Title
Tractable Global Solutions to Bayesian Optimal Experiment Design

Authors
Rodrigues, Diogo
Makrygiorgos, Georgios
Mesbah, Ali

Publication Date
2020

Peer reviewed

Tractable Global Solutions to Bayesian Optimal Experiment Design

Diogo Rodrigues, Georgios Makrygiorgos, and Ali Mesbah

Abstract— **Optimal experiment design (OED) aims to optimize the information content of experimental observations for various types of applications by designing the experimental conditions. In Bayesian OED for parameter estimation, the design selection is based on an expected utility metric that accounts for the joint probability distribution of the uncertain parameters and the observations. This work presents an approximation of the Bayesian OED problem based on Kullback–Leibler divergence that is amenable to global optimization. The experiment design adopts a parsimonious input parametrization that reduces the number of design variables. This leads to a tractable polynomial optimization problem that can be solved to global optimality via the concept of sum-of-squares polynomials.**

I. INTRODUCTION

The optimal selection of conditions under which experiments are conducted is crucial for maximizing the value of data for inference and prediction, in particular when experiments are time-consuming or resource-intensive to perform. Optimal experiment design (OED) uses a system model to systematically select experimental conditions (i.e., designs) by maximizing the information content of observations for parameter inference or model discrimination [1]–[6].

This paper focuses on OED for parameter estimation, which has been extensively studied in the classical frequentist framework. Classical OED formulations are based on scalar metrics of the Fisher information matrix (FIM) such as the alphabetic optimality criteria [7]–[9]. On the other hand, the design criteria in Bayesian OED approaches are defined in terms of *expected utility*, which is often expressed in terms of prior and posterior distributions of the parameters [10], [11]. Bayesian OED is especially advantageous when the system observations are noisy, incomplete, and indirect [12].

A common choice for the expected utility is the mutual information between parameters and observations, defined in terms of the Kullback–Leibler (KL) divergence from the prior to the posterior parameter distributions [13], [14]. As no closed-form expression exists for the expected utility for general nonlinear systems [15], a key computational challenge in Bayesian OED arises from numerical evaluation of the expected utility using Monte Carlo-based methods [16]. Due to this sample-based evaluation of the expected utility, Bayesian OED is naturally formulated as a stochastic optimization problem, which can be prohibitively expensive to solve for OED problems with large design spaces. Alternatively, gradient-based optimization approaches such as stochastic approximation [17] and sample average approximation [18] methods can be used to attain locally optimal designs. The gradient-based optimization approaches generally require fewer iterations and are potentially much less expensive than stochastic optimization approaches to Bayesian OED. However, sample-based approximations of the expected utility and its gradients can be prohibitively expensive. These challenges have been addressed by constructing surrogates for the model outputs based on polynomial chaos expansions [12], [19]. Despite these advances, gradient-based methods cannot guarantee the global optimality of the selected designs.

Hence, this paper presents a novel tractable approach for obtaining globally optimal solutions to Bayesian OED for nonlinear systems with time-varying designs. We express the expected utility in terms of the KL divergence from the prior to the posterior parameter distributions, which is approximated as D-optimality of the FIM for the special case of Gaussian prior distribution and Gaussian observation noise. A sample-based approach is then utilized for the computation of the expected utility for a given design via an optimal stochastic collocation scheme for numerical integration over the domain of uncertain parameters. The quadrature rule is built upon the notion of orthogonal polynomials, which has been extensively used in the approximation of functions of random variables [20]. It is known that the complexity of optimization problems in a nonconvex and global optimization framework scales exponentially with the number of decision variables. Thus, we look to formulate the problem in terms of as few as possible decision variables to enable tractable solutions. This is achieved by a parsimonious input parametrization [21], [22], which can be especially useful for OED problems since they are typically high-dimensional in the design variables. Then, a generic polynomial mapping of the design variables to the expected utility is established. Based on this mapping, a reformulation of the OED problem is performed, leading to a convex problem via the concept of sum-of-squares polynomials and semidefinite relaxations for which the solution can be attained with global optimality certificates [23]. The proposed approach is demonstrated on a benchmark Lotka-Volterra problem.

II. PROBLEM STATEMENT

Consider the continuous-time dynamical system given by

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}(t;\boldsymbol{\theta}) = \mathbf{f}\big(\mathbf{x}(t;\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{d}(t)\big), \quad \mathbf{x}(t_0;\boldsymbol{\theta}) = \mathbf{x}_0(\boldsymbol{\theta}), \quad (1)$$

where $\mathbf{x}(t;\boldsymbol{\theta})$ is the n_x-dimensional vector of states that depend on the n_θ-dimensional vector of uncertain parameters $\boldsymbol{\theta} \in \Theta$ and the n_d-dimensional vector of inputs $\mathbf{d}(t) \in \mathscr{D}$,

This work was supported by the Swiss National Science Foundation, project number 184521. This material is in part based upon work supported by the National Aeronautics and Space Administration (NASA) under grant number NNX17AJ31G. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NASA.

Diogo Rodrigues, Georgios Makrygiorgos, and Ali Mesbah are with the Department of Chemical and Biomolecular Engineering, University of California, Berkeley, CA 94720, USA.
{d.rodrigues,gmakr,mesbah}@berkeley.edu

and $\mathbf{f}(\mathbf{x},\boldsymbol{\theta},\mathbf{d})$ is an n_x-dimensional smooth vector function. The input set $\mathscr{D}$ may restrict $\mathbf{d}(t)$ to lie between a lower bound $\underline{\mathbf{d}}$ and an upper bound $\overline{\mathbf{d}}$. Noisy measurements $\mathbf{y} := (y(t_1),\ldots,y(t_T)) \in \mathscr{Y}$ are collected at T instants $t_1,\ldots,t_T$ as

$$y(t_k) = c\big(\mathbf{x}(t_k;\boldsymbol{\theta})\big) + e(t_k), \quad k = 1,\ldots,T, \tag{2}$$

where $\mathbf{e} := (e(t_1),\ldots,e(t_T))$ is additive measurement noise.

We aim to optimally design the inputs $\mathbf{d}(t)$ by maximizing the information content of the observations $\mathbf{y}$ for estimation of the unknown parameters $\boldsymbol{\theta}$. To this end, we adopt a Bayesian perspective. Under a given design $\mathbf{d}$ and a realization of the observations $\mathbf{y}$, the change in the information about $\boldsymbol{\theta}$ between a prior probability density function (pdf) $p(\boldsymbol{\theta})$ and a posterior pdf $p(\boldsymbol{\theta}|\mathbf{y},\mathbf{d})$ is given by Bayes' rule

$$p(\boldsymbol{\theta}|\mathbf{y},\mathbf{d}) = \tfrac{p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})p(\boldsymbol{\theta})}{p(\mathbf{y}|\mathbf{d})}, \tag{3}$$

where $p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})$ denotes a likelihood function, which results in the evidence $p(\mathbf{y}|\mathbf{d}) := \int_\Theta p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$.

In Bayesian OED, the optimal inputs $\mathbf{d}^* \in \mathscr{D}$ are designed by maximizing a so-called expected utility [11]

$$u(\mathbf{d}) := \int_\Theta U(\boldsymbol{\theta},\mathbf{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{4}$$

with the utility function defined as

$$U(\boldsymbol{\theta},\mathbf{d}) := \int_{\mathscr{Y}} G(\boldsymbol{\theta},\mathbf{y},\mathbf{d})\, p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})\mathrm{d}\mathbf{y}, \tag{5}$$

where $G(\boldsymbol{\theta},\mathbf{y},\mathbf{d})$ denotes a gain function that expresses the gain in reduction of uncertainty of the parameters $\boldsymbol{\theta}$ based on the observations $\mathbf{y}$ under the design $\mathbf{d}$ [10]. Since the goal is to design $\mathbf{d}$ so as to maximize the mutual information between $\boldsymbol{\theta}$ and $\mathbf{y}$, we define the gain function as

$$G_{KL}(\boldsymbol{\theta},\mathbf{y},\mathbf{d}) = \log\left(\tfrac{p(\boldsymbol{\theta}|\mathbf{y},\mathbf{d})}{p(\boldsymbol{\theta})}\right) = \log\left(\tfrac{p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})}{p(\mathbf{y}|\mathbf{d})}\right), \tag{6}$$

which implies that $U(\boldsymbol{\theta},\mathbf{d})$ becomes the KL divergence from the evidence to the likelihood function

$$U_{KL}(\boldsymbol{\theta},\mathbf{d}) = \int_{\mathscr{Y}} \log\left(\tfrac{p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})}{p(\mathbf{y}|\mathbf{d})}\right) p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})\mathrm{d}\mathbf{y}. \tag{7}$$

Accordingly, we formulate the Bayesian OED problem as

$$\mathbf{d}^*_{KL} := \arg\max_{\mathbf{d}\in\mathscr{D}} u_{KL}(\mathbf{d}) = \int_\Theta U_{KL}(\boldsymbol{\theta},\mathbf{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{8}$$

Remark 1: One can also show that the design $\mathbf{d}^*_{KL}$ maximizes the expected utility in terms of the KL divergence from the prior to the posterior distributions as well as the expected gain in Shannon information between the distributions.

As noted in Remark 1, the Bayesian OED problem (8) designs the inputs according to a relevant goal with respect to information content. However, the OED problem (8) is computationally intractable, as discussed in the next section. The goal of this paper is to approximate (8) as an optimization problem that can be efficiently solved to global optimality.

III. APPROXIMATION OF BAYESIAN OED

A main challenge in Bayesian OED is its high computational cost relative to classical OED approaches. This arises from the numerical evaluation of the expected utility in (8). In general, $u_{KL}(\mathbf{d})$ must be approximated using nested Monte Carlo integration over the joint observation and parameter space, which can become prohibitively expensive [19], [24].

To address this computational challenge, we approximate the Bayesian OED problem (8). This approximation leads to a tractable design criterion that involves the prior expectation of a function of the FIM [25]. To this end, the following assumptions related to normality of the likelihood function and prior pdf are required.

Assumption 1: The noise realizations $e(t_1),\ldots,e(t_T)$ are independent and identically distributed (i.i.d.) and drawn from a normal distribution with zero mean and variance σ^2. Let $h_k(\boldsymbol{\theta},\mathbf{d}) := c\big(\mathbf{x}(t_k;\boldsymbol{\theta})\big)$ for $k = 1,\ldots,T$. Since $\mathbf{y} = \mathbf{h}(\boldsymbol{\theta},\mathbf{d}) + \mathbf{e}$, the likelihood function in (3) takes the form

$$p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d}) = f(\mathbf{y}|\mathbf{h}(\boldsymbol{\theta},\mathbf{d}),\sigma^2\mathbf{I}_T), \tag{9}$$

where $f(\mathbf{x}|\bar{\mathbf{x}},\boldsymbol{\Sigma}_\mathbf{x})$ is the pdf of a multivariate normal distribution with mean $\bar{\mathbf{x}}$ and covariance $\boldsymbol{\Sigma}_\mathbf{x}$.

Assumption 2: The prior distribution of the parameters $\boldsymbol{\theta}$ follows a normal distribution with pdf

$$p(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}},\boldsymbol{\Sigma}_{\boldsymbol{\theta}}), \tag{10}$$

for some mean vector $\bar{\boldsymbol{\theta}}$ and some covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$.

Under Assumptions 1 and 2, $\mathbf{d}^*_{KL}$ can be approximated as the design that maximizes the scalar metric of the FIM for Bayes D-optimality [10]

$$\mathbf{d}^*_D := \arg\max_{\mathbf{d}\in\mathscr{D}} u_D(\mathbf{d}) = \int_\Theta U_D(\boldsymbol{\theta},\mathbf{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{11}$$

which corresponds to the utility function

$$U_D(\boldsymbol{\theta},\mathbf{d}) = \log\left(\det\left(\boldsymbol{\mathscr{I}}(\boldsymbol{\theta},\mathbf{d}) + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)\right), \tag{12}$$

where $\boldsymbol{\mathscr{I}}(\boldsymbol{\theta},\mathbf{d})$ is the FIM defined as

$$\begin{aligned}\boldsymbol{\mathscr{I}}(\boldsymbol{\theta},\mathbf{d}) &= \int_{\mathscr{Y}} \tfrac{\partial \log p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})}{\partial\boldsymbol{\theta}}^{\mathrm{T}} \tfrac{\partial \log p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})}{\partial\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})\mathrm{d}\mathbf{y} \\ &= \tfrac{\partial\mathbf{h}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})^{\mathrm{T}} \left(\sigma^2\mathbf{I}_T\right)^{-1} \tfrac{\partial\mathbf{h}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d}),\end{aligned} \tag{13}$$

and the FIM depends on the sensitivities described by

$$\begin{aligned}&\tfrac{\mathrm{d}}{\mathrm{d}t}\left(\tfrac{\partial\mathbf{x}}{\partial\boldsymbol{\theta}}(t;\boldsymbol{\theta})\right) = \tfrac{\partial\mathbf{f}}{\partial\mathbf{x}}\big(\mathbf{x}(t;\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{d}(t)\big)\tfrac{\partial\mathbf{x}}{\partial\boldsymbol{\theta}}(t;\boldsymbol{\theta}) \\ &\quad + \tfrac{\partial\mathbf{f}}{\partial\boldsymbol{\theta}}\big(\mathbf{x}(t;\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{d}(t)\big), \qquad \tfrac{\partial\mathbf{x}}{\partial\boldsymbol{\theta}}(t_0;\boldsymbol{\theta}) = \tfrac{\partial\mathbf{x}_0}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}),\end{aligned} \tag{14}$$

since

$$\tfrac{\partial h_k}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d}) = \tfrac{\partial c}{\partial\mathbf{x}}\big(\mathbf{x}(t_k;\boldsymbol{\theta})\big)\tfrac{\partial\mathbf{x}}{\partial\boldsymbol{\theta}}(t_k;\boldsymbol{\theta}), \quad k = 1,\ldots,T. \tag{15}$$

Then, the augmented dynamics of the system states and their sensitivities are described by

$$\tfrac{\mathrm{d}\mathbf{X}}{\mathrm{d}t}(t;\boldsymbol{\theta}) = \mathbf{F}\big(\mathbf{X}(t;\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{d}(t)\big), \quad \mathbf{X}(t_0;\boldsymbol{\theta}) = \mathbf{X}_0(\boldsymbol{\theta}), \tag{16}$$

with the $n_x(n_\theta+1)$ augmented states and initial conditions

$$\begin{aligned}\mathbf{X}(t;\boldsymbol{\theta}) &:= \left[\mathbf{x}(t;\boldsymbol{\theta}) \quad \tfrac{\partial\mathbf{x}}{\partial\boldsymbol{\theta}}(t;\boldsymbol{\theta})\right], \\ \mathbf{X}_0(\boldsymbol{\theta}) &:= \left[\mathbf{x}_0(\boldsymbol{\theta}) \quad \tfrac{\partial\mathbf{x}_0}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta})\right].\end{aligned} \tag{17}$$

Remark 2: Bayesian OED problems related to different Bayes alphabetic optimality criteria could be addressed by replacing $U_D(\boldsymbol{\theta},\mathbf{d})$ by other functions of the FIM $\mathscr{I}(\boldsymbol{\theta},\mathbf{d})$.

In the remainder, we aim to determine the design that maximizes the approximate expected utility $u_D(\mathbf{d})$ in (11). A computational challenge that arises from (11) is the multivariate integration over Θ, which is addressed next.

IV. TRACTABLE FORMULATION OF THE APPROXIMATE BAYESIAN OED PROBLEM

The aim of this section is to convert Problem (11) to a tractable formulation. To this end, it is first necessary to compute the expected utility $u_D(\mathbf{d})$ via multivariate integration. Then, we formulate a tractable optimal control problem.

A. Multivariate integration for computing the expected utility

For brevity, we use the shorthand notation $u(\mathbf{d}) := u_D(\mathbf{d})$ and $U(\boldsymbol{\theta},\mathbf{d}) := U_D(\boldsymbol{\theta},\mathbf{d})$. To compute $u(\mathbf{d})$ in (11) for a given $\mathbf{d}$, one needs to compute an integral of $U(\boldsymbol{\theta},\mathbf{d})$ in (12) over Θ by sampling according to the pdf $p(\boldsymbol{\theta})$. However, this integration typically requires computing $U(\boldsymbol{\theta},\mathbf{d})$ for a very large number of samples $\boldsymbol{\theta}$ to achieve accurate uncertainty propagation, which is computationally prohibitive when this procedure is repeated for different values of $\mathbf{d}$ [19].

Thus, we compute $u(\mathbf{d})$ by selecting m_θ quadrature points $\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_{m_\theta}$, which allows expressing $u(\mathbf{d})$ approximately as

$$\hat{u}(\mathbf{d}) = \mathbf{w}^{\mathrm{T}}\mathbf{p}_U(\mathbf{d}), \tag{18}$$

with the vector $\mathbf{w}$ of m_θ weight factors and

$$(\mathbf{p}_U(\mathbf{d}))_l = U(\boldsymbol{\theta}_l,\mathbf{d}), \quad l = 1,\dots,m_\theta. \tag{19}$$

We seek to construct an integration rule for the multivariate integral (11) based on as few quadrature points as possible. It is known that, even in the univariate case, methods based on Gaussian quadrature minimize the number of points needed for exact integration of polynomials of a given degree [26]. Here, we use an efficient approach that corresponds to sparse stochastic collocation and is the multivariate equivalent of Gaussian quadrature.

One can express $U(\boldsymbol{\theta},\mathbf{d})$ as

$$\begin{aligned} U(\boldsymbol{\theta},\mathbf{d}) &= \textstyle\sum_{\mathbf{k}\in\bar{\mathscr{K}}_n^{n_\theta}} (\mathbf{c}_U(\mathbf{d}))_{\mathbf{k}}\,\Psi\big(\Delta\boldsymbol{\theta}^{\mathbf{k}}\big) + R_U(\boldsymbol{\theta},\mathbf{d}) \\ &= \mathbf{a}_\theta(\boldsymbol{\theta})^{\mathrm{T}}\mathbf{c}_U(\mathbf{d}) + R_U(\boldsymbol{\theta},\mathbf{d}), \end{aligned} \tag{20}$$

where $\mathbf{c}_U(\mathbf{d})$ is the vector of polynomial coefficients of $U(\boldsymbol{\theta},\mathbf{d})$, $\Psi\big(\Delta\boldsymbol{\theta}^{\mathbf{k}}\big)$ denotes the first of the orthogonal polynomials with respect to $p(\boldsymbol{\theta})$ that contains the monomial $\Delta\boldsymbol{\theta}^{\mathbf{k}}$, which are Hermite polynomials for a normal prior pdf, with $\mathbf{k}$ the vector of monomial powers in the set $\bar{\mathscr{K}}_n^{n_\theta} \subseteq \mathscr{K}_n^{n_\theta} := \big\{(k_1,\dots,k_{n_\theta}) \in \mathbb{N}_0^{n_\theta} : 0 \le k_1+\dots+k_{n_\theta} \le n\big\}$ in the case of a polynomial of degree n, $\Delta\boldsymbol{\theta} := \boldsymbol{\theta} - \bar{\boldsymbol{\theta}}$ the deviation of $\boldsymbol{\theta}$ around $\bar{\boldsymbol{\theta}}$, $\Delta\boldsymbol{\theta}^{\mathbf{k}} := \big(\theta_1-\bar{\theta}_1\big)^{k_1}\dots\big(\theta_{n_\theta}-\bar{\theta}_{n_\theta}\big)^{k_{n_\theta}}$, $\mathbf{a}_\theta(\boldsymbol{\theta})$ is a vector with elements $(\mathbf{a}_\theta(\boldsymbol{\theta}))_{\mathbf{k}} = \Psi\big(\Delta\boldsymbol{\theta}^{\mathbf{k}}\big)$, for all $\mathbf{k}\in\bar{\mathscr{K}}_n^{n_\theta}$, and $R_U(\boldsymbol{\theta},\mathbf{d})$ is the orthogonal part with respect to $\mathbf{a}_\theta(\boldsymbol{\theta})$.

We assume that $\bar{\mathscr{K}}_n^{n_\theta}$ is a subset of $\mathscr{K}_n^{n_\theta}$ given by a maximum interaction or hyperbolic truncation scheme to introduce sparsity since this reduces the number of points needed for the integration rule when the dimension n_θ is large [27]. For example, in the case of a maximum interaction scheme with up to p_θ interaction terms, $\bar{\mathscr{K}}_n^{n_\theta} = \mathscr{K}_n^{n_\theta} \cap \big\{(k_1,\dots,k_{n_\theta}) \in \mathbb{N}_0^{n_\theta} : \lim_{q\to 0}\sum_{i=1}^{n_\theta} k_i^q \le p_\theta\big\}$. Then, since the polynomials in $\mathbf{a}_\theta(\boldsymbol{\theta})$ are orthogonal with respect to $p(\boldsymbol{\theta})$, it holds that $\mathbf{I}_{|\bar{\mathscr{K}}_n^{n_\theta}|} = \int_\Theta \mathbf{a}_\theta(\boldsymbol{\theta})\mathbf{a}_\theta(\boldsymbol{\theta})^{\mathrm{T}} p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$, where $|\bar{\mathscr{K}}_n^{n_\theta}| \le |\mathscr{K}_n^{n_\theta}| = \binom{n_\theta+n}{n_\theta}$. This implies

$$\begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}}_{|\bar{\mathscr{K}}_n^{n_\theta}|-1} \end{bmatrix} = \int_\Theta \mathbf{a}_\theta(\boldsymbol{\theta})^{\mathrm{T}} p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{21}$$

For some m_θ, one can choose a diagonal matrix $\mathbf{W}$ of dimension m_θ and points $\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_{m_\theta}$ such that

$$\mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}\mathbf{A}_\theta = \int_\Theta \mathbf{a}_\theta(\boldsymbol{\theta})^{\mathrm{T}} p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{22}$$

with $(\mathbf{A}_\theta)_{l,\mathbf{k}} = (\mathbf{a}_\theta(\boldsymbol{\theta}_l))_{\mathbf{k}}$ for $l = 1,\dots,m_\theta$ and $\mathbf{k}\in\bar{\mathscr{K}}_n^{n_\theta}$. Suppose that $(n_\theta+1)m_\theta \ge |\bar{\mathscr{K}}_n^{n_\theta}|$ and $\mathbf{W}$ and $\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_{m_\theta}$ are chosen such that they satisfy (22). Then, since

$$\begin{aligned} u(\mathbf{d}) &= \int_\Theta \mathbf{a}_\theta(\boldsymbol{\theta})^{\mathrm{T}}\mathbf{c}_U(\mathbf{d}) p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} + \int_\Theta R_U(\boldsymbol{\theta},\mathbf{d}) p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ &= \mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}\mathbf{p}_U(\mathbf{d}) - \mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}\big(\mathbf{p}_U(\mathbf{d}) - \mathbf{A}_\theta\mathbf{c}_U(\mathbf{d})\big) \\ &\quad + \int_\Theta R_U(\boldsymbol{\theta},\mathbf{d}) p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \end{aligned} \tag{23}$$

the integral $u(\mathbf{d})$ can be approximated as $\hat{u}(\mathbf{d})$ in (18) with $\mathbf{w}^{\mathrm{T}} = \mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}$ and the approximation error $\hat{u}(\mathbf{d}) - u(\mathbf{d})$ vanishes when $R_U(\boldsymbol{\theta},\mathbf{d}) = 0$.

A method based on polynomial chaos expansions could also be used [12], [19]. However, we propose the use of the approach based on Gaussian quadrature since it needs fewer quadrature points and does not require any regression.

B. Reformulation of OED as an optimal control problem

The approximate expected utility $\hat{u}(\mathbf{d})$ in (18) is an explicit function of the states $\mathbf{X}(t;\boldsymbol{\theta}_1),\dots,\mathbf{X}(t;\boldsymbol{\theta}_{m_\theta})$ from (12), (13), (15), (19). Thus, this approximation involves the dynamics

$$\mathbf{r}\big(\mathbf{s}(t),\mathbf{d}(t)\big) := \mathrm{vec}\begin{bmatrix} \mathbf{F}(\mathbf{X}(t;\boldsymbol{\theta}_1),\boldsymbol{\theta}_1,\mathbf{d}(t)) \\ \vdots \\ \mathbf{F}\big(\mathbf{X}(t;\boldsymbol{\theta}_{m_\theta}),\boldsymbol{\theta}_{m_\theta},\mathbf{d}(t)\big) \end{bmatrix}, \tag{24}$$

for the $n_r := n_x(n_\theta+1)m_\theta$ states and initial conditions

$$\mathbf{s}(t) := \mathrm{vec}\begin{bmatrix} \mathbf{X}(t;\boldsymbol{\theta}_1) \\ \vdots \\ \mathbf{X}(t;\boldsymbol{\theta}_{m_\theta}) \end{bmatrix}, \quad \mathbf{s}_0 := \mathrm{vec}\begin{bmatrix} \mathbf{X}_0(\boldsymbol{\theta}_1) \\ \vdots \\ \mathbf{X}_0(\boldsymbol{\theta}_{m_\theta}) \end{bmatrix}. \tag{25}$$

Hence, we define

$$\phi\big(\mathbf{s}(t_1),\dots,\mathbf{s}(t_T)\big) := \mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}\mathbf{p}_U(\mathbf{d}). \tag{26}$$

Accordingly, the Bayesian OED problem (11) can be approximated by the optimal control problem (OCP)

$$\hat{\mathbf{d}}^* := \arg\max_{\mathbf{d}\in\mathscr{D}} \hat{u}(\mathbf{d}) = \phi\big(\mathbf{s}(t_1),\dots,\mathbf{s}(t_T)\big), \tag{27a}$$

$$\text{s.t. } \dot{\mathbf{s}}(t) = \mathbf{r}\big(\mathbf{s}(t),\mathbf{d}(t)\big), \quad \mathbf{s}(t_0) = \mathbf{s}_0, \tag{27b}$$

$$(12),(13),(15),(19). \tag{27c}$$

The inputs that represent the solution to the OCP (27) are composed of several arcs. For each input d_j, each arc can be of type 1) input constraint-seeking, such that it is determined by an equality $d_j = \underline{d}_j$ or $d_j = \overline{d}_j$ (types 1Lower or 1Upper, respectively), or 2) sensitivity-seeking, such that it is determined by an equality that stems from the

dynamics given by $\mathbf{r}\big(\mathbf{s}(t),\mathbf{d}(t)\big)$ [22], [28]. Hence, there is a finite number of arc types from which arc sequences can be formed. If we consider as plausible arc sequences only sequences with a number of arcs no larger than some upper bound $\bar{n}_a$ and without consecutive arcs of the same type, it follows that the number of plausible sequences is also finite.

We aim to show how Bayesian OED problems reformulated as (27) can be solved efficiently to global optimality. The proposed approach for global optimality relies on determining: (i) when and how the optimal switching between arcs takes place for a given plausible arc sequence; and (ii) which sequence provides the optimal solution. Once question (i) is addressed for every plausible sequence, for example via parallel computing, it is trivial to answer question (ii).

Parsimonious input parameterization is an effective approach for describing the optimal inputs using only a few decision variables, in contrast to infinite-dimensional variables in the original OCP [21], [22]. For a given plausible arc sequence composed of n_s+1 input constraint-seeking and sensitivity-seeking arcs, the inputs $\mathbf{d}$ are defined by the following decision variables: the switching times $\bar{t}_1,\ldots,\bar{t}_{n_s}$ to arcs of all types and the initial conditions of the sensitivity-seeking arcs. The final time $\bar{t}_{n_s+1}=t_f$ is not a decision variable in this paper. Then, addressing question (i) above consists in computing the optimal values of the decision variables for the given arc sequence. For this, we describe the cost of the OCP as an explicit polynomial function, since it converts the OCP into a set of polynomial optimization problems (POPs), one for each arc sequence, as shown next.

V. REFORMULATION OF THE OCP AS POLYNOMIAL OPTIMIZATION PROBLEMS

For a given arc sequence, we describe the input in the ith time interval $[\bar{t}_{i-1},\bar{t}_i)$, for $i=1,\ldots,n_s+1$, by defining the $n_{z,i}$ states and initial conditions for this interval as $\mathbf{z}_i(t)$ and $\mathbf{z}_{i,0}$. One can then combine all the states into vectors with a dimension $n_z := n_r+n_{z,1}+\ldots+n_{z,n_s+1}$

$$\mathbf{z}(t) := \begin{bmatrix} \mathbf{s}(t)^{\mathrm{T}} & \begin{bmatrix} \mathbf{z}_1(t) \\ \vdots \\ \mathbf{z}_{n_s+1}(t) \end{bmatrix}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}, \tag{28}$$

with corresponding initial conditions $\mathbf{z}_0$.

The arc type determines the dimension and meaning of the elements of $\mathbf{z}_i(t)$, $\mathbf{z}_{i,0}$ and their effect on the inputs $\mathbf{d}(t)$ given by the control law $\tilde{\mathbf{c}}\big(\mathbf{z}(t)\big)$ and on the dynamics of $\mathbf{z}_i(t)$ given by $\mathbf{q}_i\big(\mathbf{s}(t),\mathbf{z}_i(t)\big)$. For input constraint-seeking arcs, $\mathbf{z}_i(t)$, $\mathbf{z}_{i,0}$ are of dimension 0 and $\tilde{c}_j\big(\mathbf{z}(t)\big)=\underline{d}_j$ or $\tilde{c}_j\big(\mathbf{z}(t)\big)=\overline{d}_j$ for the jth input. For sensitivity-seeking arcs, if we assume for the sake of simplicity that the jth input is approximated by a linear function, then $\mathbf{z}_i(t)=\begin{bmatrix}\tilde{d}_{j,i}(t)\\ \tilde{p}_{j,i}(t)\end{bmatrix}$, $\mathbf{z}_{i,0}=\begin{bmatrix}d^0_{j,i}\\ p_{j,i}\end{bmatrix}$ are of dimension 2, where $d^0_{j,i}$, $p_{j,i}$ are the initial value and derivative of the input and $\tilde{d}_{j,i}(t)$ is its value at time t, which implies $\tilde{c}_j\big(\mathbf{z}(t)\big)=\tilde{d}_{j,i}(t)$, $\mathbf{q}_i\big(\mathbf{s}(t),\mathbf{z}_i(t)\big)=\begin{bmatrix}\tilde{p}_{j,i}(t)\\ 0\end{bmatrix}$. The set $\big\{i : i\text{th arc of } d_j \text{ is of type } 2\big\}$ is denoted as $\mathscr{S}_j$.

Then, upon eliminating input dependencies and rewriting the OCP (27) in terms of the extended states $\mathbf{z}$, one obtains $\tilde{\phi}\big(\mathbf{z}(t_1),\ldots,\mathbf{z}(t_T)\big) := \phi\big(\mathbf{s}(t_1),\ldots,\mathbf{s}(t_T)\big)$ and the dynamics

$$\tilde{\mathbf{f}}\big(\mathbf{z}(t)\big) := \begin{bmatrix} \mathbf{r}\big(\mathbf{s}(t),\tilde{\mathbf{c}}\big(\mathbf{z}(t)\big)\big)^{\mathrm{T}} & \begin{bmatrix} \mathbf{q}_1\big(\mathbf{s}(t),\mathbf{z}_1(t)\big) \\ \vdots \\ \mathbf{q}_{n_s+1}\big(\mathbf{s}(t),\mathbf{z}_{n_s+1}(t)\big) \end{bmatrix}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}. \tag{29}$$

Since the input parameters for the given arc sequence are $\boldsymbol{\tau} := (\bar{t}_1,\ldots,\bar{t}_{n_s},\mathbf{z}_{1,0},\ldots,\mathbf{z}_{n_s+1,0})$, the OCP (27) can be reformulated in terms of these new decision variables as

$$\boldsymbol{\tau}^* := \arg\max_{\boldsymbol{\tau}} \hat{\phi}(\boldsymbol{\tau}) := \tilde{\phi}\big(\mathbf{z}(t_1),\ldots,\mathbf{z}(t_T)\big), \tag{30a}$$

$$\text{s.t. } \bar{t}_{i-1}\le\bar{t}_i, \quad i=1,\ldots,n_s+1, \tag{30b}$$

$$\underline{d}_j \le d^0_{j,s} \le \overline{d}_j,$$

$$\underline{d}_j \le d^0_{j,s}+p_{j,s}(\bar{t}_s-\bar{t}_{s-1}) \le \overline{d}_j, \quad s\in\mathscr{S}_j, \tag{30c}$$

$$\dot{\mathbf{z}}(t)=\tilde{\mathbf{f}}\big(\mathbf{z}(t)\big), \quad \mathbf{z}(t_0)=\mathbf{z}_0, \tag{30d}$$

$$(12),(13),(15),(19), \tag{30e}$$

which is convenient for numerical optimization since there are only $N := n_s+n_{z,1}+\ldots+n_{z,n_s+1}$ decision variables.

We aim to reformulate the OCP for each arc sequence as a POP that is amenable to global optimization. This entails expressing the metric $\hat{\phi}(\boldsymbol{\tau})$ as a polynomial function [29], [30]. To this end, we compute $\hat{\phi}(\boldsymbol{\tau})$ and its first-order partial derivatives with respect to $\boldsymbol{\tau}$.

For this, it is essential to consider not only the extended states $\mathbf{z}(t)$ and the extended adjoint variables

$$\boldsymbol{\zeta}(t) := \begin{bmatrix} \boldsymbol{\lambda}(t)^{\mathrm{T}} & \begin{bmatrix} \boldsymbol{\zeta}_1(t) \\ \vdots \\ \boldsymbol{\zeta}_{n_s+1}(t) \end{bmatrix}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}, \tag{31}$$

but also the concept of modified Hamiltonian function $\tilde{H}\big(\mathbf{z}(t),\boldsymbol{\zeta}(t)\big)=\tilde{\mathbf{f}}\big(\mathbf{z}(t)\big)^{\mathrm{T}}\boldsymbol{\zeta}(t)$. As shown in (30), the extended states $\mathbf{z}(t)$ are described by the differential equations

$$\tfrac{\mathrm{d}\mathbf{z}}{\mathrm{d}t}(t)=\tfrac{\partial\tilde{H}}{\partial\boldsymbol{\zeta}}\big(\mathbf{z}(t),\boldsymbol{\zeta}(t)\big)^{\mathrm{T}}=\tilde{\mathbf{f}}\big(\mathbf{z}(t)\big), \quad \mathbf{z}(t_0)=\mathbf{z}_0. \tag{32}$$

Likewise, the extended adjoint variables $\boldsymbol{\zeta}(t)$ are described by the differential equations

$$\tfrac{\mathrm{d}\boldsymbol{\zeta}}{\mathrm{d}t}(t)=-\tfrac{\partial\tilde{H}}{\partial\mathbf{z}}\big(\mathbf{z}(t),\boldsymbol{\zeta}(t)\big)^{\mathrm{T}}=-\tfrac{\partial\tilde{\mathbf{f}}}{\partial\mathbf{z}}\big(\mathbf{z}(t)\big)^{\mathrm{T}}\boldsymbol{\zeta}(t), \quad \boldsymbol{\zeta}(t_T)=\mathbf{0}_{n_z},$$
$$\boldsymbol{\zeta}(t_k^-)=\boldsymbol{\zeta}(t_k)+\tfrac{\partial\tilde{\phi}}{\partial\mathbf{z}(t_k)}\big(\mathbf{z}(t_1),\ldots,\mathbf{z}(t_T)\big)^{\mathrm{T}}, \quad k=1,\ldots,T. \tag{33}$$

With these results, one can obtain the first-order partial derivatives of $\hat{\phi}(\boldsymbol{\tau})$ with respect to $\boldsymbol{\tau}$

$$\begin{aligned}\tfrac{\partial\hat{\phi}}{\partial\bar{t}_i}(\boldsymbol{\tau}) &= \tilde{H}\big(\mathbf{z}(\bar{t}_i^-),\boldsymbol{\zeta}(\bar{t}_i^-)\big)-\tilde{H}\big(\mathbf{z}(\bar{t}_i),\boldsymbol{\zeta}(\bar{t}_i)\big)\\ &= \big(\tilde{\mathbf{f}}(\mathbf{z}(\bar{t}_i^-))-\tilde{\mathbf{f}}\big(\mathbf{z}(\bar{t}_i)\big)\big)^{\mathrm{T}}\boldsymbol{\zeta}(\bar{t}_i), \quad i=1,\ldots,n_s, \end{aligned}\tag{34}$$

$$\tfrac{\partial\hat{\phi}}{\partial\mathbf{z}_{i,0}}(\boldsymbol{\tau})=\boldsymbol{\zeta}_i(t_0)^{\mathrm{T}}, \quad i=1,\ldots,n_s+1. \tag{35}$$

An efficient approach to approximating $\hat{\phi}(\boldsymbol{\tau})$ as a polynomial function consists in (i) computing the partial derivatives of $\hat{\phi}(\boldsymbol{\tau})$ up to first order with respect to $\boldsymbol{\tau}$ and (ii) using multivariate Hermite interpolation to obtain a polynomial of degree $n>1$ that matches the value $\hat{\phi}(\boldsymbol{\tau}_l)$ and the partial derivatives $\frac{\partial\hat{\phi}}{\partial\boldsymbol{\tau}}(\boldsymbol{\tau}_l)$ at the sample points $\boldsymbol{\tau}_l$, for $l=1,\ldots,m_\tau$ [31]. Note that this requires no more than computing the extended states $\mathbf{z}(t)$ and adjoint variables $\boldsymbol{\zeta}(t)$ for $\hat{\phi}(\boldsymbol{\tau})$ that

correspond to each point $\boldsymbol{\tau}_l$, which amounts to solving two systems of n_z differential equations for each $l = 1,\ldots,m_\tau$.

The vector of polynomial coefficients is of dimension $\binom{N+n}{N}$, while the number of value vectors of dimension m_τ is $N+1$. This means that the number m_τ of sample points must be at least $\frac{(N+n)!}{n!(N+1)!}$, which is polynomial in N since n is typically bounded to avoid an overfitting polynomial. In addition, recall that N is typically small owing to the parsimonious nature of the input parameterization.

Hence, when the metric $\hat{\phi}(\boldsymbol{\tau})$ is expressed as a polynomial $p_{\hat{\phi}}(\boldsymbol{\tau})$ in the variables $\boldsymbol{\tau}$ for a given arc sequence, the OCP for that arc sequence is reformulated as a POP. This problem is solved efficiently to global optimality via reformulation as a hierarchy of convex semidefinite programs (SDPs) of increasing relaxation order using the concept of sum-of-squares polynomials [23]. Although the method to solve such problems to global optimality is out of the scope of the paper, standard methods for this purpose are described in [29], [30].

VI. CASE STUDY

The proposed OED approach is demonstrated on a Lotka-Volterra (LV) system represented by a set of nonlinear differential equations that describes the interaction of predator and prey populations. The LV system is widely used as a benchmark problem for optimal control [32] and OED [33]. The nondimensional governing equations are given as

$$\dot{x}_1(t) = \quad x_1(t) - (1+\theta_1)x_1(t)x_2(t) - 0.4x_1(t)d(t), \quad (36a)$$
$$\dot{x}_2(t) = -x_2(t) + (1+\theta_2)x_1(t)x_2(t) - 0.2x_2(t)d(t), \quad (36b)$$

where $t \in [0,t_f]$ is the integration time span, with $t_f = 12$. The differential states x_1 and x_2 describe the population of the prey and the predator, respectively. The uncertain parameters, which must be inferred from experimental observations, are denoted by θ_1 and θ_2. The system (36) is integrated using the initial conditions $\mathbf{x}_0(\boldsymbol{\theta}) = [0.5, 0.7]$. We assume that we can experimentally measure the predator population at the final time step, i.e., $y(t_f) = x_2(t_f;\boldsymbol{\theta}) + e(t_f)$, where $e(t_f)$ is the measurement error. The variance of the error is assumed to be constant and equal to $\sigma^2 = 0.1^2$. The uncertain parameters follow a bivariate normal prior pdf $f(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}},\boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ with $\bar{\boldsymbol{\theta}} = \mathbf{0}_2$, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = 0.2^2\mathbf{I}_2$, and the designed input is allowed to attain values in a predefined interval, i.e., $d(t) \in [0,1], \forall t \in [0,t_f]$.

When linear functions are used to approximate sensitivity-seeking arcs, a locally optimal solution consists of 3 arcs: the first arc is sensitivity-seeking with $\underline{d} < d^*(t) < \overline{d}$, for which a linear function is used; in the second arc, $d^*(t) = \overline{d}$; and in the third arc, $d^*(t) = \underline{d}$. This results in an input trajectory described by the 4 decision variables $\bar{t}_1$, $\bar{t}_2$, $d^0_{1,1}$, $p_{1,1}$. The optimal switching times are $\bar{t}^*_1 = 5.334$, $\bar{t}^*_2 = 9.477$. The optimal initial conditions for the first arc are the initial value and the constant derivative of the linear function that describes $d^*(t)$ in this arc: $d^{0*}_{1,1} = 0.482$, $p^*_{1,1} = -0.090$. The optimal metric is $\hat{\phi}(\boldsymbol{\tau}^*) = 11.6706$. The local optimality is indicated by the fact that the gradients (34), (35) are equal to zero and the solution satisfies the necessary conditions given by Pontryagin's maximum principle [34].

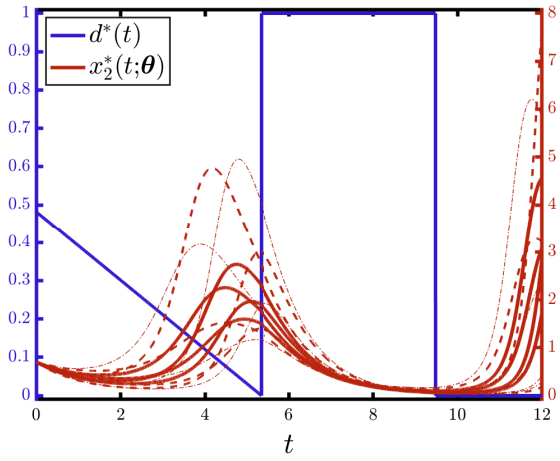

Fig. 1. Optimal input trajectory (in blue) for the Bayesian OED problem with the approximation of the sensitivity-seeking arc using a linear function. The trajectories of the measured variable (in red) are juxtaposed for the $m_\theta = 12$ realizations $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{m_\theta}$ used for multivariate integration. The relative width of the lines corresponds to the weights $\mathbf{w}$ of these realizations.

We use $m_\theta = 12$ quadrature points to compute $\hat{u}(\mathbf{d})$ via integration of $U(\boldsymbol{\theta},\mathbf{d})$. This corresponds to exact integration of Hermite polynomials up to degree 7 using the multivariate equivalent of Gaussian quadrature, with weights $\mathbf{w}$ and points $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{m_\theta}$. The input $d^*(t)$ and the measured variable $x_2^*(t;\boldsymbol{\theta})$ for the realizations $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{m_\theta}$ are shown in Fig. 1, which indicates that $x_2^*(t_f;\boldsymbol{\theta})$ is sensitive to variations of the parameters $\boldsymbol{\theta}$. The proposed approach for obtaining global solutions to Bayesian OED problems is applied by investigating all the 6 plausible arc sequences with a number of arcs no larger than $\bar{n}_a = 3$. Table I reports the execution time of the procedure on an Intel Core i5 1.8 GHz processor, the optimal metric $\hat{\phi}(\boldsymbol{\tau}^*)$, and the optimal values of the decision variables for these plausible arc sequences. The execution time includes the evaluation of $m_\tau = 2000$ sample points to obtain the polynomial representation $p_{\hat{\phi}}(\boldsymbol{\tau})$ of degree $n = 8$ and the local optimization of $\hat{\phi}(\boldsymbol{\tau})$ with initial guess $\boldsymbol{\tau}^*_p$ needed to compute $\boldsymbol{\tau}^*$ for each arc sequence. For all the arc sequences, it is possible to extract the unique solution $\boldsymbol{\tau}^*_p$ to the POP for $p_{\hat{\phi}}(\boldsymbol{\tau})$ from the solution to the SDP for the relaxation order 7 and certify the global optimality of $\boldsymbol{\tau}^*_p$. The duration of the formulation of the SDP and the extraction and certification of the global solution is much smaller than the execution time of the SDP solver MOSEK 8.1. For the design $\mathbf{d}^*_\tau$ that corresponds to $\boldsymbol{\tau}^*$ for each arc sequence such that $\hat{\phi}(\boldsymbol{\tau}^*) = \hat{u}(\mathbf{d}^*_\tau)$, accurate approximations of $u_D(\mathbf{d}^*_\tau)$ and $u_{KL}(\mathbf{d}^*_\tau)$ are also computed. One can observe that $\hat{u}(\mathbf{d}^*_\tau)$ overestimates $u_D(\mathbf{d}^*_\tau)$ and $u_{KL}(\mathbf{d}^*_\tau)$ consistently. Moreover, the execution time is below 1000 s for all arc sequences and the sequence with the best optimal metrics is 2-1Upper-1Lower, that is, the sequence of the locally optimal solution. In addition, the globally optimal values $\bar{t}^*_1$, $\bar{t}^*_2$, $d^{0*}_{1,1}$, $p^*_{1,1}$ of the decision variables for that arc sequence also correspond to the optimal values given by the locally optimal solution.

In summary, one can show that the locally optimal solution to the Bayesian OED problem shown in Fig. 1 is also the globally optimal solution with no more than $\bar{n}_a = 3$ arcs, and this only requires solving 6 problems in parallel in less than

TABLE I

EXECUTION TIME, OPTIMAL METRICS $\hat{\phi}(\boldsymbol{\tau}^*) = \hat{u}(\mathbf{d}_\tau^*)$, $u_D(\mathbf{d}_\tau^*)$, $2u_{KL}(\mathbf{d}_\tau^*) + \log\left(\det\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)\right)$, FINAL TIME t_f, AND OPTIMAL VALUES $\bar{t}_1^*$, $\bar{t}_2^*$, $d_{1,i}^{0*}$, $p_{1,i}^*$ OF THE DECISION VARIABLES FOR THE GLOBAL SOLUTION TO THE BAYESIAN OED PROBLEM FOR DIFFERENT PLAUSIBLE ARC SEQUENCES.

Arc sequence	Execution time (s)	$\hat{\phi}(\boldsymbol{\tau}^*) = \hat{u}(\mathbf{d}_\tau^*)$	$u_D(\mathbf{d}_\tau^*)$	$2u_{KL}(\mathbf{d}_\tau^*) + \log\left(\det\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)\right)$	$\bar{t}_1^*$	$\bar{t}_2^*$	t_f	$d_{1,i}^{0*}$	$p_{1,i}^*$
2-1Lower-1Upper	929	11.1308	11.0959	10.9337	2.320	12.000	12.000	1.000	0.000 ($i = 1$)
2-1Upper-1Lower	964	11.6706	11.6309	11.4050	5.334	9.477	12.000	0.482	-0.090 ($i = 1$)
1Lower-2-1Lower	966	11.5277	11.4599	11.1177	5.130	10.158	12.000	1.000	0.000 ($i = 2$)
1Lower-2-1Upper	864	10.6843	10.6181	10.4023	4.978	12.000	12.000	1.000	-0.129 ($i = 2$)
1Upper-2-1Lower	929	11.4730	11.4509	11.2201	1.794	9.260	12.000	0.000	0.134 ($i = 2$)
1Upper-2-1Upper	818	11.1308	11.0959	10.9337	2.320	12.000	12.000	0.000	0.000 ($i = 2$)

1000 s. Recall that, if we had only used local optimization to compute a local solution to (27), we could have obtained a local solution worse than $\boldsymbol{\tau}^*$ and it would not have been possible to provide any guarantee that the local solution is in any way close to the globally optimal solution.

VII. CONCLUSIONS AND FUTURE WORK

A methodology for obtaining globally optimal solutions to Bayesian OED problems for normally distributed likelihood and prior was presented. Numerical tractability was reinforced by an optimal stochastic collocation scheme that required only a few points in the parameter space for approximating the expected utility. Moreover, the execution time of the optimization procedure for each arc sequence indicates that a global solution can be obtained in a tractable way via a convex SDP.

In future work, we aim to circumvent the approximation of the expected utility for KL divergence as a Bayes D-optimality criterion and the assumption of normal distributions for the likelihood and the prior. Further extensions would also include design-dependent initial conditions and more complex path constraints.

REFERENCES

[1] A. C. Atkinson and A. N. Donev, *Optimum Experimental Designs*. Clarendon Press, 1992.

[2] E. Walter and L. Pronzato, "Qualitative and quantitative experiment design for phenomenological models - A survey," *Automatica*, vol. 26, no. 2, pp. 195 – 213, 1990.

[3] V. Fedorov, "Optimal experimental design," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 5, pp. 581–589, 2010.

[4] G. Franceschini and S. Macchietto, "Model-based design of experiments for parameter precision: State of the art," *Chem. Eng. Sci.*, vol. 63, no. 19, pp. 4846 – 4872, 2008.

[5] M. Martin-Casas and A. Mesbah, "Discrimination between competing model structures of biological systems in the presence of population heterogeneity," *IEEE Life Sci. Lett.*, vol. 2, no. 3, pp. 23–26, 2016.

[6] S. Streif, F. Petzke, A. Mesbah, R. Findeisen, and R. D. Braatz, "Optimal experimental design for probabilistic model discrimination using polynomial chaos," *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 4103–4109, 2014.

[7] A. Wald, "On the efficient design of statistical investigations," *Ann. Math. Statist.*, vol. 14, no. 2, pp. 134–140, 1943.

[8] S. D. Silvey and D. M. Titterington, "A geometric approach to optimal design theory," *Biometrika*, vol. 60, no. 1, pp. 21–32, 1973.

[9] G. Elfving, "Optimum allocation in linear regression theory," *Ann. Math. Statist.*, vol. 23, no. 2, pp. 255–262, 1952.

[10] K. Chaloner and I. Verdinelli, "Bayesian experimental design: A review," *Stat. Sci.*, vol. 10, no. 3, pp. 273–304, 1995.

[11] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 63, no. 3, pp. 425–464, 2001.

[12] X. Huan and Y. M. Marzouk, "Simulation-based optimal Bayesian experimental design for nonlinear systems," *J. Comput. Phys.*, vol. 232, no. 1, pp. 288 – 317, 2013.

[13] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[14] D. V. Lindley, "On a measure of the information provided by an experiment," *Ann. Math. Statist.*, vol. 27, no. 4, pp. 986–1005, 1956.

[15] G. E. P. Box and H. L. Lucas, "Design of experiments in non-linear situations," *Biometrika*, vol. 46, no. 1-2, pp. 77–90, 1959.

[16] E. G. Ryan, C. C. Drovandi, J. M. McGree, and A. N. Pettitt, "A review of modern computational algorithms for Bayesian optimal design," *Int. Stat. Rev.*, vol. 84, no. 1, pp. 128–154, 2016.

[17] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

[18] A. Shapiro, "Asymptotic analysis of stochastic programs," *Ann. Oper. Res.*, vol. 30, no. 1-4, pp. 169–186, 1991.

[19] J. A. Paulson, M. Martin-Casas, and A. Mesbah, "Optimal Bayesian experiment design for nonlinear dynamic systems with chance constraints," *J. Process Control*, vol. 77, pp. 155 – 171, 2019.

[20] G. Makrygiorgos, G. M. Maggioni, and A. Mesbah, "Surrogate modeling for fast uncertainty quantification: Application to 2D population balance models," *Comput. Chem. Eng.*, vol. 138, no. 106814, 2020.

[21] D. Rodrigues and D. Bonvin, "Dynamic optimization of reaction systems via exact parsimonious input parameterization," *Ind. Eng. Chem. Res.*, vol. 58, no. 26, pp. 11 199–11 212, 2019.

[22] ——, "On reducing the number of decision variables for dynamic optimization," *Optim. Control Appl. Meth.*, vol. 41, no. 1, pp. 292–311, 2020.

[23] J. B. Lasserre, *Moments, Positive Polynomials and Their Applications*. Imperial College Press, 2010.

[24] K. J. Ryan, "Estimating expected information gains for experimental designs with application to the random fatigue-limit model," *J. Comput. Graph. Stat.*, vol. 12, no. 3, pp. 585–603, 2003.

[25] S. Körkel, E. Kostina, H. G. Bock, and J. P. Schlöder, "Numerical methods for optimal control problems in design of robust optimal experiments for nonlinear dynamic processes," *Optim. Methods Soft.*, vol. 19, no. 3-4, pp. 327–338, 2004.

[26] M. Sinsbeck and W. Nowak, "An optimal sampling rule for nonintrusive polynomial chaos expansions of expensive models," *Int. J. Uncertain. Quantif.*, vol. 5, no. 3, pp. 275–295, 2015.

[27] G. Blatman and B. Sudret, "Adaptive sparse polynomial chaos expansion based on least angle regression," *J. Comput. Phys.*, vol. 230, no. 6, pp. 2345–2367, 2011.

[28] B. Srinivasan, S. Palanki, and D. Bonvin, "Dynamic optimization of batch processes: I. Characterization of the nominal solution," *Comput. Chem. Eng.*, vol. 27, no. 1, pp. 1–26, 2003.

[29] D. Henrion and J. B. Lasserre, "GloptiPoly: Global optimization over polynomials with Matlab and SeDuMi," *ACM Trans. Math. Softw.*, vol. 29, no. 2, pp. 165–194, 2003.

[30] H. Waki, S. Kim, M. Kojima, and M. Muramatsu, "Sums of squares and semidefinite program relaxations for polynomial optimization problems with structured sparsity," *SIAM J. Optim.*, vol. 17, no. 1, pp. 218–242, 2006.

[31] R. A. Lorentz, "Multivariate Hermite interpolation by algebraic polynomials: A survey," *J. Comput. Appl. Math.*, vol. 122, no. 1-2, pp. 167–201, 2000.

[32] S. Sager, H. G. Bock, M. Diehl, G. Reinelt, and J. P. Schlöder, "Numerical methods for optimal control with binary control functions applied to a Lotka-Volterra type fishing problem," in *Recent Advances in Optimization*, A. Seeger, Ed. Springer, 2006, pp. 269–289.

[33] D. Telen, F. Logist, E. Van Derlinden, I. Tack, and J. Van Impe, "Optimal experiment design for dynamic bioprocesses: A multi-objective approach," *Chem. Eng. Sci.*, vol. 78, pp. 82 – 97, 2012.

[34] R. F. Hartl, S. P. Sethi, and R. G. Vickson, "A survey of the maximum principles for optimal control problems with state constraints," *SIAM Rev.*, vol. 37, no. 2, pp. 181–218, 1995.