since the first problem has analytical expressions, we examine the performance of our method by comparing the numerical result with the exact solution. The second problem is a commonly-used test problem, and we demonstrate the efficiency of our method through it. In the third test problem, a physical system governed by the diffusion equation is considered, in which a contaminant source inversion problem is studied.

4.1. Test Problem 1: Linear Gaussian Problem

We consider the standard linear Gaussian problem in the following form,

$$G(\boldsymbol{\theta}, \boldsymbol{d}) = \boldsymbol{\theta}^T \boldsymbol{d}, \quad \boldsymbol{y} = G(\boldsymbol{\theta}, \boldsymbol{d}) + \boldsymbol{\epsilon},$$

where the noise and the prior are assumed to be Gaussian, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2)$, $\boldsymbol{\theta} \sim \mathcal{N}(0, I_{n_\theta \times n_\theta})$ and $n_\theta = n_d$. The posterior distribution is a multivariate Gaussian distribution

$$\boldsymbol{\theta}|\boldsymbol{d}, \boldsymbol{y} \sim \mathcal{N}(\overline{\boldsymbol{\theta}}, \overline{\Sigma}),$$

where the mean and the covariance are

$$\overline{\boldsymbol{\theta}} = \frac{\boldsymbol{y}}{\sigma^2}\overline{\Sigma}\boldsymbol{d}, \quad \overline{\Sigma} = \left(\frac{\boldsymbol{d}\boldsymbol{d}^T}{\sigma^2} + I\right)^{-1}.$$

The expected information gain (EIG) for $\boldsymbol{\theta}$ can then be given in a closed form. (The detailed deduction is shown in Appendix A.),

$$U(\boldsymbol{d}) = -\frac{1}{2}\log\left[\det\left(\overline{\Sigma}\right)\right]. \tag{13}$$

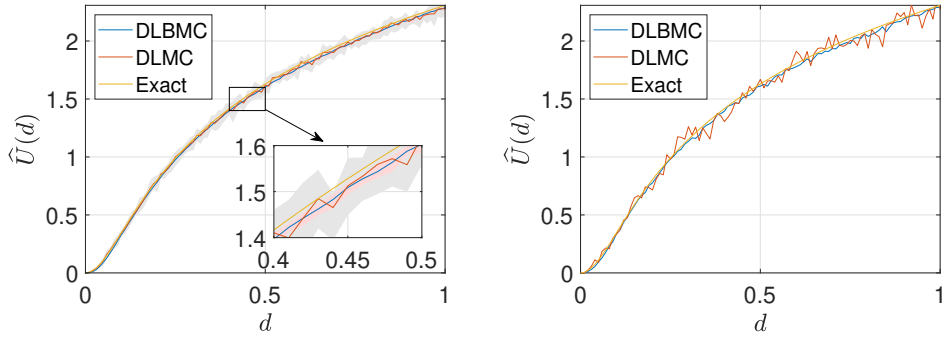
Note that maximizing EIG is equivalent to minimizing the determinant of the posterior covariance matrix [11] (the Bayesian D-optimal design).

We consider the one-dimensional case where $n_\theta = 1$, $d \in [0,1]$, and set the standard deviation of the noise to $\sigma = 0.1$. The hyperparameters of DLBMC are set to $l_{\text{in}} = 0.5$, $(\sigma_f)_{\text{in}} = 1$, and $l_{\text{out}} = 0.2$, $(\sigma_f)_{\text{out}} = 1$. Figure 1 shows the exact EIG, the estimated EIG using DLMC with 300 samples, and the estimated EIG using DLBMC with 300 samples. It can be seen that the estimated EIG of DLBMC is more accurate and more stable than that of DLMC. Besides, for $d = 0.3$, the relationship between the sample size and the bias of the DLMC and DLBMC estimators is studied. As the sample size increases, we compute the bias of the DLBMC estimator and the DLMC estimator averaging 20 trails respectively. Figure 2(Left) shows the results of the bias, where it can be seen that the DLBMC estimator converges to the true value faster than the DLMC estimator, and DLBMC is more accurate than DLMC.

As discussed in Section 3.2, the variance of the BMC estimator for the uniform distribution is not explicitly given. We use the sample variance as an alternative to compare the discrepancy of the DLMC estimator and the DLBMC estimator. Fixing the design point $d = 0.5$, for different sample sizes of DLMC and DLBMC, we repeatedly compute the estimator $\widehat{U}(d)$ n times, and denote them by $\{\widehat{U}^{(i)}\}_{i=1}^{n}$. Let $\overline{U}(d)$ denote the mean estimator, $\overline{U}(d) := \frac{1}{n}\sum_{i=1}^{n}\widehat{U}^{(i)}(d)$, and then the sample variance estimator is defined as

$$s^2 := \frac{\sum_{i=1}^{n}(\widehat{U}^{(i)}(d) - \overline{U}(d))}{n-1}.$$

As the number of repeat times n increases, we compare the sample variance of two estimators with different sample sizes for DLMC and DLBMC in Figure 2(Right). It is clear that the DLBMC estimator outperforms the DLMC estimator.



(a) Estimated EIG $\widehat{U}(d)$ with 300 samples

(b) Estimated EIG $\widehat{U}(d)$ with 300 samples (one realization)

Figure 1. Estimated expected information gain (EIG) profile over the design space for test problem 1. (**a**) Pink and gray shaded areas represent the interval containing 80% of 20 independent estimates of two estimators at each d respectively. Blue line and red line indicates the means of estimates. (**b**) One set of realizations of the two estimators.

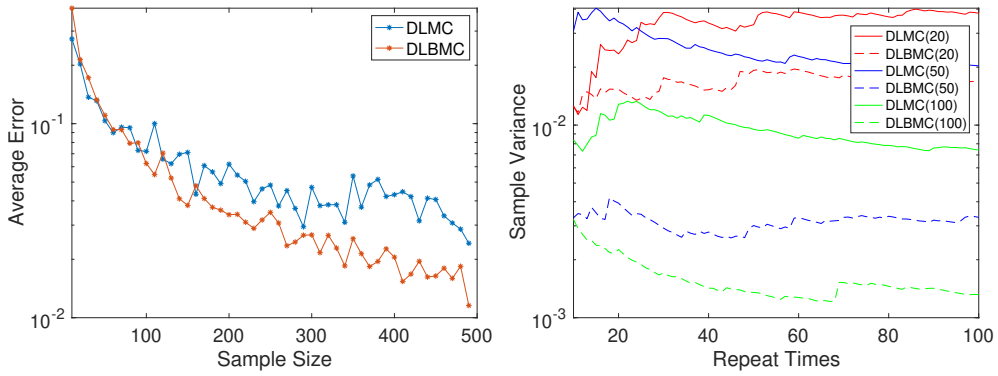


Figure 2. (**Left**) Error averaging over $n = 20$ trails versus the sample size of DLMC and DLBMC. (**Right**) Sample variance versus the repeat times for different sample sizes of double-loop Monte Carlo (DLMC) and double-loop Bayesian Monte Carlo (DLBMC) (numbers in the parenthesis indicate the sample sizes).

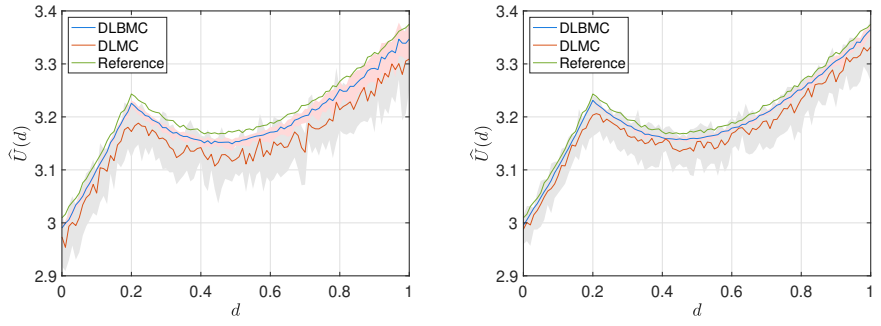
4.2. Test Problem 2: Nonlinear Simple Problem

In this section, a simple nonlinear model is tested, which is also studied in [14]. This model is written as

$$G(\theta, d) = \theta^3 d^2 + \theta \exp(-|0.2 - d|),$$

the prior is set to $\theta \sim \mathcal{U}(0, 1)$, and the standard deviation of the observation noise is set to $\sigma = 0.01$. The hyperparameters of DLBMC are set to $l_{\text{in}} = 0.01$, $(\sigma_f)_{\text{in}} = 0.2$, and $l_{\text{out}} = 0.05$, $(\sigma_f)_{\text{out}} = 0.2$.

Figure 3a,b show the estimated EIG using DLMC and DLBMC with 300 and 500 samples respectively. A reference solution using DLMC with 10^5 samples is also compared in Figure 3a,b. Here, 20 trails of the DLMC estimator and the DLBMC estimator are generated, and Figure 3 shows the mean estimates and the intervals containing 80% of the trails. It is clear that, compared to DLMC, our DLBMC estimator has smaller variances. Compared to the reference solution, DLBMC gives biased estimation. With the increasing sample size, the extent of bias is reduced as we expect.



(**a**) Estimated EIG $\widehat{U}(d)$ with 300 samples (**b**) Estimated EIG $\widehat{U}(d)$ with 500 samples

Figure 3. Estimated EIG profile over the design space for test problem 2. Pink and gray shaded areas represent the interval containing 80% of 20 independent estimates of EIG at each d for DLBMC and DLMC respectively. Blue line and red line denotes the means of these estimators. Green line denotes the reference solution given by DLMC with 10^5 samples.

To illustrate the effect of optimal design, we compare the posterior distribution given by three design points. Let design A = 1 be the optimal design point, let design B = 0.2 be the local maximizer of the EIG, and let design C = 0 which has the least information since it has the least EIG value. The MH-MCMC algorithm (Algorithm A1) with $N_{\text{iter}} = 1000$ and $\gamma = 0.2$ is used to generate samples

of the posterior distribution, and kernel density estimation is used to obtain the posterior density functions from these samples. For this test problem, the ground-truth is set to 0.75. From Figure 4, it can be seen that the posterior distribution obtained through design A is the most accurate, and it has the smallest variance.

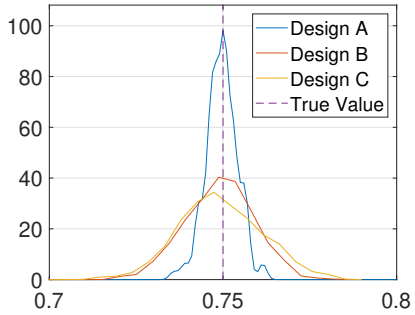


Figure 4. Posterior density functions given by different designs.

4.3. Test Problem 3: Source Inversion for the Diffusion Problem

Letting $\mathcal{D} \subset \mathbb{R}^2$ be a bounded and connected domain with a polygonal boundary $\partial\mathcal{D}$, the governing equation of the diffusion problem studied in this test problem is: find a random function $u(\boldsymbol{x}, \omega) \in D \times \Omega \to \mathbb{R}$, such that $\mathcal{P}$-a.e. in Ω,

$$-\nabla^2 u(\boldsymbol{x}, \omega) = f(\boldsymbol{x}, \omega), \quad \text{in } \mathcal{D}, \tag{14}$$

$$u(\boldsymbol{x}, \omega) = 0, \quad \text{on } \partial\mathcal{D}, \tag{15}$$

where $(\Omega, \mathcal{F}, \mathcal{P})$ is a probability space. We consider a square physical domain $\mathcal{D} = [-1, 1] \times [-1, 1] \subset \mathbb{R}^2$, and $u(\boldsymbol{x}, \omega)$ in (14)–(15) denotes the concentration of a contaminant at the point $\boldsymbol{x} \in \mathcal{D}$. Let $f(\boldsymbol{x}, \omega)$ denote the field of contaminant source. As f is usually strictly positive following the setting in [13,46–48], we set the prior distribution of $f(\boldsymbol{x}, \omega)$ to a log-normal random field, that is, $f(\boldsymbol{x}, \omega) = \exp(a(\boldsymbol{x}, \omega))$ where $a(\boldsymbol{x}, \omega)$ is a normal random field. In this study, the experimental goal is to infer the underlying contaminant field f given several observations $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^K$, where design variable $\boldsymbol{x}_i$ denotes i-th sensor placement, the response is the corresponding numerical PDE solution $u(\boldsymbol{x}_i)$ with additional noise, that is, $\boldsymbol{y}_i = u(\boldsymbol{x}_i) + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, and K is the number of sensors. In this test problem, the hyperparameters of DLBMC are set to $l_{\text{in}} = 0.02, (\sigma_f)_{\text{in}} = 0.01$, and $l_{\text{out}} = 0.005, (\sigma_f)_{\text{out}} = 0.005$.

We parameterize the permeability field $\log[f(\boldsymbol{x}, \omega)]$ by a truncated Karhunen-Loève (KL) expansion. Consider the random field $a(\boldsymbol{x}, \omega) = \log[f(\boldsymbol{x}, \omega)]$ with mean function $a_0(\boldsymbol{x})$, standard deviation σ and covariance function $C(\boldsymbol{x}, \boldsymbol{y})$,

$$C(\boldsymbol{x}, \boldsymbol{y}) = \sigma \exp\left(-\frac{|x_1 - y_1|}{c} - \frac{|x_2 - y_2|}{c}\right), \tag{16}$$

where c is the correlation length. Then the truncated KL expansion of f is expressed as

$$f(\boldsymbol{x}, \omega) \approx \exp\left(a_0(\boldsymbol{x}) + \sum_{n=0}^{M} \sqrt{\lambda_n}\xi_n a_n(\boldsymbol{x})\right), \tag{17}$$

where $a_n(\boldsymbol{x})$ and λ_n are the eigenfunctions and eigenvalues of (16) and $\{\xi_n\}_{n=1}^M$ are uncorrelated random variables. The prior of $\{\xi_n\}_{n=1}^M$ are set to be independent standard normal distributions, $\xi_n \sim \mathcal{N}(0, 1)$ for $n = 1, \ldots, M$.

In the numerical experiment, we set $a_0(x) = 1$, $c = 2$ and $\sigma = 1$. Fixing the hyperparameters in the truncated KL expansion, the response depends on the random variables $\{\xi_i\}_{i=1}^{M}$. Therefore we define the parameter of interest as $\boldsymbol{\theta} := [\xi_1, \ldots, \xi_M]$, and rewrite the governing diffusion equation as,

$$-\nabla^2 u(\boldsymbol{x}, \boldsymbol{\theta}) = f(\boldsymbol{x}, \boldsymbol{\theta}), \quad \text{in } \mathcal{D} \times \Gamma, \tag{18}$$

$$u(\boldsymbol{x}, \boldsymbol{\theta}) = 0, \quad \text{on } \partial\mathcal{D} \times \Gamma. \tag{19}$$

We use the bilinear finite element method (FEM) to discretize the diffusion equation over a 17×17 square grid and let the standard deviation of noise be 1% of the mean observed value. Supposing K sensors are placed over the design space, generally, we can perform a batch design, and write the following altered forward model

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_i \\ \vdots \\ \boldsymbol{y}_K \end{bmatrix} = \begin{bmatrix} \boldsymbol{G}(\boldsymbol{d}_1, \boldsymbol{\theta}) \\ \vdots \\ \boldsymbol{G}(\boldsymbol{d}_i, \boldsymbol{\theta}) \\ \vdots \\ \boldsymbol{G}(\boldsymbol{d}_K, \boldsymbol{\theta}) \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_i \\ \vdots \\ \boldsymbol{\epsilon}_K \end{bmatrix} = \mathcal{G}(\boldsymbol{d}_{1:K}, \boldsymbol{\theta}) + \boldsymbol{\epsilon},$$

where the subscript i denotes the i-th design variable for $i = 1, \ldots, K$. Directly maximing over the EIG with altered forward model can give optimal design in the context of batch design.

Let f_{truth} denote the underlying true permeability field. Suppose we have collected data on K sensors, and then we perform MCMC to get samples $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{N_{\text{iter}}}$ of the posterior distribution using Algorithm A1. In this test problem, we set $N_{\text{iter}} = 4000$. Two useful statistics can be obtained from the samples—the maximum a posterior (MAP) estimate $\boldsymbol{\theta}_{\text{MAP}}$ and the mean estimate $\boldsymbol{\theta}_{\text{MEAN}}$. Let f_{MAP} and f_{MEAN} be the source fields generated by $\boldsymbol{\theta}_{\text{MAP}}$ and $\boldsymbol{\theta}_{\text{MEAN}}$ respectively. To test the accuracy of the inversion, we introduce the following relative errors

$$E_{\text{MAP}} = \frac{\|f_{\text{MAP}} - f_{\text{truth}}\|_2}{\|f_{\text{truth}}\|_2},$$

$$E_{\text{MEAN}} = \frac{\|f_{\text{MEAN}} - f_{\text{truth}}\|_2}{\|f_{\text{truth}}\|_2},$$

where f_{truth}, f_{MAP} and f_{MEAN} are discretized over the FEM grids for computational simplicity and $\|\cdot\|_2$ denotes the l^2 vector norm.

Noting that performing grid search over $(17 \times 17)^K$ grid is computationally expensive, we utilize Bayesian optimization method to efficiently find the optimal designs within a few iterations. Three cases of K are considered in the following, which are $K = 1, 2, 3$.

First, for $K = 1$, we first perform Bayesian optimization over a 17×17 grid. The performance of Bayesian optimization is shown in Figure 5. In Figure 5(Left), we can see that, Bayesian optimization can find the maximum of the EIG within a few iterations. The optimal design found by Bayesian optimization is $[0, 0]$, which is expected due to the symmetry of the forward model. In this case, as the number of gird points (289 points) is not too large, we can verify that the optimal design is $[0, 0]$ through grid search. Figure 5(Right) shows that the average cumulative regret is convergent to zero.

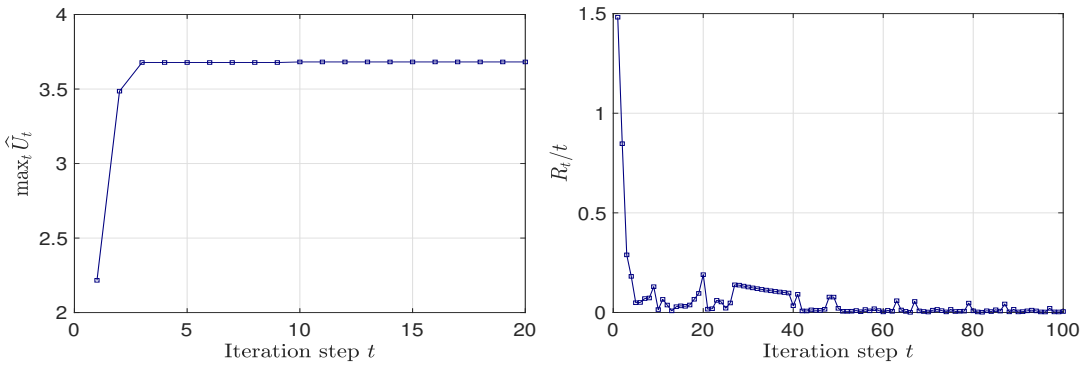


Figure 5. (**Left**) Maximum EIG value at iteration t of Bayesian optimization. (**Right**) Average cumulative regret at iteration t of Bayesian optimization. $K = 1$, test problem 3.

For comparison, we randomly generate 20 different design points and use the MH-MCMC algorithm to generate 4000 samples of the posterior distribution. Figure 6 shows the locations of the optimal design and the 20 random designs. Figure 7a,b show the relative errors of MAP and mean estimates of the source field (averaged over 20 trails) versus the values of EIG, where it is clear that as the value of EIG becomes larger, the errors of both MAP and mean estimates reduce. In addition, the optimal design point gives large EIG values and relatively small errors. Although the errors associated with the optimal design point are typically smaller than the errors associated with the random design points. They are still large, and the estimated source fields are not accurate enough. Therefore, we next consider more design points.

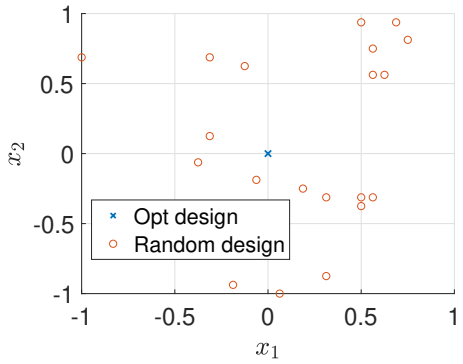
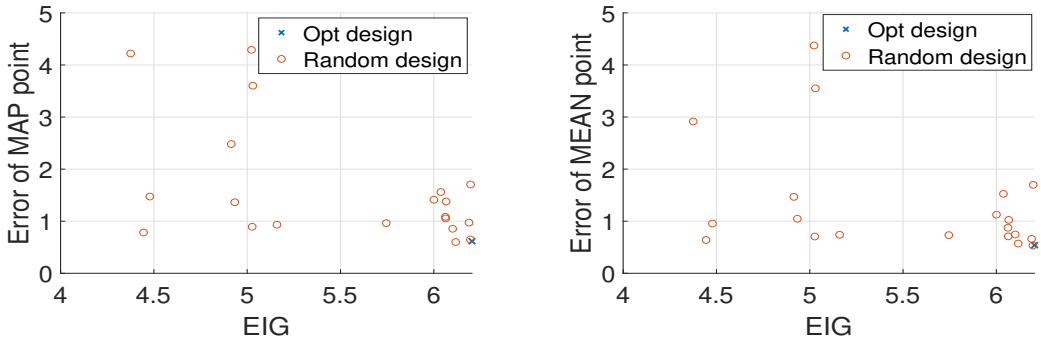


Figure 6. Sensor locations for $K = 1$, test problem 3.



(**a**) EIG versus MAP error E_{MAP} (**b**) EIG versus mean error E_{MEAN}

Figure 7. Value of EIG versus the relative errors averaged over 20 trails, $K = 1$, test problem 3.

For the multiple design cases ($K = 2, 3$), given the fact that the computational cost of batch design increases exponentially as K increases, we perform Bayesian optimization over a uniform 9×9 coarse grid. For $K \geq 2$, it can be seen that the optimal solution is not unique due to the symmetry of the forward model, for example, $[x_1; x_2]$ and $[-x_1; -x_2]$ share the same EIG value. After performing Bayesian optimization several times with BO budget $t_{\max} = 100$, the sets of optimal designs for $K = 2$ case are shown in Figure 8, where each line connecting blue circle and red circle represents a pair of optimal design. The numerical result shows that, compared with a pair of two design points that are symmetric with respect to $[0; 0]$, a pair of slightly skewed design points can provide more information.

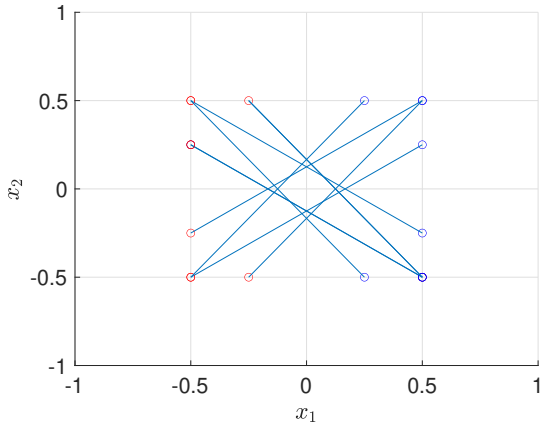
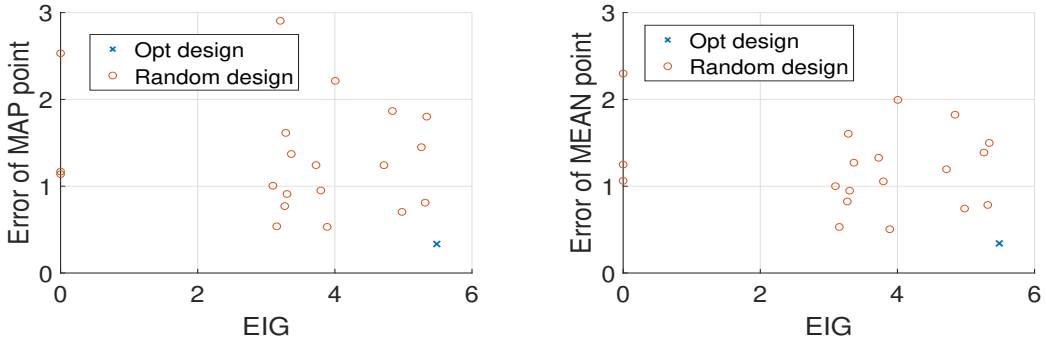


Figure 8. Optimal design (each line connecting blue circle and red circle represents a pair of optimal design points), $K = 2$, test problem 3.

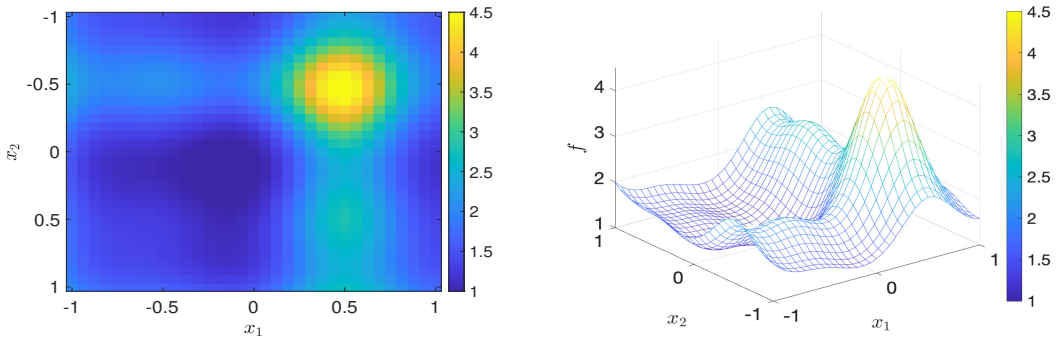
Again, we randomly generate 20 sets of design points, and compare them with the optimal design (we choose $[0.5, 0.5; -0.5, -0.25]$). Figure 9 shows the errors of MAP and mean estimates (averaged over 20 trails) of the source field versus the values of EIG, where it is clear that the optimal design leads to the largest EIG value and the smallest error. Besides, the comparison of true source field and the fields generated by MAP and mean estimates are presented in Figure 10. It can be seen that the estimated source field associated with the optimal design matches the true source field well.



(**a**) EIG versus MAP error E_{MAP} (**b**) EIG versus mean error E_{MEAN}

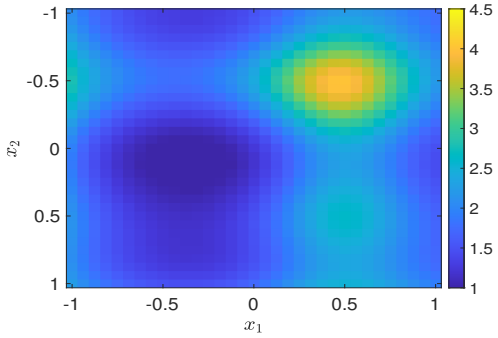
Figure 9. Value of EIG versus the relative errors averaged over 20 trails, $K = 2$, test problem 3.

For $K = 3$, let the BO budget $t_{\max} = 100$, the set of optimal design that found by Algorithm 2 is $[0.75, 0.25; 0.5, -0.25; -0.75, -0.25]$. Figure 11 shows the values of EIG and the relative errors in MAP and mean estimates (averaged over 20 trails) for the optimal design and twenty random designs, where it can be seen that the optimal design has the largest EIG value and the smallest errors which are consistent with the results for $K = 1, 2$. Figure 12 shows that the estimated source fields generated by the MAP and the mean estimates match the true source field well.

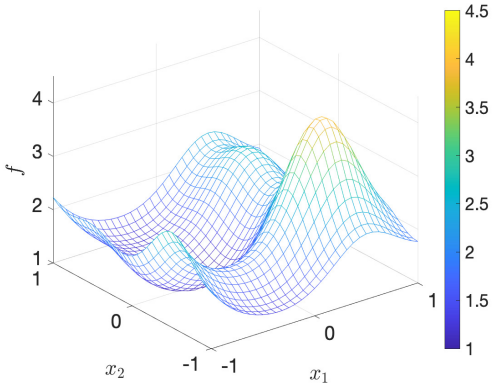


(**a**) True source field

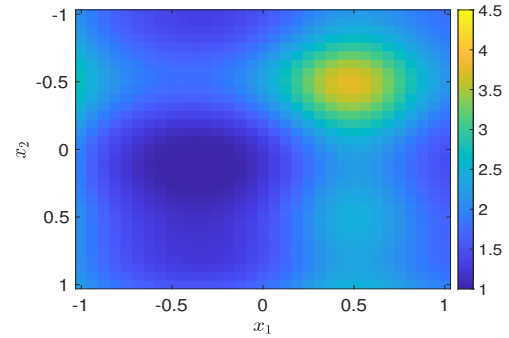
(**b**) True source field



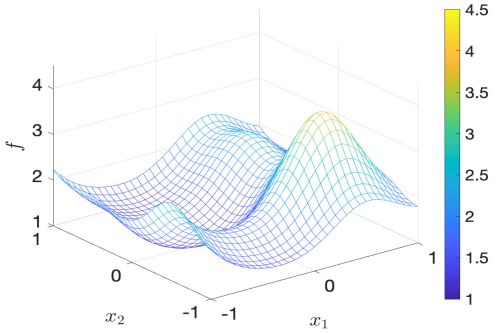
(**c**) Estimated source field by MAP



(**d**) Estimated source field by MAP



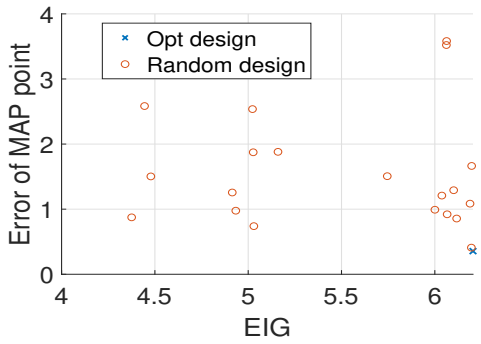
(**e**) Estimated source field by the mean estimate



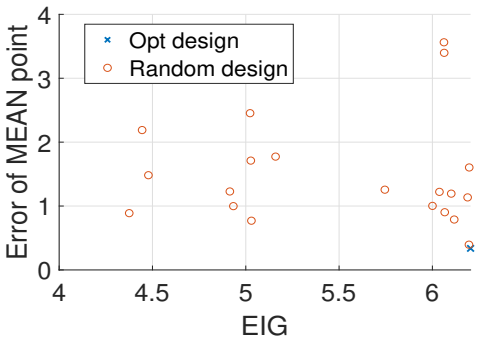
(**f**) Estimated source field by the mean estimate

Figure 10. Comparison of the true source field and estimated source fields by MAP and mean estimates for $K = 2$, test problem 3.

To further quantify the performance of optimal designs, we compute the ratio of E_{MAP} of random designs and E_{MAP} of optimal designs (denoted as $E_{\text{MAP}}^{(\text{random})}$ and $E_{\text{MAP}}^{(\text{opt})}$ respectively) for $K = 1, 2, 3$ cases. Figure 13 presents the histograms of ratio of relative errors (i.e., $E_{\text{MAP}}^{(\text{random})}/E_{\text{MAP}}^{(\text{opt})}$), where the green lines are the corresponding kernel smoothing function estimates. We can see that, in all three cases, with high probability, the optimal design can give better posterior performance than the random designs. Especially, from Figure 13c, the error of random designs can be ten times greater than that of the optimal design.

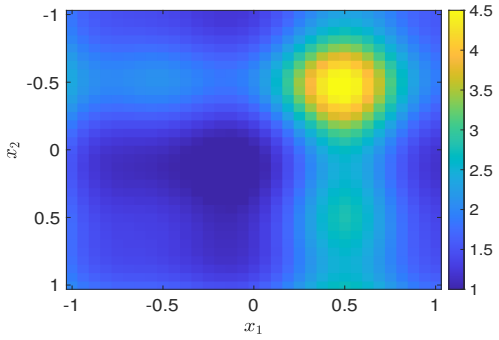


(**a**) EIG versus maximum a posterior (MAP) error E_{MAP}

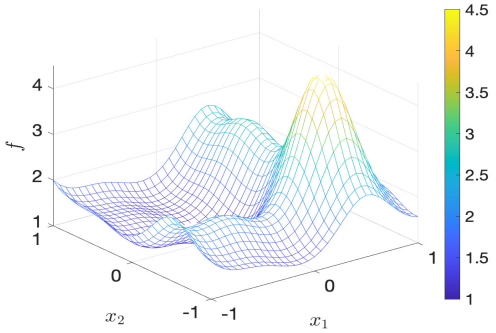


(**b**) EIG versus mean error E_{MEAN}

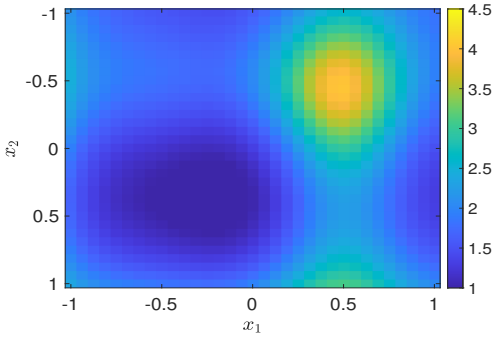
Figure 11. Value of EIG versus the relative errors averaged over 20 trails, $K = 3$, test problem 3.



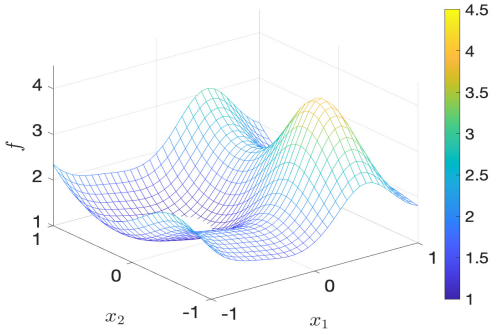
(**a**) True source field



(**b**) True source field

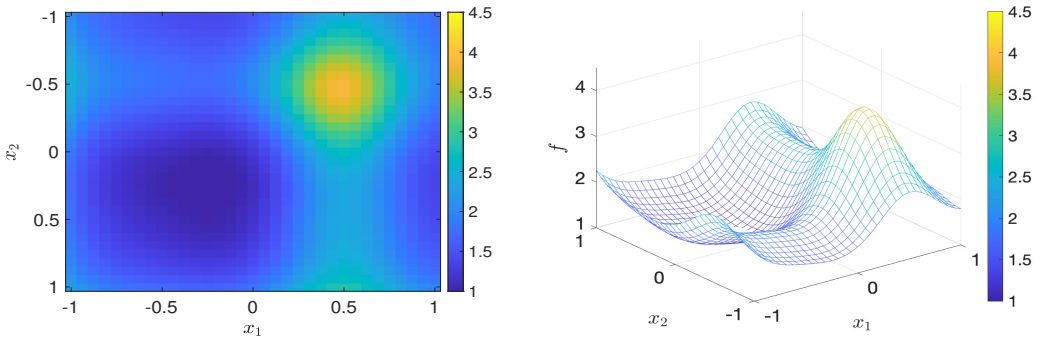


(**c**) Estimated source field by MAP



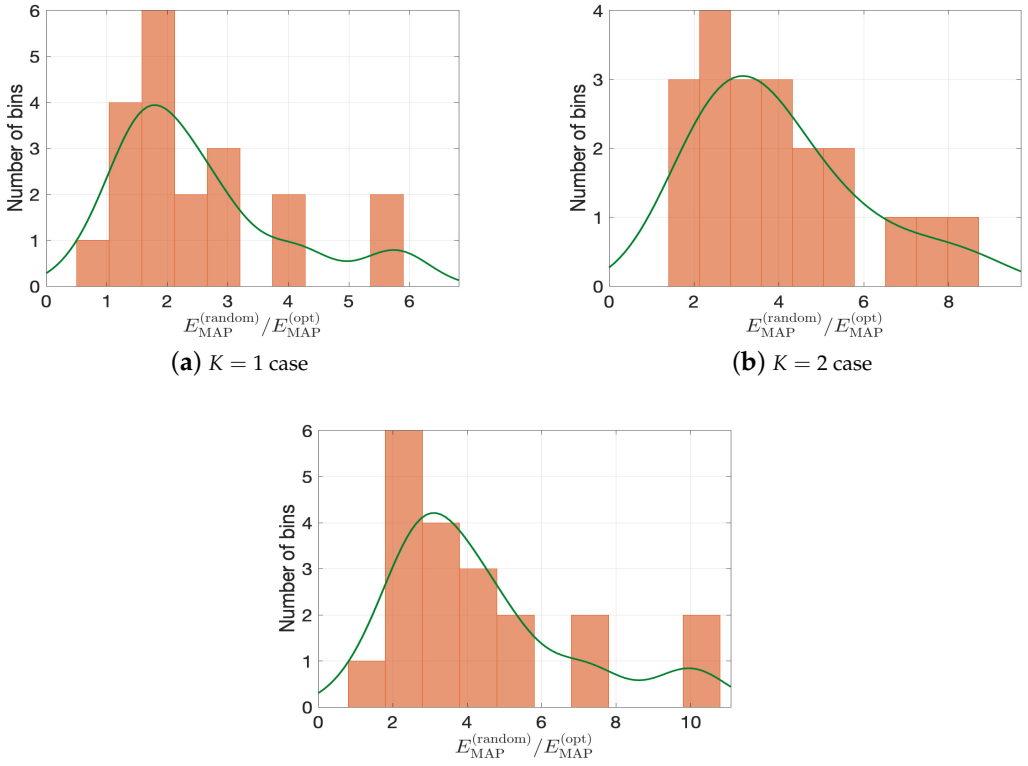
(**d**) Estimated source field by MAP

Figure 12. *Cont.*



(**e**) Estimated source field by the mean estimate (**f**) Estimated source field by the mean estimate

Figure 12. Comparison of the true source field and estimated source fields by MAP and mean estimates for $K = 3$, test problem 3.



(**a**) $K = 1$ case (**b**) $K = 2$ case

(**c**) $K = 3$ case

Figure 13. Histograms of $E_{\mathrm{MAP}}^{(\mathrm{random})}/E_{\mathrm{MAP}}^{(\mathrm{opt})}$. Green lines denotes the kernel smoothing function estimates.

5. Conclusions and Discussion

Efficiently using a small number of samples to reduce the cost of computing the expected information gain (EIG) is a fundamental concept to solve the challenging Bayesian optimal experimental design problem. Based on the Bayesian Monte Carlo (BMC) method, a novel double-loop Bayesian Monte Carlo (DLBMC) estimator is proposed for evaluating the EIG in this work. To result in an efficient overall optimization procedure to find the maximizer of the EIG, a Bayesian optimization (BO) procedure for EIG is developed. In addition, our analysis gives explicit expressions of the mean estimate of the BMC estimator and the bounds of its variance for the uniform and the normal distributions. Detailed numerical studies show that our DLBMC method can provide accurate mean

estimates with small variances, and the overall BO procedure leads to optimal designs which give efficient Bayesian inference results.

As our novel DLBMC estimator for EIG is based on Gaussian process, it is currently limited to low-dimensional problems where the number of design variables is not large. In this work, the classical BO and MCMC approaches are used, and it is not straight forward to apply them for high-dimensional problems. For high-dimensional Bayesian optimal design problems with a large number of design variables, a possible solution is to conduct a sequential design procedure, and apply DLBMC at each step in the sequential procedure. Conducting a systematic DLBMC based on the sequential design will be the focus of our future work.

Author Contributions: Conceptualization, Z.X. and Q.L.; methodology, Q.L.; software, Z.X.; validation, Z.X.; formal analysis, Z.X.; investigation, Z.X.; resources, Z.X.; data curation, Z.X.; writing—original draft preparation, Z.X.; writing—review and editing, Q.L.; visualization, Z.X.; supervision, Q.L.; project administration, Q.L.; funding acquisition, Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (No. 11601329).

Acknowledgments: The authors thank Jinglai Li for helpful discussions, and the anonymous reviewers for their thoughtful comments and suggestions that helped us improve our work and article.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Deduction of Linear Gaussian Model

We consider the standard linear Gaussian model,

$$G(\theta, d) = \theta^T d\,, \quad y = G(\theta, d) + \epsilon\,,$$

where the noise and prior are assumed to be Gaussian, $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$, and $\theta \sim \mathcal{N}(0, I_{n\times n})$.

$$p(\theta) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\theta^T\theta\right)\,.$$

then the likelihood is given by

$$\begin{aligned} p(y|d,\theta) &= \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(y - G(d,\theta))^2}{2\sigma_n^2}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(y - \theta^T d)^2}{2\sigma_n^2}\right)\,. \end{aligned}$$

The posterior is given by

$$\begin{aligned} p(\theta|y,d) &= \frac{p(y|\theta,d)p(\theta)}{p(y|d)} \propto p(y|\theta,d)p(\theta) \\ &\propto \exp\left[-\frac{(y-\theta^T d)^2}{2\sigma_n^2} - \frac{\theta^T\theta}{2}\right] \\ &= \exp\left[-\frac{1}{2}\theta^T\left(\frac{dd^T}{\sigma_n^2} + I\right)\theta + \frac{y}{\sigma_n^2}\theta^T d + \frac{y^2}{2\sigma_n^2}\right] \\ &\propto \exp\left[-\frac{1}{2}(\theta-\overline{\theta})^T\left(\frac{dd^T}{\sigma_n^2} + I\right)(\theta-\overline{\theta})\right]\,, \end{aligned}$$

where

$$\overline{\theta} = \frac{y}{\sigma_n^2}\left(\frac{dd^T}{\sigma_n^2} + I\right)^{-1} d\,, \quad \overline{\Sigma} = \left(\frac{dd^T}{\sigma_n^2} + I\right)^{-1}\,.$$

By the definition of expected information gain,

$$U(d) = \int_{\mathcal{Y}} \int_{\Theta} \log \left[\frac{p(\theta|y,d)}{p(\theta)} \right] p(\theta|y,d)\, \mathrm{d}\theta\, \mathrm{d}y\,.$$

Supposing $\theta|y,d \sim \mathcal{N}(\overline{\theta}, \overline{\Sigma})$, then we have

$$\frac{p(\theta|y,d)}{p(\theta)} = \frac{1}{(\det \overline{\Sigma})^{1/2}} \exp(-\frac{(\theta-\overline{\theta})\Sigma^{-1}(\theta-\overline{\theta})}{2} + \frac{\theta^T\theta}{2})\,.$$

Thus,

$$\log \frac{p(\theta|y,d)}{p(\theta)} = -\frac{1}{2}\ln(\det \overline{\Sigma}) - \frac{(\theta-\overline{\theta})\Sigma^{-1}(\theta-\overline{\theta})}{2} + \frac{\theta^T\theta}{2}.$$

Below, we show that

$$\begin{aligned}&\int_y \int_{\Theta} \frac{(\theta-\overline{\theta})\Sigma^{-1}(\theta-\overline{\theta})}{2} p(\theta|y,d)\, \mathrm{d}\theta p(y|d)\, \mathrm{d}y \\ &= \int_y \int_{\Theta} \frac{\theta^T\theta}{2} p(\theta|y,d)\, \mathrm{d}\theta p(y|d)\, \mathrm{d}y.\end{aligned}$$

Let $z = \Sigma^{-1/2}\theta$, which can be viewed as a linear transformation of a random vector. We have that $\overline{z} = \Sigma^{-1/2}\overline{\theta}$, and the covariance of z is

$$\begin{aligned}\mathbb{V}(z) &= \mathbb{E}[(z-\overline{z})(z-\overline{z})^T] \\ &= \Sigma^{-1/2}\mathbb{E}[(\theta-\overline{\theta})(\theta-\overline{\theta})^T]\Sigma^{-1/2} = I.\end{aligned} \tag{A1}$$

Taking (A1) at hand, the following equation can be obtained

$$\begin{aligned}&\int_{\Theta} \frac{(\theta-\overline{\theta})\Sigma^{-1}(\theta-\overline{\theta})}{2} p(\theta|y,d)\, \mathrm{d}\theta = \int_Z \frac{(z-\overline{z})^T(z-\overline{z})}{2} p(z|y,d)\, \mathrm{d}z \\ &= \sum_{i=1}^{n} E[(z_i-\overline{z}_i)^2]/2 = n/2.\end{aligned}$$

Then we also have

$$\int_y \int_{\Theta} \frac{(\theta-\overline{\theta})\Sigma^{-1}(\theta-\overline{\theta})}{2} p(\theta|y,d)\, \mathrm{d}\theta p(y|d)\, \mathrm{d}y = n/2.$$

Next, we consider the integration $\int_y \int_{\Theta} \frac{\theta^T\theta}{2} p(\theta|y,d)\, \mathrm{d}\theta p(y|d)\, \mathrm{d}y$. Since $p(\theta|y,d)p(y|d) = p(\theta)p(y|\theta,d)$,

$$\begin{aligned}&\int_y \int_{\Theta} \frac{\theta^T\theta}{2} p(\theta|y,d)\, \mathrm{d}\theta p(y|d)\, \mathrm{d}y = \int_y \int_{\Theta} \frac{\theta^T\theta}{2} p(\theta) d\theta p(y|\theta,d)\, \mathrm{d}y \\ &= \int_{\Theta} \frac{\theta^T\theta}{2} p(\theta) \int_y p(y|\theta,d)\, \mathrm{d}y\, \mathrm{d}\theta = \int_{\Theta} \frac{\theta^T\theta}{2} p(\theta)\, \mathrm{d}\theta = n/2.\end{aligned}$$

Therefore we have the analytical form of EIG,

$$U(d) = -\frac{1}{2}\log\left[\det\left(\overline{\Sigma}\right)\right]\,. \tag{A2}$$

Appendix B. Metropolis-Hastings MCMC Algorithm

Algorithm A1 The Metropolis-Hastings Markov chain Monte Carlo (MH-MCMC) algorithm

Input: Prior $p(\boldsymbol{\theta})$, stepsize γ, maximum number of iterations N_{iter}.
Initialize: $\boldsymbol{\theta}^{(0)} \sim p(\boldsymbol{\theta})$.
for $i = 1, 2, \ldots, N_{\text{iter}}$ **do**
    Propose: $\boldsymbol{\theta}^{\text{cand}} = \boldsymbol{\theta}^{(i-1)} + \gamma \mathcal{N}(0, I)$.
    Acceptance Probability:
        $$\alpha(\boldsymbol{\theta}^{\text{cand}} | \boldsymbol{\theta}^{(i-1)}) = \min \left\{ 1, \frac{p(\boldsymbol{\theta}^{\text{cand}}|D)}{p(\boldsymbol{\theta}^{(i-1)}|D)} \right\} .$$
    Draw $u \sim \mathcal{U}(0, 1)$.
    if $u < \alpha$ **then**
        Accept the proposal: $\boldsymbol{\theta}^{(i)} \leftarrow \boldsymbol{\theta}^{\text{cand}}$.
    else
        Reject the proposoal: $\boldsymbol{\theta}^{(i)} \leftarrow \boldsymbol{\theta}^{(i-1)}$.
    end if
end for

References

1. Jones, M.; Goldstein, M.; Jonathan, P.; Randell, D. Bayes linear analysis for Bayesian optimal experimental design. *J. Stat. Plan. Inference* **2016**, *171*, 115–129. [CrossRef]
2. Atkinson, A.; Donev, A.; Tobias, R. *Optimum Experimental Designs, with SAS*; Oxford University Press: Oxford, UK, 2007; Volume 34.
3. Bernardo, J.M.; Smith, A.F. *Bayesian Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2009; Volume 405.
4. Stuart, A.M. Inverse problems: A Bayesian perspective. *Acta Numer.* **2010**, *19*, 451–559. [CrossRef]
5. Tarantola, A. *Inverse Problem Theory and Methods for Model Parameter Estimation*; SIAM: Philadelphia, PA, USA, 2005; Volume 89.
6. Kaipio, J.; Somersalo, E. *Statistical and Computational Inverse Problems*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006; Volume 160.
7. Alexanderian, A.; Petra, N.; Stadler, G.; Ghattas, O. A-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems with regularized ℓ_0-sparsification. *SIAM J. Sci. Comput.* **2014**, *36*, A2122–A2148. [CrossRef]
8. Alexanderian, A.; Petra, N.; Stadler, G.; Ghattas, O. A fast and scalable method for A-optimal design of experiments for infinite-dimensional Bayesian nonlinear inverse problems. *SIAM J. Sci. Comput.* **2016**, *38*, A243–A272. [CrossRef]
9. Weaver, B.P.; Williams, B.J.; Anderson-Cook, C.M.; Higdon, D.M. Computational enhancements to Bayesian design of experiments using Gaussian processes. *Bayesian Anal.* **2016**, *11*, 191–213. [CrossRef]
10. Lindley, D.V. *Bayesian Statistics, A Review*; SIAM: Philadelphia, PA, USA, 1972, Volume 2.
11. Chaloner, K.; Verdinelli, I. Bayesian experimental design: A review. *Stat. Sci.* **1995**, *10*, 273–304. [CrossRef]
12. Müller, P.; Parmigiani, G. Optimal design via curve fitting of Monte Carlo experiments. *J. Am. Stat. Assoc.* **1995**, *90*, 1322–1330.
13. Huan, X.; Marzouk, Y. Gradient-based stochastic optimization methods in Bayesian experimental design. *Int. J. Uncertain. Quantif.* **2014**, *4*, 479–510. [CrossRef]
14. Huan, X.; Marzouk, Y.M. Simulation-based optimal Bayesian experimental design for nonlinear systems. *J. Comput. Phys.* **2013**, *232*, 288–317. [CrossRef]
15. Wang, H.; Lin, G.; Li, J. Gaussian process surrogates for failure detection: A Bayesian experimental design approach. *J. Comput. Phys.* **2016**, *313*, 247–259. [CrossRef]
16. Drovandi, C.C.; Tran, M.N. Improving the efficiency of fully Bayesian optimal design of experiments using randomised quasi-Monte Carlo. *Bayesian Anal.* **2018**, *13*, 139–162. [CrossRef]
17. Feng, C.; Marzouk, Y.M. A layered multiple importance sampling scheme for focused optimal Bayesian experimental design. *arXiv* **2019**, arXiv:1903.11187.

18. Overstall, A.M.; Woods, D.C. Bayesian design of experiments using approximate coordinate exchange. *Technometrics* **2017**, *59*, 458–470. [CrossRef]
19. Overstall, A.; McGree, J. Bayesian design of experiments for intractable likelihood models using coupled auxiliary models and multivariate emulation. *Bayesian Anal.* **2018**, *15*, 103–131. [CrossRef]
20. Ryan, E.G.; Drovandi, C.C.; McGree, J.M.; Pettitt, A.N. A review of modern computational algorithms for Bayesian optimal design. *Int. Stat. Rev.* **2016**, *84*, 128–154. [CrossRef]
21. Rasmussen, C.E. Gaussian processes in machine learning. In *Summer School on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2003, pp. 63–71.
22. Deisenroth, M.P.; Huber, M.F.; Hanebeck, U.D. Analytic moment-based Gaussian process filtering. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 225–232.
23. Santner, T.J.; Williams, B.J.; Notz, W.; Williams, B.J. *The Design and Analysis of Computer Experiments*; Springer: Berlin/Heidelberg, Germany, 2003; Volume 1.
24. O'Hagan, A. Bayes–hermite quadrature. *J. Stat. Plan. Inference* **1991**, *29*, 245–260. [CrossRef]
25. Rasmussen, C.E.; Ghahramani, Z. Bayesian Monte Carlo. In Proceedings of the 2003 Neural Information Processing Systems, Vancouver, BC, Canada, 8–13 December 2003; pp. 505–512.
26. Briol, F.X.; Oates, C.J.; Girolami, M.; Osborne, M.A.; Sejdinovic, D. Probabilistic integration: A role in statistical computation? *Stat. Sci.* **2019**, *34*, 1–22. [CrossRef]
27. Brochu, E.; Cora, V.M.; De Freitas, N. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv* **2010**, arXiv:1012.2599.
28. Snoek, J.; Larochelle, H.; Adams, R.P. Practical bayesian optimization of machine learning algorithms. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 2951–2959.
29. Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R.P.; De Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE* **2015**, *104*, 148–175. [CrossRef]
30. Ryan, K.J. Estimating expected information gains for experimental designs with application to the random fatigue-limit model. *J. Comput. Graph. Stat.* **2003**, *12*, 585–603. [CrossRef]
31. Bungartz, H.J.; Griebel, M. Sparse grids. *Acta Numer.* **2004**, *13*, 147–269. [CrossRef]
32. Xiu, D. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*; Princeton University Press: Princeton, NJ, USA, 2010.
33. Elman, H.; Liao, Q. Reduced Basis Collocation Methods for Partial Differential Equations with Random Coefficients. *SIAM/ASA J. Uncertain. Quantif.* **2013**, *1*, 192–217. [CrossRef]
34. Villemonteix, J.; Vazquez, E.; Walter, E. An informational approach to the global optimization of expensive-to-evaluate functions. *J. Glob. Optim.* **2009**, *44*, 509. [CrossRef]
35. Jones, D.R. A taxonomy of global optimization methods based on response surfaces. *J. Glob. Optim.* **2001**, *21*, 345–383. [CrossRef]
36. Lam, R.; Willcox, K.; Wolpert, D.H. Bayesian optimization with a finite budget: An approximate dynamic programming approach. In Proceedings of the 2016 NIPS, Barcelona, Spain, 5–10 December 2016; pp. 883–891.
37. Srinivas, N.; Krause, A.; Kakade, S.M.; Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv* **2009**, arXiv:0912.3995.
38. Shahriari, B.; Bouchard-Côté, A.; Freitas, N. Unbounded Bayesian optimization via regularization. In Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, Cadiz, Spain, 9–11 May 2016; pp. 1168–1176.
39. Lizotte, D.J. Practical Bayesian Optimization. Doctor's Thesis, University of Alberta, Edmonton, AB, Canada, 2008.
40. Kushner, H.J. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *J. Basic Eng. Mar.* **1964**, *86*, 97–106. [CrossRef]
41. Mockus, J.; Tiesis, V.; Zilinskas, A. The application of bayesian methods for seeking the extremum. In *Toward Global Optimization*; North-Holand: Amsterdam, The Netherlands, 1978.
42. Jones, D.R.; Schonlau, M.; Welch, W.J. Efficient global optimization of expensive black-box functions. *J. Glob. Optim.* **1998**, *13*, 455–492. [CrossRef]

43. Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092. [CrossRef]
44. Hastings, W.K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **1970**, *57*, 97–109. [CrossRef]
45. Robert, C.; Casella, G. *Monte Carlo Statistical Methods*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
46. Marzouk, Y.M.; Najm, H.N.; Rahn, L.A. Stochastic spectral methods for efficient Bayesian solution of inverse problems. *J. Comput. Phys.* **2007**, *224*, 560–586. [CrossRef]
47. Marzouk, Y.M.; Najm, H.N. Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. *J. Comput. Phys.* **2009**, *228*, 1862–1902. [CrossRef]
48. Li, J.; Marzouk, Y.M. Adaptive construction of surrogates for the Bayesian solution of inverse problems. *SIAM J. Sci. Comput.* **2014**, *36*, A1163–A1186. [CrossRef]