Title

Tractable global solutions to chance-constrained Bayesian optimal experiment design for arbitrary prior and noise distributions

Journal

Journal of Process Control, 116(1)

ISSN

0959-1524

Authors

Rodrigues, Diogo
Makrygiorgos, Georgios
Mesbah, Ali

Publication Date

2022-08-01

DOI

10.1016/j.jprocont.2022.05.008

Peer reviewed

Journal of Process Control 116 (2022) 1–18

Contents lists available at ScienceDirect

Journal of Process Control

journal homepage: www.elsevier.com/locate/jprocont

Tractable global solutions to chance-constrained Bayesian optimal experiment design for arbitrary prior and noise distributions

Diogo Rodrigues [a,b,*], Georgios Makrygiorgos [a], Ali Mesbah [a]

[a] Department of Chemical and Biomolecular Engineering, University of California, Berkeley, CA 94720, USA
[b] Centro de Química Estrutural, Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisboa, Portugal

ARTICLE INFO

Article history:
Received 27 September 2021
Received in revised form 17 May 2022
Accepted 18 May 2022
Available online 6 June 2022

Keywords:
Global optimization
Bayesian experiment design
Sum-of-squares polynomials
Stochastic collocation
Information theory

ABSTRACT

Optimal experiment design (OED) aims to optimize the information content of experimental observations 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 solution methods for two approximate formulations of the Bayesian OED problem based on Kullback–Leibler divergence for the particular case of Gaussian prior and observation noise distributions and the general case of arbitrary prior distributions and arbitrary observation noise distributions when the observation noise corresponds to arbitrary functions of the states and random variables with an arbitrary multivariate distribution. The proposed methods also allow satisfying path constraints with a specified probability. The solution approach relies on the reformulation of the approximate Bayesian OED problem as an optimal control problem (OCP), for which a parsimonious input parameterization is adopted to reduce the number of decision variables. An efficient global solution method for OCPs via sum-of-squares polynomials and parallel computing is then applied, which is based on approximating the cost of the OCP by a polynomial function of the decision variables and solving the resulting polynomial optimization problem to global optimality in a tractable way via semidefinite programming. It is established that the difference between the cost obtained by solving the polynomial optimization problem and the globally optimal cost of the OCP is bounded and depends on the polynomial approximation error.

1. Introduction

The optimal selection of conditions under which experiments are conducted is crucial for maximizing the information content 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 or designs by maximizing the information content of observations for parameter inference or model discrimination [1–6]. Other formulations include the design of experiments such that the experimental cost is minimized subject to bounds on the model uncertainty or other constraints related to application performance, typically in the context of use of the model for control [7,8].

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 [9–11]. 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 [12,13]. Generally, Bayesian OED can be useful when the system observations are noisy, incomplete, and indirect since the use of prior knowledge allows alleviating the lack of informative observations [14].

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 [15,16]. As no closed-form expression exists for the expected utility for general nonlinear systems [17], a key computational challenge in Bayesian OED arises from numerical evaluation of the expected utility using Monte Carlo-based methods [18]. Due to this sample-based evaluation of the expected utility, Bayesian OED is naturally formulated as a stochastic optimization problem, which can become prohibitively expensive to solve for OED problems with large design spaces. Alternatively, gradient-based optimization approaches such as stochastic approximation [19] and sample average approximation [20] methods can be used to attain locally optimal designs. The gradient-based optimization methods 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 via nested Monte Carlo integration over

* Corresponding author at: Centro de Química Estrutural, Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisboa, Portugal.
E-mail address: dfmr@tecnico.ulisboa.pt (D. Rodrigues).

https://doi.org/10.1016/j.jprocont.2022.05.008

the joint observation and parameter space can be prohibitively expensive, even if importance sampling is used [21,22]. Other types of approximation such as Laplace approximation are less computationally demanding but can lead to bias [23]. These challenges have been addressed by constructing surrogates for the model outputs based on polynomial chaos expansions [13,24,25]. For the case of OED for model discrimination, surrogate models based on Gaussian processes have also been proposed [26]. Despite these advances, gradient-based methods cannot guarantee the global optimality of the selected designs in a general context. For the related problems of design of the optimal input spectrum in the particular case of linear dynamical systems and design of the optimal input among the particular class of realizations of a stationary process in nonlinear dynamical systems, convex problems have been formulated [27,28]. Furthermore, the handling of chance path constraints remains a relatively unexplored topic in OED, with few exceptions [25,29]. The previous remarks show that, although OED is a well-established technique that has been abundantly addressed in the literature, including textbooks [30], significant challenges remain for general nonlinear dynamical systems subject to probabilistic constraints.

This paper presents a tractable approach for obtaining globally optimal solutions to Bayesian OED for constrained nonlinear dynamical systems with probabilistic uncertainty in model parameters. We express the expected utility in terms of the KL divergence from the prior to the posterior parameter distributions. In our recent work, we studied the approximation of the OED problem as Bayes D-optimality of the FIM for the special case of Gaussian prior and observation noise distributions [31]. Here, we extend this work to the general case of arbitrary prior and observation noise distributions via approximation of the OED problem as Monte Carlo integration in the observation space. In addition, we propose a method to ensure that the selected design satisfies path constraints with a specified probability. To this end, we propose a moment-based reformulation of chance path constraints. A sample-based approach is then utilized to compute the expected utility and the moments for a given design. To this end, a sparse stochastic collocation scheme for numerical integration over the domain of uncertain parameters is used. The quadrature rule is built upon the notion of orthogonal polynomials, which has been extensively used in the approximation of functions of random variables [32]. This novel use of sparse stochastic collocation in the context of Bayesian OED represents an improvement with respect to the performance of previous sample-based methods in terms of number of quadrature points. Moreover, it is known that the complexity of optimization problems in a nonconvex and global optimization framework scales exponentially with the number of decision variables. This is particularly difficult in OED problems for dynamical systems since the corresponding designs include time-varying inputs, which result in infinite-dimensional decision variables. Thus, we look to formulate the problem in terms of as few as possible decision variables to enable tractable solutions. This goal is achieved by the reformulation of the OED problem as an optimal control problem (OCP) and the use of a parsimonious input parameterization, which has been shown to reduce the number of decision variables in OCPs without causing any loss of optimality [33,34]. This can be especially useful for OED problems related to dynamical systems since they typically result in a large number of design variables. Owing to this parameterization, a generic polynomial mapping of the design variables to the expected utility is established. Based on this mapping, the OCP in terms of few decision variables that results from Bayesian OED is approximated as a polynomial optimization problem. The approximation method and the establishment of a quantifiable bound for the error between the solutions to both problems is a main contribution of this paper. This approximation leads to a convex problem via the concept of sum-of-squares polynomials and semidefinite relaxations for which the solution can be attained in a tractable way with global optimality certificates [35]. This method for tractable computation of global solutions is another contribution of this work with respect to previous optimization methods for Bayesian OED. The proposed approach is demonstrated on a Lotka–Volterra problem, for which different scenarios are considered, including non-Gaussian prior distribution, state-dependent observation noise, and chance path constraints.

Notation. Matrices are denoted by uppercase and boldface Latin or Greek symbols. Vectors are denoted by lowercase and boldface Latin or Greek symbols. Scalars are denoted by italic Latin or lowercase Greek symbols. Sets are denoted by uppercase Greek or calligraphic symbols. The superscripts $(\cdot)^{-1}$ and $(\cdot)^{\mathrm{T}}$ denote the matrix inverse and the matrix transpose, $\det(\cdot)$ denotes the matrix determinant, $\mathrm{tr}(\cdot)$ denotes the matrix trace, the superscript $(\cdot)^*$ denotes an optimal solution, $\|\mathbf{v}\|_{\mathbf{M}} = \sqrt{\mathbf{v}^{\mathrm{T}}\mathbf{M}\mathbf{v}}$ and $\|\mathbf{v}\| = \sqrt{\mathbf{v}^{\mathrm{T}}\mathbf{v}}$ denote norms, v_i denotes the ith element of the vector $\mathbf{v}$, $M_{i,j}$ denotes the element (i, j) of the matrix $\mathbf{M}$, $\mathbf{I}_n$ denotes the identity matrix of dimension n, $\mathbf{0}_n$ and $\mathbf{1}_n$ denote the column vectors of zeros and ones of dimension n, $p(\cdot)$ denotes a probability density function, $\mathrm{P}[\cdot]$ denotes the probability of an event, $\mathrm{E}[\cdot]$ and $\mathrm{V}[\cdot]$ denote the expected value and variance of a random variable, and $\mathbb{N}_0$ denotes the set of natural numbers including zero.

2. Problem statement

Consider the continuous-time dynamical system given by

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

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_u-dimensional vector of manipulated inputs $\mathbf{u}(t) \in \mathcal{U}$, $\mathbf{f}(\mathbf{x}, \boldsymbol{\theta}, \mathbf{u})$ is an n_x-dimensional smooth vector function, and $\mathbf{x}_0(\boldsymbol{\theta}, \mathbf{b})$ are the initial states that depend on $\boldsymbol{\theta}$. The dependence of the system on the n_b manipulated parameters $\mathbf{b} \in \mathcal{B}$, which are constant in contrast to the time-varying inputs $\mathbf{u}(t)$, is established by making the initial conditions depend on $\mathbf{b}$. The input set $\mathcal{U}$ restricts $\mathbf{u}(t)$ to lie between a lower bound $\underline{\mathbf{u}}$ and an upper bound $\overline{\mathbf{u}}$. In addition, the system is subject to the path constraints $\mathbf{h}\big(\mathbf{x}(t;\boldsymbol{\theta})\big) \leq \mathbf{0}_{n_h}$, where $\mathbf{h}(\mathbf{x})$ is an n_h-dimensional smooth vector function with $\mathbf{h}^{(1)}(\mathbf{x}, \boldsymbol{\theta}, \mathbf{u}) := \frac{\partial \mathbf{h}}{\partial \mathbf{x}}(\mathbf{x})\mathbf{f}(\mathbf{x}, \boldsymbol{\theta}, \mathbf{u})$ that depends explicitly on $\mathbf{u}$. The collection of manipulated variables that comprises the inputs $\mathbf{u}(t)$ and the parameters $\mathbf{b}$ is denoted as $\mathbf{d} \in \mathcal{D} = \mathcal{U} \times \mathcal{B}$. Noisy measurements $\mathbf{y} := (y(t_1), \ldots, y(t_T)) \in \mathcal{Y}$ 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 and $c(\mathbf{x})$ is a smooth scalar function of the states.

We aim to optimally design $\mathbf{d}$ 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 [36]

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

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

$$p(\mathbf{y}|\mathbf{d}) = \int_{\Theta} p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{4}$$

In Bayesian OED, the optimal design $\mathbf{d}^* \in \mathcal{D}$ is chosen by maximizing a so-called expected utility [13]

$$u(\mathbf{d}) \coloneqq \int_{\Theta} U(\boldsymbol{\theta}, \mathbf{d}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \tag{5}$$

with the utility function defined as

$$U(\boldsymbol{\theta}, \mathbf{d}) \coloneqq \int_{\mathcal{Y}} G(\boldsymbol{\theta}, \mathbf{y}, \mathbf{d})\, p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d}) \mathrm{d}\mathbf{y}, \tag{6}$$

where $G(\boldsymbol{\theta}, \mathbf{y}, \mathbf{d})$ denotes a gain function that quantifies the gain in reduction of uncertainty of the parameters $\boldsymbol{\theta}$ based on the observations $\mathbf{y}$ under the design $\mathbf{d}$ [12]. The optimal design is also subject to chance path constraints that specify a probability of violation $0 < \beta_k < 1$ for each path constraint

$$q_k(t; \mathbf{d}) \geq 1 - \beta_k, \quad k = 1, \ldots, n_h, \tag{7}$$

with $q_k(t; \mathbf{d}) \coloneqq \mathrm{P}\left[h_k\left(\mathbf{x}(t; \boldsymbol{\theta})\right) \leq 0\right] = \mathrm{E}\left[H\left(-h_k\left(\mathbf{x}(t; \boldsymbol{\theta})\right)\right)\right]$ and the unit step $H(x)$. The last equality implies that

$$q_k(t; \mathbf{d}) = \int_{\Theta} Q_k(t; \boldsymbol{\theta}, \mathbf{d}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \quad k = 1, \ldots, n_h, \tag{8}$$

with $Q_k(t; \boldsymbol{\theta}, \mathbf{d}) \coloneqq H\left(-h_k\left(\mathbf{x}(t; \boldsymbol{\theta})\right)\right)$. Note that $q_k(t; \mathbf{d})$ is an integral of a nonsmooth function.

Since the goal is to design $\mathbf{d}$ so as to maximize the mutual information between $\boldsymbol{\theta}$ and $\mathbf{y}$ for estimation of the unknown parameters $\boldsymbol{\theta}$, we define the gain function as

$$G_{KL}(\boldsymbol{\theta}, \mathbf{y}, \mathbf{d}) = \log\left(\frac{p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d})}{p(\boldsymbol{\theta})}\right) = \log\left(\frac{p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d})}{p(\mathbf{y}|\mathbf{d})}\right), \tag{9}$$

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

$$U_{KL}(\boldsymbol{\theta}, \mathbf{d}) = \int_{\mathcal{Y}} \log\left(\frac{p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d})}{p(\mathbf{y}|\mathbf{d})}\right) p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d}) \mathrm{d}\mathbf{y}. \tag{10}$$

Accordingly, we formulate the Bayesian OED problem as

$$\mathbf{d}^*_{KL} \coloneqq \arg\max_{\mathbf{d} \in \mathcal{D}} u_{KL}(\mathbf{d}) = \int_{\Theta} U_{KL}(\boldsymbol{\theta}, \mathbf{d}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \ \text{s.t. (7)}. \tag{11}$$

Remark 1. 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 shown in Appendix A. Hence, a large KL divergence from the prior to the posterior distributions implies that the data $\mathbf{y}$ are more informative for parameter estimation.

Although the Bayesian OED problem (11) provides a relevant design with respect to information content, the main challenge in this problem is its high computational cost relative to classical OED approaches [18]. For the general nonlinear system (1), closed-form expressions do not exist for the expected utility $u_{KL}(\mathbf{d})$ [17], which generally makes the OED problem (11) computationally intractable in its original form. This computational challenge arises from the numerical evaluation of the expected utility in (11). 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 [21,22,25]. The chance path constraints (7) are also intractable in their original form. Due to the formulation of $q_k(t; \mathbf{d})$ as an integral of a nonsmooth function, its approximation is typically performed via Monte Carlo integration, which is computationally costly [37].

The goal of this paper is to present a solution method for the Bayesian OED problem (11) by approximating it as an optimization problem that can be efficiently solved to global optimality. To this end, we first approximate the expected utility $u_{KL}(\mathbf{d})$ and the chance path constraints (7) in (11) for the two cases of Gaussian prior and observation noise distributions and arbitrary prior and observation noise distributions. We then use multivariate integration based on Gaussian quadrature for efficient sample-based evaluation of the approximate versions of both the expected utility and the chance path constraints. Finally, the resulting constrained optimal control problem is reformulated as a convex problem via polynomial optimization, which enables the tractable computation of solutions with global optimality certificates.

3. Approximation of Bayesian OED

To address the computational challenge posed by the numerical evaluation of $u_{KL}(\mathbf{d})$, we approximate the expected utility in (11) in two cases. Stricter assumptions related to normality of the prior pdf and observation noise are required in one case, while only relatively mild assumptions about the prior pdf and observation noise are required in the other case, as described in the next subsections. Lastly, we approximate the chance path constraints (7) to circumvent the numerical evaluation of $q_k(t; \mathbf{d})$.

3.1. Expected utility for Gaussian prior pdf and noise

We first address the approximation of the expected utility $u_{KL}(\mathbf{d})$ in (11) in the case of Gaussian prior pdf and observation noise according to the following assumptions.

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 $\mathbf{g}(\boldsymbol{\theta}, \mathbf{d})$ be a T-dimensional vector with $g_k(\boldsymbol{\theta}, \mathbf{d}) \coloneqq c\left(\mathbf{x}(t_k; \boldsymbol{\theta})\right)$ for $k = 1, \ldots, T$. Since $\mathbf{y} = \mathbf{g}(\boldsymbol{\theta}, \mathbf{d}) + \mathbf{e}$, the likelihood function in (3) takes the form

$$p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d}) = f(\mathbf{y}|\mathbf{g}(\boldsymbol{\theta}, \mathbf{d}), \sigma^2 \mathbf{I}_T), \tag{12}$$

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 multivariate normal distribution with pdf

$$p(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}), \tag{13}$$

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

These assumptions lead to the following remark.

Remark 2. 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

$$\mathbf{d}^*_D \coloneqq \arg\max_{\mathbf{d} \in \mathcal{D}} u_D(\mathbf{d}) = \int_{\Theta} U_D(\boldsymbol{\theta}, \mathbf{d}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \ \text{s.t. (7)}, \tag{14}$$

which corresponds to the utility function

$$U_D(\boldsymbol{\theta}, \mathbf{d}) = \log\left(\det\left(\mathcal{I}(\boldsymbol{\theta}, \mathbf{d}) + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)\right), \tag{15}$$

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

$$\begin{aligned} \mathcal{I}(\boldsymbol{\theta}, \mathbf{d}) &= \int_{\mathcal{Y}} \frac{\partial \log p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d})}{\partial \boldsymbol{\theta}}^{\mathrm{T}} \frac{\partial \log p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d})}{\partial \boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d}) \mathrm{d}\mathbf{y} \\ &= \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{d})^{\mathrm{T}} \left(\sigma^2 \mathbf{I}_T\right)^{-1} \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{d}), \end{aligned} \tag{16}$$

where the last equality results from the fact that $\frac{\partial \log p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d})}{\partial \boldsymbol{\theta}} = (\mathbf{y} - \mathbf{g}(\boldsymbol{\theta}, \mathbf{d}))^{\mathrm{T}} \left(\sigma^2 \mathbf{I}_T\right)^{-1} \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{d})$ according to Assumption 1. The approximation error due to the approximation of $\mathbf{d}^*_{KL}$ by $\mathbf{d}^*_D$ depends on the nonlinearity of $\mathbf{g}(\boldsymbol{\theta}, \mathbf{d})$ with respect to $\boldsymbol{\theta}$ and vanishes for $\mathbf{g}(\boldsymbol{\theta}, \mathbf{d})$ linear in $\boldsymbol{\theta}$, as shown in [30]. More specifically, the approximation error will be small if the following conditions hold, as shown in Appendix B:

1. The first condition is $\| \left(\int_0^1 \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta} + t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \mathbf{d}) \mathrm{d}t - \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{d}) \right) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \| / \sigma \ll 1$ for all $\boldsymbol{\theta}$ such that $\| \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-\frac{1}{2}} (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) \|^2 < F^{-1}_{\chi^2_{n_\theta}}(\alpha)$ and for all $\hat{\boldsymbol{\theta}}$ such that $\| \int_0^1 \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta} + t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \mathbf{d}) \mathrm{d}t \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \|^2 / \sigma^2 < F^{-1}_{\chi^2_T}(\alpha)$, with $F^{-1}_{\chi^2_{n_\theta}}$ and $F^{-1}_{\chi^2_T}$ denoting the inverse cumulative distribution function of the chi-squared distribution with n_θ degrees of freedom and T degrees of freedom, respectively, and α denoting a confidence level.

2. The second condition is $\| \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \left(\mathcal{I}(\boldsymbol{\theta}, \mathbf{d}) + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \right)^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \left(\mathcal{I}(\bar{\boldsymbol{\theta}}, \mathbf{d}) + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \right)^{-1} \| \ll 1$ for all $\boldsymbol{\theta}$ such that $\| \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-\frac{1}{2}} (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) \|^2 < F_{\chi^2_{n_\theta}}^{-1}(\alpha)$.

Both conditions express a mild nonlinearity of $\mathbf{g}(\boldsymbol{\theta}, \mathbf{d})$ with respect to $\boldsymbol{\theta}$. In particular, the conditions are satisfied (with $\ll 1$ replaced by $= 0$) for $\mathbf{g}(\boldsymbol{\theta}, \mathbf{d})$ linear in $\boldsymbol{\theta}$, which again implies that the approximation error vanishes, as shown in Appendix B.

The FIM depends on the sensitivities described by

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}(t; \boldsymbol{\theta}) \right) = \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \left(\mathbf{x}(t; \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}(t) \right) \frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}(t; \boldsymbol{\theta}) + \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} \left(\mathbf{x}(t; \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}(t) \right), \qquad \frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}(t_0; \boldsymbol{\theta}) = \frac{\partial \mathbf{x}_0}{\partial \boldsymbol{\theta}} (\boldsymbol{\theta}, \mathbf{b}) , \tag{17}$$

since

$$\frac{\partial g_k}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{d}) = \frac{\partial c}{\partial \mathbf{x}} \left(\mathbf{x}(t_k; \boldsymbol{\theta}) \right) \frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}(t_k; \boldsymbol{\theta}), \quad k = 1, \ldots, T. \tag{18}$$

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

$$\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t}(t; \boldsymbol{\theta}) = \mathbf{F} \left(\mathbf{X}(t; \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}(t) \right), \quad \mathbf{X}(t_0; \boldsymbol{\theta}) = \mathbf{X}_0 (\boldsymbol{\theta}, \mathbf{b}) , \tag{19}$$

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

$$\mathbf{X}(t; \boldsymbol{\theta}) := \left[\mathbf{x}(t; \boldsymbol{\theta}) \quad \frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}(t; \boldsymbol{\theta}) \right],$$
$$\mathbf{X}_0 (\boldsymbol{\theta}, \mathbf{b}) := \left[\mathbf{x}_0 (\boldsymbol{\theta}, \mathbf{b}) \quad \frac{\partial \mathbf{x}_0}{\partial \boldsymbol{\theta}} (\boldsymbol{\theta}, \mathbf{b}) \right]. \tag{20}$$

Remark 3. Bayesian OED problems related to different Bayes alphabetic optimality criteria could be solved by replacing $U_D(\boldsymbol{\theta}, \mathbf{d})$ by other functions of the FIM [12].

Remark 4. Another possibility to compute $\frac{\partial g_k}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{d})$ for $k = 1, \ldots, T$ would be to compute adjoint sensitivities instead of the forward sensitivities $\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}(t_k; \boldsymbol{\theta})$, which would require solving one system of n_x differential equations and T systems of $n_x + n_\theta$ differential equations instead of one system of $n_x (n_\theta + 1)$ differential equations. However, as described later in the paper, adjoint variables are used in this paper to compute sensitivities of the expected utility $u_D(\mathbf{d})$ with respect to the design $\mathbf{d}$, which requires expressing $u_D(\mathbf{d})$ directly in terms of a set of (augmented) states. For this reason, the forward sensitivities $\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}(t_k; \boldsymbol{\theta})$ are considered for the computation of $u_D(\mathbf{d})$.

3.2. Expected utility for arbitrary prior pdf and noise

Now, we address the approximation of the expected utility $u_{KL}(\mathbf{d})$ in (11) in the case of arbitrary prior pdf and observation noise according to the following assumptions.

Assumption 3. The noise realizations $\mathbf{e}$ are given by

$$e(t_k) = w \left(\mathbf{x}(t_k; \boldsymbol{\theta}) \right) \xi_k, \quad k = 1, \ldots, T, \tag{21}$$

with $\boldsymbol{\xi} \in \mathcal{X}$ drawn from an arbitrary multivariate distribution with pdf $\pi(\boldsymbol{\xi} | \bar{\mathbf{x}}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}})$, where $\mathcal{X}$ is a sample space and $\bar{\mathbf{x}}$ and $\boldsymbol{\Sigma}_{\mathbf{x}}$ are the mean and covariance of $\boldsymbol{\xi}$. Let $\mathbf{g}(\boldsymbol{\theta}, \mathbf{d})$ be a T-dimensional vector and $\mathbf{J}(\boldsymbol{\theta}, \mathbf{d})$ be a T-dimensional diagonal matrix, with $g_k(\boldsymbol{\theta}, \mathbf{d}) := c \left(\mathbf{x}(t_k; \boldsymbol{\theta}) \right)$ and $J_{k,k}(\boldsymbol{\theta}, \mathbf{d}) := w \left(\mathbf{x}(t_k; \boldsymbol{\theta}) \right)$ for $k = 1, \ldots, T$. Since $\mathbf{y} = \mathbf{g}(\boldsymbol{\theta}, \mathbf{d}) + \mathbf{J}(\boldsymbol{\theta}, \mathbf{d}) \boldsymbol{\xi}$, the likelihood function in (3) takes the form

$$p(\mathbf{y} | \boldsymbol{\theta}, \mathbf{d}) = \frac{\pi \left(\mathbf{J}(\boldsymbol{\theta}, \mathbf{d})^{-1} (\mathbf{y} - \mathbf{g}(\boldsymbol{\theta}, \mathbf{d})) | \bar{\mathbf{x}}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}} \right)}{\det(\mathbf{J}(\boldsymbol{\theta}, \mathbf{d}))}, \tag{22}$$

which can be replaced in the expressions for $p(\mathbf{y} | \mathbf{d})$ in (4) and $U_{KL}(\boldsymbol{\theta}, \mathbf{d})$ in (10).

Assumption 4. The prior distribution of the parameters $\boldsymbol{\theta}$ follows a multivariate distribution with mean $\bar{\boldsymbol{\theta}}$.

These assumptions lead to the following remark.

Remark 5. Under Assumptions 3 and 4, $\mathbf{d}_{KL}^*$ can be approximated as the design

$$\mathbf{d}_{MC}^* := \arg \max_{\mathbf{d} \in \mathcal{D}} u_{MC}(\mathbf{d}) = \int_\Theta U_{MC}(\boldsymbol{\theta}, \mathbf{d}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \text{ s.t. (7)}, \tag{23}$$

which corresponds to the utility function

$$U_{MC}(\boldsymbol{\theta}, \mathbf{d}) = \frac{1}{m_\xi} \sum_{k=1}^{m_\xi} \log \left(\frac{\det(\mathbf{J}(\boldsymbol{\theta}, \mathbf{d}))^{-1} \pi(\boldsymbol{\xi}_k | \bar{\mathbf{x}}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}})}{p(\mathbf{g}(\boldsymbol{\theta}, \mathbf{d}) + \mathbf{J}(\boldsymbol{\theta}, \mathbf{d}) \boldsymbol{\xi}_k | \mathbf{d})} \right), \tag{24}$$

where the points $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{m_\xi}$ are obtained by sampling in $\mathcal{X}$ according to the pdf $\pi(\boldsymbol{\xi} | \bar{\mathbf{x}}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}})$ and independently of the prior pdf $p(\boldsymbol{\theta})$, and the evidence $p(\mathbf{g}(\boldsymbol{\theta}, \mathbf{d}) + \mathbf{J}(\boldsymbol{\theta}, \mathbf{d}) \boldsymbol{\xi}_k | \mathbf{d})$ is computed as in (4) from the likelihood function in (22), that is,

$$\frac{p(\mathbf{g}(\boldsymbol{\theta}, \mathbf{d}) + \mathbf{J}(\boldsymbol{\theta}, \mathbf{d}) \boldsymbol{\xi}_k | \mathbf{d})}{\det(\mathbf{J}(\boldsymbol{\theta}, \mathbf{d}))^{-1}} = \int_\Theta \frac{p(\mathbf{g}(\boldsymbol{\theta}, \mathbf{d}) + \mathbf{J}(\boldsymbol{\theta}, \mathbf{d}) \boldsymbol{\xi}_k | \hat{\boldsymbol{\theta}}, \mathbf{d})}{\det(\mathbf{J}(\boldsymbol{\theta}, \mathbf{d}))^{-1}} p(\hat{\boldsymbol{\theta}}) \mathrm{d}\hat{\boldsymbol{\theta}}, \tag{25}$$

with

$$\frac{p(\mathbf{g}(\boldsymbol{\theta}, \mathbf{d}) + \mathbf{J}(\boldsymbol{\theta}, \mathbf{d}) \boldsymbol{\xi}_k | \hat{\boldsymbol{\theta}}, \mathbf{d})}{\det(\mathbf{J}(\boldsymbol{\theta}, \mathbf{d}))^{-1}} = \frac{\pi \left(\mathbf{J}(\hat{\boldsymbol{\theta}}, \mathbf{d})^{-1} \left(\mathbf{g}(\boldsymbol{\theta}, \mathbf{d}) + \mathbf{J}(\boldsymbol{\theta}, \mathbf{d}) \boldsymbol{\xi}_k - \mathbf{g}(\hat{\boldsymbol{\theta}}, \mathbf{d}) \right) | \bar{\mathbf{x}}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}} \right)}{\det(\mathbf{J}(\boldsymbol{\theta}, \mathbf{d}))^{-1} \det(\mathbf{J}(\hat{\boldsymbol{\theta}}, \mathbf{d}))}, \tag{26}$$

where the only approximation is the Monte Carlo integration $\int_{\mathcal{X}} f(\boldsymbol{\xi}) \pi(\boldsymbol{\xi} | \bar{\mathbf{x}}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}}) \mathrm{d}\boldsymbol{\xi} \simeq \frac{1}{m_\xi} \sum_{k=1}^{m_\xi} f(\boldsymbol{\xi}_k)$ in the observation space for some function $f(\boldsymbol{\xi})$, as shown in Appendix C.

Remark 6. As described in Assumption 3, the concept of arbitrary observation noise in this paper amounts to considering observation noise $e(t_k)$ at each instant t_k that corresponds to the product of an arbitrary function $w \left(\mathbf{x}(t_k; \boldsymbol{\theta}) \right)$ of the states $\mathbf{x}(t_k; \boldsymbol{\theta})$ and an element ξ_k of a vector-valued random variable $\boldsymbol{\xi}$ drawn from an arbitrary multivariate distribution, which is a rather general case and encompasses a special but relevant case of state-dependent noise. The slightly more general case of observation noise $e(t_k)$ at each instant t_k that corresponds to an arbitrary function $W_k \left(\mathbf{x}(t_k; \boldsymbol{\theta}), \boldsymbol{\xi} \right)$ of both the states $\mathbf{x}(t_k; \boldsymbol{\theta})$ and the vector-valued random variable $\boldsymbol{\xi}$ drawn from an arbitrary multivariate distribution, where $W_k \left(\mathbf{x}(t_k; \boldsymbol{\theta}), \boldsymbol{\xi} \right)$ is an element of a vector-valued function $\mathbf{W} \left(\mathbf{x}(t_1; \boldsymbol{\theta}), \ldots, \mathbf{x}(t_T; \boldsymbol{\theta}), \boldsymbol{\xi} \right)$, could be handled in a similar way. If $\mathbf{W}$ is a bijective and continuously differentiable function of $\boldsymbol{\xi}$ with invertible Jacobian $\frac{\mathrm{d}\mathbf{W}}{\mathrm{d}\boldsymbol{\xi}} \left(\mathbf{x}(t_1; \boldsymbol{\theta}), \ldots, \mathbf{x}(t_T; \boldsymbol{\theta}), \boldsymbol{\xi} \right)$ and inverse function $\mathbf{W}^{-1} \left(\mathbf{x}(t_1; \boldsymbol{\theta}), \ldots, \mathbf{x}(t_T; \boldsymbol{\theta}), \mathbf{e} \right)$, then $\mathbf{J}(\boldsymbol{\theta}, \mathbf{d}) \boldsymbol{\xi}$, $\mathbf{J}(\boldsymbol{\theta}, \mathbf{d})^{-1} \mathbf{e}$, $\pi \left(\boldsymbol{\xi} | \bar{\mathbf{x}}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}} \right) \det(\mathbf{J}(\boldsymbol{\theta}, \mathbf{d}))^{-1}$ in the previous equations would be replaced by $\mathbf{W} \left(\mathbf{x}(t_1; \boldsymbol{\theta}), \ldots, \mathbf{x}(t_T; \boldsymbol{\theta}), \boldsymbol{\xi} \right)$, $\mathbf{W}^{-1} \left(\mathbf{x}(t_1; \boldsymbol{\theta}), \ldots, \mathbf{x}(t_T; \boldsymbol{\theta}), \mathbf{e} \right)$, $\pi \left(\boldsymbol{\xi} | \bar{\mathbf{x}}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}} \right) \det \left(\frac{\mathrm{d}\mathbf{W}}{\mathrm{d}\boldsymbol{\xi}} \left(\mathbf{x}(t_1; \boldsymbol{\theta}), \ldots, \mathbf{x}(t_T; \boldsymbol{\theta}), \boldsymbol{\xi} \right) \right)^{-1}$, respectively. However, this more general case is not covered in more detail in this paper for improved clarity.

3.3. Chance path constraints

The chance path constraints (7) can be approximated in terms of the first two moments of $h_k \left(\mathbf{x}(t; \boldsymbol{\theta}) \right)$, that is,

$$\mathrm{E} \left[h_k \left(\mathbf{x}(t; \boldsymbol{\theta}) \right) \right] + r_k \sqrt{\mathrm{V} \left[h_k \left(\mathbf{x}(t; \boldsymbol{\theta}) \right) \right]} \leq 0, \quad k = 1, \ldots, n_h, \tag{27}$$

with the back-off parameter r_k. Since $\mathrm{V} \left[h_k \left(\mathbf{x}(t; \boldsymbol{\theta}) \right) \right] = \mathrm{E} \left[h_k \left(\mathbf{x}(t; \boldsymbol{\theta}) \right)^2 \right] - \mathrm{E} \left[h_k \left(\mathbf{x}(t; \boldsymbol{\theta}) \right) \right]^2$, and assuming that $\mathrm{E} \left[h_k \left(\mathbf{x}(t; \boldsymbol{\theta}) \right) \right] < 0$, (27) is equivalent to

$$\frac{r_k^2}{1 + r_k^2} m_k^{\mathrm{II}}(t; \mathbf{d}) - \left(m_k^{\mathrm{I}}(t; \mathbf{d}) \right)^2 \leq 0, \quad k = 1, \ldots, n_h, \tag{28}$$

with $m_k^{\mathrm{I}}(t; \mathbf{d}) := \mathrm{E} \left[h_k \left(\mathbf{x}(t; \boldsymbol{\theta}) \right) \right]$, $m_k^{\mathrm{II}}(t; \mathbf{d}) := \mathrm{E} \left[h_k \left(\mathbf{x}(t; \boldsymbol{\theta}) \right)^2 \right]$, which implies that

$$m_k^{\mathrm{I}}(t; \mathbf{d}) = \int_\Theta M_k^{\mathrm{I}}(t; \boldsymbol{\theta}, \mathbf{d}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \quad k = 1, \ldots, n_h, \tag{29}$$

$$m_k^{\mathrm{II}}(t;\mathbf{d}) = \int_\Theta M_k^{\mathrm{II}}(t;\boldsymbol{\theta},\mathbf{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \quad k = 1,\dots,n_h, \tag{30}$$

with $M_k^{\mathrm{I}}(t;\boldsymbol{\theta},\mathbf{d}) := h_k\big(\mathbf{x}(t;\boldsymbol{\theta})\big)$, $M_k^{\mathrm{II}}(t;\boldsymbol{\theta},\mathbf{d}) := h_k\big(\mathbf{x}(t;\boldsymbol{\theta})\big)^2$ [38]. Note that the moments $m_k^{\mathrm{I}}(t;\mathbf{d})$ and $m_k^{\mathrm{II}}(t;\mathbf{d})$ are integrals of smooth functions, in contrast to $q_k(t;\mathbf{d})$.

One can ensure satisfaction of the chance constraints (7) for any distribution of $h_k\big(\mathbf{x}(t;\boldsymbol{\theta})\big)$, including a heavy-tail distribution, via the Cantelli–Chebyshev inequality by enforcing the moment constraints (28) with $r_k = \sqrt{\frac{1-\beta_k}{\beta_k}}$, where β_k is the probability of violation for each constraint in (7) [39]. However, this approximation of the chance constraints is generally known to be conservative [40]. On the other hand, assuming that $h_k\big(\mathbf{x}(t;\boldsymbol{\theta})\big)$ is normally distributed, a tighter approximation is given by $r_k = \phi^{-1}(1-\beta_k)$, where ϕ^{-1} denotes the inverse cumulative distribution function of the standard normal distribution. However, the assumption of normal distribution of $h_k\big(\mathbf{x}(t;\boldsymbol{\theta})\big)$ may be too strong.

Hence, an approach to determine the correct value of the back-off parameter r_k proceeds as follows [25,41]: (i) choose an initial guess for r_k; (ii) determine the design that maximizes the approximate expected utilities $u_D(\mathbf{d})$ or $u_{MC}(\mathbf{d})$ subject to the moment constraints (28); (iii) for that design, compute $h_k\big(\mathbf{x}(t;\boldsymbol{\theta}_1)\big),\dots,h_k\big(\mathbf{x}(t;\boldsymbol{\theta}_{m_\theta})\big)$, where the points $\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_{m_\theta}$ are obtained by sampling in Θ according to the pdf $p(\boldsymbol{\theta})$, and estimate the minimal H_k such that $\mathrm{P}\big[h_k\big(\mathbf{x}(t;\boldsymbol{\theta})\big) \le H_k\big] \ge 1-\beta_k$; (iv) compute $r_k = \min_t\Big(H_k - \mathrm{E}\big[h_k\big(\mathbf{x}(t;\boldsymbol{\theta})\big)\big]\Big)\Big(\sqrt{\mathrm{V}\big[h_k\big(\mathbf{x}(t;\boldsymbol{\theta})\big)\big]}\Big)^{-1}$, and return to step (ii) if r_k changes more than a specified tolerance. Note that, even if the computation of $h_k\big(\mathbf{x}(t;\boldsymbol{\theta})\big)$ in step (iii) is performed for a large number of samples $\boldsymbol{\theta}$, this does not lead to a major computational burden since it is done for a single design, in contrast to the multiple designs that must be evaluated in any optimization procedure.

In the remainder, we aim to determine the design that maximizes the approximate expected utilities $u_D(\mathbf{d})$ or $u_{MC}(\mathbf{d})$ subject to the moment constraints (28). A computational challenge that arises from (14) and (23) is the multivariate integration over Θ, which is addressed next.

4. Tractable formulation of the approximate Bayesian OED problem

This section presents tractable formulations of Problems (14) and (23). We first express the expected utilities $u_D(\mathbf{d})$ or $u_{MC}(\mathbf{d})$ and the constraint moments $m_k^{\mathrm{I}}(t;\mathbf{d})$ and $m_k^{\mathrm{II}}(t;\mathbf{d})$ in terms of multivariate integration in the space of parameters. Then, we compute $u_D(\mathbf{d})$ or $u_{MC}(\mathbf{d})$ and $m_k^{\mathrm{I}}(t;\mathbf{d})$, $m_k^{\mathrm{II}}(t;\mathbf{d})$ and formulate tractable optimal control problems (OCPs).

4.1. Multivariate integration in the space of parameters

To compute a function given by

$$v(\mathbf{y},\mathbf{d}) = \int_\Theta V(\boldsymbol{\theta},\mathbf{y},\mathbf{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{31}$$

an integral of $V(\boldsymbol{\theta},\mathbf{y},\mathbf{d})$ over Θ is computed by sampling according to the pdf $p(\boldsymbol{\theta})$. However, this integration typically requires computing $V(\boldsymbol{\theta},\mathbf{y},\mathbf{d})$ for a very large number of samples $\boldsymbol{\theta}$ to achieve accurate uncertainty propagation, which can become computationally prohibitive when this procedure is repeated for different values of $\mathbf{y}$ or $\mathbf{d}$ [25].

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

$$v(\mathbf{y},\mathbf{d}) \approx \hat{v}(\mathbf{y},\mathbf{d}) = \textstyle\sum_{l=1}^{m_\theta} w_l V(\boldsymbol{\theta}_l,\mathbf{y},\mathbf{d}) = \mathbf{w}^{\mathrm{T}}\mathbf{p}_V(\mathbf{y},\mathbf{d}), \tag{32}$$

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

$$(\mathbf{p}_V(\mathbf{y},\mathbf{d}))_l = V(\boldsymbol{\theta}_l,\mathbf{y},\mathbf{d}), \quad l = 1,\dots,m_\theta. \tag{33}$$

We seek to construct an integration rule for the multivariate integral (31) 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 $\bar{n}$ [42]. Here, we use an efficient approach that corresponds to sparse stochastic collocation and is the multivariate equivalent of Gaussian quadrature [43].

We express $V(\boldsymbol{\theta},\mathbf{y},\mathbf{d})$ as

$$\begin{aligned} V(\boldsymbol{\theta},\mathbf{y},\mathbf{d}) &= \textstyle\sum_{\mathbf{k}\in\bar{\mathcal{K}}_{\bar{n}}^{n_\theta}} (\mathbf{c}_V(\mathbf{y},\mathbf{d}))_{\mathbf{k}}\, \Psi\left(\Delta\boldsymbol{\theta}^{\mathbf{k}}\right) + R_V(\boldsymbol{\theta},\mathbf{y},\mathbf{d}) \\ &= \mathbf{a}_\theta(\boldsymbol{\theta})^{\mathrm{T}}\mathbf{c}_V(\mathbf{y},\mathbf{d}) + R_V(\boldsymbol{\theta},\mathbf{y},\mathbf{d}), \end{aligned} \tag{34}$$

where $\mathbf{c}_V(\mathbf{y},\mathbf{d})$ is the vector of polynomial coefficients of $V(\boldsymbol{\theta},\mathbf{y},\mathbf{d})$, $\Psi\left(\Delta\boldsymbol{\theta}^{\mathbf{k}}\right)$ denotes the first polynomial that contains the monomial $\Delta\boldsymbol{\theta}^{\mathbf{k}}$ in the sequence of orthogonal polynomials with respect to the prior pdf $p(\boldsymbol{\theta})$ (that is, under the inner product $\langle f,g\rangle = \int_\Theta f(\boldsymbol{\theta})g(\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$), with $\mathbf{k}$ being the vector of monomial powers in the set $\bar{\mathcal{K}}_{\bar{n}}^{n_\theta} \subseteq \mathcal{K}_{\bar{n}}^{n_\theta} := \left\{(k_1,\dots,k_{n_\theta}) \in \mathbb{N}_0^{n_\theta} : 0 \le k_1+\cdots+k_{n_\theta} \le \bar{n}\right\}$ in the case of a polynomial of degree $\bar{n}$ and the definitions of $\Delta\boldsymbol{\theta} := \boldsymbol{\theta} - \bar{\boldsymbol{\theta}}$ as the deviation of $\boldsymbol{\theta}$ around $\bar{\boldsymbol{\theta}}$ and $\Delta\boldsymbol{\theta}^{\mathbf{k}} := \left(\theta_1-\bar{\theta}_1\right)^{k_1}\dots\left(\theta_{n_\theta}-\bar{\theta}_{n_\theta}\right)^{k_{n_\theta}}$, $\mathbf{a}_\theta(\boldsymbol{\theta})$ is a vector with elements $(\mathbf{a}_\theta(\boldsymbol{\theta}))_{\mathbf{k}} = \Psi\left(\Delta\boldsymbol{\theta}^{\mathbf{k}}\right)$ for $\mathbf{k} \in \bar{\mathcal{K}}_{\bar{n}}^{n_\theta}$, and $R_V(\boldsymbol{\theta},\mathbf{y},\mathbf{d})$ is the orthogonal part with respect to $\mathbf{a}_\theta(\boldsymbol{\theta})$. For example, in the case of statistically independent parameters $\boldsymbol{\theta}$, the polynomials in $\mathbf{a}_\theta(\boldsymbol{\theta})$ are Hermite polynomials for a normal prior pdf and Jacobi polynomials for a beta prior pdf [44]. In the general case of correlated parameters $\boldsymbol{\theta}$ or an arbitrary prior pdf, the polynomials in $\mathbf{a}_\theta(\boldsymbol{\theta})$ can be obtained via the Gram–Schmidt process [45].

The number of polynomials in $\mathbf{a}_\theta(\boldsymbol{\theta})$ is equal to the cardinality $|\bar{\mathcal{K}}_{\bar{n}}^{n_\theta}|$ of the set $\bar{\mathcal{K}}_{\bar{n}}^{n_\theta}$. As shown below, the number m_θ of quadrature points for the integration rule depends on $|\bar{\mathcal{K}}_{\bar{n}}^{n_\theta}|$. Given the fact that $\bar{\mathcal{K}}_{\bar{n}}^{n_\theta} \subseteq \mathcal{K}_{\bar{n}}^{n_\theta}$ and the definition of $\mathcal{K}_{\bar{n}}^{n_\theta}$, we know that $|\bar{\mathcal{K}}_{\bar{n}}^{n_\theta}| \le |\mathcal{K}_{\bar{n}}^{n_\theta}| = \binom{n_\theta+\bar{n}}{n_\theta}$. However, while $|\mathcal{K}_{\bar{n}}^{n_\theta}|$ grows quickly with the dimension n_θ, we aim to keep m_θ small even when n_θ is large. Hence, we assume that $\bar{\mathcal{K}}_{\bar{n}}^{n_\theta}$ is given by a maximum interaction or hyperbolic truncation scheme to introduce sparsity when n_θ is large [46]. For example, in the case of a maximum interaction scheme with up to p_θ interaction terms, $\bar{\mathcal{K}}_{\bar{n}}^{n_\theta} = \mathcal{K}_{\bar{n}}^{n_\theta} \cap \left\{(k_1,\dots,k_{n_\theta}) \in \mathbb{N}_0^{n_\theta} : \lim_{q\to 0}\sum_{i=1}^{n_\theta} k_i^q \le p_\theta\right\}$, and in the case of a hyperbolic truncation scheme with a q-norm of the monomial powers up to p_θ, $\bar{\mathcal{K}}_{\bar{n}}^{n_\theta} = \mathcal{K}_{\bar{n}}^{n_\theta} \cap \left\{(k_1,\dots,k_{n_\theta}) \in \mathbb{N}_0^{n_\theta} : \left(\sum_{i=1}^{n_\theta} k_i^q\right)^{1/q} \le p_\theta\right\}$. Such schemes for introducing sparsity can potentially eliminate many elements of $\mathcal{K}_{\bar{n}}^{n_\theta}$ from $\bar{\mathcal{K}}_{\bar{n}}^{n_\theta}$ in the case of large n_θ. Since the polynomials in $\mathbf{a}_\theta(\boldsymbol{\theta})$ are orthogonal with respect to $p(\boldsymbol{\theta})$, $\mathbf{I}_{|\bar{\mathcal{K}}_{\bar{n}}^{n_\theta}|} = \int_\Theta \mathbf{a}_\theta(\boldsymbol{\theta})\mathbf{a}_\theta(\boldsymbol{\theta})^{\mathrm{T}}p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ holds. This implies that

$$\begin{bmatrix} 1 & \mathbf{0}_{|\bar{\mathcal{K}}_{\bar{n}}^{n_\theta}|-1}^{\mathrm{T}} \end{bmatrix} = \int_\Theta \mathbf{a}_\theta(\boldsymbol{\theta})^{\mathrm{T}}p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{35}$$

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

$$\mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}\mathbf{A}_\theta = \int_\Theta \mathbf{a}_\theta(\boldsymbol{\theta})^{\mathrm{T}}p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{36}$$

with $(\mathbf{A}_\theta)_{l,\mathbf{k}} = (\mathbf{a}_\theta(\boldsymbol{\theta}_l))_{\mathbf{k}}$ for $l = 1,\dots,m_\theta$ and $\mathbf{k} \in \bar{\mathcal{K}}_{\bar{n}}^{n_\theta}$. Suppose that $(n_\theta+1)m_\theta \ge |\bar{\mathcal{K}}_{\bar{n}}^{n_\theta}|$ and $\mathbf{W}$ and $\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_{m_\theta}$ are chosen such that they satisfy (36). Then, since

$$\begin{aligned} v(\mathbf{y},\mathbf{d}) &= \int_\Theta \mathbf{a}_\theta(\boldsymbol{\theta})^{\mathrm{T}}\mathbf{c}_V(\mathbf{y},\mathbf{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} + \int_\Theta R_V(\boldsymbol{\theta},\mathbf{y},\mathbf{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ &= \mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}\mathbf{p}_V(\mathbf{y},\mathbf{d}) - \mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}\left(\mathbf{p}_V(\mathbf{y},\mathbf{d}) - \mathbf{A}_\theta\mathbf{c}_V(\mathbf{y},\mathbf{d})\right) \\ &\quad + \int_\Theta R_V(\boldsymbol{\theta},\mathbf{y},\mathbf{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ &= \mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}\mathbf{p}_V(\mathbf{y},\mathbf{d}) - \mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}\mathbf{p}_{R_V}(\mathbf{y},\mathbf{d}) + \int_\Theta R_V(\boldsymbol{\theta},\mathbf{y},\mathbf{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \end{aligned} \tag{37}$$

with $\left(\mathbf{p}_{R_V}(\mathbf{y},\mathbf{d})\right)_l = R_V(\boldsymbol{\theta}_l,\mathbf{y},\mathbf{d})$ for $l = 1,\dots,m_\theta$, the integral $v(\mathbf{y},\mathbf{d})$ can be approximated as

$$\hat{v}(\mathbf{y},\mathbf{d}) = \mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}\mathbf{p}_V(\mathbf{y},\mathbf{d}). \tag{38}$$

It follows that (32) with $\mathbf{w}^{\mathrm{T}} = \mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}$ holds and the approximation error $\hat{v}(\mathbf{y},\mathbf{d}) - v(\mathbf{y},\mathbf{d})$ vanishes when $R_V(\boldsymbol{\theta},\mathbf{y},\mathbf{d}) = 0$. More generally, the approximation error $\hat{v}(\mathbf{y},\mathbf{d}) - v(\mathbf{y},\mathbf{d})$ will be small if $R_V(\boldsymbol{\theta},\mathbf{y},\mathbf{d})$ is small, that is, if $V(\boldsymbol{\theta},\mathbf{y},\mathbf{d})$ is well approximated by a polynomial $\mathbf{a}_\theta(\boldsymbol{\theta})^{\mathrm{T}}\mathbf{c}_V(\mathbf{y},\mathbf{d})$ of degree $\bar{n}$ for some vector $\mathbf{c}_V(\mathbf{y},\mathbf{d})$ of polynomial coefficients. This can be the case if $V(\boldsymbol{\theta},\mathbf{y},\mathbf{d})$ is a sufficiently smooth function of $\boldsymbol{\theta}$, but not if $V(\boldsymbol{\theta},\mathbf{y},\mathbf{d})$ is a nonsmooth function of $\boldsymbol{\theta}$. Hence, this method for multivariate integration based on Gaussian quadrature is effective for smooth functions of $\boldsymbol{\theta}$ but less appropriate for nonsmooth functions of $\boldsymbol{\theta}$.

Remark 7. A method based on polynomial chaos expansions could also be used for approximation of multivariate integrals [13, 25]. Then, the question is how the approximation of $v(\mathbf{y},\mathbf{d})$ in (38) compares to the one obtained by integrating a polynomial approximation of $V(\boldsymbol{\theta},\mathbf{y},\mathbf{d})$. It turns out that this approximation requires $m_\theta \geq |\bar{\mathcal{K}}_{\bar{n}}^{n_\theta}|$ quadrature points. Hence, we propose the use of the approach based on Gaussian quadrature since it needs fewer quadrature points and does not require any regression.

4.2. Approximation of expected utility and constraint moments

We now use the previous results to approximate the expected utilities $u_D(\mathbf{d})$ and $u_{MC}(\mathbf{d})$, as well as the constraint moments $m_k^{\mathrm{I}}(t;\mathbf{d})$ and $m_k^{\mathrm{II}}(t;\mathbf{d})$, that will be used in the remainder of the paper. We first apply the approximation in (32) for multivariate integrals of the form in (31) to $u_D(\mathbf{d})$ in (14), which yields

$$u_D(\mathbf{d}) \approx \hat{u}_D(\mathbf{d}) = \textstyle\sum_{l=1}^{m_\theta} w_l U_D(\boldsymbol{\theta}_l,\mathbf{d}) = \mathbf{w}^{\mathrm{T}}\mathbf{p}_{U_D}(\mathbf{d}). \tag{39}$$

Also, we apply the approximation in (32) for multivariate integrals of the form in (31) to $u_{MC}(\mathbf{d})$ in (23), which yields

$$u_{MC}(\mathbf{d}) \approx \hat{u}_{MC}(\mathbf{d}) = \textstyle\sum_{l=1}^{m_\theta} w_l U_{MC}(\boldsymbol{\theta}_l,\mathbf{d}) = \mathbf{w}^{\mathrm{T}}\mathbf{p}_{U_{MC}}(\mathbf{d}). \tag{40}$$

To compute the function $U_{MC}(\boldsymbol{\theta},\mathbf{d})$, one needs to compute the evidence $p(\mathbf{y}|\mathbf{d})$ for different $\mathbf{y}$. Again, we apply the approximation in (32) for multivariate integrals of the form in (31) to $p(\mathbf{y}|\mathbf{d})$ in (4), which yields

$$p(\mathbf{y}|\mathbf{d}) \approx \hat{p}(\mathbf{y}|\mathbf{d}) = \textstyle\sum_{l=1}^{m_\theta} w_l p(\mathbf{y}|\boldsymbol{\theta}_l,\mathbf{d}). \tag{41}$$

Note that this approximation of $p(\mathbf{y}|\mathbf{d})$ is particularly accurate in the case of large observation noise or small amount of data for which Bayesian OED is most useful, since the likelihood function $p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})$ is not concentrated in a small region of the parameter space in this case. This results in the approximation of the utility function

$$\begin{aligned} U_{MC}(\boldsymbol{\theta},\mathbf{d}) \approx U_Q(\boldsymbol{\theta},\mathbf{d}) &= \tfrac{1}{m_\xi}\textstyle\sum_{k=1}^{m_\xi} \log\left(\frac{\det(\mathbf{J}(\boldsymbol{\theta},\mathbf{d}))^{-1}\pi(\boldsymbol{\xi}_k|\bar{\mathbf{x}},\boldsymbol{\Sigma}_{\boldsymbol{\xi}})}{\hat{p}(\mathbf{g}(\boldsymbol{\theta},\mathbf{d})+\mathbf{J}(\boldsymbol{\theta},\mathbf{d})\boldsymbol{\xi}_k|\mathbf{d})}\right) \\ &= \tfrac{1}{m_\xi}\textstyle\sum_{k=1}^{m_\xi} \log\left(\frac{\det(\mathbf{J}(\boldsymbol{\theta},\mathbf{d}))^{-1}\pi(\boldsymbol{\xi}_k|\bar{\mathbf{x}},\boldsymbol{\Sigma}_{\boldsymbol{\xi}})}{\sum_{l=1}^{m_\theta} w_l p(\mathbf{g}(\boldsymbol{\theta},\mathbf{d})+\mathbf{J}(\boldsymbol{\theta},\mathbf{d})\boldsymbol{\xi}_k|\boldsymbol{\theta}_l,\mathbf{d})}\right), \end{aligned} \tag{42}$$

where $\frac{p(\mathbf{g}(\boldsymbol{\theta},\mathbf{d})+\mathbf{J}(\boldsymbol{\theta},\mathbf{d})\boldsymbol{\xi}_k|\boldsymbol{\theta}_l,\mathbf{d})}{\det(\mathbf{J}(\boldsymbol{\theta},\mathbf{d}))^{-1}}$ is obtained by replacing $\hat{\boldsymbol{\theta}}$ with $\boldsymbol{\theta}_l$ in (26) and computing the pdf $\pi(\cdot|\bar{\mathbf{x}},\boldsymbol{\Sigma}_{\boldsymbol{\xi}})$, and the approximation

$$u_{MC}(\mathbf{d}) \approx \hat{u}_Q(\mathbf{d}) = \textstyle\sum_{l=1}^{m_\theta} w_l U_Q(\boldsymbol{\theta}_l,\mathbf{d}) = \mathbf{w}^{\mathrm{T}}\mathbf{p}_{U_Q}(\mathbf{d}). \tag{43}$$

Note that the computation of $\hat{u}_Q(\mathbf{d})$ implies computing the pdf $\pi(\cdot|\bar{\mathbf{x}},\boldsymbol{\Sigma}_{\boldsymbol{\xi}})$ $m_\theta m_\xi(1+m_\theta)$ times but uses only m_θ different values of $\boldsymbol{\theta}$.

Lastly, we apply the approximation in (32) to $m_k^{\mathrm{I}}(t;\mathbf{d})$ and $m_k^{\mathrm{II}}(t;\mathbf{d})$ in (29) and (30), which yields the constraints

$$\tfrac{r_k^2}{1+r_k^2}\hat{m}_k^{\mathrm{II}}(t;\mathbf{d}) - \left(\hat{m}_k^{\mathrm{I}}(t;\mathbf{d})\right)^2 \leq 0, \quad k = 1,\dots,n_h, \tag{44}$$

with the approximations of the moments

$$\begin{aligned} m_k^{\mathrm{I}}(t;\mathbf{d}) \approx \hat{m}_k^{\mathrm{I}}(t;\mathbf{d}) &= \textstyle\sum_{l=1}^{m_\theta} w_l M_k^{\mathrm{I}}(t;\boldsymbol{\theta}_l,\mathbf{d}) = \mathbf{w}^{\mathrm{T}}\mathbf{p}_{M_k^{\mathrm{I}}}(t;\mathbf{d}), \\ &\qquad k = 1,\dots,n_h, \end{aligned} \tag{45}$$

$$\begin{aligned} m_k^{\mathrm{II}}(t;\mathbf{d}) \approx \hat{m}_k^{\mathrm{II}}(t;\mathbf{d}) &= \textstyle\sum_{l=1}^{m_\theta} w_l M_k^{\mathrm{II}}(t;\boldsymbol{\theta}_l,\mathbf{d}) = \mathbf{w}^{\mathrm{T}}\mathbf{p}_{M_k^{\mathrm{II}}}(t;\mathbf{d}), \\ &\qquad k = 1,\dots,n_h. \end{aligned} \tag{46}$$

Remark 8. In theory, one could directly apply the approximation in (32) to $q_k(t;\mathbf{d})$ in (8) since $q_k(t;\mathbf{d})$ is a multivariate integral of the form in (31). However, since $Q_k(t;\boldsymbol{\theta},\mathbf{d})$ is a nonsmooth function of $\boldsymbol{\theta}$, the method for multivariate integration based on Gaussian quadrature is not appropriate for $Q_k(t;\boldsymbol{\theta},\mathbf{d})$. In contrast, the method based on Gaussian quadrature is effective for smooth functions of $\boldsymbol{\theta}$ such as $M_k^{\mathrm{I}}(t;\boldsymbol{\theta},\mathbf{d})$ and $M_k^{\mathrm{II}}(t;\boldsymbol{\theta},\mathbf{d})$. For this reason, we propose to first reformulate the chance path constraints (7) with (8) as (28) with (29) and (30) and then apply the approximation in (32) to $m_k^{\mathrm{I}}(t;\mathbf{d})$ and $m_k^{\mathrm{II}}(t;\mathbf{d})$ in (29) and (30).

4.3. Reformulation of OED as an optimal control problem

This subsection proposes the explicit reformulation of the approximate Bayesian OED problem as OCPs for the two cases considered previously. In the case of Gaussian prior pdf and noise, we maximize the approximate expected utility $\hat{u}_D(\mathbf{d})$ in (39) subject to the path constraints (44) with the approximate constraint moments $\hat{m}_k^{\mathrm{I}}(t;\mathbf{d})$ and $\hat{m}_k^{\mathrm{II}}(t;\mathbf{d})$ in (45) and (46). On the other hand, in the case of arbitrary prior pdf and noise, we maximize the approximate expected utility $\hat{u}_Q(\mathbf{d})$ in (43) subject to the path constraints (44) with the approximate constraint moments $\hat{m}_k^{\mathrm{I}}(t;\mathbf{d})$ and $\hat{m}_k^{\mathrm{II}}(t;\mathbf{d})$ in (45) and (46). As shown below, these two cases involve a different number of differential equations and states in the OCP. In both cases, the reformulation of the approximate Bayesian OED problem as OCPs with a finite number of states is enabled by the approximation in (32) for multivariate integrals of the form in (31).

4.3.1. Reformulation for Gaussian prior pdf and noise

The approximate expected utility $\hat{u}_D(\mathbf{d})$ in (39) can be written as an explicit function of the states $\mathbf{X}(t;\boldsymbol{\theta}_1),\dots,\mathbf{X}(t;\boldsymbol{\theta}_{m_\theta})$ from (15), (16), (18), and the same is valid for the approximate constraint moments $\hat{m}_k^{\mathrm{I}}(t;\mathbf{d})$ and $\hat{m}_k^{\mathrm{II}}(t;\mathbf{d})$ in (45) and (46). Thus, these approximations involve the dynamics

$$\mathbf{R}\big(\mathbf{S}(t),\mathbf{u}(t)\big) := \mathrm{vec}\begin{bmatrix} \mathbf{F}\left(\mathbf{X}(t;\boldsymbol{\theta}_1),\boldsymbol{\theta}_1,\mathbf{u}(t)\right) \\ \vdots \\ \mathbf{F}\left(\mathbf{X}(t;\boldsymbol{\theta}_{m_\theta}),\boldsymbol{\theta}_{m_\theta},\mathbf{u}(t)\right) \end{bmatrix}, \tag{47}$$

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

$$\mathbf{S}(t) := \mathrm{vec}\begin{bmatrix} \mathbf{X}(t;\boldsymbol{\theta}_1) \\ \vdots \\ \mathbf{X}(t;\boldsymbol{\theta}_{m_\theta}) \end{bmatrix}, \quad \mathbf{S}_0(\mathbf{b}) := \mathrm{vec}\begin{bmatrix} \mathbf{X}_0\left(\boldsymbol{\theta}_1,\mathbf{b}\right) \\ \vdots \\ \mathbf{X}_0\left(\boldsymbol{\theta}_{m_\theta},\mathbf{b}\right) \end{bmatrix}. \tag{48}$$

Hence, we define

$$\phi_D\big(\mathbf{S}(t_1),\dots,\mathbf{S}(t_T)\big) := \mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}\mathbf{p}_{U_D}(\mathbf{d}), \tag{49}$$

$$\begin{aligned} \mu_{D,k}\big(\mathbf{S}(t)\big) &:= \tfrac{r_k^2}{1+r_k^2}\mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}\mathbf{p}_{M_k^{\mathrm{II}}}(t;\mathbf{d}) - \left(\mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}\mathbf{p}_{M_k^{\mathrm{I}}}(t;\mathbf{d})\right)^2, \\ &\qquad k = 1,\dots,n_h. \end{aligned} \tag{50}$$

Accordingly, the Bayesian OED problem (14) can be approximated by the OCP

$$\hat{\mathbf{d}}_D^* := \arg\max_{\mathbf{d}\in\mathcal{D}} \hat{u}_D(\mathbf{d}) = \phi_D\big(\mathbf{S}(t_1),\dots,\mathbf{S}(t_T)\big), \tag{51a}$$

$$\text{s.t. } \dot{\mathbf{S}}(t) = \mathbf{R}\big(\mathbf{S}(t), \mathbf{u}(t)\big), \quad \mathbf{S}(t_0) = \mathbf{S}_0(\mathbf{b}), \tag{51b}$$

$$\tfrac{r_k^2}{1+r_k^2}\hat{m}_k^{\mathrm{II}}(t;\mathbf{d}) - \big(\hat{m}_k^{\mathrm{I}}(t;\mathbf{d})\big)^2 = \mu_{D,k}\big(\mathbf{S}(t)\big) \leq 0, \quad k = 1,\dots,n_h, \tag{51c}$$

(15), (16), (18), (51d)

where $\boldsymbol{\mu}_D^{(1)}(\mathbf{S},\mathbf{u}) := \frac{\partial \boldsymbol{\mu}_D}{\partial \mathbf{S}}(\mathbf{S})\mathbf{R}(\mathbf{S},\mathbf{u})$ depends explicitly on $\mathbf{u}$.

4.3.2. Reformulation for arbitrary prior pdf and noise

Since the sample points $\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_{m_\xi}$ can be chosen in advance, the approximate expected utility $\hat{u}_Q(\mathbf{d})$ in (43) can be written as an explicit function of the states $\mathbf{x}(t;\boldsymbol{\theta}_1),\dots,\mathbf{x}(t;\boldsymbol{\theta}_{m_\theta})$ from (26), (42), and the same is valid for the approximate constraint moments $\hat{m}_k^{\mathrm{I}}(t;\mathbf{d})$ and $\hat{m}_k^{\mathrm{II}}(t;\mathbf{d})$ in (45) and (46). Thus, these approximations involve the dynamics

$$\mathbf{r}\big(\mathbf{s}(t),\mathbf{u}(t)\big) := \begin{bmatrix} \mathbf{f}\left(\mathbf{x}(t;\boldsymbol{\theta}_1),\boldsymbol{\theta}_1,\mathbf{u}(t)\right) \\ \vdots \\ \mathbf{f}\left(\mathbf{x}(t;\boldsymbol{\theta}_{m_\theta}),\boldsymbol{\theta}_{m_\theta},\mathbf{u}(t)\right) \end{bmatrix}, \tag{52}$$

for the $n_r := n_x m_\theta$ states and initial conditions

$$\mathbf{s}(t) := \begin{bmatrix} \mathbf{x}(t;\boldsymbol{\theta}_1) \\ \vdots \\ \mathbf{x}(t;\boldsymbol{\theta}_{m_\theta}) \end{bmatrix}, \quad \mathbf{s}_0(\mathbf{b}) := \begin{bmatrix} \mathbf{x}_0\left(\boldsymbol{\theta}_1,\mathbf{b}\right) \\ \vdots \\ \mathbf{x}_0\left(\boldsymbol{\theta}_{m_\theta},\mathbf{b}\right) \end{bmatrix}. \tag{53}$$

Hence, we define

$$\phi_Q\big(\mathbf{s}(t_1),\dots,\mathbf{s}(t_T)\big) := \mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}\mathbf{p}_{U_Q}(\mathbf{d}), \tag{54}$$

$$\mu_{Q,k}\big(\mathbf{s}(t)\big) := \tfrac{r_k^2}{1+r_k^2}\mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}\mathbf{p}_{M_k^{\mathrm{II}}}(t;\mathbf{d}) - \big(\mathbf{1}_{m_\theta}^{\mathrm{T}}\mathbf{W}\mathbf{p}_{M_k^{\mathrm{I}}}(t;\mathbf{d})\big)^2, \quad k = 1,\dots,n_h. \tag{55}$$

Accordingly, the Bayesian OED problem (23) can be approximated by the OCP

$$\hat{\mathbf{d}}_Q^* := \arg\max_{\mathbf{d}\in\mathcal{D}} \hat{u}_Q(\mathbf{d}) = \phi_Q\big(\mathbf{s}(t_1),\dots,\mathbf{s}(t_T)\big), \tag{56a}$$

$$\text{s.t. } \dot{\mathbf{s}}(t) = \mathbf{r}\big(\mathbf{s}(t),\mathbf{u}(t)\big), \quad \mathbf{s}(t_0) = \mathbf{s}_0(\mathbf{b}), \tag{56b}$$

$$\tfrac{r_k^2}{1+r_k^2}\hat{m}_k^{\mathrm{II}}(t;\mathbf{d}) - \big(\hat{m}_k^{\mathrm{I}}(t;\mathbf{d})\big)^2 = \mu_{Q,k}\big(\mathbf{s}(t)\big) \leq 0, \quad k = 1,\dots,n_h, \tag{56c}$$

(26), (42), (56d)

where $\boldsymbol{\mu}_Q^{(1)}(\mathbf{s},\mathbf{u}) := \frac{\partial \boldsymbol{\mu}_Q}{\partial \mathbf{s}}(\mathbf{s})\mathbf{r}(\mathbf{s},\mathbf{u})$ depends explicitly on $\mathbf{u}$.

For the sake of clarity, in the remainder we use the notation $\hat{u}(\mathbf{d})$, $\phi\big(\mathbf{s}(t_1),\dots,\mathbf{s}(t_T)\big)$, $\mu_k\big(\mathbf{s}(t)\big)$ in lieu of $\hat{u}_D(\mathbf{d})$, $\phi_D\big(\mathbf{S}(t_1),\dots,\mathbf{S}(t_T)\big)$, $\mu_{D,k}\big(\mathbf{S}(t)\big)$ as well as $\hat{u}_Q(\mathbf{d})$, $\phi_Q\big(\mathbf{s}(t_1),\dots,\mathbf{s}(t_T)\big)$, $\mu_{Q,k}\big(\mathbf{s}(t)\big)$, but all the following developments are valid for both OCPs (51) and (56).

5. Reformulation of the OCP as polynomial optimization problems

This section defines the proposed reformulation of the OCPs (51) and (56), with the purpose of transforming these OCPs into polynomial optimization problems that are amenable to global optimality.

5.1. Solution approach

The inputs that represent the solution to the OCPs (51) and (56) are composed of several arcs. For each input u_j, each arc can be of type (1) bang–bang, such that it is determined by an equality $u_j = \underline{u}_j$ or $u_j = \overline{u}_j$, (2) active-state constraint, such that it is determined by an equality $\mu_k^{(1)}(\mathbf{s},\mathbf{u}) = 0$ for some $k = 1,\dots,n_h$, or (3) free, such that it is determined by an equality that stems from the dynamics given by $\mathbf{r}\big(\mathbf{s}(t),\mathbf{u}(t)\big)$, also labeled as singular in the relevant case of input-affine OCPs with $\mathbf{r}\big(\mathbf{s}(t),\mathbf{u}(t)\big)$ affine in $\mathbf{u}(t)$ [34,47]. 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. Suppose that we denote the bang–bang arcs as 1L or 1U, depending on whether they are determined by $u_j = \underline{u}_j$ or $u_j = \overline{u}_j$. Then, note that: (i) sequences with fewer than $\bar{n}_a$ arcs are particular cases of the sequences with $\bar{n}_a$ arcs where some arcs vanish, and (ii) the sequences that end with an arc of type 3 are not plausible in input-affine OCPs according to Pontryagin's maximum principle [48,49]. Hence, by recalling that plausible arc sequences do not have consecutive arcs of the same type, the branching factor is 2 for each arc in a plausible sequence for a single-input OCP, and the number of plausible sequences is equal to $2^{\bar{n}_a}$ for an input-affine OCP or $\frac{3}{2}2^{\bar{n}_a}$ otherwise. In addition, one can note that the arcs of types 1L and 1U can be seen as particular cases of arcs of type 3 since in arcs of types 1L and 1U the input is constant, while in arcs of type 3 the input is assumed to be approximated by a linear function. This implies that certain arc sequences do not need to be considered. For example, for $\bar{n}_a = 3$, the sequence 1U-1L-1U does not need to considered because it is a particular case of the sequences 1U-3-1U and 3-1L-1U. However, it is not recommendable to express all the arc sequences as particular cases of a hypothetical sequence with $\bar{n}_a$ arcs of type 3 since the number of decision variables for that sequence would be excessive.

Remark 9. The effect of non-optimal inputs is different for arcs of different types: a non-optimal input in bang–bang and active-state constraint arcs has an important effect on the cost of the OCP, while a non-optimal input in free/singular arcs has a negligible effect on the cost [47]. Hence, we use the fact that a non-optimal input in free/singular arcs has a negligible effect on the cost to assume that free/singular arcs are approximated by linear functions of time throughout the paper. Even if the true optimal input in a free/singular arc is not a linear function, its approximation by a linear function does not have a significant effect on the cost and leads to a small loss of optimality. On the other hand, since a non-optimal input in bang–bang and active-state constraint arcs has an important effect on the cost, constraint handling is emphasized in the paper.

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 [33,34]. For a given plausible arc sequence composed of n_s+1 bang–bang and free/singular arcs, the inputs are defined by the following decision variables: the switching times $\bar{t}_1,\dots,\bar{t}_{n_s}$ to arcs of types 1 and 3 and the initial conditions of the free/singular arcs. The difference with respect to a switching point optimization approach is precisely the inclusion of the initial conditions of the free/singular arcs as decision variables, which allow representing singular arcs in input-affine OCPs related to complex nonlinear systems [50]. The final time $\bar{t}_{n_s+1} = t_f$ is not a decision variable in this paper. The entry points in arcs of type 2 are given by the n_η-dimensional vector $\boldsymbol{\eta} = (\eta_1,\dots,\eta_{n_\eta})$, but the switching to these arcs cannot occur at arbitrary times since it depends on the states $\mathbf{s}$. In this paper, we assume that $\boldsymbol{\mu}^{(1)}(\mathbf{s},\mathbf{u})$ explicitly depends on $\mathbf{u}$ because otherwise it would be impossible to ensure that the state constraint $\mu_k\big(\mathbf{s}(t)\big) \leq 0$ remains active for $t > \eta$ once an entry point η is reached such that $\mu_k\big(\mathbf{s}(\eta)\big) = 0$ for

some $k = 1, \dots, n_h$. For example, suppose that $\boldsymbol{\mu}^{(1)}(\mathbf{s})$ does not explicitly depend on $\mathbf{u}$ but $\boldsymbol{\mu}^{(2)}(\mathbf{s}, \mathbf{u}) := \frac{\partial \boldsymbol{\mu}^{(1)}}{\partial \mathbf{s}}(\mathbf{s})\mathbf{r}(\mathbf{s}, \mathbf{u})$ explicitly depends on $\mathbf{u}$. Then, once an entry point η is reached such that $\mu_k\big(\mathbf{s}(\eta)\big) = 0$ and $\mu_k^{(1)}\big(\mathbf{s}(\eta)\big) > 0$ for some $k = 1, \dots, n_h$, there exists no $\mathbf{u}(t)$ that guarantees that $\mu_k\big(\mathbf{s}(t)\big) \leq 0$ for $t > \eta$. In contrast, if $\boldsymbol{\mu}^{(1)}(\mathbf{s}, \mathbf{u})$ explicitly depends on $\mathbf{u}$ as assumed, once an entry point η is reached such that $\mu_k\big(\mathbf{s}(\eta)\big) = 0$ for some $k = 1, \dots, n_h$, it is possible to choose $\mathbf{u}(t)$ such that $\mu_k^{(1)}\big(\mathbf{s}(t), \mathbf{u}(t)\big) = 0$, which ensures that the state constraint $\mu_k\big(\mathbf{s}(t)\big) \leq 0$ remains active for $t > \eta$. Also, we assume that the optimal sequence of arcs of types 1, 2, and 3 is known for each given sequence of arcs of types 1 and 3 for clarity and convenience, that is, to simplify the exposition in the remainder of the paper, although this assumption is not a requirement.

We aim to apply the parsimonious input parameterization approach and show how Bayesian OED problems reformulated as (51) and (56) can be solved efficiently to global optimality. The proposed approach for global optimality relies on determining: (i) when and how the globally optimal switching between arcs takes place for a given plausible arc sequence; and (ii) which sequence provides the globally optimal solution. Then, addressing question (i) consists in computing the globally optimal values of the decision variables for the given arc sequence. For this, we represent the cost of the OCP as an explicit polynomial function since that converts the OCP into a set of polynomial optimization problems (POPs), one for each arc sequence, as shown next. Once question (i) is addressed for each sequence via parallel computing, it is trivial to answer question (ii) efficiently.

Hence, the remainder of this section shows how to reformulate the OCPs (51) and (56) as a set of POPs, one for each plausible arc sequence.

5.2. OCP with new decision variables

For a given arc sequence, we describe the input in the ith time interval $\big[\bar{t}_{i-1}, \bar{t}_i\big)$, for $i = 1, \dots, n_s + 1$, by defining $n_{z,i}$ new 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 a vector with a dimension $n_z := n_r + n_{z,1} + \cdots + n_{z,n_s+1}$

$$\mathbf{z}(t) := \begin{bmatrix} \mathbf{s}(t)^{\mathrm{T}} & \begin{bmatrix} \mathbf{z}_1(t) \\ \vdots \\ \mathbf{z}_{n_s+1}(t) \end{bmatrix}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}, \tag{57}$$

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

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 $\mathbf{u}(t)$ given by the control law $\mathbf{u}(t) = \tilde{\mathbf{c}}\big(\mathbf{z}(t)\big)$ and on the dynamics of $\mathbf{z}_i(t)$ given by $\dot{\mathbf{z}}_i(t) = \mathbf{q}_i\big(\mathbf{s}(t), \mathbf{z}_i(t)\big)$. For bang–bang arcs, $\mathbf{z}_i(t)$, $\mathbf{z}_{i,0}$ are of dimension 0 and $\tilde{c}_j\big(\mathbf{z}(t)\big) := \underline{u}_j$ or $\tilde{c}_j\big(\mathbf{z}(t)\big) := \overline{u}_j$ for the jth input. For active-state constraint arcs, $\mathbf{z}_i(t)$, $\mathbf{z}_{i,0}$ are not needed and $\tilde{c}_j\big(\mathbf{z}(t)\big)$ is such that $\mu_k^{(1)}\big(\mathbf{s}(t), \tilde{\mathbf{c}}\big(\mathbf{z}(t)\big)\big) = 0$ for some $k = 1, \dots, n_h$. For free/singular arcs, since we assume that the jth input is approximated by a linear function, then $\mathbf{z}_i(t) = \begin{bmatrix} \tilde{u}_{j,i}(t) \\ \tilde{p}_{j,i}(t) \end{bmatrix}$ and $\mathbf{z}_{i,0} = \begin{bmatrix} u_{j,i}^0 \\ p_{j,i} \end{bmatrix}$ are of dimension 2, where $u_{j,i}^0$ and $p_{j,i}$ are the initial value and slope of the input and $\tilde{u}_{j,i}(t)$ is its value at time t, which implies that $\tilde{c}_j\big(\mathbf{z}(t)\big) := \tilde{u}_{j,i}(t)$ and $\mathbf{q}_i\big(\mathbf{s}(t), \mathbf{z}_i(t)\big) := \begin{bmatrix} \tilde{p}_{j,i}(t) \\ 0 \end{bmatrix}$. The set $\big\{i : i\text{th arc of } u_j \text{ is of type } 3\big\}$ is denoted as $\mathcal{S}_j$, which implies that $n_{z,1} + \cdots + n_{z,n_s+1} = 2\sum_{j=1}^{n_u} |\mathcal{S}_j|$. This way to deal with active-state constraint arcs avoids the need to check that path constraints are satisfied at a finite number of points t, which typically requires solving a sequence of approximate problems where the points and the approximations are updated at each step [51].

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

$$\tilde{\mathbf{f}}\big(\mathbf{z}(t)\big) := \begin{bmatrix} \mathbf{r}\big(\mathbf{s}(t), \tilde{\mathbf{c}}\big(\mathbf{z}(t)\big)\big)^{\mathrm{T}} & \begin{bmatrix} \mathbf{q}_1\big(\mathbf{s}(t), \mathbf{z}_1(t)\big) \\ \vdots \\ \mathbf{q}_{n_s+1}\big(\mathbf{s}(t), \mathbf{z}_{n_s+1}(t)\big) \end{bmatrix}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}. \tag{58}$$

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

$$\boldsymbol{\tau}^* := \arg\max_{\boldsymbol{\tau}} \hat{\phi}(\boldsymbol{\tau}) := \tilde{\phi}\big(\mathbf{z}(t_1), \dots, \mathbf{z}(t_T)\big), \tag{59a}$$

$$\text{s.t. } \bar{t}_{i-1} \leq \bar{t}_i, \quad i = 1, \dots, n_s + 1, \tag{59b}$$

$$\underline{u}_j \leq u_{j,s}^0 \leq \overline{u}_j,$$

$$\underline{u}_j \leq u_{j,s}^0 + p_{j,s}\big(\bar{t}_s - \bar{t}_{s-1}\big) \leq \overline{u}_j, \quad s \in \mathcal{S}_j, \tag{59c}$$

$$\dot{\mathbf{z}}(t) = \tilde{\mathbf{f}}\big(\mathbf{z}(t)\big), \quad \mathbf{z}(t_0) = \mathbf{z}_0(\mathbf{b}), \tag{59d}$$

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

For each entry point $\hat{\eta}_j(\boldsymbol{\tau}) := \eta_j$, there exists $k = 1, \dots, n_h$ such that $\tilde{\mu}_k\big(\mathbf{z}(\hat{\eta}_j(\boldsymbol{\tau})^-)\big) < 0$, $\tilde{\mu}_k\big(\mathbf{z}(\hat{\eta}_j(\boldsymbol{\tau}))\big) = 0$, which means that $\tilde{\mu}_k\big(\mathbf{z}(t)\big) := \mu_k\big(\mathbf{s}(t)\big) \leq 0$ becomes active at $t = \eta_j$.

5.3. Reformulation as polynomial optimization problems

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 [52,53]. 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

$$\boldsymbol{\zeta}(t) := \begin{bmatrix} \boldsymbol{\lambda}(t)^{\mathrm{T}} & \begin{bmatrix} \boldsymbol{\zeta}_1(t) \\ \vdots \\ \boldsymbol{\zeta}_{n_s+1}(t) \end{bmatrix}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}, \tag{60}$$

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 (59), the extended states $\mathbf{z}(t)$ are described by the differential equations

$$\frac{\mathrm{d}\mathbf{z}}{\mathrm{d}t}(t) = \frac{\partial \tilde{H}}{\partial \boldsymbol{\zeta}}\big(\mathbf{z}(t), \boldsymbol{\zeta}(t)\big)^{\mathrm{T}} = \tilde{\mathbf{f}}\big(\mathbf{z}(t)\big), \quad \mathbf{z}(t_0) = \mathbf{z}_0(\mathbf{b}). \tag{61}$$

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

$$\frac{\mathrm{d}\boldsymbol{\zeta}}{\mathrm{d}t}(t) = -\frac{\partial \tilde{H}}{\partial \mathbf{z}}\big(\mathbf{z}(t), \boldsymbol{\zeta}(t)\big)^{\mathrm{T}} = -\frac{\partial \tilde{\mathbf{f}}}{\partial \mathbf{z}}\big(\mathbf{z}(t)\big)^{\mathrm{T}}\boldsymbol{\zeta}(t), \quad \boldsymbol{\zeta}(t_T) = \mathbf{0}_{n_z},$$

$$\boldsymbol{\zeta}(t_k^-) = \boldsymbol{\zeta}(t_k) + \frac{\partial \tilde{\phi}}{\partial \mathbf{z}(t_k)}\big(\mathbf{z}(t_1), \dots, \mathbf{z}(t_T)\big)^{\mathrm{T}}, \quad k = 1, \dots, T, \tag{62}$$

and in addition, for each entry point η such that $\tilde{\mu}_k\big(\mathbf{z}(t)\big) \leq 0$ becomes active at $t = \eta$ for some $k = 1, \dots, n_h$, it holds that

$$\boldsymbol{\zeta}(\eta^-) = \boldsymbol{\zeta}(\eta) - \frac{\partial \tilde{\mu}_k}{\partial \mathbf{z}}\big(\mathbf{z}(\eta^-)\big)^{\mathrm{T}} \frac{\big(\tilde{\mathbf{f}}\big(\mathbf{z}(\eta^-)\big) - \tilde{\mathbf{f}}\big(\mathbf{z}(\eta)\big)\big)^{\mathrm{T}}\boldsymbol{\zeta}(\eta)}{\tilde{\mu}_k^{(1)}\big(\mathbf{z}(\eta^-)\big)}, \tag{63}$$

where the last expression is known for the case of state constraints of first order, that is, if $\boldsymbol{\mu}^{(1)}(\mathbf{s}, \mathbf{u}) := \frac{\partial \boldsymbol{\mu}}{\partial \mathbf{s}}(\mathbf{s})\mathbf{r}(\mathbf{s}, \mathbf{u})$ depends explicitly on $\mathbf{u}$ as assumed, but is unknown for state constraints of higher order, to the best of our knowledge.

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

$$\frac{\partial\hat{\phi}}{\partial\bar{t}_i}(\boldsymbol{\tau}) = \tilde{H}\big(\mathbf{z}(\bar{t}_i^-), \boldsymbol{\zeta}(\bar{t}_i^-)\big) - \tilde{H}\big(\mathbf{z}(\bar{t}_i), \boldsymbol{\zeta}(\bar{t}_i)\big) = \Big(\tilde{\mathbf{f}}\big(\mathbf{z}(\bar{t}_i^-)\big) - \tilde{\mathbf{f}}\big(\mathbf{z}(\bar{t}_i)\big)\Big)^{\mathrm{T}} \boldsymbol{\zeta}(\bar{t}_i), \quad i = 1, \dots, n_s, \tag{64}$$

$$\frac{\partial\hat{\phi}}{\partial\mathbf{z}_{i,0}}(\boldsymbol{\tau}) = \boldsymbol{\zeta}_i(t_0)^{\mathrm{T}}, \quad i = 1, \dots, n_s + 1, \tag{65}$$

$$\frac{\partial\hat{\phi}}{\partial\mathbf{b}}(\boldsymbol{\tau}) = \boldsymbol{\lambda}(t_0)^{\mathrm{T}} \frac{\partial\mathbf{z}_0}{\partial\mathbf{b}}(\mathbf{b}). \tag{66}$$

Then, suppose that there exists $\bar{\boldsymbol{\tau}}$ such that, for all $\Delta\boldsymbol{\tau} \in \mathcal{R}$,

$$\hat{\phi}(\boldsymbol{\tau}) = \sum_{\mathbf{k}\in\mathcal{K}_n^N} \big(\mathbf{c}_{\hat{\phi}}\big)_{\mathbf{k}} \Delta\boldsymbol{\tau}^{\mathbf{k}} + R_{\hat{\phi}}(\boldsymbol{\tau}), \tag{67}$$

where $\mathbf{c}_{\hat{\phi}}$ is the vector of polynomial coefficients of $\hat{\phi}(\boldsymbol{\tau})$, with $\big(\mathbf{c}_{\hat{\phi}}\big)_{\mathbf{k}} := \frac{1}{\mathbf{k}!}\frac{\partial^{\mathbf{k}}\hat{\phi}}{\partial\boldsymbol{\tau}^{\mathbf{k}}}(\bar{\boldsymbol{\tau}})$, $\mathbf{k}$ the vector of monomial powers in the set $\mathcal{K}_n^N := \big\{(k_1, \dots, k_N) \in \mathbb{N}_0^N : 0 \le k_1 + \cdots + k_N \le n\big\}$ in the case of a polynomial of degree n, $\Delta\boldsymbol{\tau} := \boldsymbol{\tau} - \bar{\boldsymbol{\tau}}$ the deviation of $\boldsymbol{\tau}$ around $\bar{\boldsymbol{\tau}}$, $\mathbf{k}! := k_1! \dots k_N!$, $\Delta\boldsymbol{\tau}^{\mathbf{k}} := (\tau_1 - \bar{\tau}_1)^{k_1} \dots (\tau_N - \bar{\tau}_N)^{k_N}$, $\frac{\partial^{\mathbf{k}}}{\partial\boldsymbol{\tau}^{\mathbf{k}}} := \frac{\partial^{k_1+\cdots+k_N}}{\partial\tau_1^{k_1}\dots\partial\tau_N^{k_N}}$, and $R_{\hat{\phi}}(\boldsymbol{\tau})$ is the orthogonal part with respect to the polynomial basis.

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 fits 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 $l = 1, \dots, m_\tau$ [54]. 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 $l = 1, \dots, m_\tau$.

Remark 10. One could also avoid computing the partial derivatives $\frac{\partial\hat{\phi}}{\partial\boldsymbol{\tau}}(\boldsymbol{\tau}_l)$ and obtain a polynomial that fits only the value $\hat{\phi}(\boldsymbol{\tau}_l)$ at the sample points $\boldsymbol{\tau}_l$, for $l = 1, \dots, m_\tau$. This would require no more than computing the extended states $\mathbf{z}(t)$ that correspond to each point $\boldsymbol{\tau}_l$, which would amount to solving one system of n_z differential equations for each $l = 1, \dots, m_\tau$. Hence, this would entail solving m_τ systems of n_z differential equations to obtain m_τ values for interpolation. In contrast, the approach proposed in this paper requires solving only $2m_\tau$ systems of n_z differential equations to obtain $(N+1)m_\tau$ values and partial derivatives for interpolation, owing to the computation of the extended adjoint variables $\boldsymbol{\zeta}(t)$. For this reason, the latter approach was chosen.

Hence, one can compute the coefficient vector $\hat{\mathbf{c}}_{\hat{\phi}}$ that minimizes $\sum_{\boldsymbol{\kappa}\in\mathcal{K}_1^N} \|\mathbf{p}_{\hat{\phi},\boldsymbol{\kappa}} - \mathbf{A}_{\tau,\boldsymbol{\kappa}}\hat{\mathbf{c}}_{\hat{\phi}}\|^2$, where $\big(\hat{\mathbf{c}}_{\hat{\phi}}\big)_{\mathbf{k}}$ is an approximation of $\big(\mathbf{c}_{\hat{\phi}}\big)_{\mathbf{k}}$, for all $\mathbf{k} \in \mathcal{K}_n^N$, and

$$\big(\mathbf{p}_{\hat{\phi},\boldsymbol{\kappa}}\big)_l = \frac{\partial^{\boldsymbol{\kappa}}\hat{\phi}}{\partial\boldsymbol{\tau}^{\boldsymbol{\kappa}}}(\boldsymbol{\tau}_l), \quad \boldsymbol{\kappa} \in \mathcal{K}_1^N, \quad l = 1, \dots, m_\tau, \tag{68}$$

$$\big(\mathbf{A}_{\tau,\boldsymbol{\kappa}}\big)_{l,\mathbf{k}} = \begin{cases} \frac{\mathbf{k}!}{(\mathbf{k}-\boldsymbol{\kappa})!}\Delta\boldsymbol{\tau}_l^{\mathbf{k}-\boldsymbol{\kappa}}, & \mathbf{k} \ge \boldsymbol{\kappa} \\ 0, & \text{otherwise}, \end{cases} \quad \boldsymbol{\kappa} \in \mathcal{K}_1^N, \quad l = 1, \dots, m_\tau, \quad \mathbf{k} \in \mathcal{K}_n^N. \tag{69}$$

The vector of polynomial coefficients $\hat{\mathbf{c}}_{\hat{\phi}}$ is of dimension $\binom{N+n}{N}$, while the number of value vectors $\mathbf{p}_{\hat{\phi},\boldsymbol{\kappa}}$ 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. This way, although only the partial derivatives of $\hat{\phi}(\boldsymbol{\tau})$ up to first order with respect to $\boldsymbol{\tau}$ at the points $\boldsymbol{\tau}_l$ are fitted, for $l = 1, \dots, m_\tau$, the interpolating polynomial approximates the partial derivatives of $\hat{\phi}(\boldsymbol{\tau})$ of higher order (up to n) with respect to $\boldsymbol{\tau}$ in (67).

This yields the polynomial representation of $\hat{\phi}(\boldsymbol{\tau})$

$$p_{\hat{\phi}}(\boldsymbol{\tau}) = \sum_{\mathbf{k}\in\mathcal{K}_n^N} \big(\hat{\mathbf{c}}_{\hat{\phi}}\big)_{\mathbf{k}} \Delta\boldsymbol{\tau}^{\mathbf{k}}. \tag{70}$$

Remark 11. The polynomial function $p_{\hat{\phi}}(\boldsymbol{\tau})$ is used to approximate a mapping between the decision variables $\boldsymbol{\tau}$ and a function of the states $\mathbf{s}(t)$ at a finite number of times $t_1, \dots, t_T$ that do not include the switching times $\bar{t}_1, \dots, \bar{t}_{n_s}$ in $\boldsymbol{\tau}$. In other words, no switching time $\bar{t}_i$ is simultaneously related to the inputs and outputs of the mapping that is approximated by the polynomial function $p_{\hat{\phi}}(\boldsymbol{\tau})$. Hence, the polynomial function does not approximate the dependence of any function of the states $\mathbf{s}(t)$ on the generic time $t < t_f$.

Remark 12. To avoid non-smoothness of $\hat{\phi}(\boldsymbol{\tau})$ due to the existence of different sequences of arcs of types 1, 2, and 3 for the given sequence of arcs of types 1 and 3, the sample points $\boldsymbol{\tau}_l$ must be restricted to the ones that correspond to the optimal sequence of arcs of types 1, 2, and 3. The procedure is as follows. For the given sequence of arcs of types 1 and 3, sample points $\boldsymbol{\tau}_l$ are chosen. Different points $\boldsymbol{\tau}_l$ will lead to different sequences of arcs of types 1, 2, and 3, depending on (i) which state constraints become active and result in arcs of type 2 and (ii) the order of these arcs of type 2 with respect to the arcs of types 1 and 3. For example, suppose that, for a given sequence 1U-3-1U of arcs of types 1 and 3, it is known that the optimal sequence of arcs of types 1, 2, and 3 is 1U-3-1U-2, where the arc of type 2 is an arc with an active state constraint. In this example, some points $\boldsymbol{\tau}_l$ lead to the optimal sequence of arcs 1U-3-1U-2, while other points may lead to other sequences such as 1U-3-1U (without active state constraints) or 1U-2-3-1U (with a different order of the arcs of type 2 with respect to the arcs of types 1 and 3), among others. Then, only the points $\boldsymbol{\tau}_l$ that correspond to the optimal sequence of arcs of types 1, 2, and 3 (the sequence 1U-3-1U-2 in the example above) are used for the computation of the polynomial approximation $p_{\hat{\phi}}(\boldsymbol{\tau})$ in (70). This is done to avoid the non-smoothness of $\hat{\phi}(\boldsymbol{\tau})$ that would occur if all the points $\boldsymbol{\tau}_l$ were used to construct the polynomial approximation regardless of their sequences of arcs of types 1, 2, and 3. Hence, we consider the problem only for the optimal sequence of arcs of types 1, 2, and 3. To this end, we use a support vector machine $p_{\hat{\eta}_j}(\boldsymbol{\tau})$ with polynomial kernel to decide whether the points $\boldsymbol{\tau}$ are such that each entry point $\hat{\eta}_j(\boldsymbol{\tau})$ in arcs of type 2 is placed with respect to $\bar{t}_1, \dots, \bar{t}_{n_s}$ according to the optimal sequence of arcs of types 1, 2, and 3. To construct the support vector machine, the points $\boldsymbol{\tau}_l$ are classified in two groups: the points $\boldsymbol{\tau}_l$ that correspond to the optimal sequence of arcs of types 1, 2, and 3 (the sequence 1U-3-1U-2 in the example above) are labeled with the value 1, and the remaining points are labeled with the value -1. This is done to prevent the POP from searching values of $\boldsymbol{\tau}$ for which the corresponding sequence of arcs of types 1, 2, and 3 is not the optimal one.

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 the POP

$$\min_{\boldsymbol{\tau}} \quad -p_{\hat{\phi}}(\boldsymbol{\tau}), \quad \text{s.t.} \quad \mathbf{p}_{\hat{\boldsymbol{\gamma}}}(\boldsymbol{\tau}) \ge \mathbf{0}_{n_\gamma},$$

$$\left(\mathbf{p}_{\hat{\boldsymbol{y}}}(\boldsymbol{\tau})\right)_j := \begin{cases} h^{\eta}_j(\boldsymbol{\tau}),\ j=1,\dots,n_\eta, \\ h^{t}_{j-n_\eta}(\boldsymbol{\tau}),\ j=n_\eta+1,\dots,\bar{n}_\psi, \\ \underline{h}^{b}_{j-\bar{n}_\psi}(\boldsymbol{\tau}),\ j=\bar{n}_\psi+1,\dots,\bar{n}_\psi+|\mathcal{S}|, \\ \overline{h}^{b}_{j-\bar{n}_\psi-|\mathcal{S}|}(\boldsymbol{\tau}),\ j=\bar{n}_\psi+|\mathcal{S}|+1,\dots,\bar{n}_\psi+2|\mathcal{S}|, \\ \underline{h}^{e}_{j-\bar{n}_\psi-2|\mathcal{S}|}(\boldsymbol{\tau}),\ j=\bar{n}_\psi+2|\mathcal{S}|+1,\dots,\bar{n}_\psi+3|\mathcal{S}|, \\ \overline{h}^{e}_{j-\bar{n}_\psi-3|\mathcal{S}|}(\boldsymbol{\tau}),\ j=\bar{n}_\psi+3|\mathcal{S}|+1,\dots,n_\gamma, \end{cases} \tag{71}$$

where $\bar{n}_\psi := n_\eta + n_s + 1$, $n_\gamma := \bar{n}_\psi + 4|\mathcal{S}|$,

$$h^{\eta}_j(\boldsymbol{\tau}) := p_{\hat{\eta}_j}(\boldsymbol{\tau}), \quad j=1,\dots,n_\eta, \tag{72a}$$

$$h^{t}_i(\boldsymbol{\tau}) := \bar{t}_i - \bar{t}_{i-1}, \quad i=1,\dots,n_s+1, \tag{72b}$$

$$\begin{aligned} &\underline{h}^{b}_i(\boldsymbol{\tau}) := u^0_s - \underline{u}, \quad \overline{h}^{b}_i(\boldsymbol{\tau}) := \overline{u} - u^0_s, \\ &\underline{h}^{e}_i(\boldsymbol{\tau}) := u^0_s + p_s\left(\bar{t}_s - \bar{t}_{s-1}\right) - \underline{u}, \\ &\overline{h}^{e}_i(\boldsymbol{\tau}) := \overline{u} - u^0_s - p_s\left(\bar{t}_s - \bar{t}_{s-1}\right), \quad s=\mathcal{S}_i,\ s\in\mathcal{S}, \end{aligned} \tag{72c}$$

and $\hat{\boldsymbol{y}}(\boldsymbol{\tau})$ is defined as $\mathbf{p}_{\hat{\boldsymbol{y}}}(\boldsymbol{\tau})$. Note that the case of a single input is considered above for clarity.

Remark 13. The differential equations and initial conditions in (59d) are removed from (71) since the approximated function $\hat{\phi}(\boldsymbol{\tau})$ is replaced by its polynomial approximation $p_{\hat{\phi}}(\boldsymbol{\tau})$, which no longer depends on any differential equations or initial conditions.

The POP (71) 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 [35]. 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 [52,53]. To provide some key properties of this reformulation, we introduce the following definitions:

$$\varphi(\boldsymbol{\tau}) := J(\boldsymbol{\tau}) - \xi, \quad J(\boldsymbol{\tau}) := -p_{\hat{\phi}}(\boldsymbol{\tau}), \tag{73a}$$

$$g_j(\boldsymbol{\tau}) := \left(\mathbf{p}_{\hat{\boldsymbol{y}}}(\boldsymbol{\tau})\right)_j, \quad j=1,\dots,n_\gamma, \tag{73b}$$

$$g_{n_c}(\boldsymbol{\tau}) := r^{2v} - \textstyle\sum_{k=1}^{N}(\tau_k - \bar{\tau}_k)^{2v}, \tag{73c}$$

where r is a constant, $\varphi(\boldsymbol{\tau})$ is of degree $2v_0$ or $2v_0-1$ and $g_j(\boldsymbol{\tau})$ is of degree $2v_j$ or $2v_j-1$, with $c_d := \max_{j=1,\dots,n_c} v_j$, and the relaxation order $d \geq v := \max_{j=0,1,\dots,n_c} v_j$.

Then, the POP (71) is equivalent to the problem of computing the maximum ξ such that $\varphi(\boldsymbol{\tau})$ is strictly positive $\forall \boldsymbol{\tau} \in \mathbb{K} = \left\{\boldsymbol{\tau} : g_j(\boldsymbol{\tau}) \geq 0, \forall j = 1,\dots,n_\gamma\right\}$. The problem of computing the global minimum of $J(\boldsymbol{\tau})$ subject to $g_j(\boldsymbol{\tau}) \geq 0$, for $j=1,\dots,n_c$, can be formulated as an SDP for some relaxation order $d \geq v = \lceil n/2 \rceil$. A certificate of the representation in terms of sum-of-squares polynomials for the order d is obtained upon convergence of the SDP, which is a certificate of global optimality of the solution $\boldsymbol{\tau}^*_p$ and the cost $\xi^* = J^*$.

As for the complexity of this method, suppose that $c_d \geq 1$ and a global optimum is computed and certified for the relaxation order $d = 5$. This implies that an SDP has been solved with $\binom{N+2d}{N} = \frac{(N+10)\dots(N+1)}{3628800}$ equality constraints, one linear matrix inequality (LMI) of size $\binom{N+d}{N} = \frac{(N+5)(N+4)(N+3)(N+2)(N+1)}{120}$, and $n_c = n_\gamma + 1$ LMIs of size $\binom{N+d-v_j}{N} \leq \frac{(N+4)(N+3)(N+2)(N+1)}{24}$. Since the complexity of SDPs is polynomial in their input size, that is, the number of constraints and the size of the LMIs, it means that a global solution $\boldsymbol{\tau}^*_p$ is computed and certified in polynomial time.

When the globally optimal cost is known for each arc sequence, one can check which sequence is the best one. As mentioned in Section 5.1, the number of plausible arc sequences in single-input input-affine OCPs is less than $2^{\bar{n}_a}$. The number of decision variables for each arc sequence is $N = n_s + 2|\mathcal{S}| \leq 2\bar{n}_a$. This means that, even for a relatively large upper bound $\bar{n}_a = 5$, less than $2^{\bar{n}_a} = 32$ arc sequences would be considered, and the problem for each sequence can be solved in parallel and involves only $N \leq 2\bar{n}_a = 10$ decision variables.

5.4. Error due to polynomial approximation

Since the solution to the POP (71) is not exactly the same as the solution to Problem (59) due to the fact that the polynomial function $-p_{\hat{\phi}}(\boldsymbol{\tau})$ is an approximation of the cost function $-\hat{\phi}(\boldsymbol{\tau})$, the question arises as to whether one can quantify the error in the optimal solution and the optimal value of the cost of the POP.

Suppose that the global solution to the POP (71) is $\boldsymbol{\tau}^*_p$, for which n_a constraints $-\mathbf{p}_{\hat{\boldsymbol{y}}}(\boldsymbol{\tau}) \leq \mathbf{0}_{n_\gamma}$ given by a selection matrix $\mathbf{S}_a$ are active with Lagrange multipliers $\boldsymbol{\nu}^*_p \geq \mathbf{0}_{n_a}$. The Karush–Kuhn–Tucker (KKT) conditions for $\boldsymbol{\tau}^*_p$ are

$$-\tfrac{\partial p_{\hat{\phi}}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}^*_p)^{\mathrm{T}} - \tfrac{\partial \mathbf{p}_{\hat{\boldsymbol{y}}}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}^*_p)^{\mathrm{T}}\mathbf{S}_a^{\mathrm{T}}\boldsymbol{\nu}^*_p = \mathbf{0}_N, \tag{74a}$$

$$-\mathbf{S}_a\mathbf{p}_{\hat{\boldsymbol{y}}}(\boldsymbol{\tau}^*_p) = \mathbf{0}_{n_a}. \tag{74b}$$

We aim to obtain explicit expressions for (i) the difference $\boldsymbol{\delta\tau}$ between $\boldsymbol{\tau}^*_p$ and $\boldsymbol{\tau}^*$, the KKT point of Problem (59) that corresponds to $\boldsymbol{\tau}^*_p$, and (ii) the difference $\delta\hat{\phi}$ between $-p_{\hat{\phi}}(\boldsymbol{\tau}^*_p)$ and $-\hat{\phi}(\boldsymbol{\tau}^*)$, the cost of Problem (59) at $\boldsymbol{\tau}^*$. It is impossible to obtain exact and explicit expressions for these differences since that would involve infinite series expansions around $\boldsymbol{\tau}^*_p$ and would imply explicit solutions to high-degree polynomials for $\boldsymbol{\delta\tau}$ and $\delta\hat{\phi}$ and the Abel–Ruffini theorem states that there is no closed-form algebraic expression for the solution to general polynomial equations of degree five or higher with arbitrary coefficients [55]. However, one can obtain explicit expressions for the approximations of $\boldsymbol{\delta\tau}$ and $\delta\hat{\phi}$, as well as exact and implicit expressions that consider the variations of the second-order derivatives of the cost and Lagrangian functions and of the first-order derivatives of the constraint functions, which is done in the following theorem.

Theorem 1. *For a first-order approximation of the KKT conditions for Problem (59) and a second-order approximation of its cost, the explicit difference between the KKT points is*

$$\boldsymbol{\delta\tau} \simeq \left(\mathbf{L}_p - \mathbf{L}_p\mathbf{Z}_p^{\mathrm{T}}\left(\mathbf{Z}_p\mathbf{L}_p\mathbf{Z}_p^{\mathrm{T}}\right)^{-1}\mathbf{Z}_p\mathbf{L}_p\right)\tfrac{\partial \epsilon_{\hat{\phi}}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}^*_p)^{\mathrm{T}}, \tag{75}$$

while the explicit difference between the costs is

$$\delta\hat{\phi} \simeq \epsilon_{\hat{\phi}}(\boldsymbol{\tau}^*_p) - \tfrac{\partial \hat{\phi}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}^*_p)\boldsymbol{\delta\tau} - \tfrac{1}{2}\boldsymbol{\delta\tau}^{\mathrm{T}}\mathbf{H}_p\boldsymbol{\delta\tau}, \tag{76}$$

with the Lagrangian $\hat{\mathcal{L}}(\boldsymbol{\tau}, \boldsymbol{\nu}) := -\hat{\phi}(\boldsymbol{\tau}) - \boldsymbol{\nu}^{\mathrm{T}}\mathbf{S}_a\hat{\boldsymbol{y}}(\boldsymbol{\tau})$, the approximation error $\epsilon_{\hat{\phi}}(\boldsymbol{\tau}) := -p_{\hat{\phi}}(\boldsymbol{\tau}) + \hat{\phi}(\boldsymbol{\tau})$ for the cost, and the definitions $\mathbf{L}_p := -\tfrac{\partial^2 \hat{\mathcal{L}}}{\partial \boldsymbol{\tau}^2}(\boldsymbol{\tau}^*_p, \boldsymbol{\nu}^*_p)^{-1}$, $\mathbf{Z}_p := \mathbf{S}_a\tfrac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}^*_p)$, and $\mathbf{H}_p := -\tfrac{\partial^2 \hat{\phi}}{\partial \boldsymbol{\tau}^2}(\boldsymbol{\tau}^*_p)$.
Implicitly, the exact difference between the KKT points is

$$\boldsymbol{\delta\tau} = \left(\mathbf{L} - \mathbf{L}\mathbf{Z}^{\mathrm{T}}\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathrm{T}}\right)^{-1}\mathbf{Z}\mathbf{L}\right)\tfrac{\partial \epsilon_{\hat{\phi}}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}^*_p)^{\mathrm{T}}, \tag{77}$$

while the exact difference between the costs is

$$\delta\hat{\phi} = \epsilon_{\hat{\phi}}(\boldsymbol{\tau}^*_p) - \tfrac{\partial \hat{\phi}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}^*_p)\boldsymbol{\delta\tau} - \tfrac{1}{2}\boldsymbol{\delta\tau}^{\mathrm{T}}\mathbf{H}\boldsymbol{\delta\tau}, \tag{78}$$

with the definitions

$$\mathbf{L} := -\left(\textstyle\int_0^1 \tfrac{\partial^2 \hat{\mathcal{L}}}{\partial \boldsymbol{\tau}^2}(\boldsymbol{\tau}^*_p - \xi\boldsymbol{\delta\tau}, \boldsymbol{\nu}^*_p - \xi\boldsymbol{\delta\nu})\mathrm{d}\xi\right)^{-1}, \tag{79}$$

$$\mathbf{Z} := \textstyle\int_0^1 \mathbf{S}_a\tfrac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}^*_p - \xi\boldsymbol{\delta\tau})\mathrm{d}\xi, \tag{80}$$

$$\mathbf{H} := \textstyle\int_0^1 -2(1-\xi)\tfrac{\partial^2 \hat{\phi}}{\partial \boldsymbol{\tau}^2}(\boldsymbol{\tau}^*_p - \xi\boldsymbol{\delta\tau})\mathrm{d}\xi. \tag{81}$$

Proof. The KKT conditions for the solution $\boldsymbol{\tau}_p^* - \boldsymbol{\delta\tau}$ to Problem (59) are given by

$$\tfrac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}_p^*, \boldsymbol{\nu}_p^*)^{\mathrm{T}} + \mathbf{L}^{-1}\boldsymbol{\delta\tau} + \mathbf{Z}^{\mathrm{T}}\boldsymbol{\delta\nu} = \mathbf{0}_N, \tag{82a}$$

$$-\mathbf{S}_a\hat{\boldsymbol{\gamma}}(\boldsymbol{\tau}_p^*) + \mathbf{Z}\boldsymbol{\delta\tau} = \mathbf{0}_{n_a}. \tag{82b}$$

Upon using the first-order approximation of these KKT conditions, they become

$$\tfrac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}_p^*, \boldsymbol{\nu}_p^*)^{\mathrm{T}} + \mathbf{L}_p^{-1}\boldsymbol{\delta\tau} + \mathbf{Z}_p^{\mathrm{T}}\boldsymbol{\delta\nu} \simeq \mathbf{0}_N, \tag{83a}$$

$$-\mathbf{S}_a\hat{\boldsymbol{\gamma}}(\boldsymbol{\tau}_p^*) + \mathbf{Z}_p\boldsymbol{\delta\tau} \simeq \mathbf{0}_{n_a}. \tag{83b}$$

Hence, from (74), it holds that

$$\mathbf{L}^{-1}\boldsymbol{\delta\tau} + \mathbf{Z}^{\mathrm{T}}\boldsymbol{\delta\nu} = \tfrac{\partial \epsilon_{\hat{\phi}}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}_p^*)^{\mathrm{T}}, \tag{84a}$$

$$\mathbf{Z}\boldsymbol{\delta\tau} = \mathbf{0}_{n_a}, \tag{84b}$$

which yields (77) by using the blockwise inversion formula, while the approximation

$$\mathbf{L}_p^{-1}\boldsymbol{\delta\tau} + \mathbf{Z}_p^{\mathrm{T}}\boldsymbol{\delta\nu} \simeq \tfrac{\partial \epsilon_{\hat{\phi}}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}_p^*)^{\mathrm{T}}, \tag{85a}$$

$$\mathbf{Z}_p\boldsymbol{\delta\tau} \simeq \mathbf{0}_{n_a}, \tag{85b}$$

yields the explicit expression for $\boldsymbol{\delta\tau}$ in (75) by using the blockwise inversion formula.

Then, since the cost of Problem (59) at $\boldsymbol{\tau}^*$ is given by

$$-\hat{\phi}(\boldsymbol{\tau}_p^* - \boldsymbol{\delta\tau}) = -\hat{\phi}(\boldsymbol{\tau}_p^*) + \tfrac{\partial \hat{\phi}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}_p^*)\boldsymbol{\delta\tau} + \tfrac{1}{2}\boldsymbol{\delta\tau}^{\mathrm{T}}\mathbf{H}\boldsymbol{\delta\tau}, \tag{86}$$

(78) holds and one can use the second-order approximation

$$-\hat{\phi}(\boldsymbol{\tau}_p^* - \boldsymbol{\delta\tau}) \simeq -\hat{\phi}(\boldsymbol{\tau}_p^*) + \tfrac{\partial \hat{\phi}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}_p^*)\boldsymbol{\delta\tau} + \tfrac{1}{2}\boldsymbol{\delta\tau}^{\mathrm{T}}\mathbf{H}_p\boldsymbol{\delta\tau} \tag{87}$$

to obtain the explicit expression for $\delta\hat{\phi}$ in (76). □

Remark 14. Theorem 1 only provides an explicit expression for the first-order approximation of the difference $\boldsymbol{\delta\tau}$ between $\boldsymbol{\tau}_p^*$, the global solution to the POP (71), and $\boldsymbol{\tau}^*$, the KKT point of Problem (59) that corresponds to $\boldsymbol{\tau}_p^*$. This means that $\boldsymbol{\tau}_p^* - \boldsymbol{\delta\tau}$ is a good approximation for $\boldsymbol{\tau}^*$ with an explicit expression. One can obtain the exact KKT point $\boldsymbol{\tau}^*$ of Problem (59) that corresponds to $\boldsymbol{\tau}_p^*$ via local optimization of Problem (59) with initial guess $\boldsymbol{\tau}_p^* - \boldsymbol{\delta\tau}$.

Moreover, one can assess the quality of the solution $\boldsymbol{\tau}^*$ obtained by solving the POP (71) to global optimality followed by local optimization of Problem (59). To this end, the following theorem shows that the difference between the cost $-\hat{\phi}(\boldsymbol{\tau}^*)$ obtained from (71) and the globally optimal cost of (59) is bounded and depends on the polynomial approximation error $\epsilon_{\hat{\phi}}$ defined in Theorem 1.

Theorem 2. *If $\delta\hat{\phi}_{max}^{\mathrm{KKT}}$ is the maximum difference between the costs of any KKT point of the POP (71) and any corresponding KKT point of Problem (59), then the difference between $-\hat{\phi}(\boldsymbol{\tau}^*)$ and the cost of Problem (59) at its globally optimal solution is at most $\delta\hat{\phi}_{max}^{\mathrm{KKT}} - \delta\hat{\phi}$ and is bounded if the error $\epsilon_{\hat{\phi}}$ is bounded.*

Proof. The globally optimal solution to Problem (59) is a KKT point $\boldsymbol{\tau}^{\mathrm{KKT}}$ of Problem (59) that corresponds to some KKT point $\boldsymbol{\tau}_p^{\mathrm{KKT}}$ of the POP (71), and $\hat{\phi}(\boldsymbol{\tau}^{\mathrm{KKT}}) - \hat{\phi}(\boldsymbol{\tau}^*) - \delta\hat{\phi}_{max}^{\mathrm{KKT}} + \delta\hat{\phi} \leq p_{\hat{\phi}}(\boldsymbol{\tau}_p^{\mathrm{KKT}}) - p_{\hat{\phi}}(\boldsymbol{\tau}_p^*) \leq 0$ since $\boldsymbol{\tau}_p^*$ is the globally optimal solution to the POP (71). In addition, from Theorem 1, $\delta\hat{\phi}_{max}^{\mathrm{KKT}}$ and $\delta\hat{\phi}$ depend on $\epsilon_{\hat{\phi}}$. □

6. Design requirements and computational complexity

In this paper, several steps are taken to allow obtaining a tractable formulation of the Bayesian OED problem. In particular, several functions are used as expected utilities:

- The original expected utility $u_{KL}(\mathbf{d})$ based on the utility function $U_{KL}(\boldsymbol{\theta}, \mathbf{d})$ that corresponds to a KL divergence.
- The approximate expected utilities $u_D(\mathbf{d})$ or $u_{MC}(\mathbf{d})$ based on the utility function $U_D(\boldsymbol{\theta}, \mathbf{d})$ that corresponds to a function of the FIM or the utility function $U_{MC}(\boldsymbol{\theta}, \mathbf{d})$ that corresponds to Monte Carlo integration in the observation space.
- The approximation $\hat{u}(\mathbf{d})$ that is obtained when the multivariate integrals $u_D(\mathbf{d})$, $u_{MC}(\mathbf{d})$, or $p(\mathbf{y}|\mathbf{d})$ in the space of parameters $\boldsymbol{\theta}$ are computed via Gaussian quadrature, which results in a cost $-\phi\big(\mathbf{s}(t_1), \ldots, \mathbf{s}(t_T)\big)$ for the reformulation as an OCP.
- The metric $\hat{\phi}(\boldsymbol{\tau})$ that is obtained when $\hat{u}(\mathbf{d})$ is restricted to designs $\mathbf{d}$ that correspond to a certain arc sequence.
- The polynomial approximation $p_{\hat{\phi}}(\boldsymbol{\tau})$ of $\hat{\phi}(\boldsymbol{\tau})$.

Also, several functions related to each chance path constraint are used:

- The original function $q_k(t; \mathbf{d})$ that specifies the probability of satisfaction of the path constraint $h_k\big(\mathbf{x}(t; \boldsymbol{\theta})\big) \leq 0$.
- The first two moments $m_k^{\mathrm{I}}(t; \mathbf{d})$ and $m_k^{\mathrm{II}}(t; \mathbf{d})$ of $h_k\big(\mathbf{x}(t; \boldsymbol{\theta})\big)$ that express the moment-based approximation of the chance path constraint.
- The approximations $\hat{m}_k^{\mathrm{I}}(t; \mathbf{d})$ and $\hat{m}_k^{\mathrm{II}}(t; \mathbf{d})$ that are obtained when the multivariate integrals $m_k^{\mathrm{I}}(t; \mathbf{d})$ and $m_k^{\mathrm{II}}(t; \mathbf{d})$ in the space of parameters $\boldsymbol{\theta}$ are computed via Gaussian quadrature, which result in a path constraint $\mu_k\big(\mathbf{s}(t)\big) \leq 0$ for the reformulation as an OCP.

Accordingly, the computational procedure can be summarized as follows:

1. A set of plausible arc sequences is chosen for the solution to the OCP that results from reformulation of the OED problem, which includes only sequences with a number of arcs no larger than some upper bound $\bar{n}_a$ and without consecutive arcs of the same type. Less than $2^{\bar{n}_a}$ arc sequences would be considered for a single-input input-affine OCP, and the problem for each sequence can be solved in parallel. Note that, even for a relatively large upper bound $\bar{n}_a = 5$, the problem involves only $N \leq 2\bar{n}_a = 10$ decision variables for a single-input OCP.
2. To compute a global solution $\boldsymbol{\tau}_p^*$ to the POP for a given arc sequence, the polynomial approximation $p_{\hat{\phi}}(\boldsymbol{\tau})$ must be obtained. To this end, the value $\hat{\phi}(\boldsymbol{\tau}_l)$ and partial derivatives $\frac{\partial \hat{\phi}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}_l)$ are fitted at the sample points $\boldsymbol{\tau}_l$, for $l = 1, \ldots, m_\tau$. The number m_τ of sample points must be at least $\frac{(N+n)!}{n!(N+1)!}$, which is polynomial in N since the degree n of the polynomial approximation is typically bounded.
3. For each sample point $\boldsymbol{\tau}_l$, the extended states $\mathbf{z}(t)$ and adjoint variables $\boldsymbol{\zeta}(t)$ for $\hat{\phi}(\boldsymbol{\tau})$ are computed. This amounts to solving two systems of n_z differential equations for each $l = 1, \ldots, m_\tau$, where $n_z = n_r + n_{z,1} + \cdots + n_{z,n_s+1}$, $n_{z,1} + \cdots + n_{z,n_s+1} = 2|\mathcal{S}| \leq \bar{n}_a$ for a single-input OCP, and $n_r = n_x m_\theta$, where n_x is the number of states of the dynamical system. Once the differential equations are solved for a given $\boldsymbol{\tau}_l$, $\hat{\phi}(\boldsymbol{\tau}_l)$ and $\frac{\partial \hat{\phi}}{\partial \boldsymbol{\tau}}(\boldsymbol{\tau}_l)$ are computed by evaluating the pdf $\pi(\cdot|\bar{\mathbf{x}}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}})$ $m_\theta m_\xi(1 + m_\theta)$ times, where m_ξ is the number of samples used for Monte Carlo integration in the observation space.
4. The number m_θ of quadrature points must be equal to $\left\lceil \frac{|\bar{\mathcal{K}}_{\bar{n}}^{n_\theta}|}{n_\theta+1} \right\rceil \leq \left\lceil \frac{|\mathcal{K}_{\bar{n}}^{n_\theta}|}{n_\theta+1} \right\rceil = \left\lceil \frac{(n_\theta+\bar{n})!}{\bar{n}!(n_\theta+1)!} \right\rceil$ for exact integration of polynomials of a given degree $\bar{n}$ using the multivariate equivalent of Gaussian quadrature, where truncation schemes can be used to introduce sparsity and to eliminate many elements of $\mathcal{K}_{\bar{n}}^{n_\theta}$ from $\bar{\mathcal{K}}_{\bar{n}}^{n_\theta}$ in the case of a large number n_θ of parameters.

5. Once the polynomial approximation $p_{\hat{\phi}}(\boldsymbol{\tau})$ is obtained, the global solution to the POP is computed by solving a hierarchy of convex SDPs of increasing relaxation order, where each SDP is solved in polynomial time with respect to N and the relaxation order.
6. Finally, the exact solution $\boldsymbol{\tau}^*$ of the OCP that corresponds to $\boldsymbol{\tau}_p^*$ is obtained via local optimization of the OCP with initial guess $\boldsymbol{\tau}_p^* - \boldsymbol{\delta\tau}$, which requires an additional sample point for each iteration of the local optimization solver.

From the procedure above, one can observe that the main computational bottleneck is the solution of differential equations. In total, at least $2\frac{(N+n)!}{n!(N+1)!}n_x m_\theta$ differential equations must be solved for each arc sequence, where $N \leq 2\bar{n}_a = 10$ for $\bar{n}_a = 5$. On the other hand, the use of an efficient method for multivariate integration in the space of parameters based on sparse stochastic collocation and Gaussian quadrature is an important feature of the proposed approach since it reduces the number m_θ of quadrature points, particularly in the case of large n_θ, and ensures that $m_\theta \leq \left\lceil \frac{(n_\theta+\bar{n})!}{\bar{n}!(n_\theta+1)!} \right\rceil$. In particular, m_θ is much smaller than what it would be if Monte Carlo integration had been used for multivariate integration in the space of parameters. Also, the number of differential equations for each arc sequence does not depend on the number T of time instants at which measurements are available since T may only affect the number m_ξ of samples used for Monte Carlo integration in the observation space. In turn, m_ξ only affects the number $m_\theta m_\xi(1 + m_\theta)$ of times that the pdf $\pi(\cdot|\bar{\mathbf{x}}, \boldsymbol{\Sigma}_\xi)$ is evaluated for each sample point $\boldsymbol{\tau}_l$, which is cheaper than the solution of $2n_x m_\theta$ differential equations. In summary, the computational complexity for each arc sequence is polynomial in N, n_x, and n_θ for fixed n and $\bar{n}$, and the problem for each sequence can be solved in parallel, which makes the computational procedure tractable.

7. Case study

The proposed methods for approximate Bayesian OED are 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 has been used as a testbed for optimal control [56,57] and OED [25,58]. The nondimensional governing equations are given as

$$\dot{x}_1(t) = \quad x_1(t) - (1+\theta_1)\, x_1(t)x_2(t) - 0.4x_1(t)u_1(t), \tag{88a}$$

$$\dot{x}_2(t) = -x_2(t) + (1+\theta_2)\, x_1(t)x_2(t) - 0.2x_2(t)u_1(t), \tag{88b}$$

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 (88) is integrated using the initial conditions $\mathbf{x}_0(\boldsymbol{\theta}) = [0.5, 0.7]$. It is assumed that we can measure the predator population at the final time step, i.e., $y(t_f) = x_2(t_f; \boldsymbol{\theta}) + e(t_f)$. The designed input is allowed to attain values in a predefined interval, i.e., $u_1(t) \in [0, 1], \forall t \in [0, t_f]$. Two cases are studied: in the first one, the measurement error is state-independent and normally distributed, the parameters follow a normal prior pdf, and no chance constraint is considered; in the second one, the measurement error is state-dependent, the parameters follow a beta prior pdf, and a chance constraint is considered.

7.1. Gaussian distributions with no chance constraint

In this first case, the measurement error $e(t_f)$ is state-independent and normally distributed, and its variance is assumed to be constant and equal to $\sigma^2 = 0.1^2$. The uncertain parameters follow a bivariate normal prior pdf $p(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ with mean $\bar{\boldsymbol{\theta}} = \mathbf{0}_2$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = 0.2^2\mathbf{I}_2$. For this case, which corresponds to Assumptions 1 and 2, we consider the OCP (51) and use the notation $\hat{\phi}(\boldsymbol{\tau}) \coloneqq \phi_D\big(\mathbf{S}(t_1), \ldots, \mathbf{S}(t_T)\big)$.

When linear functions are used to approximate free/singular arcs, a locally optimal solution consists of 3 arcs: the first arc is free/singular with $\underline{u}_1 < u_1^*(t) < \bar{u}_1$, for which an approximation by a linear function is used; in the second arc, $u_1^*(t) = \bar{u}_1$; and in the third arc, $u_1^*(t) = \underline{u}_1$. This results in an input trajectory described by the 4 decision variables $\bar{t}_1$, $\bar{t}_2$, $u_{1,1}^0$, $p_{1,1}$. 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 slope of the linear function that describes $u_1^*(t)$ in this arc: $u_{1,1}^{0*} = 0.482$, $p_{1,1}^* = -0.090$. The optimal metric is $\hat{\phi}(\boldsymbol{\tau}^*) = 11.6706$. The local optimality is indicated by the fact that the gradients (64), (65) are equal to zero and the solution satisfies approximately the necessary conditions given by Pontryagin's maximum principle [49].

We use the $m_\theta = 12$ quadrature points and weights shown in Fig. 1 to compute $\hat{u}_D(\mathbf{d})$ via integration of $U_D(\boldsymbol{\theta}, \mathbf{d})$. This corresponds to exact integration of Hermite polynomials (the orthogonal polynomials with respect to the normal distribution of $p(\boldsymbol{\theta})$) up to degree $\bar{n} = 7$ using the multivariate equivalent of Gaussian quadrature, with weights $\mathbf{w}$ and points $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{m_\theta}$. Note that, even for this problem with $n_\theta = 2$ parameters, the use of a univariate Gaussian quadrature followed by a tensor product rule for the choice of points and weights would already require 16 quadrature points. The input $u_1^*(t)$ and the measured state $x_2^*(t; \boldsymbol{\theta})$ for the realizations $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{m_\theta}$ are shown in Fig. 2, which indicates that $x_2^*(t_f; \boldsymbol{\theta})$ is sensitive to variations of $\boldsymbol{\theta}$.

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 1 reports the execution time of the procedure on an Intel Core i7 3.4 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 a 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 the very last step of local optimization with an initial guess that is already close to the solution $\boldsymbol{\tau}^*$, MATLAB's *fmincon* function with optimality tolerance of 10^{-12} and step tolerance of 10^{-15} is used. 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 a low relaxation order 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 9.2. The table is read as: for each arc sequence, the optimal values $\boldsymbol{\tau}^* = (\bar{t}_1^*, \bar{t}_2^*, u_{1,i}^{0*}, p_{1,i}^*)$ of the decision variables allow obtaining the optimal metric $\hat{\phi}(\boldsymbol{\tau}^*) = \hat{u}_D(\mathbf{d}_\tau^*)$, where the design $\mathbf{d}_\tau^*$ corresponds to $\boldsymbol{\tau}^*$ for each arc sequence. For the design $\mathbf{d}_\tau^*$, accurate approximations of $u_D(\mathbf{d}_\tau^*)$ and $u_{KL}(\mathbf{d}_\tau^*)$ are also computed. One can observe that $\hat{u}_D(\mathbf{d}_\tau^*)$ overestimates $u_D(\mathbf{d}_\tau^*)$ and $u_{KL}(\mathbf{d}_\tau^*)$ consistently. For comparison, a design $\mathbf{d}_{PRBS}$ corresponding to a pseudorandom binary sequence (PRBS) of size 31 would allow obtaining significantly worse results: $\hat{u}_D(\mathbf{d}_{PRBS}) = 8.9247$, $u_D(\mathbf{d}_{PRBS}) = 8.9200$, $2u_{KL}(\mathbf{d}_{PRBS}) + K = 8.8370$. Moreover, the execution time is below 900 s for all arc sequences, and the sequence with the best optimal metrics is 3-1U-1L (highlighted in bold in the table), that is, the sequence of the locally optimal solution. In addition, the globally optimal values $\bar{t}_1^*$, $\bar{t}_2^*$, $u_{1,1}^{0*}$, $p_{1,1}^*$ of the decision variables for that arc sequence also correspond

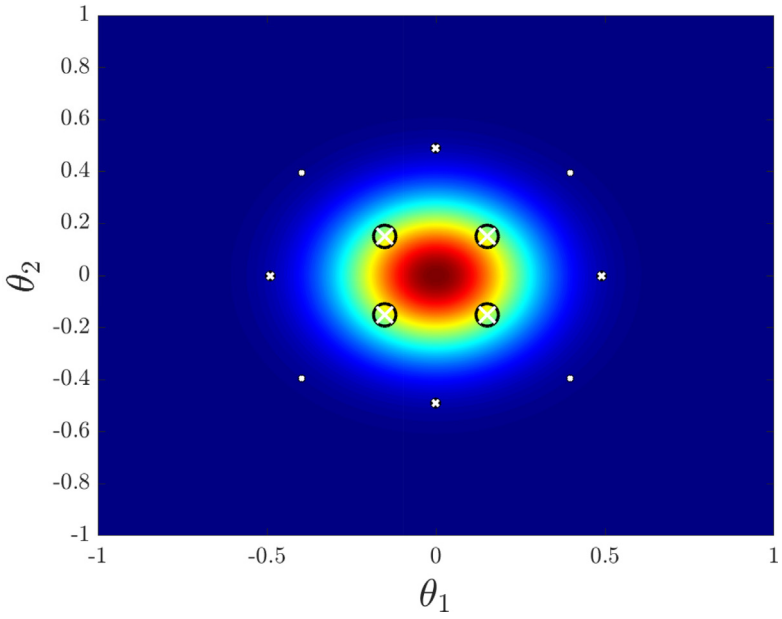

Fig. 1. Contour plot of the prior pdf $p(\boldsymbol{\theta})$ in Section 7.1 and location of the $m_\theta = 12$ quadrature points (indicated by markers that consist of white crosses inside black circles) used for integration via Gaussian quadrature. The marker size represents the weight of each point.

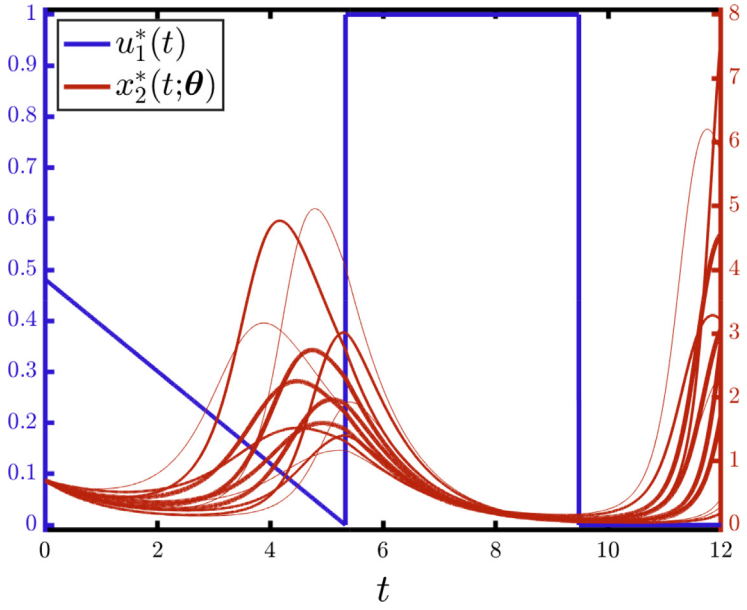

Fig. 2. Optimal input trajectory (in blue) for the Bayesian OED problem in Section 7.1 with the approximation of the free/singular arc using a linear function. The trajectories of the measured state (in red) are juxtaposed for the $m_\theta = 12$ realizations $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{m_\theta}$ used for multivariate integration. The relative width of the lines corresponds to the weights $\mathbf{w}$ of these realizations.

to the optimal values given by the locally optimal solution. For this optimal design, the covariance matrix of the posterior parameter distribution computed as in Appendix D becomes $\begin{bmatrix} 0.01850 & 0.00938 \\ 0.00938 & 0.01569 \end{bmatrix}$, while for the design based on a PRBS the covariance matrix would become $\begin{bmatrix} 0.01898 & -0.01665 \\ -0.01665 & 0.02566 \end{bmatrix}$, which is worse.

In summary, one can show that the locally optimal solution to the Bayesian OED problem shown in Fig. 2 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 900 s. Based on the computational complexity reported in Section 6, if the method for multivariate integration in the space of parameters based on sparse stochastic collocation and Gaussian quadrature is chosen such that the number m_θ of quadrature points depends quadratically on the number n_θ of parameters, a similar problem with $n_x = 10$ states of the dynamical system and $n_\theta = 10$ parameters could still be considered tractable on a standard laptop. Recall that, if we had only used local optimization to compute a local solution to (51), 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. For example, note that, since even the approximation and reformulation of the Bayesian OED problem results in an optimization problem with 4 decision variables, it is impossible to show that $\boldsymbol{\tau}^*$ or any other local solution is the globally optimal solution by simply plotting the cost $-\hat{\phi}(\boldsymbol{\tau})$ in the

Table 1
Execution time, optimal metrics $\hat{\phi}(\boldsymbol{\tau}^*) = \hat{u}_D(\mathbf{d}_\tau^*)$, $u_D(\mathbf{d}_\tau^*)$, $2u_{KL}(\mathbf{d}_\tau^*) + K$, where $K = \log\left(\det\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)\right)$, and optimal values $\bar{t}_1^*$, $\bar{t}_2^*$, $u_{1,i}^{0*}$, $p_{1,i}^*$ of the decision variables for the global solution to the Bayesian OED problem in Section 7.1, for different arc sequences and the final time $t_f = 12$. The arc sequence with the best optimal value of $\hat{\phi}(\boldsymbol{\tau}^*)$ is highlighted in bold.

Arc sequence	Execution time (s)	$\hat{u}_D(\mathbf{d}_\tau^*)$	$u_D(\mathbf{d}_\tau^*)$	$2u_{KL}(\mathbf{d}_\tau^*) + K$	$\bar{t}_1^*$	$\bar{t}_2^*$	$u_{1,i}^{0*}$	$p_{1,i}^*$
3-1L-1U	866	11.1308	11.0959	10.9337	2.320	12.000	1.000	0.000 ($i = 1$)
3-1U-1L	**876**	**11.6706**	**11.6309**	**11.4050**	**5.334**	**9.477**	**0.482**	−0.090 ($i = 1$)
1L-3-1L	716	11.5277	11.4599	11.1177	5.130	10.158	1.000	0.000 ($i = 2$)
1L-3-1U	835	10.6843	10.6181	10.4023	4.978	12.000	1.000	−0.129 ($i = 2$)
1U-3-1L	690	11.4730	11.4509	11.2201	1.794	9.260	0.000	0.134 ($i = 2$)
1U-3-1U	779	11.1308	11.0959	10.9337	2.320	12.000	0.000	0.000 ($i = 2$)

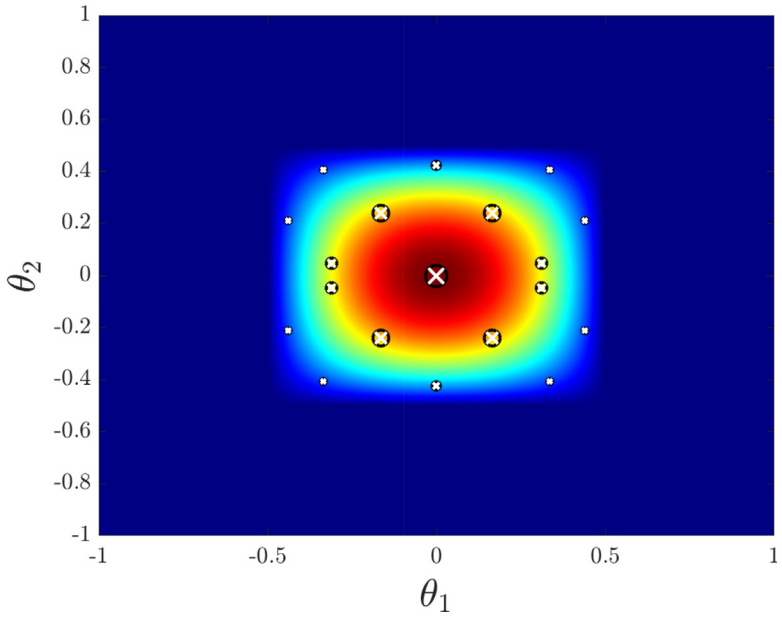

Fig. 3. Contour plot of the prior pdf $p(\boldsymbol{\theta})$ in Section 7.2 and location of the $m_\theta = 19$ quadrature points (indicated by markers that consist of white crosses inside black circles) used for integration via Gaussian quadrature. The marker size represents the weight of each point.

4-dimensional space of decision variables, but the methods in this paper guarantee that $\boldsymbol{\tau}^*$ is indeed the globally optimal solution.

7.2. Arbitrary distributions with a chance constraint

In this second case, the measurement error $e(t_f) = \left(1 + 0.1x_2(t_f;\boldsymbol{\theta})\right)\xi_1$ is state-dependent, ξ_1 is normally distributed, and its variance is assumed to be constant and equal to $\sigma^2 = 0.1^2$. The uncertain parameters follow a bivariate prior pdf $p(\boldsymbol{\theta}) = \prod_{i=1}^{2} B(1/2 - \theta_i|a, b)$, with $a = 2$ and $b = 2$, where $B(x|a, b)$ is the pdf of a univariate beta distribution with parameters a and b, which corresponds to a prior distribution with mean $\bar{\boldsymbol{\theta}} = \mathbf{0}_2$ and covariance $0.05\mathbf{I}_2$. Furthermore, the approximate chance constraint $\mathrm{E}\left[h_1\left(\mathbf{x}(t;\boldsymbol{\theta})\right)\right] + r_1\sqrt{\mathrm{V}\left[h_1\left(\mathbf{x}(t;\boldsymbol{\theta})\right)\right]} \leq 0$ is enforced for $h_1\left(\mathbf{x}(t;\boldsymbol{\theta})\right) = x_1(t;\boldsymbol{\theta}) - \bar{x}_1$, with $\bar{x}_1 = 6.6$ and $r_1 = \phi^{-1}(0.95)$. For this case, which corresponds to Assumptions 3 and 4, we consider the OCP (56) and use the notation $\hat{\phi}(\boldsymbol{\tau}) := \phi_Q\left(\mathbf{s}(t_1), \ldots, \mathbf{s}(t_T)\right)$.

When linear functions are used to approximate free/singular arcs, a locally optimal solution consists of the same 3 arcs as in the first case. This results in an input trajectory described by the 4 decision variables $\bar{t}_1$, $\bar{t}_2$, $u_{1,1}^0$, $p_{1,1}$. The optimal switching times are $\bar{t}_1^* = 5.206$, $\bar{t}_2^* = 9.330$. The optimal initial conditions for the first arc are the initial value and the constant slope of the linear function that describes $u_1^*(t)$ in this arc: $u_{1,1}^{0*} = 0.538$, $p_{1,1}^* = -0.103$. The optimal metric is $\hat{\phi}(\boldsymbol{\tau}^*) = 2.1724$. The local optimality is indicated by the fact that the gradients (64), (65) are equal to zero and the solution satisfies approximately the necessary conditions given by Pontryagin's maximum principle [49].

We use the $m_\theta = 19$ quadrature points and weights shown in Fig. 3 to compute $\hat{u}_Q(\mathbf{d})$ via integration of $U_Q(\boldsymbol{\theta}, \mathbf{d})$, which requires the computation of $\hat{p}(\mathbf{y}|\mathbf{d})$ via integration of $p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d})$, as well as $\hat{m}_1^{\mathrm{I}}(t;\mathbf{d})$ and $\hat{m}_1^{\mathrm{II}}(t;\mathbf{d})$ via integration of $M_1^{\mathrm{I}}(t;\boldsymbol{\theta},\mathbf{d})$ and $M_1^{\mathrm{II}}(t;\boldsymbol{\theta},\mathbf{d})$. This corresponds to exact integration of Jacobi polynomials (the orthogonal polynomials with respect to the beta distribution of $p(\boldsymbol{\theta})$) up to degree $\bar{n} = 9$ using the multivariate equivalent of Gaussian quadrature, with weights $\mathbf{w}$ and points $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{m_\theta}$. Note that, even for this problem with $n_\theta = 2$ parameters, the use of a univariate Gaussian quadrature followed by a tensor product rule for the choice of points and weights would already require 25 quadrature points. Also, note that a larger polynomial degree is used in this case due to the nested computation of multivariate integrals in (43). The input $u_1^*(t)$ and the measured state $x_2^*(t;\boldsymbol{\theta})$ for the realizations $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{m_\theta}$ are shown in Fig. 4, which indicates that $x_2^*(t_f;\boldsymbol{\theta})$ is sensitive to variations of $\boldsymbol{\theta}$. The mean $\pm\, r_1$ standard deviations of the state $x_1^*(t;\boldsymbol{\theta})$ subject to the upper bound $\bar{x}_1 = 6.6$ is also shown in Fig. 4. One can observe that the upper bound is marginally satisfied. As mentioned in Section 3.3, enforcing the approximate chance constraint $\mathrm{E}\left[h_1\left(\mathbf{x}(t;\boldsymbol{\theta})\right)\right] + r_1\sqrt{\mathrm{V}\left[h_1\left(\mathbf{x}(t;\boldsymbol{\theta})\right)\right]} \leq 0$ for the chosen value of r_1 does not ensure satisfaction of the chance path constraint (7) for $h_1\left(\mathbf{x}(t;\boldsymbol{\theta})\right)$. To achieve this goal, one would have to apply the iterative procedure proposed in Section 3.3 to determine the correct value of r_1. However, here we only show the result of the approach for obtaining global solutions to Bayesian OED problems for the chosen value of r_1, and the result of the iterative procedure in Section 3.3 is not shown for the sake of brevity.

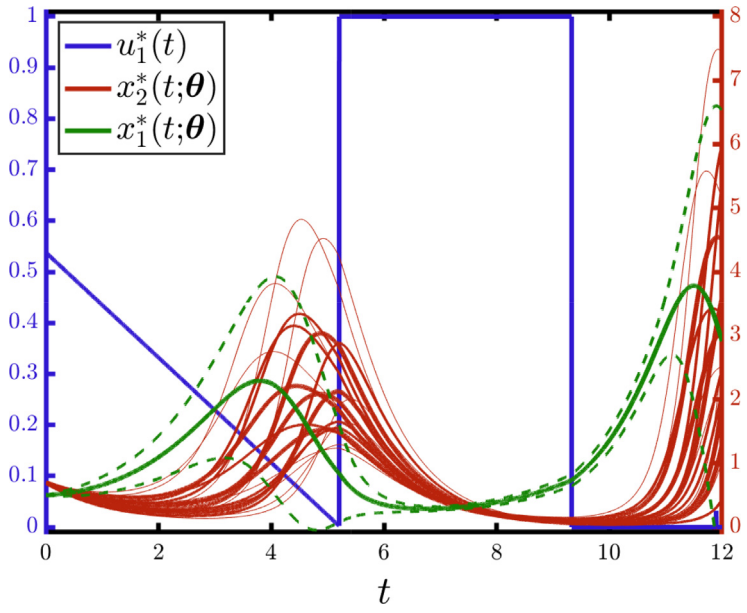

Fig. 4. Optimal input trajectory (in blue) for the Bayesian OED problem in Section 7.2 with the approximation of the free/singular arc using a linear function. The trajectories of the measured state (in red) are juxtaposed for the $m_\theta = 19$ realizations $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{m_\theta}$ used for multivariate integration. The relative width of the lines corresponds to the weights $\mathbf{w}$ of these realizations. The trajectory of the state subject to an upper bound (in green) is represented in terms of its mean (solid line) $\pm\, r_1$ standard deviations (dashed lines).

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 2 reports the execution time of the procedure on an Intel Core i7 3.4 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 $m_\tau = 6000$ sample points to obtain a polynomial representation $p_{\hat{\phi}}(\boldsymbol{\tau})$ of degree $n = 8$ with an approximation error similar to the previous case 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 the very last step of local optimization with an initial guess that is already close to the solution $\boldsymbol{\tau}^*$, MATLAB's *fmincon* function with optimality tolerance of 10^{-12} and step tolerance of 10^{-15} is used. 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 a low relaxation order 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 9.2. The table is read as: for each arc sequence, the optimal values $\boldsymbol{\tau}^* = (\bar{t}_1^*, \bar{t}_2^*, u_{1,i}^{0*}, p_{1,i}^*)$ of the decision variables allow obtaining the optimal metric $\hat{\phi}(\boldsymbol{\tau}^*) = \hat{u}_Q(\mathbf{d}_\tau^*)$, where the design $\mathbf{d}_\tau^*$ corresponds to $\boldsymbol{\tau}^*$ for each arc sequence. For the design $\mathbf{d}_\tau^*$, accurate approximations of $u_{KL}(\mathbf{d}_\tau^*)$ are also computed. One can observe that $\hat{u}_Q(\mathbf{d}_\tau^*)$ underestimates $u_{KL}(\mathbf{d}_\tau^*)$ consistently. For comparison, a design $\mathbf{d}_{PRBS}$ corresponding to a pseudorandom binary sequence (PRBS) of size 31 would allow obtaining significantly worse results: $\hat{u}_Q(\mathbf{d}_{PRBS}) = 1.1113$, $u_{KL}(\mathbf{d}_{PRBS}) = 1.2467$. Moreover, the execution time is below 3600 s for all arc sequences, and the sequence with the best optimal metrics is 3-1U-1L (highlighted in bold in the table), that is, the sequence of the locally optimal solution. In addition, the globally optimal values $\bar{t}_1^*$, $\bar{t}_2^*$, $u_{1,1}^{0*}$, $p_{1,1}^*$ of the decision variables for that arc sequence also correspond to the optimal values given by the locally optimal solution. For this optimal design, the covariance matrix of the posterior parameter distribution computed as in Appendix D becomes $\begin{bmatrix} 0.02030 & 0.00812 \\ 0.00812 & 0.01683 \end{bmatrix}$, while for the design based on a PRBS the covariance matrix would become $\begin{bmatrix} 0.02378 & -0.02153 \\ -0.02153 & 0.03162 \end{bmatrix}$, which is worse.

In summary, one can show that the locally optimal solution to the Bayesian OED problem shown in Fig. 4 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 3600 s. Based on the computational complexity reported in Section 6, if the method for multivariate integration in the space of parameters based on sparse stochastic collocation and Gaussian quadrature is chosen such that the number m_θ of quadrature points depends quadratically on the number n_θ of parameters, a similar problem with $n_x = 10$ states of the dynamical system and $n_\theta = 10$ parameters could still be considered tractable on a standard laptop. Recall that, if we had only used local optimization to compute a local solution to (56), 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. For example, note that, since even the approximation and reformulation of the Bayesian OED problem results in an optimization problem with 4 decision variables, it is impossible to show that $\boldsymbol{\tau}^*$ or any other local solution is the globally optimal solution by simply plotting the cost $-\hat{\phi}(\boldsymbol{\tau})$ in the 4-dimensional space of decision variables, but the methods in this paper guarantee that $\boldsymbol{\tau}^*$ is indeed the globally optimal solution.

8. Conclusions

This paper presented solution methods for obtaining globally optimal solutions to approximate Bayesian OED problems for two cases: normal prior and observation noise distributions, and arbitrary prior and observation noise distributions. The expected utility for KL divergence was approximated as a Bayes D-optimality criterion in the former case and as Monte Carlo integration in the observation space in the latter case. The proposed methods also deal with designed initial conditions and chance path constraints. Numerical tractability was reinforced by a sparse stochastic collocation scheme that required only a few points in the parameter space for approximating the expected

Table 2
Execution time, optimal metrics $\hat{\phi}(\boldsymbol{\tau}^*) = \hat{u}_Q(\mathbf{d}_\tau^*)$, $u_{KL}(\mathbf{d}_\tau^*)$, and optimal values $\bar{t}_1^*$, $\bar{t}_2^*$, $u_{1,i}^{0*}$, $p_{1,i}^*$ of the decision variables for the global solution to the Bayesian OED problem in Section 7.2, for different arc sequences and the final time $t_f = 12$. The arc sequence with the best optimal value of $\hat{\phi}(\boldsymbol{\tau}^*)$ is highlighted in bold.

Arc sequence	Execution time (s)	$\hat{u}_Q(\mathbf{d}_\tau^*)$	$u_{KL}(\mathbf{d}_\tau^*)$	$\bar{t}_1^*$	$\bar{t}_2^*$	$u_{1,i}^{0*}$	$p_{1,i}^*$
3-1L-1U	2666	1.9924	2.1845	2.251	12.000	1.000	0.000 ($i = 1$)
3-1U-1L	**3254**	**2.1724**	**2.3451**	**5.206**	**9.330**	**0.538**	**−0.103** ($i = 1$)
1L-3-1L	2559	2.1053	2.2603	5.116	10.341	1.000	0.000 ($i = 2$)
1L-3-1U	2563	1.8626	1.9318	5.378	12.000	1.000	−0.051 ($i = 2$)
1U-3-1L	2701	2.1318	2.2703	1.556	9.324	0.104	0.115 ($i = 2$)
1U-3-1U	2613	2.0046	2.1118	2.189	12.000	0.236	−0.024 ($i = 2$)

utility as the cost of an OCP. The approximation of this cost as an explicit polynomial function of the decision variables allows for reformulation of the OCP as a set of polynomial optimization problems, which can be solved to global optimality in a tractable way via the concept of sum-of-squares polynomials. The paper showed that the difference between the cost obtained by the proposed approach and the globally optimal cost of the OCP is bounded and depends on the polynomial approximation errors. Moreover, the execution time of the optimization procedure for each arc sequence considered for the OCP in a relevant case study indicates that a global solution can be obtained in a tractable way via a convex SDP.

CRediT authorship contribution statement

Diogo Rodrigues: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Georgios Makrygiorgos:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Ali Mesbah:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

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), USA 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.

Appendix A. KL divergence and Shannon information

In this appendix we present some evidence about the statement in Remark 1.

The posterior expected gain is defined as

$$\upsilon(\mathbf{y}, \mathbf{d}) \coloneqq \int_\Theta G(\boldsymbol{\theta}, \mathbf{y}, \mathbf{d})\, p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d}) \mathrm{d}\boldsymbol{\theta}, \tag{A.1}$$

while the definition of the expected utility (5) based on the posterior expected gain is given by

$$\begin{aligned} u(\mathbf{d}) &= \int_\Theta \int_\mathcal{Y} G(\boldsymbol{\theta}, \mathbf{y}, \mathbf{d})\, p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d}) \mathrm{d}\mathbf{y} p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\ &= \int_\mathcal{Y} \int_\Theta G(\boldsymbol{\theta}, \mathbf{y}, \mathbf{d})\, p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d}) \mathrm{d}\boldsymbol{\theta} p(\mathbf{y}|\mathbf{d}) \mathrm{d}\mathbf{y} \\ &= \int_\mathcal{Y} \upsilon(\mathbf{y}, \mathbf{d}) p(\mathbf{y}|\mathbf{d}) \mathrm{d}\mathbf{y}. \end{aligned} \tag{A.2}$$

The posterior expected gain that corresponds to the gain function $G_{KL}(\boldsymbol{\theta}, \mathbf{y}, \mathbf{d})$ in (9) is the KL divergence from the prior distribution to the posterior distribution

$$\upsilon_{KL}(\mathbf{y}, \mathbf{d}) = \int_\Theta \log\left(\frac{p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d})}{p(\boldsymbol{\theta})}\right) p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d}) \mathrm{d}\boldsymbol{\theta}. \tag{A.3}$$

Then we show that the design $\mathbf{d}_{KL}^*$ that maximizes the expected utility for the KL divergence $\upsilon_{KL}(\mathbf{y}, \mathbf{d})$ also does so for the expected gain in Shannon information (SI)

$$\begin{aligned} \upsilon_{SI}(\mathbf{y}, \mathbf{d}) &= \int_\Theta \log\left(p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d})\right) p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d}) \mathrm{d}\boldsymbol{\theta} \\ &\quad - \int_\Theta \log\left(p(\boldsymbol{\theta})\right) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \end{aligned} \tag{A.4}$$

since $\int_\mathcal{Y} p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d})\mathrm{d}\mathbf{y} = 1$ and $\int_\mathcal{Y} p(\mathbf{y}|\mathbf{d})\mathrm{d}\mathbf{y} = 1$ regardless of $\mathbf{d}$, which implies that

$$\begin{aligned} \mathbf{d}_{KL}^* &= \arg\max_{\mathbf{d}\in\mathcal{D}} \int_\mathcal{Y} \int_\Theta \log\left(p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d})\right) p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d}) \mathrm{d}\boldsymbol{\theta} p(\mathbf{y}|\mathbf{d}) \mathrm{d}\mathbf{y} \\ &\quad - \int_\mathcal{Y} \int_\Theta \log\left(p(\boldsymbol{\theta})\right) p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d}) \mathrm{d}\boldsymbol{\theta} p(\mathbf{y}|\mathbf{d}) \mathrm{d}\mathbf{y}, \text{ s.t. (7)} \\ &= \arg\max_{\mathbf{d}\in\mathcal{D}} \int_\mathcal{Y} \int_\Theta \log\left(p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d})\right) p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d}) \mathrm{d}\boldsymbol{\theta} p(\mathbf{y}|\mathbf{d}) \mathrm{d}\mathbf{y} \\ &\quad - \int_\Theta \int_\mathcal{Y} p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d}) \mathrm{d}\mathbf{y} \log\left(p(\boldsymbol{\theta})\right) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \text{ s.t. (7)} \\ &= \arg\max_{\mathbf{d}\in\mathcal{D}} \int_\mathcal{Y} \int_\Theta \log\left(p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d})\right) p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d}) \mathrm{d}\boldsymbol{\theta} p(\mathbf{y}|\mathbf{d}) \mathrm{d}\mathbf{y} \\ &\quad - \int_\Theta \log\left(p(\boldsymbol{\theta})\right) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \text{ s.t. (7)} \\ &= \arg\max_{\mathbf{d}\in\mathcal{D}} \int_\mathcal{Y} \int_\Theta \log\left(p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d})\right) p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{d}) \mathrm{d}\boldsymbol{\theta} p(\mathbf{y}|\mathbf{d}) \mathrm{d}\mathbf{y} \\ &\quad - \int_\mathcal{Y} \int_\Theta \log\left(p(\boldsymbol{\theta})\right) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} p(\mathbf{y}|\mathbf{d}) \mathrm{d}\mathbf{y}, \text{ s.t. (7)} \\ &= \arg\max_{\mathbf{d}\in\mathcal{D}} \int_\mathcal{Y} \upsilon_{SI}(\mathbf{y}, \mathbf{d}) p(\mathbf{y}|\mathbf{d}) \mathrm{d}\mathbf{y}, \text{ s.t. (7)}. \end{aligned} \tag{A.5}$$

Appendix B. Approximation of design for Gaussian prior and noise distributions

Under Assumptions 1 and 2, for a given $\boldsymbol{\theta}$ obtained by sampling in Θ according to the pdf $p(\boldsymbol{\theta})$ and $\mathbf{y}$ obtained by sampling in $\mathcal{Y}$ according to the pdf $p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d})$, $p(\mathbf{y}|\mathbf{d})$ can be approximated as

$$\begin{aligned} p(\mathbf{y}|\mathbf{d}) &= \int_\Theta f(\mathbf{y}|\mathbf{g}(\boldsymbol{\theta}, \mathbf{d}) + \int_0^1 \tfrac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta} + t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \mathbf{d}) \mathrm{d}t \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right), \\ &\qquad \sigma^2 \mathbf{I}_T) p(\hat{\boldsymbol{\theta}}) \mathrm{d}\hat{\boldsymbol{\theta}} \\ &= \int_\Theta f(\mathbf{e}| \int_0^1 \tfrac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta} + t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \mathbf{d}) \mathrm{d}t \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right), \\ &\qquad \sigma^2 \mathbf{I}_T) p(\hat{\boldsymbol{\theta}}) \mathrm{d}\hat{\boldsymbol{\theta}} \\ &\simeq \int_\Theta f(\mathbf{e}| \tfrac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{d})\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right), \sigma^2 \mathbf{I}_T) p(\hat{\boldsymbol{\theta}}) \mathrm{d}\hat{\boldsymbol{\theta}} \\ &= f(\mathbf{y}|\mathbf{g}(\boldsymbol{\theta}, \mathbf{d}) + \tfrac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{d})\left(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}\right), \\ &\qquad \sigma^2 \mathbf{I}_T + \tfrac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{d}) \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \tfrac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{d})^{\mathrm{T}}), \end{aligned} \tag{B.1}$$

which results in a small approximation error if $\| \left(\int_0^1 \tfrac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta} + t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \mathbf{d}) \mathrm{d}t - \tfrac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{d}) \right) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \| / \sigma \ll 1$ for all $\boldsymbol{\theta}$ such that $\| \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-\frac{1}{2}} \left(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\right) \|^2 < F_{\chi^2_{n_\theta}}^{-1}(\alpha)$ and for all $\hat{\boldsymbol{\theta}}$ such that $\| \int_0^1 \tfrac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta} + t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \mathbf{d}) \mathrm{d}t \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \|^2 / \sigma^2 < F_{\chi^2_T}^{-1}(\alpha)$, with $F_{\chi^2_{n_\theta}}^{-1}$ and $F_{\chi^2_T}^{-1}$ denoting the inverse cumulative distribution function of the chi-squared distribution with n_θ degrees of freedom and T degrees of freedom, respectively, and α denoting a confidence level, and no approximation error if $\mathbf{g}(\boldsymbol{\theta}, \mathbf{d})$

is linear in $\boldsymbol{\theta}$ since $\|\left(\int_0^1 \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}+t(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}),\mathbf{d})\mathrm{d}t-\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})\right)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\|/\sigma = 0$ in this case. Furthermore,

$$\begin{aligned}
&\frac{\det\left(\sigma^2\mathbf{I}_T+\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})^{\mathrm{T}}\right)}{\det\left(\sigma^2\mathbf{I}_T\right)}\\
&=\det\left(\mathbf{I}_T+\tfrac{1}{\sigma^2}\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})^{\mathrm{T}}\right)\\
&=\det\left(\mathbf{I}_{n_\theta}+\tfrac{1}{\sigma^2}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})^{\mathrm{T}}\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})\right)\\
&=\frac{\det\left(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta},\mathbf{d})+\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)}{\det\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)},
\end{aligned}\tag{B.2}$$

and

$$\begin{aligned}
&\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})^{\mathrm{T}}\left(\sigma^2\mathbf{I}_T+\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})^{\mathrm{T}}\right)^{-1}\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})\\
&=\tfrac{1}{\sigma^2}\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})^{\mathrm{T}}\left(\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\tfrac{1}{\sigma^2}\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})^{\mathrm{T}}+\mathbf{I}_T\right)^{-1}\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})\\
&=\left(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta},\mathbf{d})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}+\mathbf{I}_{n_\theta}\right)^{-1}\tfrac{1}{\sigma^2}\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})^{\mathrm{T}}\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})\\
&=\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta},\mathbf{d})+\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)^{-1}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta},\mathbf{d})\\
&=\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}-\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta},\mathbf{d})+\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}.
\end{aligned}\tag{B.3}$$

Then, for a given $\boldsymbol{\theta}$ obtained by sampling in Θ according to the pdf $p(\boldsymbol{\theta})$, $U_{KL}(\boldsymbol{\theta},\mathbf{d})$ in (10) can be approximated as

$$\begin{aligned}
&U_{KL}(\boldsymbol{\theta},\mathbf{d})\\
&\simeq\int_{\mathcal{Y}}\log\left(\frac{f(\mathbf{y}|\mathbf{g}(\boldsymbol{\theta},\mathbf{d}),\sigma^2\mathbf{I}_T)}{f(\mathbf{y}|\mathbf{g}(\boldsymbol{\theta},\mathbf{d})+\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}),\sigma^2\mathbf{I}_T+\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})^{\mathrm{T}})}\right)\\
&\quad f(\mathbf{y}|\mathbf{g}(\boldsymbol{\theta},\mathbf{d}),\sigma^2\mathbf{I}_T)\mathrm{d}\mathbf{y}\\
&=\tfrac{1}{2}\log\left(\frac{\det\left(\sigma^2\mathbf{I}_T+\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})^{\mathrm{T}}\right)}{\det\left(\sigma^2\mathbf{I}_T\right)}\right)\\
&+\tfrac{1}{2}\|\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\|^2_{\left(\sigma^2\mathbf{I}_T+\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})^{\mathrm{T}}\right)^{-1}}\\
&-\tfrac{1}{2}\mathrm{tr}\left(\mathbf{I}_T-\left(\mathbf{I}_T+\tfrac{1}{\sigma^2}\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{d})^{\mathrm{T}}\right)^{-1}\right)\\
&=\tfrac{1}{2}\log\left(\frac{\det\left(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta},\mathbf{d})+\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)}{\det\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)}\right)\\
&+\tfrac{1}{2}\|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-\frac{1}{2}}\left(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}\right)\|^2_{\mathbf{I}_{n_\theta}-\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-\frac{1}{2}}\left(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta},\mathbf{d})+\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-\frac{1}{2}}}\\
&-\tfrac{1}{2}\mathrm{tr}\left(\mathbf{I}_{n_\theta}-\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-\frac{1}{2}}\left(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta},\mathbf{d})+\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-\frac{1}{2}}\right)\\
&\simeq\tfrac{1}{2}\log\left(\frac{\det\left(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta},\mathbf{d})+\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)}{\det\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)}\right)\\
&+\tfrac{1}{2}\mathrm{tr}\Big(\left(\mathbf{I}_{n_\theta}-\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\mathcal{I}}(\bar{\boldsymbol{\theta}},\mathbf{d})+\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)^{-1}\right)\\
&\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}\right)^{\mathrm{T}}-\mathbf{I}_{n_\theta}\right)\Big),
\end{aligned}\tag{B.4}$$

where the first approximation follows from the approximation of $p(\mathbf{y}|\mathbf{d})$ in (B.1) and the second one results in a small approximation error if $\|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta},\mathbf{d})+\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)^{-1}-\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\mathcal{I}}(\bar{\boldsymbol{\theta}},\mathbf{d})+\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)^{-1}\| \ll 1$ for all $\boldsymbol{\theta}$ such that $\|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-\frac{1}{2}}\left(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}\right)\|^2 < F^{-1}_{\chi^2_{n_\theta}}(\alpha)$, with $F^{-1}_{\chi^2_{n_\theta}}$ denoting the inverse cumulative distribution function of the chi-squared distribution with n_θ degrees of freedom and α denoting a confidence level, and no approximation error if $\mathbf{g}(\boldsymbol{\theta},\mathbf{d})$ is linear in $\boldsymbol{\theta}$ since $\|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta},\mathbf{d})+\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)^{-1}-\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\mathcal{I}}(\bar{\boldsymbol{\theta}},\mathbf{d})+\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)^{-1}\| = 0$ in this case. Thus, the corresponding OED can be approximated as

$$\begin{aligned}
\mathbf{d}^*_{KL}&\simeq\arg\max_{\mathbf{d}\in\mathcal{D}}\int_\Theta\tfrac{1}{2}\log\left(\frac{\det\left(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta},\mathbf{d})+\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)}{\det\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)}\right)p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta},\ \text{s.t. (7)}\\
&=\arg\max_{\mathbf{d}\in\mathcal{D}}\int_\Theta\log\left(\det\left(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta},\mathbf{d})+\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)\right)p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta},\ \text{s.t. (7)}\\
&=\mathbf{d}^*_D.
\end{aligned}\tag{B.5}$$

Appendix C. Approximation of design for arbitrary prior and noise distributions

Under Assumptions 3 and 4, $U_{KL}(\boldsymbol{\theta},\mathbf{d})$ in (10) becomes

$$U_{KL}(\boldsymbol{\theta},\mathbf{d})=\int_{\mathcal{X}}\log\left(\frac{\det(\mathbf{J}(\boldsymbol{\theta},\mathbf{d}))^{-1}\pi(\boldsymbol{\xi}|\bar{\mathbf{x}},\boldsymbol{\Sigma}_{\boldsymbol{\xi}})}{p(\mathbf{g}(\boldsymbol{\theta},\mathbf{d})+\mathbf{J}(\boldsymbol{\theta},\mathbf{d})\boldsymbol{\xi}|\mathbf{d})}\right)\pi(\boldsymbol{\xi}|\bar{\mathbf{x}},\boldsymbol{\Sigma}_{\boldsymbol{\xi}})\mathrm{d}\boldsymbol{\xi}.\tag{C.1}$$

The utility function $U_{KL}(\boldsymbol{\theta},\mathbf{d})$ in (C.1) is a multivariate integral that can be approximated via Monte Carlo integration $\int_{\mathcal{X}}f(\boldsymbol{\xi})\pi(\boldsymbol{\xi}|\bar{\mathbf{x}},\boldsymbol{\Sigma}_{\boldsymbol{\xi}})\mathrm{d}\boldsymbol{\xi}\simeq\frac{1}{m_\xi}\sum_{k=1}^{m_\xi}f(\boldsymbol{\xi}_k)$ in the observation space for some function $f(\boldsymbol{\xi})$ using the points $\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{m_\xi}$ that are obtained by sampling in $\mathcal{X}$ according to the pdf $\pi(\boldsymbol{\xi}|\bar{\mathbf{x}},\boldsymbol{\Sigma}_{\boldsymbol{\xi}})$ and independently of the prior pdf $p(\boldsymbol{\theta})$. Then, $U_{KL}(\boldsymbol{\theta},\mathbf{d})$ in (C.1) can be approximated as

$$\begin{aligned}
U_{KL}(\boldsymbol{\theta},\mathbf{d})&\simeq\tfrac{1}{m_\xi}\sum_{k=1}^{m_\xi}\log\left(\frac{\det(\mathbf{J}(\boldsymbol{\theta},\mathbf{d}))^{-1}\pi(\boldsymbol{\xi}_k|\bar{\mathbf{x}},\boldsymbol{\Sigma}_{\boldsymbol{\xi}})}{p(\mathbf{g}(\boldsymbol{\theta},\mathbf{d})+\mathbf{J}(\boldsymbol{\theta},\mathbf{d})\boldsymbol{\xi}_k|\mathbf{d})}\right)\\
&=U_{MC}(\boldsymbol{\theta},\mathbf{d}),
\end{aligned}\tag{C.2}$$

where the evidence $p(\mathbf{g}(\boldsymbol{\theta},\mathbf{d})+\mathbf{J}(\boldsymbol{\theta},\mathbf{d})\boldsymbol{\xi}_k|\mathbf{d})$ is computed as in (4) from the likelihood function in (22), and the corresponding OED can be approximated as

$$\begin{aligned}
\mathbf{d}^*_{KL}&\simeq\arg\max_{\mathbf{d}\in\mathcal{D}}\int_\Theta U_{MC}(\boldsymbol{\theta},\mathbf{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta},\ \text{s.t. (7)}\\
&=\mathbf{d}^*_{MC}.
\end{aligned}\tag{C.3}$$

Appendix D. Computation of the posterior covariance

The covariance matrix of the posterior parameter distribution is computed as follows:

$$\begin{aligned}
&\int_{\mathcal{Y}}p(\mathbf{y}|\mathbf{d})\int_\Theta\left(\boldsymbol{\theta}-\left(\int_\Theta\boldsymbol{\theta}p(\boldsymbol{\theta}|\mathbf{y},\mathbf{d})\mathrm{d}\boldsymbol{\theta}\right)\right)\\
&\quad\times\left(\boldsymbol{\theta}-\left(\int_\Theta\boldsymbol{\theta}p(\boldsymbol{\theta}|\mathbf{y},\mathbf{d})\mathrm{d}\boldsymbol{\theta}\right)\right)^{\mathrm{T}}p(\boldsymbol{\theta}|\mathbf{y},\mathbf{d})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{y}\\
&=\int_{\mathcal{Y}}p(\mathbf{y}|\mathbf{d})\left(\int_\Theta\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}p(\boldsymbol{\theta}|\mathbf{y},\mathbf{d})\mathrm{d}\boldsymbol{\theta}\right.\\
&\quad\left.-\left(\int_\Theta\boldsymbol{\theta}p(\boldsymbol{\theta}|\mathbf{y},\mathbf{d})\mathrm{d}\boldsymbol{\theta}\right)\left(\int_\Theta\boldsymbol{\theta}p(\boldsymbol{\theta}|\mathbf{y},\mathbf{d})\mathrm{d}\boldsymbol{\theta}\right)^{\mathrm{T}}\right)\mathrm{d}\mathbf{y}\\
&=\int_{\mathcal{Y}}p(\mathbf{y}|\mathbf{d})\left(\int_\Theta\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}\frac{p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})p(\boldsymbol{\theta})}{p(\mathbf{y}|\mathbf{d})}\mathrm{d}\boldsymbol{\theta}\right.\\
&\quad\left.-\left(\int_\Theta\boldsymbol{\theta}\frac{p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})p(\boldsymbol{\theta})}{p(\mathbf{y}|\mathbf{d})}\mathrm{d}\boldsymbol{\theta}\right)\left(\int_\Theta\boldsymbol{\theta}\frac{p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})p(\boldsymbol{\theta})}{p(\mathbf{y}|\mathbf{d})}\mathrm{d}\boldsymbol{\theta}\right)^{\mathrm{T}}\right)\mathrm{d}\mathbf{y}\\
&=\int_{\mathcal{Y}}\int_\Theta\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}-\int_{\mathcal{Y}}\left(\int_\Theta p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\right)^{-1}\\
&\left(\int_\Theta\boldsymbol{\theta}p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\right)\left(\int_\Theta\boldsymbol{\theta}p(\mathbf{y}|\boldsymbol{\theta},\mathbf{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\right)^{\mathrm{T}}\mathrm{d}\mathbf{y}\\
&=\int_{\mathcal{Y}}\int_\Theta\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}\frac{\pi\left(\mathbf{J}(\boldsymbol{\theta},\mathbf{d})^{-1}(\mathbf{y}-\mathbf{g}(\boldsymbol{\theta},\mathbf{d}))|\bar{\mathbf{x}},\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\right)}{\det(\mathbf{J}(\boldsymbol{\theta},\mathbf{d}))}p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\\
&\quad-\int_{\mathcal{Y}}\left(\int_\Theta\frac{\pi\left(\mathbf{J}(\boldsymbol{\theta},\mathbf{d})^{-1}(\mathbf{y}-\mathbf{g}(\boldsymbol{\theta},\mathbf{d}))|\bar{\mathbf{x}},\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\right)}{\det(\mathbf{J}(\boldsymbol{\theta},\mathbf{d}))}p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\right)^{-1}\\
&\left(\int_\Theta\boldsymbol{\theta}\frac{\pi\left(\mathbf{J}(\boldsymbol{\theta},\mathbf{d})^{-1}(\mathbf{y}-\mathbf{g}(\boldsymbol{\theta},\mathbf{d}))|\bar{\mathbf{x}},\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\right)}{\det(\mathbf{J}(\boldsymbol{\theta},\mathbf{d}))}p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\right)\\
&\quad\times\left(\int_\Theta\boldsymbol{\theta}\frac{\pi\left(\mathbf{J}(\boldsymbol{\theta},\mathbf{d})^{-1}(\mathbf{y}-\mathbf{g}(\boldsymbol{\theta},\mathbf{d}))|\bar{\mathbf{x}},\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\right)}{\det(\mathbf{J}(\boldsymbol{\theta},\mathbf{d}))}p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\right)^{\mathrm{T}}\mathrm{d}\mathbf{y}.
\end{aligned}\tag{D.1}$$

References

[1] A.C. Atkinson, A.N. Donev, Optimum Experimental Designs, Clarendon Press, 1992.
[2] E. Walter, L. Pronzato, Qualitative and quantitative experiment design for phenomenological models - a survey, Automatica 26 (2) (1990) 195–213.
[3] V. Fedorov, Optimal experimental design, Wiley Interdiscip. Rev. Comput. Stat. 2 (5) (2010) 581–589.
[4] G. Franceschini, S. Macchietto, Model-based design of experiments for parameter precision: State of the art, Chem. Eng. Sci. 63 (19) (2008) 4846–4872.
[5] M. Martin-Casas, A. Mesbah, Discrimination between competing model structures of biological systems in the presence of population heterogeneity, IEEE Life Sci. Lett. 2 (3) (2016) 23–26.

[6] S. Streif, F. Petzke, A. Mesbah, R. Findeisen, R.D. Braatz, Optimal experimental design for probabilistic model discrimination using polynomial chaos, IFAC Proc. Vol. 47 (3) (2014) 4103–4109.
[7] X. Bombois, G. Scorletti, M. Gevers, P. Van den Hof, R. Hildebrand, Least costly identification experiment for control, Automatica 42 (10) (2006) 1651–1662.
[8] M. Annergren, C.A. Larsson, H. Hjalmarsson, X. Bombois, B. Wahlberg, Application-oriented input design in system identification: Optimal input design for control [applications of control], IEEE Control Syst. Mag. 37 (2) (2017) 31–56.
[9] A. Wald, On the efficient design of statistical investigations, Ann. Math. Stat. 14 (2) (1943) 134–140.
[10] S.D. Silvey, D.M. Titterington, A geometric approach to optimal design theory, Biometrika 60 (1) (1973) 21–32.
[11] G. Elfving, Optimum allocation in linear regression theory, Ann. Math. Stat. 23 (2) (1952) 255–262.
[12] K. Chaloner, I. Verdinelli, Bayesian experimental design: A review, Stat. Sci. 10 (3) (1995) 273–304.
[13] X. Huan, Y.M. Marzouk, Simulation-based optimal Bayesian experimental design for nonlinear systems, J. Comput. Phys. 232 (1) (2013) 288–317.
[14] M.C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, J. R. Stat. Soc. Ser. B Stat. Methodol. 63 (3) (2001) 425–464.
[15] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1) (1951) 79–86.
[16] D.V. Lindley, On a measure of the information provided by an experiment, Ann. Math. Stat. 27 (4) (1956) 986–1005.
[17] G.E.P. Box, H.L. Lucas, Design of experiments in non-linear situations, Biometrika 46 (1–2) (1959) 77–90.
[18] E.G. Ryan, C.C. Drovandi, J.M. McGree, A.N. Pettitt, A review of modern computational algorithms for Bayesian optimal design, Int. Stat. Rev. 84 (1) (2016) 128–154.
[19] H.J. Kushner, G.G. Yin, Stochastic Approximation and Recursive Algorithms and Applications, Springer, 2003.
[20] A. Shapiro, Asymptotic analysis of stochastic programs, Ann. Oper. Res. 30 (1–4) (1991) 169–186.
[21] K.J. Ryan, Estimating expected information gains for experimental designs with application to the random Fatigue-limit model, J. Comput. Graph. Stat. 12 (3) (2003) 585–603.
[22] J. Beck, B.M. Dia, L.F.R. Espath, Q. Long, R. Tempone, Fast Bayesian experimental design: Laplace-based importance sampling for the expected information gain, Comput. Methods Appl. Mech. Engrg. 334 (2018) 523–553.
[23] Q. Long, M. Scavino, R. Tempone, S. Wang, Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations, Comput. Methods Appl. Mech. Engrg. 259 (2013) 24–39.
[24] X. Huan, Y.M. Marzouk, Gradient-based stochastic optimization methods in Bayesian experimental design, Int. J. Uncertain. Quantif. 4 (2014) 479–510.
[25] J.A. Paulson, M. Martin-Casas, A. Mesbah, Optimal Bayesian experiment design for nonlinear dynamic systems with chance constraints, J. Process Control 77 (2019) 155–171.
[26] S. Olofsson, L. Hebing, S. Niedenführ, M.P. Deisenroth, R. Misener, GPdoemd: A python package for design of experiments for model discrimination, Comput. Chem. Eng. 125 (2019) 54–70.
[27] H. Jansson, H. Hjalmarsson, Input design via LMIs admitting frequency-wise model specifications in confidence regions, IEEE Trans. Automat. Control 50 (10) (2005) 1534–1549.
[28] P.E. Valenzuela, J. Dahlin, C.R. Rojas, T.B. Schön, On robust input design for nonlinear dynamical models, Automatica 77 (2017) 268–278.
[29] A. Mesbah, S. Streif, A probabilistic approach to robust optimal experiment design with chance constraints, IFAC-PapersOnLine 48 (8) (2015) 100–105.
[30] L. Pronzato, A. Pázman, Design of Experiments in Nonlinear Models: Asymptotic Normality, Optimality Criteria and Small-Sample Properties, Springer New York, New York, NY, 2013.
[31] D. Rodrigues, G. Makrygiorgos, A. Mesbah, Tractable global solutions to Bayesian optimal experiment design, in: Proc. 59th IEEE Conference on Decision and Control, CDC, Jeju Island, Republic of Korea, 2020, pp. 1614–1619.
[32] G. Makrygiorgos, G.M. Maggioni, A. Mesbah, Surrogate modeling for fast uncertainty quantification: Application to 2D population balance models, Comput. Chem. Eng. 138 (106814) (2020).
[33] D. Rodrigues, D. Bonvin, Dynamic optimization of reaction systems via exact parsimonious input parameterization, Ind. Eng. Chem. Res. 58 (26) (2019) 11199–11212.
[34] D. Rodrigues, D. Bonvin, On reducing the number of decision variables for dynamic optimization, Optim. Control Appl. Meth. 41 (1) (2020), 292–311.
[35] J.B. Lasserre, Moments, Positive Polynomials and Their Applications, Imperial College Press, 2010.
[36] J. Berger, Statistical Decision Theory and Bayesian Analysis, second ed., Springer Verlag, 1985.
[37] B.K. Pagnoncelli, S. Ahmed, A. Shapiro, Sample average approximation method for chance constrained programming: Theory and applications, J. Optim. Theory Appl. 142 (2009) 399–416.
[38] A. Nemirovski, A. Shapiro, Convex approximations of chance constrained programs, SIAM J. Optim. 17 (4) (2006) 969–996.
[39] A. Ben-Tal, L.E. Ghaoui, A. Nemirovski, Robust Optimization, Princeton University Press, 2009.
[40] J.A. Paulson, A. Mesbah, An efficient method for stochastic optimal control with joint chance constraints for nonlinear systems, Int. J. Robust Nonlin. 29 (15) (2019) 5017–5037.
[41] J.A. Paulson, A. Mesbah, Nonlinear model predictive control with explicit backoffs for stochastic systems under arbitrary uncertainty, IFAC-PapersOnLine 51 (20) (2018) 523–534.
[42] W. Gautschi, Orthogonal Polynomials: Computation and Approximation, Oxford University Press, 2004.
[43] M. Sinsbeck, W. Nowak, An optimal sampling rule for nonintrusive polynomial chaos expansions of expensive models, Int. J. Uncertain. Quantif. 5 (3) (2015) 275–295.
[44] D. Xiu, G.E. Karniadakis, The Wiener-Askey polynomial chaos for stochastic differential equations, SIAM J. Sci. Comput. 24 (2) (2002) 619–644.
[45] J.A. Paulson, E.A. Buehler, A. Mesbah, Arbitrary polynomial chaos for uncertainty propagation of correlated random variables in dynamic systems, IFAC-PapersOnLine 50 (1) (2017) 3548–3553.
[46] G. Blatman, B. Sudret, Adaptive sparse polynomial chaos expansion based on least angle regression, J. Comput. Phys. 230 (6) (2011) 2345–2367.
[47] B. Srinivasan, S. Palanki, D. Bonvin, Dynamic optimization of batch processes: I. Characterization of the nominal solution, Comput. Chem. Eng. 27 (1) (2003) 1–26.
[48] L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze, E.F. Mishchenko, The Mathematical Theory of Optimal Processes, Interscience, New York, 1962.
[49] R.F. Hartl, S.P. Sethi, R.G. Vickson, A survey of the maximum principles for optimal control problems with state constraints, SIAM Rev. 37 (2) (1995) 181–218.
[50] X. Xu, P.J. Antsaklis, Optimal control of switched systems based on parameterization of the switching instants, IEEE Trans. Autom. Contr. 49 (1) (2004) 2–16.
[51] E.S. Schultz, R. Hannemann-Tamás, A. Mitsos, Polynomial approximation of inequality path constraints in dynamic optimization, Comput. Chem. Eng. 135 (2020) 106732.
[52] D. Henrion, J.B. Lasserre, GloptiPoly: Global optimization over polynomials with matlab and SeDuMi, ACM Trans. Math. Software 29 (2) (2003) 165–194.
[53] H. Waki, S. Kim, M. Kojima, M. Muramatsu, Sums of squares and semidefinite program relaxations for polynomial optimization problems with structured sparsity, SIAM J. Optim. 17 (1) (2006) 218–242.
[54] R.A. Lorentz, Multivariate Hermite interpolation by algebraic polynomials: A survey, J. Comput. Appl. Math. 122 (1–2) (2000) 167–201.
[55] V.B. Alekseev, Abel's Theorem in Problems and Solutions, Springer, 2004.
[56] S. Sager, H.G. Bock, M. Diehl, G. Reinelt, J.P. Schlöder, Numerical methods for optimal control with binary control functions applied to a Lotka-Volterra type fishing problem, in: A. Seeger (Ed.), Recent Advances in Optimization, Springer, 2006, pp. 269–289.
[57] A. Ibañez, Optimal control of the Lotka–Volterra system: turnpike property and numerical simulations, J. Biol. Dyn. 11 (1) (2017) 25–41.
[58] D. Telen, F. Logist, E. Van Derlinden, I. Tack, J. Van Impe, Optimal experiment design for dynamic bioprocesses: A multi-objective approach, Chem. Eng. Sci. 78 (2012) 82–97.