Journal of Computational Physics 513 (2024) 113141

Contents lists available at ScienceDirect

Journal of Computational Physics

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

Tikhonov regularization as a nonparametric method for uncertainty quantification in aggregate data problems

Elena Villalón [a], Qian Yang [b], Carlos A. Sing Long [a,c,*]

[a] Institute for Mathematical and Computational Engineering, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Santiago, 7820436, Chile

[b] School of Computing, University of Connecticut, 371 Fairfield Way, Unit 4155, Storrs, 06269, CT, USA

[c] Institute for Biological and Medical Engineering, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Santiago, 7820436, Chile

ARTICLE INFO

Keywords:
Uncertainty quantification
Dynamical systems
Convex optimization
Aggregate data problems
Regularization
Nonparametric methods

ABSTRACT

Dynamical systems are widely used in mathematical models in engineering and in the applied sciences. However, the parameters of these systems are usually unknown and they must be estimated from measured data. Performing parameter estimation and quantifying the uncertainty in these estimates becomes critical to realize the predictive power of dynamical systems. In some applications, only measurements about the states of an ensemble of systems are available, where each one evolves according to the same model but for different parameter values, leading to *aggregate data* problems. In this case, an approach proposed in the literature is to estimate a density that is consistent with some estimates of the expected value of some quantities of interest. To solve the problem in practice, the density is discretized and, to solve the resulting ill-posed problem, Tikhonov regularization is used. In this work, we propose a Bayesian model that shows this approach can be interpreted as a *maximum a posteriori* estimate for the density. We show that the infinite-dimensional problem defining the MAP can be reformulated as a finite-dimensional problem that does not require discretizing the density. In several cases of interest, this problem is convex, unconstrained, and the objective function is smooth. Thus, it can be solved using algorithms with optimal convergence rates. The trade-off is that the objective is defined by an integral. However, our results characterize the regularity of the integrand, allowing the use of tailored numerical schemes to approximate it. Furthermore, our theoretical results characterize the form of the optimal density, whereas our numerical results illustrate the performance of our method and confirm our theoretical findings.

1. Introduction

Dynamical systems are a cornerstone of mathematical modeling in engineering and in the applied sciences. Models based on dynamical systems are broadly used due to their predictive power [1,2] and due to the availability of numerical methods that accurately simulate them [3–5]. However, in practical applications these models depend on unknown parameters that must be estimated from some observed trajectories of the system. Thus, to realize the predictive power of dynamical systems in practice, it

* Corresponding author.
E-mail address: casinglo@uc.cl (C.A. Sing Long).

https://doi.org/10.1016/j.jcp.2024.113141
Received 25 April 2023; Received in revised form 16 May 2024; Accepted 23 May 2024

Available online 3 June 2024

is not only important to develop methods to estimate these parameters from experimental measurements, but also to quantify the uncertainty in these estimates, and their impact on the predictions obtained from numerical simulations [6–9].

Typically, several *quantities of interest* (QoIs) are measured, possibly with errors, from observed trajectories. These measurements are then used to estimate the unknown parameters and to quantify the uncertainty on these estimates, i.e., the objective is to perform *model calibration* and/or *inverse uncertainty quantification* [8, Chapter 1]. If we have measurements of the QoIs from several trajectories generated by the same system, i.e., a system for which there is no uncertainty in the parameters, we obtain a *longitudinal data problem* [10]. In this case, the likelihood can usually be characterized and Bayesian methods are well-suited to simultaneously estimate the unknown parameters, e.g., using the posterior expectation or the maximum *a posteriori* (MAP), and to quantify the uncertainty in these estimates (see, e.g., [8, Chapter 8]). However, in some applications one can only measure the QoIs from trajectories generated by an ensemble of systems that evolve according to the same model, but may do so according to different values of the unknown parameters. As a concrete example, estimating the unknown parameters in the CAR T-Cell cancer model proposed by Schacht et al. [11] requires measuring the amount of T-cells in mice. As this measurement cannot be performed in the same mouse at different times, a cohort is used, and one measurement is performed for each mouse. Thus, we have observations about systems that evolve according to the same model, but for different values of the unknown parameters. In this case, the parameters are assumed to be random variables with an unknown distribution, and we obtain what is called an *aggregate data problem* (see, e.g., [12] or [13, Chapter 5]). Thus, the natural object to estimate is the unknown probability distribution of the parameters, and to quantify the uncertainty in this estimate.

Banks et al. [14] propose to estimate the probability density by enforcing consistency with the expected value of a QoI estimated from several measurements made about an ensemble of systems. Since the object to be estimated is a density, and thus an infinite-dimensional object, this is a *nonparametric method*. To compute this density in practice, they propose to discretize it, e.g., using a mesh, splines, or other approximations, to then solve a least-squares problem enforcing consistency with the measured data (see, e.g., [13, Chapter 5] or [15, Chapter 14]). The discretization leads to a parametric method which may lead to determined or underdetermined problems depending on the number of constraints imposed on the density and the number of degrees of freedom used to discretize it. The authors in [16] extended this approach by enforcing additional constraints on the unknown density, such as constraining the value of its higher-order moments, or the dispersion of the state of the system around a given point at specific times. To compute a discretized density satisfying these constraints efficiently, the authors leverage convex programming and the Koopman operator. This relies on the fact that the method proposed by Banks et al. leads to a convex least-squares problem, and on the fact that a single probability density at the initial time that can be pushed forward in time by the system dynamics using the Koopman operator. In fact, when the constraints are linear, they become constraints on the expected value of the Koopman operator applied to a QoI [17]. To regularize the solution to a possibly underdetermined system, which occurs when the degrees of freedom of the discretization exceed the number of constraints, the authors propose penalizing the values of the density at its discretization points by its ℓ_2-norm squared, i.e., by applying Tikhonov regularization [18]. The resulting problem can be interpreted as *Tikhonov regularization for moment constraints* on the unknown discretized density.

Although this approach has appealing properties, some open questions remain about its practical implementation. On one hand, solving the problem by discretizing the density can implicitly restrict its support, and, when the discretization is over a mesh as suggested in [17], it can preclude the use of the method in several practical applications. There is a growing number of engineering problems where the input to the model and its parameters are high-dimensional. For instance, the quantification of uncertainties in Mars' atmospheric entry, descent, and landing processes due to the uncertainty present in the atmosphere density, the initial states, and other model parameters [19], wind-turbine simulators that take time series of simulated wind speed as the input to predict mechanical loads on its sub-structures and fatigue [20], and the study of variations on fabrication processes on nano-scale chip design, such as the surface roughness of interconnects and the random doping effects in transistors, that influence the performance of chips [21].

On the other hand, Tikhonov regularization has several computational benefits, particularly its use of a smooth and strongly convex regularizer, and it is widely used to solve ill-posed or ill-conditioned inverse problems. However, its use as a regularization strategy to estimate a probability density lacks an intuitive statistical interpretation, particularly in contrast to other regularizers, such at the KL divergence [22]. This limits the appeal of the approach, and makes the interpretation of its results challenging. Consequently, both developing a theoretical framework that allows us to interpret this method as a classical estimation procedure, and developing implementations of this method that are computationally efficient, along with suitable theoretical guarantees, would foster its adoption for uncertainty quantification in aggregate data problems involving high-dimensional dynamical systems.

1.1. Contributions

To introduce the method, consider the case of a dynamical system in $\mathbb{R}^d$ given by

$$\begin{cases} \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(t, \boldsymbol{x}(t)) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0 \end{cases}$$

where there is uncertainty in the initial condition. For simplicity, suppose that we have an estimate $\overline{\boldsymbol{x}}_{t_0}$ of the expected state at a time t_0. Then, to estimate the uncertainty on the initial condition from observations about the state of the system, we may look for a density ρ that satisfies the constraint

$$\int \boldsymbol{x}(t_0, \boldsymbol{x}_0)\rho(\boldsymbol{x}_0)\, d\boldsymbol{x}_0 \approx \overline{\boldsymbol{x}}_{t_0},$$

where $\boldsymbol{x} = \boldsymbol{x}(t, \boldsymbol{x}_0)$ is the state at time t for the initial condition $\boldsymbol{x}_0$. There is an infinite number of densities satisfying this constraint exactly. The method proposed in [16] involves solving the convex problem

$$\begin{aligned} \underset{\rho}{\text{minimize}} \quad & \lambda \int \rho(\boldsymbol{x}_0)^2 d\boldsymbol{x}_0 + \frac{1}{2}\left\| \int \boldsymbol{x}(t_0, \boldsymbol{x}_0)\rho(\boldsymbol{x}_0)\, d\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{t_0} \right\|_2^2 \\ \text{subject to} \quad & \rho \geq 0, \quad \int \rho(\boldsymbol{x}_0) d\boldsymbol{x}_0 = 1, \end{aligned} \tag{1}$$

for some regularization parameter $\lambda > 0$ and over a suitable functional space. This problem can be interpreted as Tikhonov regularization on moment constraints. A direct approach to solve this problem numerically in low-dimensions is to discretize the integrals using $N := n^d$ equispaced nodes $\{\boldsymbol{x}_{0,i}\}$. Then, by defining

$$\boldsymbol{\rho} = \begin{bmatrix} \rho(\boldsymbol{x}_{0,1}) \\ \vdots \\ \rho(\boldsymbol{x}_{0,N}) \end{bmatrix} \quad \text{and} \quad \boldsymbol{a}_i = \begin{bmatrix} x_i(t_0, \boldsymbol{x}_{0,1}) \\ \vdots \\ x_i(t_0, \boldsymbol{x}_{0,N}) \end{bmatrix}$$

we may solve

$$\begin{aligned} \underset{\rho}{\text{minimize}} \quad & \frac{\lambda \Delta x^d}{2}\|\boldsymbol{\rho}\|_2^2 + \frac{1}{2}\sum_{i=1}^{d} (\Delta x^d \langle \boldsymbol{a}_i, \boldsymbol{\rho}\rangle_2 - \overline{x}_{t,i})^2 \\ \text{subject to} \quad & \boldsymbol{\rho} \geq 0, \quad \Delta x^d \langle \mathbf{1}, \boldsymbol{\rho}\rangle_2 = 1 \end{aligned}$$

where $\mathbf{1}$ is the vector with all its entries equal to one and Δx is the discretization step. On one hand, this is a convex optimization problem with a number of variables that is exponential in the number of dimensions, precluding the use of this method for high-dimensional problems. On the other hand, it is not intuitive why (1) is an efficient estimation procedure for the density, nor what is the class of possible densities that can be selected by this procedure.

In this work, we introduce a Bayesian model that provides a theoretical framework to interpret (1) as a MAP estimate for the density. This not only relates Tikhonov regularization to nonparametric estimation, in the sense that the parameter lies on an infinite-dimensional space [10], but possibly with Bayesian nonparametric methods [23]. Furthermore, we provide a method to solve (1) without needing to discretize the density. By leveraging convex duality, we prove that, even in general cases, the problem (1) can be reduced to a smooth, finite-dimensional convex optimization problem in $d + 1$ variables which involves the computation of a d-dimensional integral. Furthermore, the dual problem allows us to provide a closed-form expression for the optimal density, a result that has both computational and statistical implications.

We consider the following as the main contributions of our work.

(i) **Tikhonov regularization as a MAP estimate:** We propose a Bayesian model that provides a rigorous justification for using Tikhonov regularization on moment constraints to perform input uncertainty quantification, and allows us to interpret the computed density as a MAP estimate for this model.
(ii) **Connections to regularized regression:** When we have incomplete linear measurements about some QoIs, we show that solving (1) is equivalent to solving a regularized regression problem for the expected values of the QoIs themselves.
(iii) **Geometry and support:** We explicitly study the effect of restricting the support of ρ to a convex set Ω, showing through theoretical results and numerical experiments that it has a significant impact on the estimated density.
(iv) **Closed-form expressions:** We provide closed-form expressions for the optimal probability density, showing the explicit dependence of the probability density both on the QoIs and on the dynamics of the system.
(v) **High-dimensional problems:** We show that by solving a finite-dimensional convex problem with a number of variables that is proportional to the number of QoIs, we may recover the optimal solution to (1), enabling the efficient implementation of the method for high-dimensional problems.

1.2. Organization

The paper is organized as follows. In Section 2 we develop a Bayesian model that allows us to interpret Tikhonov regularization as a MAP estimate for the density. In Section 3 we show that the infinite-dimensional convex problem defining the MAP leads to an equivalent finite-dimensional problem. In Section 4 we explore the connections between our method and a regularized regression problem for the expected values of the QoIs. In Section 5 we discuss efficient methods to find an optimal solution to the optimization problem we introduce. In Section 6 we show study the performance of the method in terms of the kind of densities it selects, and the effect of the support on them. We conclude in Section 7 with a discussion of our results.

2. Uncertainty quantification by Tikhonov regularization

2.1. The model

In this work, we consider *random differential equations* [8, Chapter 3] in $\mathbb{R}^d$ of the form

$$\begin{cases} \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(t, \boldsymbol{x}(t)) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0 \end{cases} \qquad \text{where} \quad \boldsymbol{x}_0 \sim \Pi_{x_0} \tag{2}$$

for some unknown probability measure Π_{x_0} on $\mathbb{R}^d$. Typically, it will have a density ρ_{x_0} with respect to a reference measure μ such as the Lebesgue measure. We assume that this equation accurately models the physical phenomenon of interest, that is, that any *model uncertainty* [8, Chapter 2] is negligible. In practical applications, we only have available sample trajectories $\boldsymbol{x} = \boldsymbol{x}(t, \boldsymbol{x}_0)$ of the system for a few realizations of $\boldsymbol{x}_0$ and for each one of these realizations we typically measure a small number of fixed *quantities of interest* (QoIs) $u_1, \dots, u_q$. Since the fluctuations in the values of these QoIs are only due to the fluctuations in the initial condition, we represent these QoIs as a function

$$\boldsymbol{u} = \boldsymbol{U}(\boldsymbol{x}_0).$$

We shall assume that the QoIs can be computed to high accuracy and precision, so that there are no significant *numerical uncertainties* [8, Chapter 2] associated to this model. The measurement of the QoIs is typically corrupted by additive noise, which is statistically independent of $\boldsymbol{x}_0$. Therefore, instead of $\boldsymbol{u}$ we observe

$$\boldsymbol{y} = \boldsymbol{U}(\boldsymbol{x}_0) + \boldsymbol{\varepsilon} \tag{3}$$

for some centered random vector $\boldsymbol{\varepsilon}$. Thus, a sample $\boldsymbol{x}_0^{(1)}, \dots, \boldsymbol{x}_0^{(n)}$ of unknown initial conditions leads to a sample of QoIs $\boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(n)}$ and measurements $\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(n)}$ from which we want to quantify the uncertainty on the initial condition, that is, to quantify the *input uncertainty* [8, Chapter 3] in (2). This is sometimes called an *aggregate data inverse problem* [10,12]. In this case, it is not sufficient to provide a point estimate for the initial condition, as it is a random variable. Hence, it is necessary to find a suitable probability distribution that approximates Π_{x_0} and thus the uncertainty on $\boldsymbol{x}_0$.

Instead of formulating a Bayesian model directly from (3) we proceed as follows. We assume throughout that $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_q)$. Then

$$\boldsymbol{y} \,|\, \boldsymbol{u} \sim \mathcal{N}(\boldsymbol{u}, \sigma^2 \boldsymbol{I}_q).$$

To model the distribution of $\boldsymbol{u}$ we decompose its behavior in terms of an average value $\bar{\boldsymbol{u}}$ and fluctuations around this average that do not depend on exogenous parameters, that is, we assume that

$$\boldsymbol{u} \,|\, \bar{\boldsymbol{u}} \sim \Pi_{\boldsymbol{u}\,|\,\bar{\boldsymbol{u}}}$$

for some probability measure $\Pi_{\boldsymbol{u}\,|\,\bar{\boldsymbol{u}}}$. To model the behavior of the average, let $\{\Pi_{x_0\,|\,\theta} : \boldsymbol{\theta} \in \Theta\}$ be a parametric model for the distribution of $\boldsymbol{x}_0$ with parameter space $\Theta \subset \mathbb{R}^s$ and let

$$\bar{\boldsymbol{u}} \,|\, \boldsymbol{\theta} = \mathbf{E}_{\boldsymbol{x}_0 \sim \Pi_{x_0\,|\,\theta}}[\boldsymbol{U}(\boldsymbol{x}_0)].$$

Finally, we let Π_θ be a prior for $\boldsymbol{\theta}$. This leads to the Bayesian model

$$\begin{cases} \boldsymbol{y} \,|\, \boldsymbol{u} \sim \mathcal{N}(\boldsymbol{u}, \sigma^2 \boldsymbol{I}_q), \\ \boldsymbol{u} \,|\, \bar{\boldsymbol{u}} \sim \Pi_{u\,|\,\bar{u}}, \\ \bar{\boldsymbol{u}} \,|\, \boldsymbol{\theta} = \mathbf{E}_{\boldsymbol{x}_0 \sim \Pi_{x_0\,|\,\theta}}[\boldsymbol{U}(\boldsymbol{x}_0)], \\ \quad \boldsymbol{\theta} \sim \Pi_\theta. \end{cases} \tag{4}$$

If $\Pi_{\theta\,|\,y}$ is the posterior obtained from this model, then the posterior for $\boldsymbol{x}_0$ is

$$\mathbf{P}[\boldsymbol{x}_0 \in A] = \int_\Theta \Pi_{x_0\,|\,\theta}(A)\, d\Pi_{\theta\,|\,y}(\theta)$$

The dimension s of the parameter space Θ constrains the flexibility of the model we use for $\bar{\boldsymbol{u}}$ which, in turn, constraints the flexibility of the posterior for $\boldsymbol{x}_0$. For this reason, we will introduce a parameter space Θ and a suitable prior measure Π_θ on Θ that allows us to control the flexibility of the model, focusing on the asymptotic regime for $s \gg 1$. Furthermore, instead of aiming to find, or sample from, the posterior distribution, we will focus on characterizing the *maximum a posteriori* (MAP) for the parameter $\boldsymbol{\theta}$. However, before proceeding, we discuss some extensions of this model and its connections with previous work.

2.1.1. Partial measurements of QoIs

In some applications, we cannot measure the full set of QoIs for the sample trajectories. The results we will present extend naturally to the case in which we observe *incomplete linear measurements* of the QoIs. In other words, we assume that there exists a $m \times q$ matrix $\boldsymbol{A}$ such that we are only able to measure $\boldsymbol{Au}$. This leads to the *partial measurement model*

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{U}(\boldsymbol{x}_0) + \boldsymbol{\varepsilon}.$$

In this case, for $\boldsymbol{u}' = \boldsymbol{Au}$ the model (4) simply becomes

$$\begin{cases} \boldsymbol{y} \mid \boldsymbol{u}' \sim \mathcal{N}(\boldsymbol{u}', \sigma^2 \boldsymbol{I}_m) \\ \boldsymbol{u}' \mid \bar{\boldsymbol{u}}' \sim \Pi_{u' \mid \bar{u}'} \\ \bar{\boldsymbol{u}}' \mid \boldsymbol{\theta} = \boldsymbol{A}\, \mathbf{E}_{\boldsymbol{x}_0 \sim \Pi_{x_0 \mid \theta}}[\boldsymbol{U}(\boldsymbol{x}_0)] \\ \quad \boldsymbol{\theta} \sim \Pi_\theta \end{cases}$$

and the same results that hold for (4) will hold for this model. It is straightforward to see this is equivalent to (3) if we define $\boldsymbol{U}' = \boldsymbol{AU}$. However, in Section 4 we discuss the advantages of treating the measurement matrix $\boldsymbol{A}$ separately from the QoIs.

2.1.2. Connection with the Koopman operator

A special case of QoIs has the form

$$\boldsymbol{U}(\boldsymbol{x}_0) = \begin{bmatrix} U_1(\boldsymbol{x}(t_1, \boldsymbol{x}_0)) \\ \vdots \\ U_q(\boldsymbol{x}(t_n, \boldsymbol{x}_0)) \end{bmatrix}$$

for some known instants $t_1, \dots, t_n$ and functions $U_1, \dots, U_q$. In this case, we may represent the QoIs in terms of the *Koopman operator* [24,25]. If we denote as $\Phi^t(\boldsymbol{x}_0) := \boldsymbol{x}(t, \boldsymbol{x}_0)$ the flow of the system [2, Chapter 1] then the Koopman operator K_t maps $f : \mathbb{R}^d \to \mathbb{R}$ to

$$K_t f(\boldsymbol{x}_0) := f(\Phi^t(\boldsymbol{x}_0)).$$

This allows us to write

$$\boldsymbol{U}(\boldsymbol{x}_0) = \begin{bmatrix} K_{t_1} U_1(\boldsymbol{x}_0) \\ \vdots \\ K_{t_q} U_q(\boldsymbol{x}_0) \end{bmatrix}.$$

The expected value of $\boldsymbol{U}$ can also be represented in terms of the pushforward of Π_{x_0} under Φ^t at each $t_1, \dots, t_q$. In fact,

$$\int K_t f(\boldsymbol{x}_0) d\Pi_{x_0}(\boldsymbol{x}_0) = \int f(\Phi^t(\boldsymbol{x}_0)) d\Pi_{x_0}(\boldsymbol{x}_0) = \int f(\boldsymbol{x}_t) d\Phi^t_* \Pi_{x_0}(\boldsymbol{x}_t)$$

or, equivalently, $\boldsymbol{x}(t, \boldsymbol{x}_0) \sim \Phi^t_* \Pi_{x_0}$ when $\boldsymbol{x}_0 \sim \Pi_{x_0}$. Hence,

$$\mathbf{E}_{\boldsymbol{x}_0 \sim \Pi_{x_0}}[\boldsymbol{U}(\boldsymbol{x}_0)] = \begin{bmatrix} \mathbf{E}_{\boldsymbol{x}_0 \sim \Pi_{x_0}}[K_{t_1} U_1(\boldsymbol{x}_0)] \\ \vdots \\ \mathbf{E}_{\boldsymbol{x}_0 \sim \Pi_{x_0}}[K_{t_q} U_q(\boldsymbol{x}_0)] \end{bmatrix} = \begin{bmatrix} \mathbf{E}_{\boldsymbol{x}_{t_1} \sim \Phi^{t_1}_* \Pi_{x_0}}[U_1(\boldsymbol{x}_{t_1})] \\ \vdots \\ \mathbf{E}_{\boldsymbol{x}_{t_q} \sim \Phi^{t_q}_* \Pi_{x_0}}[U_q(\boldsymbol{x}_{t_q})] \end{bmatrix}.$$

This identity has been used previously in [16,17] to reduce the computational burden of computing the expected values in the right-hand side, which requires sampling from the measures $\Phi^{t_1}_* \Pi_{x_0}, \dots, \Phi^{t_q}_* \Pi_{x_0}$, by sampling from Π_{x_0} and then applying the maps $\Phi^{t_1}, \dots, \Phi^{t_q}$.

2.1.3. Parametric systems

In the random differential equation (2) only $\boldsymbol{x}_0$ is assumed to be uncertain. However, random differential equations in $\mathbb{R}^d$ of the form

$$\begin{cases} \dot{\boldsymbol{y}}(t) = \boldsymbol{g}(t, \boldsymbol{y}(t), \boldsymbol{\vartheta}_0) \\ \boldsymbol{y}(0) = \boldsymbol{y}_0 \end{cases} \qquad \text{where} \quad (\boldsymbol{y}_0, \boldsymbol{\vartheta}_0) \sim \Pi_{(y_0, \vartheta_0)} \tag{5}$$

where *both* the initial condition $\boldsymbol{y}_0$ and a p-dimensional parameter ϑ_0 distribute according to an unknown probability measure $\Pi_{(y_0, \vartheta_0)}$ are frequent in practice. An equation of this form can be represented as the equation of the form (2) in $\mathbb{R}^d \times \mathbb{R}^p$ by defining

$$\boldsymbol{x} := \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{\vartheta} \end{bmatrix}, \quad \boldsymbol{x}_0 := \begin{bmatrix} \boldsymbol{y}_0 \\ \boldsymbol{\vartheta}_0 \end{bmatrix}, \quad \boldsymbol{f}(t, \boldsymbol{x}) = \begin{bmatrix} \boldsymbol{g}(t, \boldsymbol{y}, \boldsymbol{\vartheta}) \\ 0 \end{bmatrix} \quad \text{and} \quad \Pi_{x_0} = \Pi_{(y_0, \vartheta_0)},$$

and then considering the random differential equation

$$\begin{cases} \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(t, \boldsymbol{x}(t)) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0 \end{cases} \qquad \text{where} \quad \boldsymbol{x}_0 \sim \Pi_{x_0}$$

The first d state variables of a trajectory of this system correspond to a trajectory of (5). In this form, the uncertainty is on the initial condition and thus our methods can be applied to quantify the uncertainty in any *finite-dimensional parameter* of the equation.

2.2. Tikhonov regularization as non-parametric maximum a posteriori

To define a family $\{\Pi_{x_0|\theta} : \boldsymbol{\theta} \in \Theta\}$ for $\boldsymbol{x}_0$ that can adapt to complex data we proceed as follows. First, in most practical applications we can usually assume that the initial condition $\boldsymbol{x}_0$ must belong to a sufficiently large *bounded* set. Therefore, we choose a compact set $\Omega \subset \mathbb{R}^d$ containing the possible values of $\boldsymbol{x}_0$, and we *fix* a finite positive Borel measure μ_Ω on Ω that will act as a *dominating measure*; we further assume, without loss of generality, that Ω is the support of μ_Ω. We will restrict ourselves to probability measures on Ω that have a density ρ with respect to μ_Ω, i.e., measures for which we have $\Pi = \rho\mu_\Omega$, and for which the densities belong to the space of (classes of equivalence of μ_Ω-almost everywhere equal) μ_Ω-square-integrable functions

$$L^2_\mu(\Omega) := \left\{ \varphi : \Omega \to \mathbb{R} : \int_\Omega \varphi(\boldsymbol{x}_0)^2 d\mu_\Omega(\boldsymbol{x}_0) < \infty \right\},$$

which becomes a Hilbert space when endowed with the inner product

$$\langle \varphi, \psi \rangle_{L^2_\mu} = \int_\Omega \varphi(\boldsymbol{x}_0)\psi(\boldsymbol{x}_0) d\mu_\Omega(\boldsymbol{x}_0).$$

To ensure that the integral expressions that arise are well-defined, we assume that the QoIs also belong to this space, whence

$$\int_\Omega \|\boldsymbol{U}(\boldsymbol{x}_0)\|_2^2 \, d\mu_\Omega(\boldsymbol{x}_0) < \infty. \tag{6}$$

Since μ_Ω is finite, the functions in $L^2_\mu(\Omega)$ are absolutely integrable and $1 \in L^2_\mu(\Omega)$. As a consequence, the subset of probability densities

$$P_\mu := \left\{ \rho \in L^2_\mu(\Omega) : \ \rho \geq 0, \int_\Omega \rho(\boldsymbol{x}_0) d\mu_\Omega(\boldsymbol{x}_0) = 1 \right\}$$

in $L^2_\mu(\Omega)$ is closed and convex.

Restricting ourselves to densities in P_μ allows us to leverage standard techniques to construct finite-dimensional approximations to approximate a density ρ. Let $V_n \subset L^2_\mu(\Omega)$ be a finite dimensional subspace with $\dim(V_n) = n$. Since not all elements of V_n are densities, we define

$$P_{\mu,n} := V_n \cap P_\mu \tag{7}$$

Let $\varphi_1, \ldots, \varphi_n$ be an orthonormal basis for V_n and, to simplify notation, define

$$\boldsymbol{\varphi} = \begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_n \end{bmatrix}.$$

When $\rho \in P_{\mu,n}$ we can represent the density as

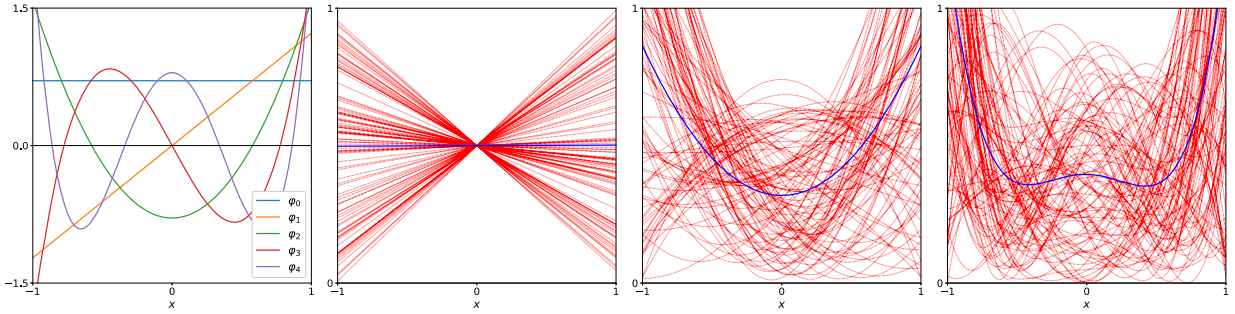
$$\rho = \langle \boldsymbol{a}, \boldsymbol{\varphi} \rangle = \sum_{i=1}^n a_i \varphi_i \tag{8}$$

for a suitable coordinate vector $\boldsymbol{a} \in \mathbb{R}^n$. An appealing property of this representation is that the expected value of any function $f \in L^2_\mu(\Omega)$ with respect to this density depends only on the orthogonal projection f_{V_n} of f onto V_n. In fact,

$$\mathbf{E}_{\boldsymbol{x}_0 \sim \rho}[f(\boldsymbol{x}_0)] = \sum_{i=1}^n a_i \int_\Omega \varphi_i(\boldsymbol{x}_0) f(\boldsymbol{x}_0) d\mu_\Omega(\boldsymbol{x}_0) = \sum_{i=1}^n a_i \int_\Omega \varphi_i(\boldsymbol{x}_0) f_{V_n}(\boldsymbol{x}_0) d\mu_\Omega(\boldsymbol{x}_0) = \mathbf{E}_{\boldsymbol{x}_0 \sim \rho}[f_{V_n}(\boldsymbol{x}_0)].$$

This becomes useful, for instance, when the distance from f to V_n is small, or when V_n encodes some desirable property of ρ, such as smoothness or localization.

It is not clear *a priori* if $P_{\mu,n}$ is empty. To ensure it is not, and to parameterize densities of the form (8) in terms of their coordinates, we need to first restrict the possible values of this vector to ensure that the expansion defines a density. We defer the proof of this proposition to Appendix A.1.



(a) Normalized Legendre polynomials (b) Random densities of degree 1 (c) Random densities of degree 2 (d) Random densities of degree 4

Fig. 1. Random densities on $\Omega = [-1, 1]$ with μ_Ω the Lebesgue measure sampled from the set $P_{\mu,n}$ generated by the normalized Legendre polynomials and the prior in (9). The blue line shows the expected density, whereas the red lines show sampled densities. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Proposition 1. *Suppose that $\varphi_1, \dots, \varphi_n$ are bounded and that φ_1 is strictly positive on Ω. Then there exists a convex set $\Theta_{V_n} \subset \mathbb{R}^{n-1}$ with non-empty interior, a vector $\boldsymbol{a}_n \in \mathbb{R}^n$, and a $n \times (n-1)$ matrix $\boldsymbol{V}$ with $\boldsymbol{V}^*\boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{V}^*\boldsymbol{a}_n = 0$ such that*

$$\langle \boldsymbol{a}, \boldsymbol{\varphi} \rangle \in P_{\mu,n} \quad \Leftrightarrow \quad \exists\, \boldsymbol{\theta} \in \Theta_{V_n} \,:\, \boldsymbol{a} = \boldsymbol{a}_n + \boldsymbol{V}\boldsymbol{\theta}.$$

In particular, $P_{\mu,n}$ is non-empty.

Proposition 1 and the representation (8) allows us to identify the set of parameters Θ_{V_n} with the set of densities $P_{\mu,n}$ whence a prior on $\boldsymbol{\theta} \in \Theta_{V_n}$ will induce a prior on $\rho \in P_{\mu,n}$. Fix $\lambda > 0$ and define on $\mathbb{R}^{n-1}$ the probability density with respect to the Lebesgue measure given by

$$\rho_\theta(\boldsymbol{\theta}) = \frac{\chi_{\Theta_{V_n}}(\boldsymbol{\theta}) e^{-\lambda\|\theta\|_2^2/2}}{\int_{\Theta_{V_n}} e^{-\lambda\|\theta\|_2^2/2}\, d\boldsymbol{\theta}}, \tag{9}$$

where the normalizing factor is strictly positive as Θ_{V_n} has non-empty interior, and where $\chi_{\Theta_{V_n}}$ is the indicator function of the set Θ_{V_n}. In Fig. 1 we see how this approach can help us draw random densities from $P_{\mu,n}$.

Suppose that $\Pi_{u\,|\,\bar{u}}$ has density

$$\rho_{\boldsymbol{u}\,|\,\bar{\boldsymbol{u}}}(\boldsymbol{u}) \propto e^{-\psi(\boldsymbol{u},\bar{\boldsymbol{u}})}$$

for some function ψ on $\mathbb{R}^d \times \mathbb{R}^d$. If we use the prior for $\boldsymbol{\theta}$ in (9) in the model (4) then the likelihood is proportional to

$$\chi_{\Theta_{V_n}}(\boldsymbol{\theta}) \exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{u}\|_2^2}{2\sigma^2}\right) \exp\left(-\psi\left(\boldsymbol{u}, \int_\Omega \boldsymbol{U}(\boldsymbol{x}_0)\rho_\theta(\boldsymbol{x}_0)\, d\mu_\Omega(\boldsymbol{x}_0)\right)\right) \exp\left(-\frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right)$$

where the constant factors depend on λ and σ. Instead of attempting to compute, or to sample from, the posterior we will compute the maximum *a posteriori* (MAP) estimate for $\boldsymbol{\theta}$, which is the solution to

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^s, \boldsymbol{u}\in\mathbb{R}^q,}{\text{maximize}} \quad \log(\chi_{\Theta_{V_n}}(\boldsymbol{\theta})) - \frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{u}\|_2^2 - \psi\left(\boldsymbol{u}, \int_\Omega \boldsymbol{U}(\boldsymbol{x}_0)\rho_\theta(\boldsymbol{x}_0) d\mu_\Omega(\boldsymbol{x}_0)\right) - \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$

The term $\log(\chi_{\Theta_{V_n}}(\boldsymbol{z}))$ essentially imposes the constraint $\boldsymbol{\theta} \in \Theta_{V_n}$ whereas $\boldsymbol{u}$ remains unconstrained. Hence, we write

$$\underset{\boldsymbol{\theta}\in\Theta_{V_n}}{\text{maximize}} \quad -L\left(\sum_{i=1}^s \theta_i \int_\Omega \boldsymbol{U}(\boldsymbol{x}_0)\varphi_i(\boldsymbol{x}_0) d\mu_\Omega(\boldsymbol{x}_0), \boldsymbol{y}\right) - \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$

where

$$L(\bar{\boldsymbol{u}}, \boldsymbol{y}) := \inf_{\boldsymbol{u}\in\mathbb{R}^q} \left(\psi(\boldsymbol{u}, \bar{\boldsymbol{u}}) + \frac{1}{2\sigma^2}\|\boldsymbol{u}-\boldsymbol{y}\|_2^2\right)$$

is the *Moreau envelope* [26, Definition 1.22] of ψ on its first argument. The envelope enjoys some regularity, namely, if $\boldsymbol{u} \mapsto \psi(\boldsymbol{u}, \bar{\boldsymbol{u}})$ is lower semicontinuous and majorizes a quadratic for a fixed $\bar{\boldsymbol{u}}$ then $\boldsymbol{y} \mapsto L(\bar{\boldsymbol{u}}, \boldsymbol{y})$ is continuous [26, Example 1.24 and Theorem 1.25].

This problem is the classical problem of *ridge regression* or *Tikhonov regularization* to estimate $\boldsymbol{\theta}$. It is clearly a finite-dimensional problem and, when L is convex on $\bar{\boldsymbol{u}}$, it is also convex. We can leverage Proposition 1 to represent this problem in terms of ρ itself. In fact, since $\varphi_1, \dots, \varphi_n$ are orthonormal, we have the isometry

$$\|\rho\|^2_{L^2_\mu} = \|\boldsymbol{a}\|_2^2 = \|\boldsymbol{a}_n + \boldsymbol{V}\boldsymbol{\theta}\|_2^2 = \|\boldsymbol{a}_n\|_2^2 + \|\boldsymbol{\theta}\|_2^2.$$

Since the constraint $\boldsymbol{\theta} \in \Theta_{V_n}$ is equivalent to $\rho_{\boldsymbol{\theta}} \in P_{\mu,n}$ and the norm of $\boldsymbol{a}_n$ is independent of $\boldsymbol{\theta}$, finding the MAP for $\boldsymbol{\theta}$ is equivalent to finding the *density* that solves the problem

$$\underset{\rho \in P_{\mu,n}}{\text{maximize}} \quad -L\left(\int_\Omega \boldsymbol{U}(\boldsymbol{x}_0)\rho(\boldsymbol{x}_0)d\mu_\Omega(\boldsymbol{x}_0), \boldsymbol{y}\right) - \frac{\lambda}{2}\|\rho\|^2_{L^2_\mu}. \tag{10}$$

This problem is also a form of regularization, but for the density itself, which emphasizes the choice of V_n over the particular choice of basis $\varphi_1, \dots, \varphi_n$. If $\{V_n\}_{n\in\mathbb{N}}$ is an increasing sequence of subspaces for which

$$\overline{\bigcup_{n\in\mathbb{N}} V_n} = L^2_\mu(\Omega) \tag{11}$$

then, from (7), we conclude that the asymptotic regime when $n \gg 1$ leads to the *infinite-dimensional problem*

$$\underset{\rho \in P_{\mu}}{\text{maximize}} \quad -L\left(\int_\Omega \boldsymbol{U}(\boldsymbol{x}_0)\rho(\boldsymbol{x}_0)d\mu_\Omega(\boldsymbol{x}_0), \boldsymbol{y}\right) - \frac{\lambda}{2}\|\rho\|^2_{L^2_\mu}. \tag{12}$$

This is a *nonparametric method* in that the parameter, i.e., the density ρ, lies on an infinite-dimensional space. We can prove that, under suitable conditions and in a suitable sense, the sequence of MAP estimates found by solving (10) for increasingly large values of n converges to a density found by solving (12). We defer the proof of the proposition to A.2

Proposition 2. *Suppose that $\{V_n\}_{n\in\mathbb{N}}$ is an increasing sequence of finite-dimensional subspaces of $L^2_\mu(\Omega)$ for which (11) holds, and suppose L is non-negative and continuous on its first argument. Furthermore, suppose that for every $n \in \mathbb{N}$ the problem (10) has an optimal solution $\rho^{\star,n}$ and that (12) has at least one solution. Then a subsequence of $\{\rho^{\star,n}\}_{n\in\mathbb{N}}$ converges weakly to a solution to (12).*

The proposition formalizes the intuition that a solution to (12) represents the limit of MAP estimates for the density. However, to be truly considered a MAP estimate, we would need to introduce a suitable prior on the set P_μ. This would result in a Bayesian nonparametric method [23]. We do not address the question whether there exists such a prior. Instead, we focus on analyzing the properties of (12), that is, the existence, uniqueness, and features of its optimal solution.

3. Analysis of the variational problem

We study the family of problems induced by (12) given by

$$\begin{aligned} \underset{\rho\in L^2_\mu(\Omega),\, \bar{\boldsymbol{u}}\in C:\, \rho\geq 0}{\text{minimize}} \quad & \frac{1}{2}\|\rho\|^2_{L^2_\mu} + \frac{1}{\lambda}L(\boldsymbol{A}\bar{\boldsymbol{u}}, \boldsymbol{y}) \\ \text{subject to} \quad & \int_\Omega \rho(\boldsymbol{x}_0)d\mu_\Omega(\boldsymbol{x}_0) = 1, \quad \int_\Omega \rho(\boldsymbol{x}_0)\boldsymbol{U}(\boldsymbol{x}_0)d\mu_\Omega(\boldsymbol{x}_0) = \bar{\boldsymbol{u}}, \end{aligned} \tag{13}$$

where $C \subset \mathbb{R}^q$ is a closed convex set representing restrictions on $\bar{\boldsymbol{u}}$ induced by the prior $\Pi_{u\,|\,\bar{u}}$, e.g., non-negativity constraints, bound constraints or norm constraints, and L is a loss function which we assume is convex on its first argument. The regularization parameter $\lambda > 0$ controls the trade-off between the loss and the squared-norm penalization term on ρ.

Proposition 3. *There exists a unique minimizer to (13).*

We defer the proof of this proposition to Appendix B.1. Observe that the method proposed in [16] is equivalent to solving

$$\begin{aligned} \underset{\rho\in L^2_\mu(\Omega):\, \rho\geq 0}{\text{minimize}} \quad & \frac{1}{2}\|\rho\|^2_{L^2_\mu} \\ \text{subject to} \quad & \int_\Omega \rho(\boldsymbol{x}_0)d\mu_\Omega(\boldsymbol{x}_0) = 1, \quad \int_\Omega \rho(\boldsymbol{x}_0)\boldsymbol{U}(\boldsymbol{x}_0)d\mu_\Omega(\boldsymbol{x}_0) = \bar{\boldsymbol{u}} \end{aligned} \tag{14}$$

and that this is a particular case of (13).

From now on, we let $\rho^\star$ be the optimal solution to (13) and we let $\Pi^\star := \rho^\star \mu_\Omega$ be the probability measure associated to the optimal density $\rho^\star$. Since (13) is an infinite-dimensional problem, it is necessary to find numerical approximations to its solution. Although this may be achieved by constructing suitable spaces V_n, we leverage convex duality instead to reformulate the problem in an equivalent form that is amenable to accurate approximations of its solution.

3.1. The dual problem

Instead of attempting to solve or to approximate (13) directly, we leverage convex duality to find an equivalent formulation that is amenable to numerical methods for high-dimensional optimization problems. We briefly review the arguments behind convex duality. First, rewrite (13) as the equivalent problem,

$$\begin{aligned} \underset{\rho\in L^2_\mu(\Omega),\,\bar{\boldsymbol{u}}\in\mathbb{R}^q,\,\boldsymbol{v}\in\mathbb{R}^m}{\text{minimize}} \quad & \frac{1}{2}\|\rho\|^2_{L^2_\mu} + \mathbb{I}_{L^{2+}_\mu}(\rho) + \frac{1}{\lambda}L(\boldsymbol{v},\boldsymbol{y}) + \mathbb{I}_C(\bar{\boldsymbol{u}}) \\ \text{subject to} \quad & \int_\Omega \rho(\boldsymbol{x}_0)\,d\mu_\Omega(\boldsymbol{x}_0) = 1, \quad \int_\Omega \rho(\boldsymbol{x}_0)\boldsymbol{U}(\boldsymbol{x}_0)\,d\mu_\Omega(\boldsymbol{x}_0) = \bar{\boldsymbol{u}}, \quad \boldsymbol{A}\bar{\boldsymbol{u}} = \boldsymbol{v}, \end{aligned} \tag{15}$$

where $\mathbb{I}_{L^{2+}_\mu}$ is the indicator function of the closed and convex cone of non-negative functions in $L^2_\mu(\Omega)$, and $\mathbb{I}_C$ is the indicator function of the convex set $C\subset\mathbb{R}^q$. If we define the Hilbert spaces

$$X_P := L^2_\mu(\Omega)\times\mathbb{R}^q\times\mathbb{R}^m \quad \text{and} \quad X_D := \mathbb{R}\times\mathbb{R}^q\times\mathbb{R}^m,$$

the linear map $\mathcal{B}: X_P\to X_D$ as

$$\mathcal{B}(\rho,\bar{\boldsymbol{u}},\boldsymbol{v}) = \begin{bmatrix} -\int_\Omega \rho(\boldsymbol{x}_0)d\mu_\Omega(\boldsymbol{x}_0) \\ -\int_\Omega \rho(\boldsymbol{x}_0)\boldsymbol{U}(\boldsymbol{x}_0)d\mu_\Omega(\boldsymbol{x}_0) + \bar{\boldsymbol{u}} \\ -\boldsymbol{A}\bar{\boldsymbol{u}} + \boldsymbol{v} \end{bmatrix},$$

the vector $\boldsymbol{r}\in X_D$ as

$$\boldsymbol{r} = \begin{bmatrix} -1 \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix},$$

and the convex functions $f: X_P\to\mathbb{R}$ and $g: X_D\to\mathbb{R}$ as

$$f(\rho,\bar{\boldsymbol{u}},\boldsymbol{v}) = \frac{1}{2}\|\rho\|^2_{L^2_\mu} + \mathbb{I}_{L^{2+}_\mu}(\rho) + \frac{1}{\lambda}L(\boldsymbol{v},\boldsymbol{y}) + \mathbb{I}_C(\bar{\boldsymbol{u}}),$$

$$g(\eta,\boldsymbol{\omega},\boldsymbol{\xi}) = \mathbb{I}_0((\eta,\boldsymbol{\omega},\boldsymbol{\xi}) - \boldsymbol{r}),$$

then (15) can be written equivalently as the unconstrained problem

$$\underset{(\rho,\bar{\boldsymbol{u}},\boldsymbol{v})\in X_P}{\text{minimize}} \quad f(\rho,\bar{\boldsymbol{u}},\boldsymbol{v}) + g(\mathcal{B}(\rho,\bar{\boldsymbol{u}},\boldsymbol{v}) - \boldsymbol{r}). \tag{16}$$

We call this the *primal problem* and its optimal solution $(\rho^\star,\bar{\boldsymbol{u}}^\star,\boldsymbol{v}^\star)$ the *optimal primal variables*. Its *dual problem* is

$$\underset{(\eta,\boldsymbol{\omega},\boldsymbol{\xi})\in X_D}{\text{maximize}} \quad -f^\star(-\mathcal{B}^*(\eta,\boldsymbol{\omega},\boldsymbol{\xi})) - g^\star(\eta,\boldsymbol{\omega},\boldsymbol{\xi}) \tag{17}$$

where $f^\star$ and $g^\star$ are the convex conjugates [27, Definition 13.1] of f and g and $\mathcal{B}^*: X_D\to X_P$ is the adjoint of $\mathcal{B}$. The optimal solution $(\eta^\star,\boldsymbol{\omega}^\star,\boldsymbol{\xi}^\star)$ yields the *optimal dual variables*. In some cases, solving the dual problem may be much easier than solving the primal problem. However, this is useful only if we are able to recover the optimal primal variables by solving the dual problem. When strong duality holds, the optimal values for (16) and (17) coincide, and we may obtain the optimal primal variables from the optimal dual variables. In the literature, this is also referred as *total duality*. We defer the proof of the following theorem to Appendix B.2.

Theorem 1. *Strong duality holds for* (13)*. The dual problem is the finite-dimensional problem*

$$\underset{\eta,\boldsymbol{\omega},\boldsymbol{\xi}}{\textit{maximize}} \quad -H(\eta,\boldsymbol{\omega}) + \eta - \frac{1}{\lambda}L^\star(-\lambda\boldsymbol{\xi},\boldsymbol{y}) - \mathbb{I}^\star_C(\boldsymbol{A}^t\boldsymbol{\xi} - \boldsymbol{\omega}),$$

where $L^\star$ is the convex dual of L on its first variable, $\mathbb{I}^\star_C$ is the convex dual of $\mathbb{I}_C$ and $H:\mathbb{R}\times\mathbb{R}^q\to\mathbb{R}$ is the convex function

$$H(\alpha,\boldsymbol{\beta}) = \frac{1}{2}\int_\Omega (\alpha + \langle\boldsymbol{\beta},\boldsymbol{U}(\boldsymbol{x}_0)\rangle_2)^2_+\,d\mu_\Omega(\boldsymbol{x}_0). \tag{18}$$

Furthermore, if $(\eta^\star,\boldsymbol{\omega}^\star,\boldsymbol{\xi}^\star)$ are the optimal dual variables then the optimal primal variables are

$$\begin{aligned} \rho^\star(\boldsymbol{x}_0) &:= (\eta^\star + \langle\boldsymbol{\omega}^\star,\boldsymbol{U}(\boldsymbol{x}_0)\rangle_2)_+, \\ \bar{\boldsymbol{u}}^\star &:= \int_\Omega (\eta^\star + \langle\boldsymbol{\omega}^\star,\boldsymbol{U}(\boldsymbol{x}_0)\rangle_2)_+\boldsymbol{U}(\boldsymbol{x}_0)\,d\mu_\Omega(\boldsymbol{x}_0), \\ \boldsymbol{v}^\star &:= \boldsymbol{A}\bar{\boldsymbol{u}}^\star. \end{aligned}$$

There are several consequences of this result. First, the theorem fully characterizes class of density functions that can be obtained by solving (13). Hence, when the optimal dual variables $\eta^\star, \boldsymbol{\omega}^\star$ are known, we may evaluate the optimal probability density to compute any statistics about the initial condition and about the QoIs. Second, the class of density functions is no more complex than that defined by the QoIs. To be concrete, if $U_1, \dots, U_q$ are the components of $\boldsymbol{U}$, then the optimal probability density can only have the form

$$\rho^\star(\boldsymbol{x}_0) = \left(\eta^\star + \sum_{i=1}^p \omega_i^\star U_i(\boldsymbol{x}_0)\right)_+$$

or, equivalently, it must be the positive part of some function in the span of $1, U_1, \dots, U_p \in L^2_\mu(\Omega)$. Third, we may change the properties of the optimal density found by solving (13) by modifying the reference measure μ_Ω. Fourth, it shows that we can solve (13) by solving its dual problem, which is equivalent to the convex minimization problem

$$\underset{\eta,\boldsymbol{\omega},\boldsymbol{\xi}}{\text{minimize}} \quad H(\eta,\boldsymbol{\omega}) - \eta + \frac{1}{\lambda} L^\star(-\lambda\boldsymbol{\xi}, \boldsymbol{y}) + \mathbb{I}_C^\star(\boldsymbol{A}^t\boldsymbol{\xi} - \boldsymbol{\omega}). \tag{19}$$

This problem is not only finite-dimensional, having $1 + q + m$ variables, but also unconstrained. Each one of the terms in the dual objective encodes the information about the problem: the first about the QoI $\boldsymbol{U}$, the second about the loss L, and the third about the *a priori* constraints we impose on $\bar{\boldsymbol{u}}$. Usually, the last two terms reflect modeling choices and there is some freedom to select them. In particular, we may choose the priors in such a way that the loss L and the set of constraints C have a small computational burden when solving (19). In contrast, the first term depends on the QoIs and the available data. As a consequence, it is typically the first term that will contribute most to the computational cost.

The complexity of solving an unconstrained optimization problem depends on the regularity of its objective function. An important property of H is that it is differentiable, with Lipschitz continuous derivatives. We defer the proof to Appendix B.4.

Proposition 4. *The function H defined in (18) is continuously differentiable, with partial derivatives*

$$\partial_\alpha H(\alpha,\boldsymbol{\beta}) = \int_\Omega (\alpha + \langle \boldsymbol{\beta}, \boldsymbol{U}(\boldsymbol{x}_0)\rangle_2)_+ d\mu_\Omega(\boldsymbol{x}_0),$$

$$\nabla_{\boldsymbol{\beta}} H(\alpha,\boldsymbol{\beta}) = \int_\Omega \boldsymbol{U}(\boldsymbol{x}_0)(\alpha + \langle \boldsymbol{\beta}, \boldsymbol{U}(\boldsymbol{x}_0)\rangle_2)_+ d\mu_\Omega(\boldsymbol{x}_0).$$

In particular, ∇H is Lipschitz continuous with

$$\mathrm{Lip}(\nabla H) \le \int_\Omega \max\{1, \|\boldsymbol{U}(\boldsymbol{x}_0)\|_2^2\}\, d\mu_\Omega(\boldsymbol{x}_0).$$

Therefore, when $L^\star$ and $\mathbb{I}_C^\star$ are differentiable, we may solve (19) using gradient descent or projected gradient descent. These algorithms are simple to implement, and their accelerated versions attain optimal convergence rates on the class of convex functions.

A natural question is whether we can leverage similar arguments to solve (13) when ρ is restricted to a finite-dimensional subspace V_n of $L^2_\mu(\Omega)$. Interestingly, a similar result to Theorem 1 holds in this case.

Theorem 2. *Let $V_n \subset L^2_\mu(\Omega)$ be a finite-dimensional subspace containing at least one strictly positive function. Let $1_{V_n}, U_{1,V_n}, \dots, U_{q,V_n}$ be the orthogonal projections of $1, U_1, \dots, U_n$ onto V_n and define $\boldsymbol{U}_{V_n}$ accordingly. Consider the problem*

$$\begin{aligned} \underset{\rho\in V_n,\, \bar{\boldsymbol{u}}\in\mathbb{R}^q,\, \boldsymbol{v}\in\mathbb{R}^m}{\text{minimize}} \quad & \frac{1}{2}\|\rho\|^2_{L^2_\mu} + \mathbb{I}_{V_n^+}(\rho) + \frac{1}{\lambda} L(\boldsymbol{v},\boldsymbol{y}) + \mathbb{I}_C(\bar{\boldsymbol{u}}) \\ \text{subject to} \quad & \int_\Omega \rho(\boldsymbol{x}_0)\, d\mu_\Omega(\boldsymbol{x}_0) = 1, \quad \int_\Omega \rho(\boldsymbol{x}_0)\boldsymbol{U}(\boldsymbol{x}_0)\, d\mu_\Omega(\boldsymbol{x}_0) = \bar{\boldsymbol{u}}, \quad \boldsymbol{A}\bar{\boldsymbol{u}} = \boldsymbol{v}. \end{aligned}$$

where V_n^+ is the closed convex cone of non-negative functions in V_n. Then strong duality holds for this problem, and its dual problem is the finite-dimensional problem

$$\underset{\eta,\boldsymbol{\omega},\boldsymbol{\xi}}{\text{maximize}} \quad -H_n(\eta,\boldsymbol{\omega}) + \eta - \frac{1}{\lambda} L^\star(-\lambda\boldsymbol{\xi},\boldsymbol{y}) - \mathbb{I}_C^\star(\boldsymbol{A}^t\boldsymbol{\xi} - \boldsymbol{\omega}).$$

where $H_n : \mathbb{R}\times\mathbb{R}^q \to \mathbb{R}$ is the convex function

$$H_n(\alpha,\boldsymbol{\beta}) = \frac{1}{2}\int_\Omega \left(\alpha 1_{V_n}(\boldsymbol{x}_0) + \langle \boldsymbol{\beta}, \boldsymbol{U}_{V_n}(\boldsymbol{x}_0)\rangle_2\right)_+^2 d\mu_\Omega(\boldsymbol{x}_0).$$

Furthermore, if the primal problem is feasible, and $(\eta^\star, \boldsymbol{\omega}^\star, \boldsymbol{\xi}^\star)$ are the optimal dual variables then the optimal primal variables are

$$\rho^\star(\boldsymbol{x}_0) := \left(\eta^\star 1_{V_n}(\boldsymbol{x}_0) + \langle \boldsymbol{\omega}^\star, \boldsymbol{U}_{V_n}(\boldsymbol{x}_0)\rangle_2\right)_+,$$
$$\bar{\boldsymbol{u}}^\star := \int_\Omega \left(\eta^\star 1_{V_n}(\boldsymbol{x}_0) + \langle \boldsymbol{\omega}^\star, \boldsymbol{U}_{V_n}(\boldsymbol{x}_0)\rangle_2\right)_+ \boldsymbol{U}(\boldsymbol{x}_0)\, d\mu_\Omega(\boldsymbol{x}_0),$$
$$\boldsymbol{v}^\star := \boldsymbol{A}\bar{\boldsymbol{u}}^\star.$$

3.2. Covariance terms

A critical condition to apply duality for (13) is that the constraints on the expectation are linear in the density, i.e.,

$$\int_\Omega \boldsymbol{U}(\boldsymbol{x}_0)\rho(\boldsymbol{x}_0) d\mu_\Omega(\boldsymbol{x}_0) = \bar{\boldsymbol{u}}.$$

In general, this constraint could be replaced by any other convex constraint and our arguments would apply with minor modifications. An exception would be any explicit constraints on the covariance matrix, as these would take the form

$$\boldsymbol{\mu} := \int_\Omega \boldsymbol{x}_0\rho(\boldsymbol{x}_0) d\mu_\Omega(\boldsymbol{x}_0) \quad \text{and} \quad \boldsymbol{\Sigma} := \int_\Omega (\boldsymbol{x}_0 - \boldsymbol{\mu})(\boldsymbol{x}_0 - \boldsymbol{\mu})^T \rho(\boldsymbol{x}_0) d\mu_\Omega(\boldsymbol{x}_0);$$

while the former is convex, the second is not. We can overcome this limitation as follows. From the linear equality constraints

$$\boldsymbol{\mu} := \int_\Omega \boldsymbol{x}_0\rho(\boldsymbol{x}_0) d\mu_\Omega(\boldsymbol{x}_0) \quad \text{and} \quad \boldsymbol{X} := \int_\Omega \boldsymbol{x}_0\boldsymbol{x}_0^T \rho(\boldsymbol{x}_0) d\mu_\Omega(\boldsymbol{x}_0)$$

we recover the covariance as $\boldsymbol{\Sigma} = \boldsymbol{X} - \boldsymbol{\mu}\boldsymbol{\mu}^T$. Then, a convex loss or a convex penalization term for $\boldsymbol{X}$ and $\boldsymbol{\mu}$ can be added to the objective function in (13). Similarly, some constraints can be imposed by choosing a suitable set C. In both cases, the same arguments used in Theorem 1 allow us to compute the dual of the problem.

3.3. Special problem instances

3.3.1. Quadratic loss

A typical choice for the loss in problems of the form (13) is the *quadratic loss*. In this case,

$$L(\boldsymbol{v}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{y}\|_2^2 \quad \text{and} \quad C = \mathbb{R}^q.$$

This loss is obtained, for instance, when using the prior

$$\rho_{\boldsymbol{u}\,|\,\bar{\boldsymbol{u}}}(\boldsymbol{u}) \propto e^{-\frac{1}{2\tau^2}\|\boldsymbol{u}-\bar{\boldsymbol{u}}\|_2}.$$

From this,

$$L^\star(\boldsymbol{v}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{\xi}\|_2^2 + \langle \boldsymbol{\xi}, \boldsymbol{y}\rangle \quad \text{and} \quad \mathbb{I}_C^\star(\boldsymbol{\omega}) = \mathbb{I}_0(\boldsymbol{\omega}).$$

Since

$$\mathbb{I}_0(\boldsymbol{A}^t\boldsymbol{\xi} - \boldsymbol{\omega}) < \infty \;\Leftrightarrow\; \boldsymbol{\omega} = \boldsymbol{A}^t\boldsymbol{\xi},$$

this instance of the problem becomes

$$\underset{\eta,\boldsymbol{\xi}}{\text{minimize}} \quad H(\eta, \boldsymbol{A}^t\boldsymbol{\xi}) - \eta + \frac{\lambda}{2}\|\boldsymbol{\xi}\|_2^2 - \langle \boldsymbol{\xi}, \boldsymbol{y}\rangle_2.$$

This instance of (23) is unconstrained, and the objective is differentiable with Lipschitz gradient. Hence, it can be solved using gradient descent with acceleration [28–30]. The computational burden at each iteration of the algorithm is then dominated by the computing the value of H and of its derivatives.

3.3.2. Lebesgue measure and star domains

In practice, the computational burden of solving (19) will be dominated by the cost of evaluating H and its derivatives. However, for some special choices of Ω and μ_Ω we can represent H as an integral over the closed Euclidean ball $\bar{B}_2$. In this case, instead of developing quadrature rules for every Ω we may use quadrature methods over the fixed set $\bar{B}_2$; the trade-off is that the regularity of the integrand may decrease.

Recall that a closed set $\Omega \subset \mathbb{R}^d$ is *star-shaped* with respect to some $\boldsymbol{x}_c \in \Omega$ if it contains a neighborhood of $\boldsymbol{x}_c$ and if for any $\boldsymbol{x} \in \Omega$ the ray

$$\{\boldsymbol{x}_c + s(\boldsymbol{x} - \boldsymbol{x}_c) : \; s > 0\}$$

does not intersect the boundary of Ω more than once [31, Section 2]. Its Minkowski functional

$$\gamma_0(\boldsymbol{x}) := \inf\{t > 0 : t\boldsymbol{x} \in \Omega_0\},$$

is positively homogeneous, non-negative and continuous and [31, Theorem 1]

$$\Omega := \{\boldsymbol{x} \in \mathbb{R}^q : \gamma_0(\boldsymbol{x}) \leq 1\}.$$

Suppose that Ω is star-shaped with respect to some

$$\boldsymbol{x}_c \in \mathbf{int}(\Omega).$$

By translating Ω we have

$$H(\alpha, \boldsymbol{\beta}) = \frac{1}{2} \int_{\Omega_0} (\alpha + \langle \boldsymbol{\beta}, \boldsymbol{U}(\boldsymbol{x}_0 + \boldsymbol{x}_c)\rangle_2)_+^2 d\boldsymbol{x}_0$$

where $\Omega_0 := \Omega - \boldsymbol{x}_c$ contains a neighborhood of the origin. Let γ_0 be the Minkowski functional associated to Ω_0 and let

$$T(\boldsymbol{x}) = \begin{cases} \frac{\|\boldsymbol{x}\|_2}{\gamma_0(\boldsymbol{x})}\boldsymbol{x} & \boldsymbol{x} \neq 0 \\ \mathbf{0} & \boldsymbol{x} = 0. \end{cases} \tag{20}$$

Then

$$T : \bar{B}_2 \to \Omega_0$$

is an homeomorphism. First, T is continuous by inspection and

$$\gamma_0(T(\boldsymbol{x})) = \|\boldsymbol{x}\|_2$$

implies $T : \bar{B}_2 \to \Omega_0$. Its inverse is

$$T^{-1}(\boldsymbol{x}) = \frac{\gamma_0(\boldsymbol{x})}{\|\boldsymbol{x}\|_2}\boldsymbol{x}. \tag{21}$$

Since the same arguments show $T^{-1} : \Omega_0 \to \bar{B}_2$ is continuous we conclude T is an homeomorphism. This map allows us to represent H as an integral over the closed unit ball by changing the integrand. We defer the proof of the following result to Appendix B.5.

Theorem 3. *Let Ω be a star shaped set with respect to some $\boldsymbol{x}_c \in \mathbf{int}(\Omega)$ and let μ_Ω be the Lebesgue measure. Under the change of variables $\boldsymbol{x} \mapsto T(\boldsymbol{x})$ we have*

$$H(\alpha, \boldsymbol{\beta}) = \frac{1}{2} \int_{\bar{B}_2} \left(\alpha + \left\langle \boldsymbol{\beta}, \boldsymbol{U}\left(\frac{\|\boldsymbol{x}_0\|_2}{\gamma_0(\boldsymbol{x}_0)}\boldsymbol{x}_0 + \boldsymbol{x}_c\right)\right\rangle_2\right)_+^2 \left(\frac{\|\boldsymbol{x}_0\|_2}{\gamma_0(\boldsymbol{x}_0)}\right)^d d\boldsymbol{x}_0. \tag{22}$$

The representation in (22) nicely shows the influence of $\boldsymbol{U}$ and Ω on the dual objective. The integrand is a product of two factors. The first one is essentially the same as the integrand in (18). The second is a weight that quantifies how different is Ω_0 from $\bar{B}_2$. This weight is homogeneous of degree 0 and it depends only on the *direction* of the integrand, but not on its magnitude. Remark the dependence on the quotient is *exponential* on the dimension, whence the points that will contribute the most to the integral are those where the quotient is larger.

4. Interpretation as regularized regression

The problem (13) can be interpreted both as using MAP to estimate a density ρ, or as a regression problem to estimate the expected value $\bar{\boldsymbol{u}}$. In fact, if we let $R_U = R_U(\bar{\boldsymbol{u}})$ be the optimal value to (14) then (13) becomes

$$\underset{\bar{\boldsymbol{u}} \in C}{\text{minimize}} \quad R_U(\bar{\boldsymbol{u}}) + \frac{1}{\lambda} L(\boldsymbol{A}\bar{\boldsymbol{u}}, \boldsymbol{y}). \tag{23}$$

This is a *regularized regression problem* to estimate $\bar{\boldsymbol{u}}$. Observe that it is finite-dimensional, and that the number of variables is equal to the number q of QoIs. The terms that depend on the density ρ in (13) act implicitly as a regularizer on the estimate of $\bar{\boldsymbol{u}}$. In this context, it is of interest to determine the properties of the regularizer. We defer the proof of the proposition to Appendix C.1.

Proposition 5. *Suppose*

$$\int_\Omega \|\boldsymbol{U}(\boldsymbol{x}_0)\|_2^2 \, d\mu_\Omega(\boldsymbol{x}_0) < \infty.$$

Then, the function $R_U : \mathbb{R}^q \to \mathbb{R}$ is proper and convex. Furthermore, its domain is the convex set

$$\mathbf{dom}(R_U) := \left\{ \int_\Omega \rho(\boldsymbol{x}_0)\boldsymbol{U}(\boldsymbol{x}_0) d\mu_\Omega(\boldsymbol{x}_0) : \rho \in L^2_\mu(\Omega),\ \rho \geq 0,\ \int_\Omega \rho(\boldsymbol{x}_0)\, d\mu_\Omega(\boldsymbol{x}_0) = 1 \right\} \tag{24}$$

Under mild assumptions (23) has at least one optimal solution [27, Proposition 11.13]. One important consequence of this result is that (13) *implicitly* enforces the constraint $\bar{\boldsymbol{u}} \in \mathbf{dom}(R_U)$. In turn, this constraint depends both on the choice of $\boldsymbol{U}$ and the set Ω. In fact,

$$\mathbf{dom}(R_U) \subset \mathbf{cvx}\{\boldsymbol{U}(\boldsymbol{x}_0) : \boldsymbol{x}_0 \in \Omega\}.$$

In the special case when Ω is a convex body, we can show that at least in some special cases $\mathbf{dom}(R_U)$ is open and hence this inclusion may be strict. We defer the proof to Appendix C.2.

Proposition 6. *Suppose that Ω is a convex body, that μ_Ω is the Lebesgue measure, and that*

$$\boldsymbol{U}(\boldsymbol{x}_0) = \boldsymbol{x}_0.$$

Then $\mathbf{dom}(R_U) = \mathbf{int}(\Omega)$. Furthermore, $R_U(\bar{\boldsymbol{u}}) \to \infty$ as $\bar{\boldsymbol{u}} \to \mathbf{bd}(\Omega)$.

The statement extends with minor modifications to other choices of measure μ_Ω on Ω. Remark that, in general, if μ_Ω is absolutely continuous with respect to the Lebesgue measure then the same arguments in the proposition yield

$$\mathbf{dom}(R_U) \subset \mathbf{supp}\left(\frac{d\mu_\Omega}{d\boldsymbol{x}_0}\right).$$

The formulation (23) shows a clear distinction between the full and partial measurement models. The partial measurement model allows us to codify the effect of $\boldsymbol{U}$ in the regularization term, whereas the measurement matrix $\boldsymbol{A}$ appears only in the loss. There is an extensive literature on the performance of estimators based on regularized regression. Results of this type can be applied to understand how close is the optimal solution to (23) to the true expectation $\bar{\boldsymbol{u}}_0$.

4.1. Linear systems and convex sets

When the system (2) is linear and the QoIs are also linear in the state, we obtain a special case of (23). By a slight abuse of notation, we write

$$\boldsymbol{U}(\boldsymbol{x}_0) = \boldsymbol{U}\boldsymbol{x}_0$$

for a $q \times d$ matrix $\boldsymbol{U}$. Suppose that Ω is a convex body. Then, we may decompose the regularization function R_U as follows. Let $R^\circ_U = R^\circ_U(\boldsymbol{\mu})$ be the optimal value of the problem

$$\begin{aligned} &\underset{\rho \in L^2_\mu(\Omega):\, \rho \geq 0}{\text{minimize}} && \frac{1}{2}\int_\Omega \rho(\boldsymbol{x}_0)^2 d\mu_\Omega(\boldsymbol{x}_0) \\ &\text{subject to} && \int_\Omega \rho(\boldsymbol{x}_0) d\mu_\Omega(\boldsymbol{x}_0) = 1, \quad \int_\Omega \rho(\boldsymbol{x}_0)\boldsymbol{x}_0 d\mu_\Omega(\boldsymbol{x}_0) = \boldsymbol{\mu}. \end{aligned}$$

By Proposition 5 and 6 we deduce that R°_U is proper, convex and $\mathbf{dom}(R^\circ_U) = \mathbf{int}(\Omega)$. Then, the value of $R_U = R_U(\bar{\boldsymbol{u}})$ is the optimal value of the problem

$$\underset{\boldsymbol{\mu} \in \mathbb{R}^d}{\text{minimize}} \quad R^\circ_U(\boldsymbol{\mu}) \quad \text{subject to} \quad \boldsymbol{U}\boldsymbol{\mu} = \bar{\boldsymbol{u}}.$$

Therefore, we can rewrite (23) equivalently as

$$\underset{\boldsymbol{\mu} \in \mathbb{R}^d, \bar{\boldsymbol{u}} \in C}{\text{minimize}} \quad R^\circ_U(\boldsymbol{\mu}) + \frac{1}{\lambda} L(\boldsymbol{A}\bar{\boldsymbol{u}}, \boldsymbol{y}) \quad \text{subject to} \quad \boldsymbol{U}\boldsymbol{\mu} = \bar{\boldsymbol{u}}. \tag{25}$$

In some cases, this problem can be further reduced to a standard form. We defer the proof to Appendix C.3.

Proposition 7. *Suppose that Ω is a closed Euclidean ball of radius $r > 0$ and center $\boldsymbol{\mu}_0$ and that μ_Ω is the Lebesgue measure. Then there exists a non-decreasing convex function $\varphi : [0, 1] \to \mathbb{R}$ such that*

$$R^\circ_U(\boldsymbol{\mu}) = \varphi\left(\frac{\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|_2}{r}\right)$$

and $\varphi(s) \to \infty$ *as* $s \to 1$. *Furthermore,* (25) *becomes*

$$\underset{\boldsymbol{\mu}\in\mathbb{R}^d,\, \bar{\boldsymbol{u}}\in C,\, s\geq 0}{\text{minimize}} \quad \varphi(s) + \frac{1}{\lambda} L(\boldsymbol{A}\bar{\boldsymbol{u}}, \boldsymbol{y}) \quad \text{subject to} \quad \boldsymbol{U}\boldsymbol{\mu} = \bar{\boldsymbol{u}}, \qquad \frac{\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|_2}{r} \leq s.$$

Therefore, for linear systems and observables, and in the Euclidean ball and (a scaling of) the Lebesgue measure, the regularization effect on the expectation is very similar to that of Tikhonov regularization. This once again shows the implicit effect that the choice of Ω and of reference measure μ_Ω has on the behavior of the method.

5. Implementation

We now discuss some methods that can be used to solve numerically a special instance of (19). We will assume that $L^\star$ is differentiable with Lipschitz gradient, that $C = \mathbb{R}^q$ and that the measure μ_Ω is a probability measure on Ω; to make this assumption explicit, from now on we write Π_Ω. From these assumptions, it follows from the arguments in Section 3.3.1 that the dual problem has the form

$$\underset{\eta,\boldsymbol{\xi}}{\text{minimize}} \quad H(\eta, \boldsymbol{A}^t\boldsymbol{\xi}) - \eta + \frac{1}{\lambda} L^\star(\lambda\boldsymbol{\xi}, \boldsymbol{y}) \tag{26}$$

Furthermore, if we define

$$h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0) := \frac{1}{2}(\alpha + \langle \boldsymbol{\beta}, \boldsymbol{U}(\boldsymbol{x}_0)\rangle_2)_+^2$$

then H can be represented as

$$H(\alpha, \boldsymbol{\beta}) = \int_\Omega h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0)\, d\Pi_\Omega(\boldsymbol{x}_0) = \mathbf{E}_{\boldsymbol{x}_0\sim\Pi_\Omega}[h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0)].$$

Therefore, we can interpret (26) as either a deterministic or stochastic convex optimization problem. Depending on this choice, different algorithms can be implemented in practice.

5.1. Gradient descent methods

As (26) is unconstrained with a convex and smooth objective function, perhaps the simplest and most efficient method to solve this problem is accelerated gradient descent (GD) [28,32,29,30]. However, to simplify the exposition we will focus on gradient descent without acceleration. Given (η_0, ξ_0) and a sequence of steps $\{s_n\}_{n\in\mathbb{N}}$ we define for $n \geq 0$ the iterates

$$\eta_{n+1} = \eta_n - s_n(\partial_\alpha H(\eta_n, \boldsymbol{A}^t\boldsymbol{\xi}_n) - 1),$$
$$\boldsymbol{\xi}_{n+1} = \boldsymbol{\xi}_n - s_n(\boldsymbol{A}\nabla_{\boldsymbol{\beta}} H(\eta_n, \boldsymbol{A}^t\boldsymbol{\xi}_n) - \boldsymbol{y} + \lambda\boldsymbol{\xi}_n),$$

In practice, the steps can be selected using backtracking, or they can be constant. In this case, from Proposition 4, the step s must satisfy

$$\frac{1}{s} > \frac{1}{2}(1 + \|\boldsymbol{A}\|_{\text{op}}) \int_\Omega \max\{1, \|\boldsymbol{U}(\boldsymbol{x}_0)\|_2^2\}\, d\Pi_\Omega(\boldsymbol{x}_0) + \frac{1}{2}\lambda\,\text{Lip}(\nabla L^\star)$$

to ensure convergence.

Gradient descent requires evaluating at least the gradient of H at each iteration. Both H and ∇H are defined in terms of integrals that need to be approximated numerically for possibly large d. We discuss cubature formulæ and the *sample average approximation (SAA)*. Both methods essentially construct an approximation $\widehat{H}$ of H from which $\widehat{H}$ and $\nabla\widehat{H}$ can be computed efficiently. In practice, one can construct a low-accuracy approximation $\widehat{H}$ using few nodes or samples, to first find an approximation $(\widehat{\eta}^\star, \widehat{\boldsymbol{\xi}}^\star)$ to the solution to (26) in a few iterations, and then use this approximation as an initial iterate to solve (26) using a more accurate approximation.

5.1.1. Cubature formulæ

A first approach is to approximate H in low dimensions is to use cubature formulæ. They can be designed to exploit the structure of the integrand to obtain theoretical guarantees about their approximation properties [33,34]. A disadvantage is that they need to be tailored to the class of functions over which one wants to achieve an exact approximation, i.e., the *degree* of the formula, and to the domain of integration, e.g., the hypercube [35,36], the simplex [36,37] or the ball [36]. Their complexity, i.e., the number of nodes required for a desired accuracy, increases exponentially as the dimension increases (however, see [35,38]). Furthermore, their accuracy may vary according to the choice of Π_Ω.

When Π_Ω is a scalar multiple of the Lebesgue measure on Ω we can leverage Theorem 3 to represent H as the integral over the unit ball. In this case, we can use cubature formulæ for the Euclidean ball. Let

$$h^\circ(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0) = \frac{1}{2}\left(\alpha + \left\langle \boldsymbol{\beta}, \boldsymbol{U}\left(\frac{\|\boldsymbol{x}_0\|_2}{\gamma_0(\boldsymbol{x}_0)}\boldsymbol{x}_0 + \boldsymbol{x}_c\right)\right\rangle_2\right)_+^2 \left(\frac{\|\boldsymbol{x}_0\|_2}{\gamma_0(\boldsymbol{x}_0)}\right)^d.$$

Theorem 3 allows us to write

$$H(\alpha,\boldsymbol{\beta}) = \frac{|\partial\bar{B}_2(0,1)|}{|\Omega|}\int_0^1\left(\frac{1}{|\partial B_2(0,1)|}\int_{\partial\bar{B}_2(0,1)} h^\circ(\alpha,\boldsymbol{\beta},r\boldsymbol{\omega})\,dS(\boldsymbol{\omega})\right)r^{d-1}dr$$
$$=: \frac{|\partial\bar{B}_2(0,1)|}{|\Omega|}\int_0^1 \bar{h}^\circ(\alpha,\boldsymbol{\beta},r)\,r^{d-1}dr$$

Since there are terms in the definition of h° that are homogeneous of degree 0, we have

$$h^\circ(\alpha,\boldsymbol{\beta},r\boldsymbol{\omega}) = \frac{1}{2}\left(\alpha + \left\langle\boldsymbol{\beta},\, \boldsymbol{U}\left(\frac{\|\boldsymbol{\omega}\|_2}{\gamma_0(\boldsymbol{\omega})}r\boldsymbol{\omega}+\boldsymbol{x}_c\right)\right\rangle_2\right)_+^2\left(\frac{\|\boldsymbol{\omega}\|_2}{\gamma_0(\boldsymbol{\omega})}\right)^d,$$

whence the regularity of $\bar{h}^\circ(\alpha,\boldsymbol{\beta},r)$ is controlled by that of $\boldsymbol{U}$. Standard cubature formulæ can be used to approximate the integral on the radial variable, e.g., Gauss-Jacobi formulæ for $\alpha = d-1$ and $\beta = 0$. To evaluate $\bar{h}^\circ$ at the nodes, we may use cubature formulæ on the unit sphere, e.g., see [39,40].

5.1.2. Sample average approximation

Although cubature formulæ provide reasonable approximations, their limitations make them effective only in very low dimensions. As the systems under study become high-dimensional systems, Monte Carlo (MC) and Quasi-Monte Carlo (QMC) methods become more appropriate. In MC the nodes are sampled according to the probability measure Π_Ω whereas in QMC they are chosen from a *low-discrepancy* sequence [41–43]; in both, the nodes are assigned the same weight. In the optimization literature, e.g., see [44], this approach leads to the sample average approximation (SAA) method to solve (26).

In MC the key idea is to use the approximation

$$\widehat{H}(\alpha,\boldsymbol{\beta}) := \frac{1}{N}\sum_{i=1}^N h(\alpha,\boldsymbol{\beta},\boldsymbol{x}_0^{(i)}) \tag{27}$$

where $\boldsymbol{x}_0^{(1)},\dots,\boldsymbol{x}_0^{(N)}$ are i.i.d. samples from Π_Ω. Remark the samples are not modified once drawn. The gradient can be readily computed as

$$\partial_\alpha\widehat{H}(\alpha,\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^N(\alpha+\langle\boldsymbol{\beta},\,\boldsymbol{U}(\boldsymbol{x}_0^{(i)})\rangle_2)_+,$$
$$\nabla_{\boldsymbol{\beta}}\widehat{H}(\alpha,\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^N(\alpha+\langle\boldsymbol{\beta},\,\boldsymbol{U}(\boldsymbol{x}_0^{(i)})\rangle_2)_+\boldsymbol{U}(\boldsymbol{x}_0^{(i)}).$$

Therefore, it suffices to pre-compute $\{\boldsymbol{U}(\boldsymbol{x}_0^{(i)})\}_{i=1}^N$ to evaluate $\widehat{H}$ and its gradient efficiently.

The key quantity to control the accuracy is the variance of the approximation (27) which itself depends on $(\alpha,\boldsymbol{\beta})$. The following provides a bound on the variance as a function of $(\alpha,\boldsymbol{\beta})$. We defer its proof to Appendix D.1.

Proposition 8. *Let $(\alpha,\boldsymbol{\beta})\in\mathbb{R}\times\mathbb{R}^q$ and let $\boldsymbol{x}_0\sim\Pi_\Omega$. Let $\sigma^2=\sigma^2(\alpha,\boldsymbol{\beta})$ be the variance of the random variable $h(\alpha,\boldsymbol{\beta},\boldsymbol{x}_0)$. Then*

$$\sigma^2(\alpha,\boldsymbol{\beta}) \le 2|\alpha|^2\|\boldsymbol{\beta}\|_2^2\,\mathbf{E}_{\boldsymbol{x}_0',\boldsymbol{x}_0\overset{iid}{\sim}\Pi_\Omega}[\|\boldsymbol{U}(\boldsymbol{x}_0)-\boldsymbol{U}(\boldsymbol{x}_0')\|^2] + \|\boldsymbol{\beta}\|_2^4\,\mathbf{E}_{\boldsymbol{x}_0',\boldsymbol{x}_0\overset{iid}{\sim}\Pi_\Omega}[\|\boldsymbol{U}(\boldsymbol{x}_0)\|_2^2\|\boldsymbol{U}(\boldsymbol{x}_0)-\boldsymbol{U}(\boldsymbol{x}_0')\|_2^2].$$

Therefore, the approximation will tend to be accurate for small values of α and $\|\boldsymbol{\beta}\|_2$. This can be used as a way to increase the number of samples in (27) depending on the magnitude of the approximate solution found.

In QMC the approximation in (27) is constructed from points $\boldsymbol{x}_0^{(1)},\dots,\boldsymbol{x}_0^{(N)}$ that belong to a low-discrepancy sequence. The approximations converge faster than those obtained with MC, and significant improvements can be obtained in some cases; in others, the performance might be comparable to that of QMC [45]. There are standard constructions for low-discrepancy sequences in the hypercube, e.g., Halton, Sobol or Faure sequences. These can be used in the approximation (27) by scaling Ω and extending h by zero outside Ω. However, this leads to a possibly discontinuous integrand, which may substantially impact the performance of QMC. For this reason, the strategy introduced for the cubature formulæ can be appropriate, in which we leverage Theorem 3 to represent H as an integral over the unit ℓ^2-ball. In this case, we can use a classical one-dimensional sequence over the interval $[0,1]$ to integrate $\bar{h}^\circ$ on r and then use the low-discrepancy sequence proposed in [46] to approximate the values of $\bar{h}^\circ$.

5.2. Stochastic approximation methods

One of the drawbacks of (27) is that, similarly to cubature formulæ, we may need a very large number of samples to approximate the objective function accurately. An alternative approach is to use *stochastic approximation (SA)* methods [44]. A popular SA method

is stochastic gradient descent (SGD) which uses the gradient of the integrand on a random sample at each iteration. Let $\{\boldsymbol{x}_0^{(i)}\}_{i\in\mathbb{N}}$ be an i.i.d. sequence of Π_Ω random vectors. Given an initial iterate $(\eta_0, \boldsymbol{\xi}_0)$ and a sequence of steps $\{s_n\}_{n\in\mathbb{N}}$ we define the iterates

$$\eta_{n+1} = \eta_n - s_n(-1 + (\eta_n + \langle \boldsymbol{A}^t\boldsymbol{\xi}_n, \boldsymbol{U}(\boldsymbol{x}_0^{(n)})\rangle_2)_+),$$
$$\boldsymbol{\xi}_{n+1} = \boldsymbol{\xi}_n - s_n(\nabla L^\star(\lambda\boldsymbol{\xi}_n, \boldsymbol{y}) + (\eta_n + \langle \boldsymbol{A}^t\boldsymbol{\xi}_n, \boldsymbol{U}(\boldsymbol{x}_0^{(n)})\rangle_2)_+),$$

for $n \geq 0$. In contrast to SAA, the samples change at each iteration. Hence, the computational cost is dominated by the evaluation of $\boldsymbol{U}$ at a single point. However, the cost-per-iteration decreases at the expense of a decreased convergence rate.

One important factor impacting the performance of SGD is the choice of steps. It is known that if H is *strongly convex* then the steps may be chosen so that $s_n \sim n^{-1}$. However, as we do not know if the objective is strongly convex beforehand, we consider steps such that $s_n \sim n^{-1/2}$ [44].

In general, we may use a batch to estimate the gradient of H. Define the *unbiased* gradient estimate

$$\boldsymbol{g}_n(\alpha, \boldsymbol{\beta}) = \frac{1}{M}\sum_{m=1}^{M} \nabla h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0^{(n,m)}) \tag{28}$$

where $\boldsymbol{x}_0^{(n,1)}, \ldots, \boldsymbol{x}_0^{(n,M)} \overset{iid}{\sim} \Pi_\Omega$. One of the factors that impact the performance of SGD is the variance of this estimate. We defer the proof of the following proposition to Appendix D.2.

Proposition 9. *We have the bound*

$$\mathbf{E}_{\boldsymbol{x}_0^{(n,1)},\ldots,\boldsymbol{x}_0^{(n,M)}\sim\Pi_\Omega}[\|\boldsymbol{g}_n(\alpha,\boldsymbol{\beta}) - \nabla H(\alpha,\boldsymbol{\beta})\|_2^2] \leq \frac{1}{M}(|\alpha| + 6\|\boldsymbol{\beta}\|_2^2)\,\mathbf{E}_{\boldsymbol{x}_0',\boldsymbol{x}_0\sim\Pi_\Omega}[\|\boldsymbol{U}(\boldsymbol{x}_0') - \boldsymbol{U}(\boldsymbol{x}_0)\|_2^2]$$
$$+ \frac{2}{M}\|\boldsymbol{\beta}\|_2^2\,\mathbf{E}_{\boldsymbol{x}_0',\boldsymbol{x}_0\sim\Pi_\Omega}[\|\boldsymbol{U}(\boldsymbol{x}_0')\|_2^2\|\boldsymbol{U}(\boldsymbol{x}_0') - \boldsymbol{U}(\boldsymbol{x}_0)\|_2^2].$$

Once again, the variance approximation becomes accurate for small values of α and $\|\boldsymbol{\beta}\|_2$. This can be used as a way to select the size of the batch in (28) depending on the size of η and $\boldsymbol{A}^t\boldsymbol{\xi}$.

5.3. Samplers

Both in SAA and SGD it is necessary to be able to sample from Π_Ω. When Ω is convex and Π_Ω admits a density with respect to the Lebesgue measure, we can use *hit-and-run* [47] to sample efficiently from Π_Ω. This method is particularly efficient when Π_Ω is the uniform probability measure $\text{UNIF}(\Omega)$. One of its advantages is that the distribution of $\boldsymbol{x}_0^{(n)}$ converges efficiently to $\text{UNIF}(\Omega)$ in total variation, namely, in $O(d^3)$ steps [48, Corollary 1.2]. Furthermore, each iteration is computationally inexpensive: to sample uniformly from the unit sphere it suffices to sample from $\boldsymbol{g} \sim N(\boldsymbol{0}, \boldsymbol{I}_d)$ and let $\boldsymbol{\omega} = \boldsymbol{g}/\|\boldsymbol{g}\|_2$, and for the sets Ω that we consider, finding the intersection between a line and its boundary is efficient. Remark that we may get good approximations for the integral even though the distribution of the current iterate may still be far from the uniform distribution.

Another alternative when Ω is simple, or the system is low-dimensional, is to use rejection sampling. This is particularly useful when Ω is an ℓ_p-ball as in this case we may use either an ℓ_∞-ball or an ℓ_2-ball containing it to sample from the corresponding uniform measure.

6. Experiments

6.1. A linear oscillator

As an illustrative example we consider a damped harmonic oscillator

$$\ddot{x} + \gamma\dot{x} + \omega^2 x = 0$$

with $x(0) = 1$ and $\dot{x}(1) = 0$. The reduction to a first-order system yields

$$\begin{cases} \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(t, \boldsymbol{x}(t)) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0 \end{cases} \quad \text{for} \quad \boldsymbol{F} = \begin{bmatrix} 0 & 1 \\ -\omega^2 & -\gamma \end{bmatrix} \quad \text{and} \quad \boldsymbol{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{29}$$

This system has the form (2) for $\Pi_{x_0} = \delta_{[0\ 1]^*}$ and it is instructive to study the probability densities that are obtained by the method in this case. Using an atomic measure to model uncertainty allows us to study whether the optimal density $\rho^\star$ is able to concentrate around this point. Assume that the QoIs are

$$u_1(\boldsymbol{x}_0) = x_1(2, \boldsymbol{x}_0), \quad u_2(\boldsymbol{x}_0) = x_2(4, \boldsymbol{x}_0), \quad u_3(\boldsymbol{x}_0) = x_1(6, \boldsymbol{x}_0), \quad u_4(\boldsymbol{x}_0) = x_2(8, \boldsymbol{x}_0).$$

Then $\boldsymbol{U}$ can be written explicitly in terms of $\boldsymbol{x}_0$ as

Table 1
Bias and feasibility for the optimal probability measure.

Ω	r	$\|\boldsymbol{x}_0 - \boldsymbol{E}_{\boldsymbol{x}_0\sim\mu^\star}[\boldsymbol{x}_0]\|_2$				$\|\boldsymbol{U}(\boldsymbol{x}_0) - \boldsymbol{E}_{\boldsymbol{x}_0\sim\mu^\star}[\boldsymbol{U}(\boldsymbol{x}_0)]\|_2$			
		λ				λ			
		10^0	10^{-1}	10^{-2}	10^{-3}	10^0	10^{-1}	10^{-2}	10^{-3}
$\bar{B}_1$	0.5	9.76E-01	8.55E-01	7.07E-01	6.22E-01	7.32E-01	6.01E-01	4.88E-01	4.47E-01
	1.0	9.18E-01	6.58E-01	4.00E-01	2.26E-01	6.68E-01	4.26E-01	2.25E-01	1.64E-01
	2.0	7.60E-01	3.34E-01	8.98E-02	4.40E-02	5.19E-01	1.90E-01	3.42E-02	2.02E-02
$\bar{B}_2$	0.5	9.67E-01	8.03E-01	6.61E-01	5.82E-01	7.24E-01	5.59E-01	4.26E-01	3.81E-01
	1.0	8.87E-01	5.61E-01	3.17E-01	1.82E-01	6.35E-01	3.67E-01	1.83E-01	1.32E-01
	2.0	6.91E-01	2.26E-01	4.60E-02	6.76E-03	4.49E-01	1.37E-01	4.82E-02	1.39E-02
$\bar{B}_\infty$	0.5	9.56E-01	7.64E-01	6.19E-01	5.63E-01	7.09E-01	5.16E-01	3.84E-01	3.14E-01
	1.0	8.58E-01	4.89E-01	2.91E-01	1.15E-01	6.13E-01	2.95E-01	1.64E-01	8.82E-02
	2.0	6.34E-01	1.75E-01	2.02E-02	1.49E-02	4.04E-01	1.03E-01	9.87E-03	1.56E-02
$\bar{B}_1^+$	0.5	8.40E-01	7.61E-01	6.16E-01	5.56E-01	7.02E-01	6.16E-01	4.85E-01	4.03E-01
	1.0	6.92E-01	4.80E-01	2.99E-01	1.87E-01	6.38E-01	4.31E-01	2.42E-01	1.34E-01
	2.0	5.79E-01	3.02E-01	2.20E-01	7.57E-02	5.49E-01	2.55E-01	1.15E-01	6.66E-02
$\bar{B}_2^+$	0.5	8.04E-01	7.14E-01	5.86E-01	5.48E-01	6.95E-01	6.07E-01	5.03E-01	4.00E-01
	1.0	6.63E-01	4.40E-01	2.60E-01	1.84E-01	6.41E-01	4.28E-01	2.58E-01	1.82E-01
	2.0	6.92E-01	4.08E-01	2.12E-01	9.71E-02	5.94E-01	2.77E-01	1.32E-01	6.29E-02
$\bar{B}_\infty^+$	0.5	7.79E-01	6.99E-01	5.83E-01	5.81E-01	6.94E-01	6.02E-01	4.88E-01	4.62E-01
	1.0	6.56E-01	4.40E-01	2.62E-01	1.85E-01	6.50E-01	4.42E-01	2.32E-01	1.58E-01
	2.0	8.26E-01	4.59E-01	2.27E-01	1.14E-01	6.35E-01	3.15E-01	1.46E-01	7.23E-02

$$\boldsymbol{U}(\boldsymbol{x}_0) = \begin{bmatrix} \langle \boldsymbol{e}_1, e^{\boldsymbol{F}}\boldsymbol{x}_0\rangle_2 \\ \langle \boldsymbol{e}_2, e^{2\boldsymbol{F}}\boldsymbol{x}_0\rangle_2 \\ \langle \boldsymbol{e}_1, e^{4\boldsymbol{F}}\boldsymbol{x}_0\rangle_2 \\ \langle \boldsymbol{e}_2, e^{6\boldsymbol{F}}\boldsymbol{x}_0\rangle_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1^* e^{\boldsymbol{F}} \\ \boldsymbol{e}_2^* e^{2\boldsymbol{F}} \\ \boldsymbol{e}_1^* e^{4\boldsymbol{F}} \\ \boldsymbol{e}_2^* e^{6\boldsymbol{F}} \end{bmatrix} \boldsymbol{x}_0 =: \boldsymbol{U}\boldsymbol{x}_0. \tag{30}$$

Fig. 2a shows how x and $\dot{x}$ evolve in time, whereas Fig. 2b shows the trajectory of the system in state space. Fig. 2c shows the QoIs being measured. The QoI (30) can be computed efficiently and with high accuracy, i.e., the matrix $\boldsymbol{F}$ is non-defective and its exponential can be computed from its eigendecomposition, allowing us to focus on the density obtained with the method. To compare the effect that Ω has on the results, we consider 6 possible convex bodies: the closed ℓ_1-ball $\bar{B}_1$, the closed ℓ_2-ball $\bar{B}_2$, the closed ℓ_∞-ball $\bar{B}_\infty$, and their intersections with the positive orthant, which we denote as $\bar{B}_1^+$, $\bar{B}_2^+$ and $\bar{B}_\infty^+$. In all cases we assume that μ_Ω is the uniform measure. These choices will help us evaluate the impact of incorporating *a priori* information on Ω, i.e., the fact that $\boldsymbol{x}_0 \geq 0$. To approximate H we use Monte Carlo and we draw 10,000 samples from μ_Ω in each case using rejection sampling.

In Fig. 3 we show the densities obtained for $\bar{B}_2$. As $\boldsymbol{U}$ is a linear map, in all cases the density is the positive part of an affine function. For $r = 1/2$ we have that $\boldsymbol{x}_0 \notin \Omega$ and the problem becomes infeasible as $\lambda \to 0$. In Figs. 3a-3d we see that the density concentrates in the boundary as $\lambda \to 0$. In addition, there seems to be a small bias towards negative values of $\dot{x}$. This is a consequence of the particular structure of the matrix $\boldsymbol{U}$. For $r = 1$ we have that $\boldsymbol{x}_0 \in \mathbf{bd}(\Omega)$. As shown in Proposition 6, this implies that the problem becomes infeasible as $\lambda \to 0$. Figs. 3e-3h show the optimal density concentrating near the boundary. Observe that, as $\lambda \to 0$, the expected value of $\Pi^\star$ approaches the true $\boldsymbol{x}_0$. Finally, for $r = 2.0$ we have $\boldsymbol{x}_0 \in \mathbf{int}(\Omega)$. As Figs. 3i-3l show, when $\lambda \to 0$ the density somewhat concentrates around $\boldsymbol{x}_0$ and the expected value quickly approaches the true $\boldsymbol{x}_0$. However, as $\boldsymbol{U}$ is linear, the set of possible densities is not flexible enough to truly concentrate around the true $\boldsymbol{x}_0$, generating instead a large dispersion around it.

In Fig. 4 we show the densities obtained for $r = 1$ and all of the 6 choices of Ω. Remark that in all cases we have that $\boldsymbol{x}_0 \in \mathbf{int}(\Omega)$. In general, the behavior as $\lambda \to 0$ is the same observed for the ℓ_2-ball, and the density concentrates near the boundary and near $\boldsymbol{x}_0$. However, the behavior of the optimal density behaves quite differently depending on the choice of Ω. In fact, as $\lambda \to 0$ the dispersion on the marginal for $\dot{x}$ depends almost exclusively on the choice of Ω, yielding the least dispersion for $\bar{B}_1$, and the largest for $\bar{B}_\infty$. By considering the closed ℓ_p-balls intersected with the positive orthant, we are adding *a priori* information about the values of $x(0)$ and $\dot{x}(0)$. However, this leads to a bias in the expected value of $\Pi^\star$. The magnitude of this bias and its behavior as $\lambda \to 0$ depends on the structure of the boundary of Ω at $\boldsymbol{x}_0$.

In Table 1 we report the ℓ_2-bias when estimating $\boldsymbol{x}_0$ by the expected value of $\Pi^\star$ and the feasibility error between the true expected value $\boldsymbol{U}(\boldsymbol{x}_0)$ and that obtained using $\Pi^\star$.

This example also allows us to compare the densities obtained by solving the finite-dimensional approximations discussed in Sections 2.2 and 3.1. For this comparison we consider $\Omega = \bar{B}_\infty$ and $\lambda = 10^{-1}$, and we use the prior induced by the Haar wavelet basis on $[-1,1]^2$. A crucial property of the Haar wavelet basis is that it is piecewise constant, enabling the efficient characterization of the set Θ_{V_n}. The results obtained from the finite-dimensional approximations up to a level $L \in \{1,2,3\}$ can be seen in Fig. 5; recall that there are 4^{L+1} basis functions for each choice of L. As the decomposition level increases, the finite-dimensional approximation begins to resemble the solution shown in Fig. 4i. In Fig. 5d we see that the L^2_μ-distance decreases, as expected.

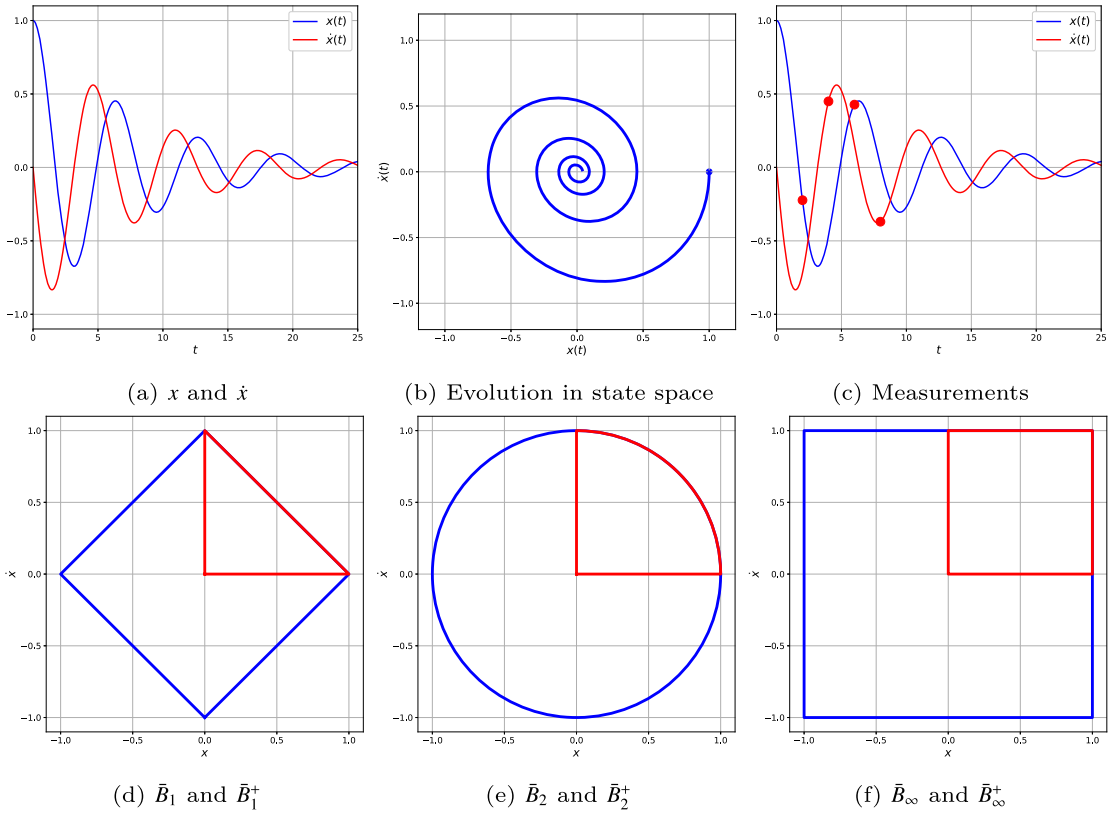


Fig. 2. (a) Evolution of the state variables x and $\dot{x}$ on the time interval $[0, 20]$ for the damped oscillator (29) for $\gamma = 1/4$ and $\omega = 1$. (b) Trajectory in phase space for $t \in [0, 20]$. (c) Observed values of the state variables. (d, e, f) Diagram of $\bar{B}_1$, the $\bar{B}_2$ and $\bar{B}_\infty$ in blue, and $\bar{B}_1^+$, $\bar{B}_2^+$ and $\bar{B}_\infty^+$ in red.

Therefore, the finite-dimensional approximations do converge to the solution to the infinite-dimensional problem. This validates our interpretation of the problem as the limit of the finite-dimensional models.

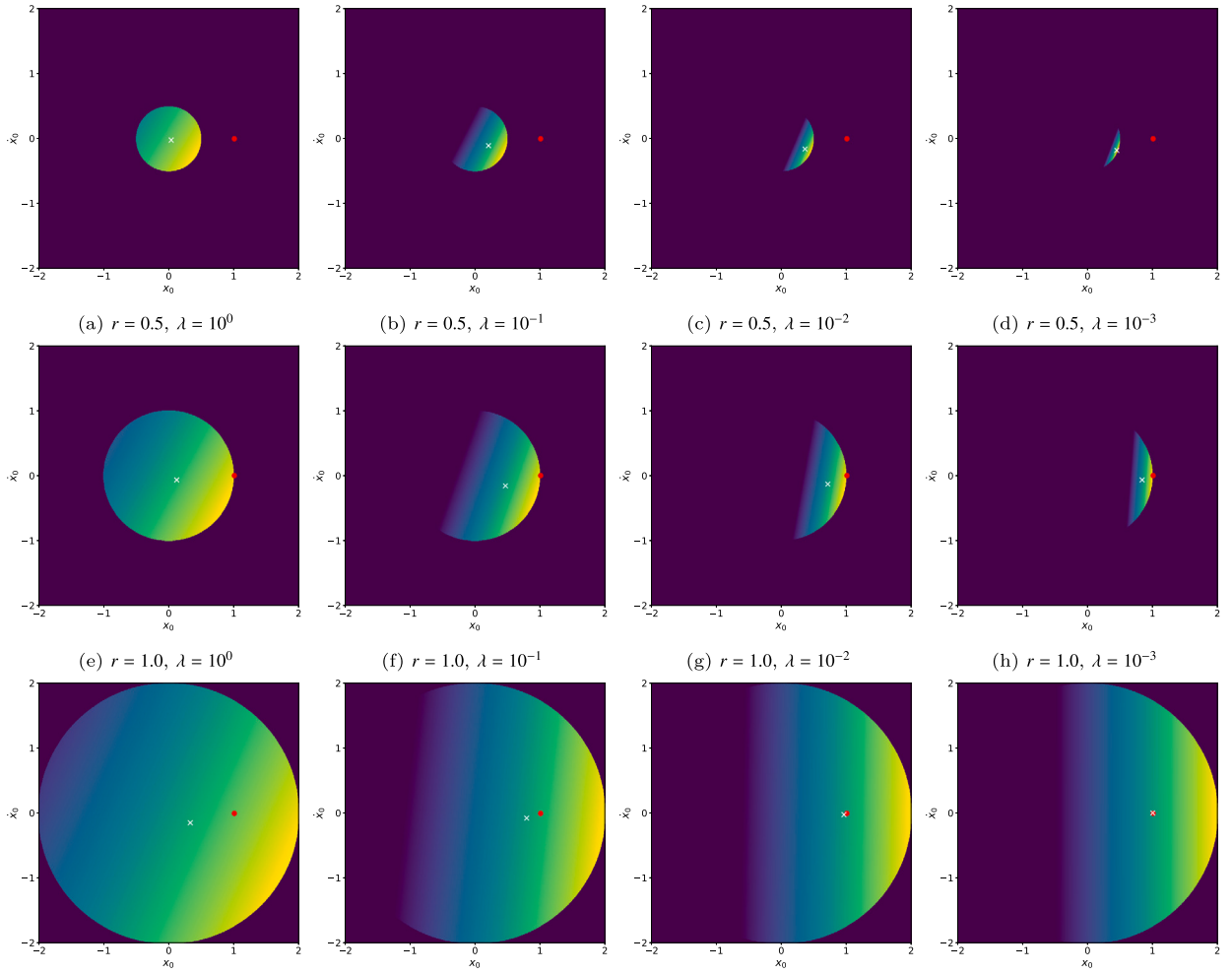
6.2. *Failure on a linear flow system*

Another illustrative example is to consider a linear flow system on a graph (Fig. 6). At the initial time the first node always has concentration equal to 1 whereas all other nodes always have concentration equal to zero. The *normal state* of the system occurs when the flow goes through node 2 and ends up at node 3, i.e., the capacity of the edges $2 \to 4$ and $4 \to 2$ are zero (Fig. 6a). The *failure state* occurs when the capacity of the edges $2 \to 4$ and $4 \to 2$ are non-zero and, as a consequence, there is a leak from node 2 to node 4 (Fig. 6b). The goal of this problem is to quantify the uncertainty on both on the initial condition and on the parameters $c_{2\to4}$ and $c_{4\to2}$ from measurements of the concentration at nodes 1 and 3 at times $\{1.0, 2.0, \ldots, 9.0, 10.0\}$.

We consider three scenarios of failure. In *Scenario A* the trajectories are randomly generated according to $c_{2\to4} \sim \Pi_{2\to4}$ with $\Pi_{2\to4} = \text{BETA}(30, 10)$ and $c_{4\to2} \sim \Pi_{4\to2}$ for $\Pi_{2\to4} = \text{BETA}(15/0.85, 10)$; in this case $\mathbf{E}[c_{2\to4}] = 0.75$ and $\mathbf{E}[c_{4\to2}] = 0.15$ (Fig. 6c). The sample trajectories and the average measurements are shown in Fig. 6d. In *Scenario B* the trajectories are generated from a mixture with probability $1/2$ between the normal state, i.e., $c_{2\to4} = c_{4\to2} = 0$, and the trajectories in Scenario B (Fig. 6e). Finally, in *Scenario C* we consider that the system operates normally for $t \leq 2.0$, but then fails for $t > 2.0$ where the parameters distribute as in Scenario A (Fig. 6f). In each case, we use the average observations over 100 samples.

To use the method we represent the model as the system of ODEs

$$\begin{aligned}
\dot{v}_1(t) &= -c_{1\to2} v_1(t) \\
\dot{v}_2(t) &= -(c_{2\to3} + c_{2\to4}(t)) v_2(t) + c_{1\to2} v_1(t) + c_{4\to2}(t) v_4(t) \\
\dot{v}_3(t) &= c_{2\to3} v_2(t) \\
\dot{v}_4(t) &= -c_{4\to2}(t) v_4(t) + c_{2\to4}(t) v_2(t) \\
\dot{c}_{2\to4}(t) &= 0 \\
\dot{c}_{4\to2}(t) &= 0.
\end{aligned}$$



(a) $r = 0.5,\ \lambda = 10^{0}$ (b) $r = 0.5,\ \lambda = 10^{-1}$ (c) $r = 0.5,\ \lambda = 10^{-2}$ (d) $r = 0.5,\ \lambda = 10^{-3}$

(e) $r = 1.0,\ \lambda = 10^{0}$ (f) $r = 1.0,\ \lambda = 10^{-1}$ (g) $r = 1.0,\ \lambda = 10^{-2}$ (h) $r = 1.0,\ \lambda = 10^{-3}$

(i) $r = 2.0,\ \lambda = 10^{0}$ (j) $r = 2.0,\ \lambda = 10^{-1}$ (k) $r = 2.0,\ \lambda = 10^{-2}$ (l) $r = 2.0,\ \lambda = 10^{-3}$

Fig. 3. Optimal density $\rho^\star$ for $\bar{B}_2$ for radii $r \in \{0.5, 1.0, 2.0\}$ and $\lambda \in \{10^0, 10^{-1}, 10^{-2}, 10^{-3}\}$. In each case, the values of the density have been normalized to the interval $[0, 1]$, the true $\boldsymbol{x}_0$ is represented as a red dot, and the expected value for $\rho^\star$ is shown as a white dot.

If we let

$$\boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}, \quad \boldsymbol{\vartheta} = \begin{bmatrix} c_{2\to 4} \\ c_{4\to 2} \end{bmatrix} \quad \text{and} \quad \boldsymbol{x} = \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{\vartheta} \end{bmatrix}$$

then we have the non-linear system

$$\begin{cases} \dot{\boldsymbol{x}}(t) = \boldsymbol{C}(\boldsymbol{x}(t))\boldsymbol{x}(t) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0 \end{cases} \tag{31}$$

where

$$\boldsymbol{C}(\boldsymbol{x}) = \begin{bmatrix} -c_{1\to 2} & 0 & 0 & 0 & 0 & 0 \\ c_{1\to 2} & -(c_{2\to 3} + x_5) & 0 & x_6 & 0 & 0 \\ 0 & c_{2\to 3} & 0 & 0 & 0 & 0 \\ 0 & x_5 & 0 & -x_6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

We consider the QoIs $\boldsymbol{U} : \mathbb{R}^6 \to \mathbb{R}^{20}$ given by

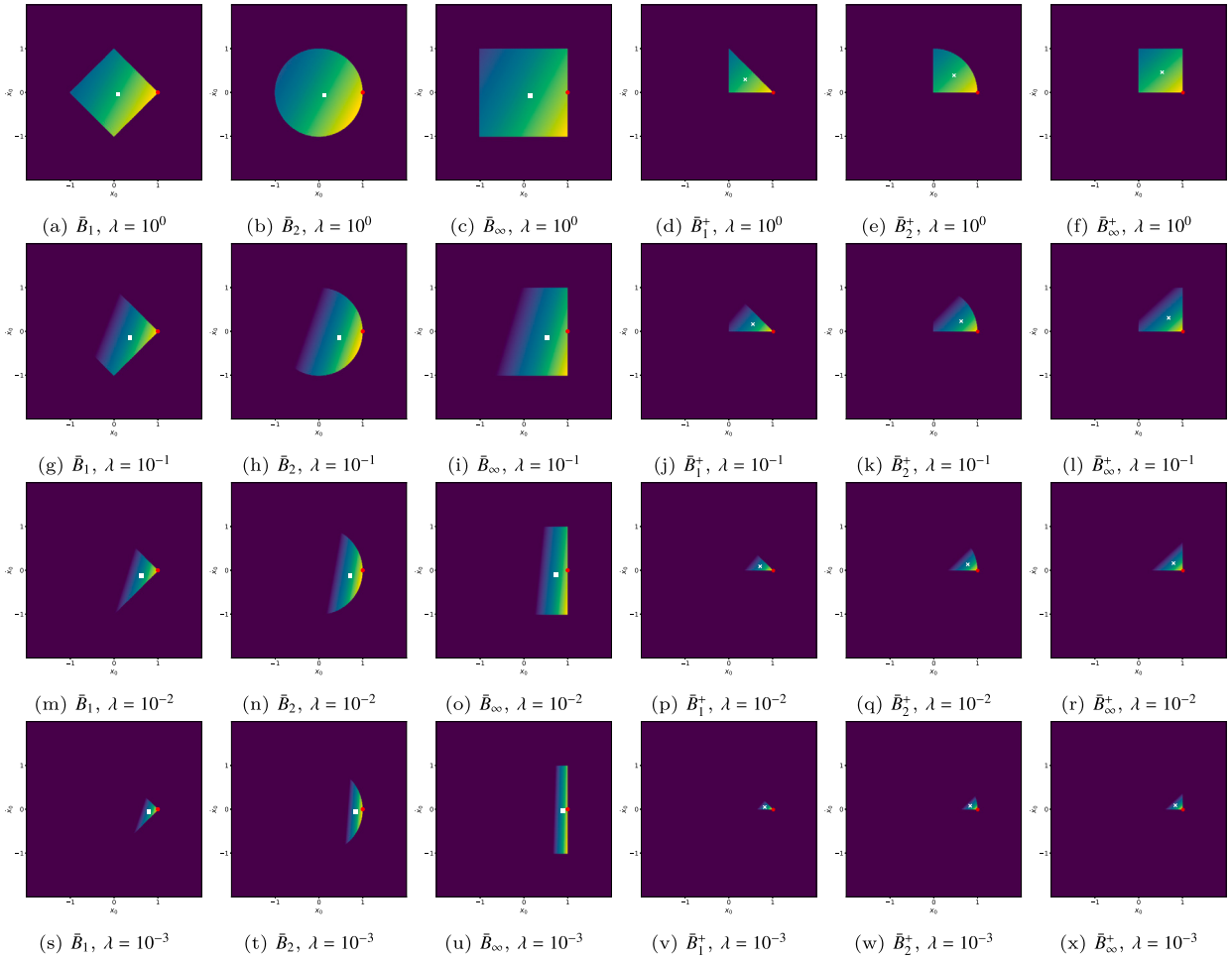
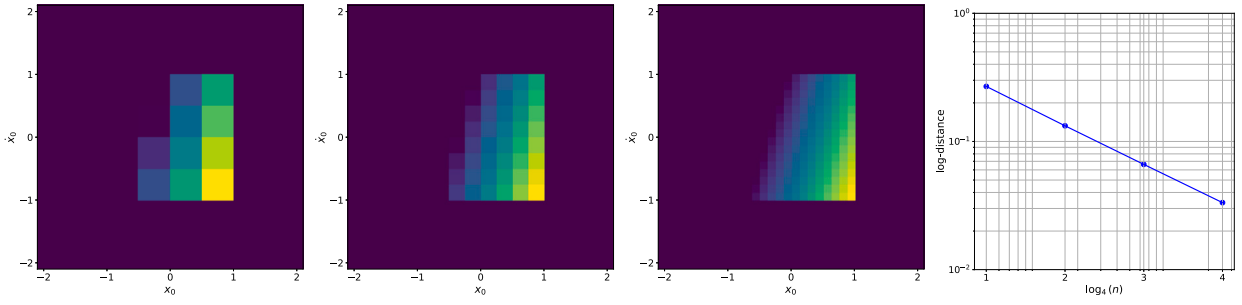


Fig. 4. Optimal density $\rho^\star$ for $r = 1.0$ and $\lambda \in \{1^0, 10^{-1}, 10^{-2}, 10^{-3}\}$. In each case, the values of the density have been normalized to the interval $[0, 1]$, the true $\boldsymbol{x}_0$ is represented as a red dot, and the expected value for $\rho^\star$ is shown as a white dot.



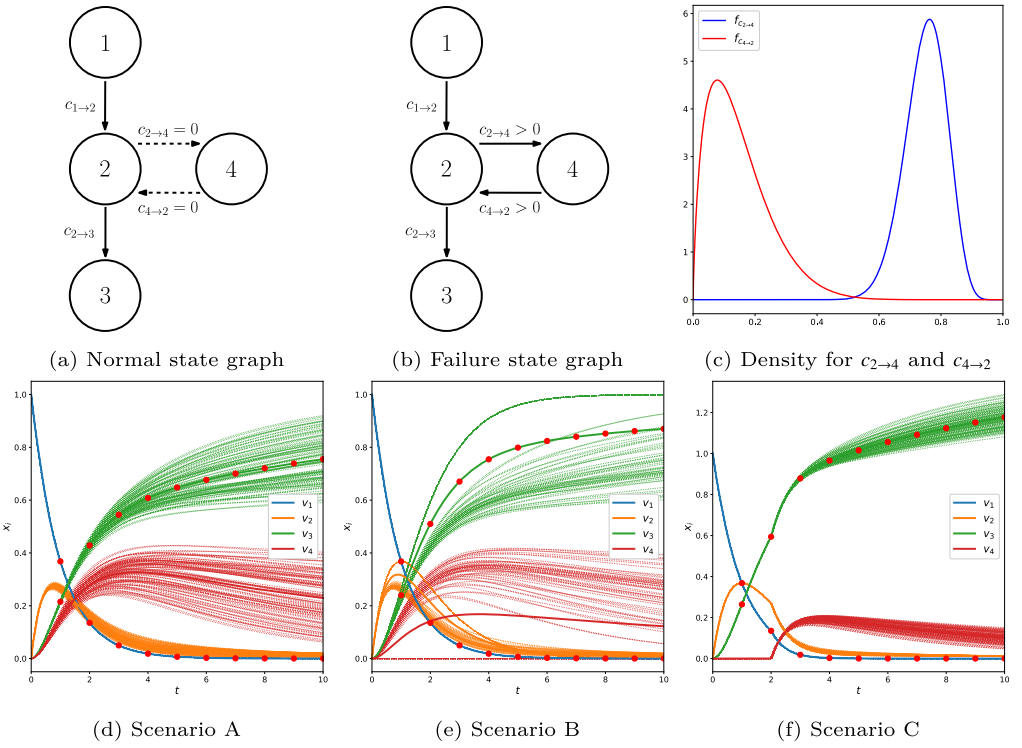
(a) Decomposition up to level $L = 1$ (b) Decomposition up to level $L = 2$ (c) Decomposition up to level $L = 3$ (d) L^2_μ-distance

Fig. 5. Solutions to the finite-dimensional problems using the Haar wavelet basis on $\bar{B}_\infty$. For each level L of the decomposition the corresponding subspace V_n has dimension $n = 4^{L+1}$. The L^2_μ-distance between the optimal density $\rho^\star$ obtained by solving the infinite-dimensional problem and the finite-dimensional problems decreases as expected.

$$U_1(\boldsymbol{x}_0) = v_1(1, \boldsymbol{x}_0), \quad \ldots, \quad U_{10}(\boldsymbol{x}_0) = v_1(10, \boldsymbol{x}_0), \quad U_{11}(\boldsymbol{x}_0) = v_3(1, \boldsymbol{x}_0), \quad \ldots, \quad U_{20}(\boldsymbol{x}_0) = v_3(10, \boldsymbol{x}_0)$$

and the measure Π_{x_0} on $\mathbb{R}^6$ for $\boldsymbol{x}_0$ is

$$\Pi_{x_0} = \delta_1 \otimes \delta_0 \otimes \delta_0 \otimes \delta_0 \otimes \pi_{2\to 4} \otimes \pi_{4\to 2}.$$



(a) Normal state graph (b) Failure state graph (c) Density for $c_{2\to4}$ and $c_{4\to2}$

(d) Scenario A (e) Scenario B (f) Scenario C

Fig. 6. (a, b) Graph representing the normal and failure states of the system. (c) Probability densities for $c_{2\to4}$ and $c_{4\to2}$. (d, e, f) Trajectories generated from 100 samples of $c_{2\to4}$ and $c_{4\to2}$ under each of the failure scenarios. In each case, the solid line represents the average trajectory whereas the red dots represent the average measurements.

For the numerical experiments, we use $c_{1\to2} = c_{2\to3} = 1$. As the solutions to the system diverge exponentially if $c_{2\to4} < 0$ or $c_{4\to2} < 0$ it is reasonable to impose that the support of the measure on the last 2 variables is contained in the positive orthant. Consequently, to perform the experiments we choose Ω as either $r\bar{B}_1^+, r\bar{B}_2^+$ or $r\bar{B}_\infty^+$, or as Cartesian products of the form $r\bar{B}_{p_1} \times r\bar{B}_{p_2}^+$ for $p_1, p_2 \in \{1, 2, \infty\}$ and some radius $r > 0$. To avoid the boundary effects shown in Section 6.1 we choose $r = 5$. To approximate H we use Monte Carlo and we sample 10,000 from μ_Ω in each case using rejection sampling. In all cases we let μ_Ω be the uniform measure on Ω and we let $\lambda = 10^{-3}$. Finally, we solve numerically (31) using SciPy's implementation of the RK45 method with absolute tolerance 10^{-4} and relative tolerance 10^{-6}.

If Fig. 7 we show numerical approximations for the marginals for $\Pi^\star$ for the selected choices $5\bar{B}_1^+$, $5\bar{B}_2 \times 5\bar{B}_1^+$ and $5\bar{B}_1^+ \times 5\bar{B}_2^+$ for Ω. To approximate the marginals for variables taking values in $[0, 5]$ we divided the interval in 10 bins, whereas for variables taking values in $[-5, 5]$ we divided the interval in 20 bins. The probability of each bin under $\Pi^\star$ was approximated with 20,000 Monte Carlo samples independent from those used to approximate H.

The marginals obtained for $5\bar{B}_1^+$ for the components of $\boldsymbol{v}_0$ concentrate near zero, except the marginal for $v_{0,1}$ (rows 1 to 3 of Fig. 7). However, the marginals cannot concentrate sufficiently near zero. As a consequence the expected values of $v_{0,2}$ and $v_{0,3}$ tend to be too large. This may also be the reason why the marginals for $v_{0,4}$ assign large probability to large values of $v_{0,4}$. In contrast, the marginals for $c_{2\to4}$ and $c_{4\to2}$ capture the behavior of these variables reasonably well, as the marginal for $c_{4\to2}$ is more concentrated toward the origin than the marginal for $c_{2\to4}$. This behavior is consistent for all scenarios. Finally, in Figs. 8a-8c we see the trajectory associated to the expected value of $\boldsymbol{v}_0$, $c_{2\to4}$ and $c_{4\to2}$. Although they capture reasonably well the measurements of v_1 in time, they tend to overestimate the average measurements for v_3. This is itself a consequence of overestimating $v_{0,4}$.

Observe that the marginals obtained for $5\bar{B}_2 \times 5\bar{B}_1^+$ for the components of $\boldsymbol{v}_0$ spread over the interval $[-5, 5]$ but concentrate near the origin (rows 4 to 6 of Fig. 7). As a consequence, the probability of observing a negative value is large. However, the expected values for $v_{0,2}$ and $v_{0,3}$ are closer to the initial condition of the system for Scenarios A and B; for Scenario C the expected value for $v_{0,3}$ is negative and that for $v_{0,4}$ is much larger. This suggests the dynamics of the system constrain these two variables. In all cases, the marginals for $c_{2\to4}$ and $c_{4\to2}$ behave in a similar manner, suggesting that in this setup the method cannot differentiate between the effect of these two variables. Figs. 8d-8f show the trajectory associated to the expected values of $\boldsymbol{x}_0$, $c_{2\to4}$ and $c_{4\to2}$. Choosing Ω as a Cartesian product allows for more flexibility, and as a consequence, the trajectories seem to be a good fit to the observations in all scenarios. However, note that the initial condition may have negative entries. This effect can be mitigated by choosing a suitable set C in (13).

Finally, the marginals obtained for $5\bar{B}_1^+ \times 5\bar{B}_2^+$ for the components of $\boldsymbol{x}_0$ behave similarly to those obtained for $5\bar{B}_1^+$ (rows 7 to 9 in Fig. 7). Remark that for this choice the marginals for $v_{0,1}$ and $v_{0,4}$ behave similarly and their expected values are also similar. As for $5\bar{B}_1^+$ we also see that the densities for $c_{2\to4}$ and $c_{4\to2}$ then to capture the overall relation between $c_{2\to4}$ and $c_{4\to2}$. In Figs. 8g-8i the

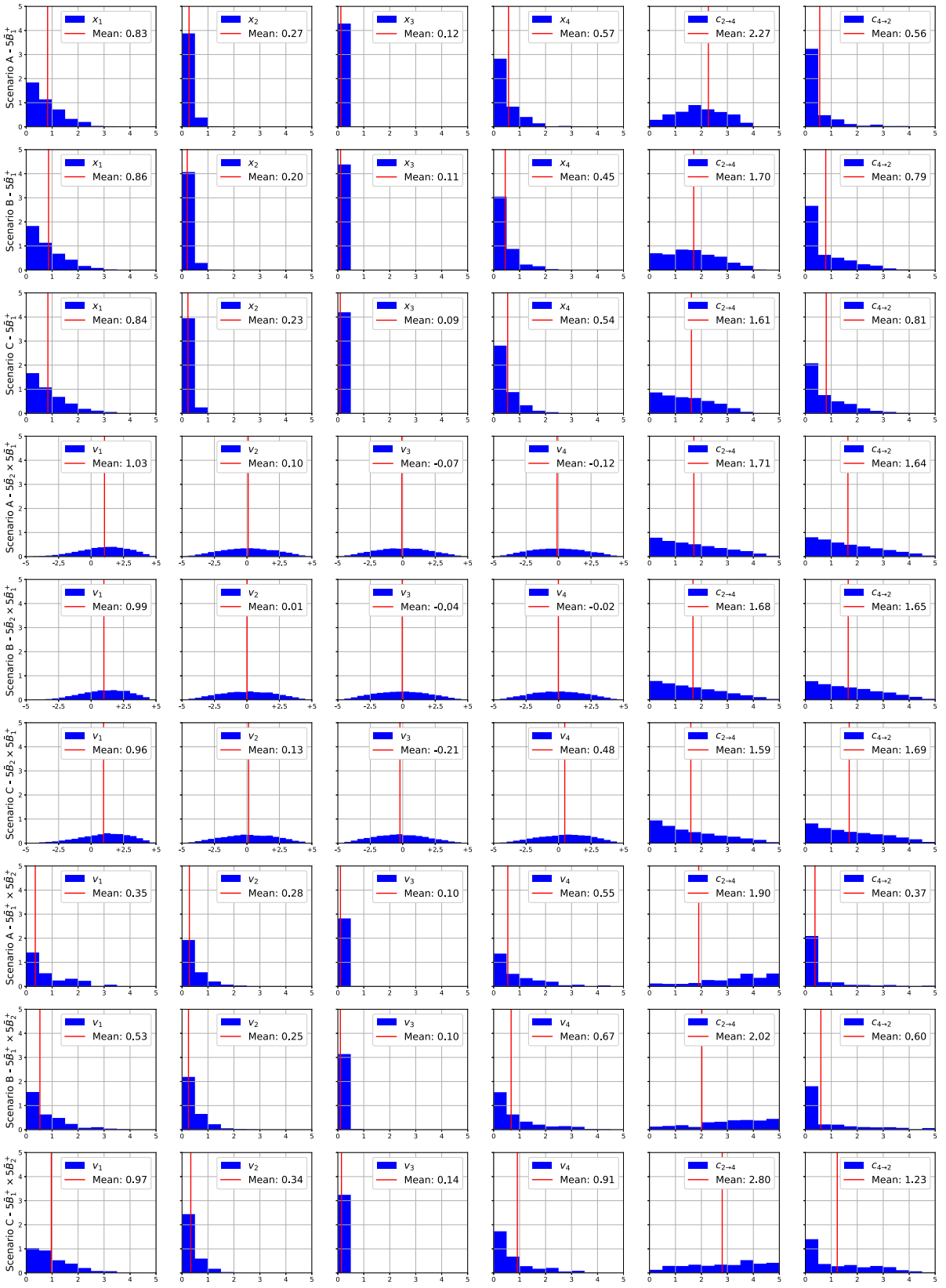
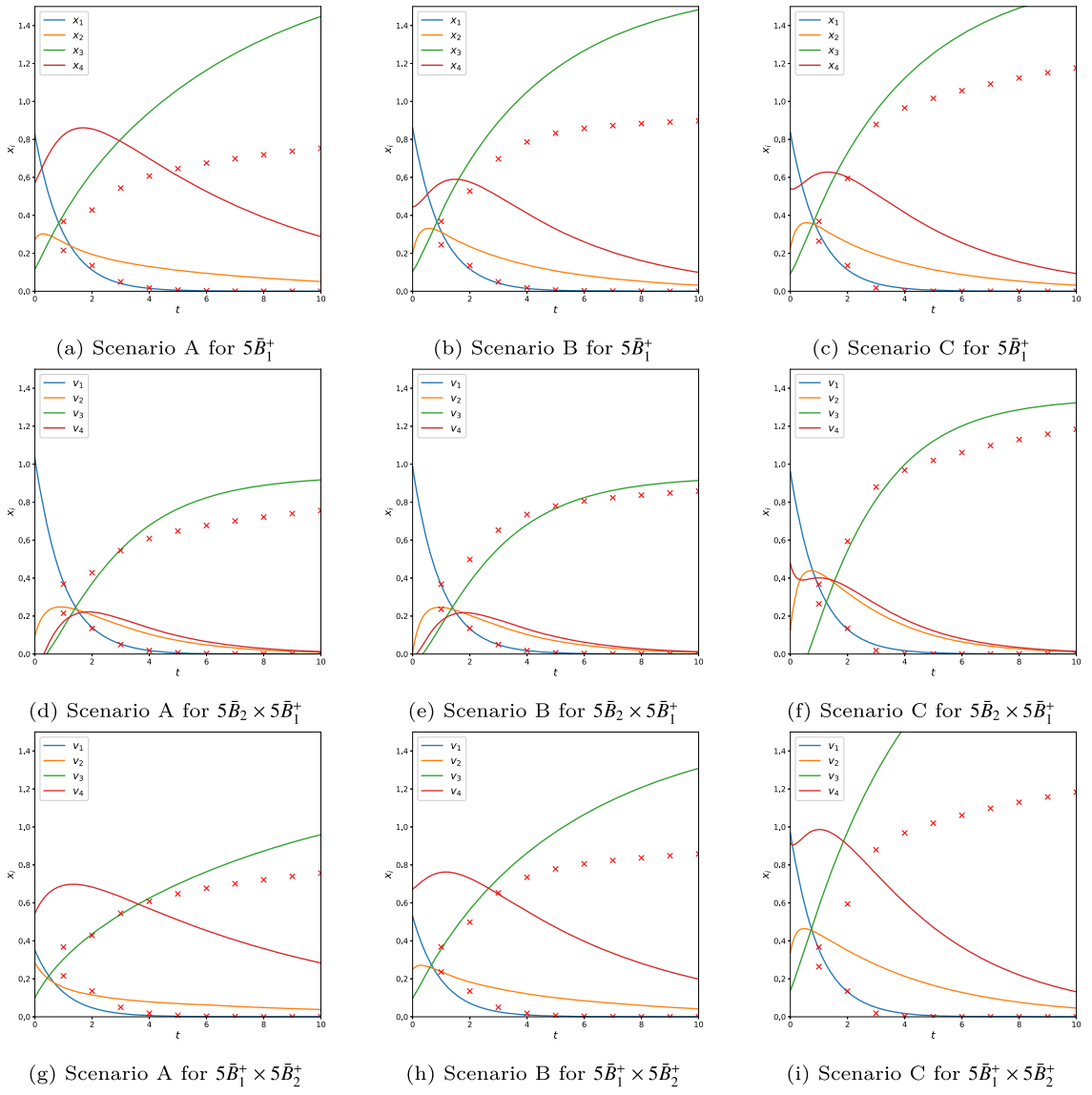


Fig. 7. Marginals for $\Pi^\star$ for selected choices of Ω under each of the failure scenarios. Rows 1 to 3 show the results for $5\bar{B}_1^+$, rows 4 to 6 for $5\bar{B}_2 \times 5\bar{B}_1^+$ and rows 7 to 9 for $5\bar{B}_1^+ \times 5\bar{B}_2^+$. The vertical line indicates the expected value.



(a) Scenario A for $5\bar{B}_1^+$ (b) Scenario B for $5\bar{B}_1^+$ (c) Scenario C for $5\bar{B}_1^+$

(d) Scenario A for $5\bar{B}_2 \times 5\bar{B}_1^+$ (e) Scenario B for $5\bar{B}_2 \times 5\bar{B}_1^+$ (f) Scenario C for $5\bar{B}_2 \times 5\bar{B}_1^+$

(g) Scenario A for $5\bar{B}_1^+ \times 5\bar{B}_2^+$ (h) Scenario B for $5\bar{B}_1^+ \times 5\bar{B}_2^+$ (i) Scenario C for $5\bar{B}_1^+ \times 5\bar{B}_2^+$

Fig. 8. Trajectories for the expected value $\mathbf{E}_{\boldsymbol{x}_0 \sim \Pi^*}[\boldsymbol{x}_0]$.

trajectories for the expected value show similarities with those obtained for $5\bar{B}_1^+$. However, for Scenario C the differences between this trajectory and the average observations for v_3 are quite large. This is due to the fact that an overestimate of $v_{0,2}$ has a strong impact on the limiting value of v_3.

6.3. Chemical reaction network

To illustrate the application of our method to a high-dimensional, non-linear system, we consider a chemical reaction network. These can be described by the reaction rate equations, which are a non-linear system of ODEs that model the time-evolution of the concentrations of the chemical species in the system [49]. If the system has n species and we let $\boldsymbol{c} \in \mathbb{R}^n = \boldsymbol{c}(t)$ denote the concentrations, then the reaction rate equations become

$$\dot{\boldsymbol{c}}(t) = \sum_{i=1}^{p} k_i h_i(\boldsymbol{c}(t)) \boldsymbol{r}_i \tag{32}$$

where p is the number of chemical reactions in the system, $k_1, \dots, k_p > 0$ are the reaction *rate constants*, $h_1, \dots, h_p : \mathbb{R}^n \to \mathbb{R}$ are the *propensity functions* and $\boldsymbol{r}_1, \dots, \boldsymbol{r}_p \in \mathbb{R}^n$ are the *stoichiometry vectors*. In this case, we shall assume the initial concentrations $\boldsymbol{c}_0$ are fixed. However, the rate constants will be random. We assume $k_i := \delta_i k_i^*$ where $k_i^* > 0$ is a known reference value for the rate constant of the i-th reaction and $\delta_1, \dots, \delta_p$ are i.i.d. $\text{LOGNORMAL}(-\sigma^2/2, \sigma)$ for $\sigma = \log(10)/2$. This ensures the rates vary roughly across 2 orders of magnitude and that the expected value of k_i is close to its reference value k_i^*.

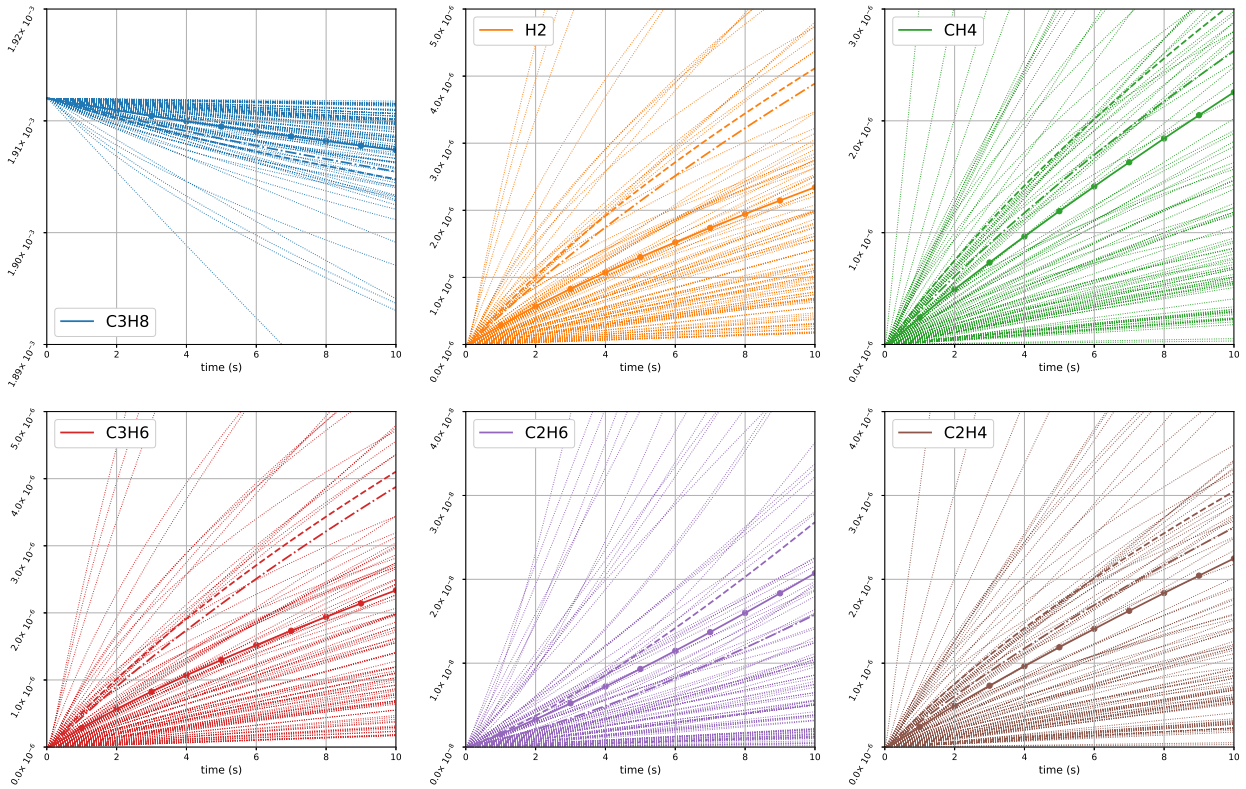


Fig. 9. Concentrations for the observed species C_3H_8, H_2, CH_4, C_3H_6, C_2H_6 and C_2H_4. The dotted lines represent sample trajectories. The solid line represents the average trajectory, while the circles represent the average value of the observations. The dashed line represents the trajectory associated to the reference values k^*. The dashed-dotted line represents the trajectory associated to the average rates $\mathbf{E}_{\boldsymbol{k}\sim\mu_0}[\boldsymbol{k}]$.

We choose $\Omega = [10^{-2}k_1^*, 10^{+2}k_1^*] \times \ldots \times [10^{-2}k_p^*, 10^{+2}k_p^*]$ and μ_Ω as the joint density of the random variables $10^{\alpha_1}k_1^*, \ldots, 10^{\alpha_p}k_p^*$ for $\alpha_1, \ldots, \alpha_p \overset{iid}{\sim} \text{UNIF}([-2,2])$. To perform the simulations, we consider a medium-sized reaction system modeling propane pyrolysis with 38 chemical species and 98 reactions [50, Section IV.B]. We simulate the system for $t \in [0,10]$ and observe every $\Delta t = 1$ starting at $t = 1$ the concentrations of the 6 most chemically relevant species in the system: C_3H_8, H_2, CH_4, C_3H_6, C_2H_6 and C_2H_4 (Fig. 9). We sample 100 trajectories to compute the average observation. The trajectories are found by solving (32) numerically using SciPy's implementation of the Radau method with absolute tolerance 10^{-8} and relative tolerance 10^{-10}. To approximate the objective function, we use 10,000 samples from μ_Ω.

In Fig. 10 we show the marginals obtained for 6 chemically relevant reactions $C_3H_8 \to CH_3 + C_2H_5$, $CH_3 + H_2 \to H + CH_4$, $C_3H_5 + H_2 \to H + C_3H_6$, $H + C_3H_6 \to C_3H_5 + H_2$, $C_2H_5 + H_2 \to H + C_2H_6$ and $CH_3 + C_2H_6 \to C_2H_5 + CH_4$ for $\lambda = 10^{-2}$. As we can see, the marginal remains almost flat over the interval. This may be a consequence of the fact that the system is ill-conditioned, that the variations of $\boldsymbol{U}$ are small, and that the amount of measurements is not sufficient to accurately characterize the uncertainty. Furthermore, there is a bias in the estimate. In Fig. 11 we see the sample trajectories from the optimal density $\Pi^\star$.

7. Discussion

Our experiments show that the method we proposed to compute the optimal density can be implemented efficiently. For typical choices of Ω for low-dimensional problems, rejection sampling and stochastic approximation methods are sufficient to obtain good results. For high-dimensional problems, stochastic gradient descent may perform well. In this case, the main cost is evaluating the map $\boldsymbol{U}$. Furthermore, the closed-form expression for the optimal density allows us to compute statistics that may be relevant when quantifying the uncertainty in the quantities of interest. In particular, this allows us to compute the marginal distributions of these quantities.

We have seen that the choice of Ω has a strong impact on the results. Although choosing a "small" Ω could have computational advantages when approximating H it may render the problem ill-conditioned as was shown in Section 6.1. When the density accumulates near the boundary of Ω its geometry will strongly influence the structure of, e.g., the marginals. The results in Section 6.2 show that even if the densities concentrate near the boundary, the expected value may be biased. Furthermore, the geometry of the boundary of Ω will determine if the density is such that the marginals concentrate for all or only some of the state variables.

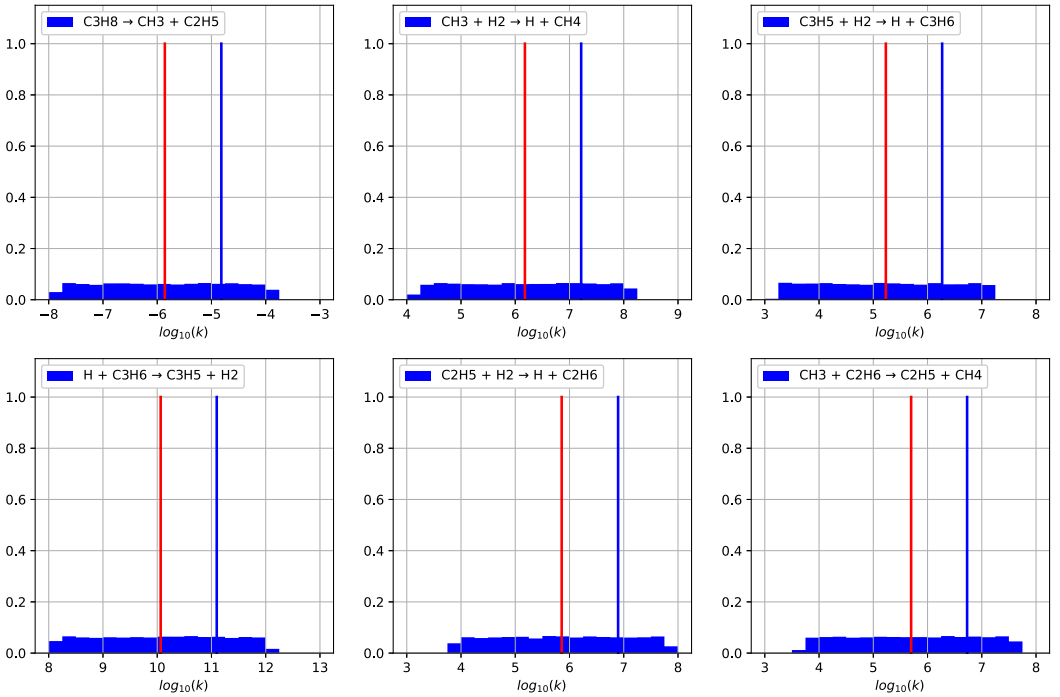


Fig. 10. Histograms for the optimal marginal densities for $\lambda = 10^{-2}$ and the selected reactions $C_3H_8 \rightarrow CH_3 + C_2H_5$, $CH_3 + H_2 \rightarrow H + CH_4$, $C_3H_5 + H_2 \rightarrow H + C_3H_6$, $H + C_3H_6 \rightarrow C_3H_5 + H_2$, $C_2H_5 + H_2 \rightarrow H + C_2H_6$ and $CH_3 + C_2H_6 \rightarrow C_2H_5 + CH_4$. The densities are in the logarithm of the rate. As we can see, the density remains almost uniform over the interval of interest.

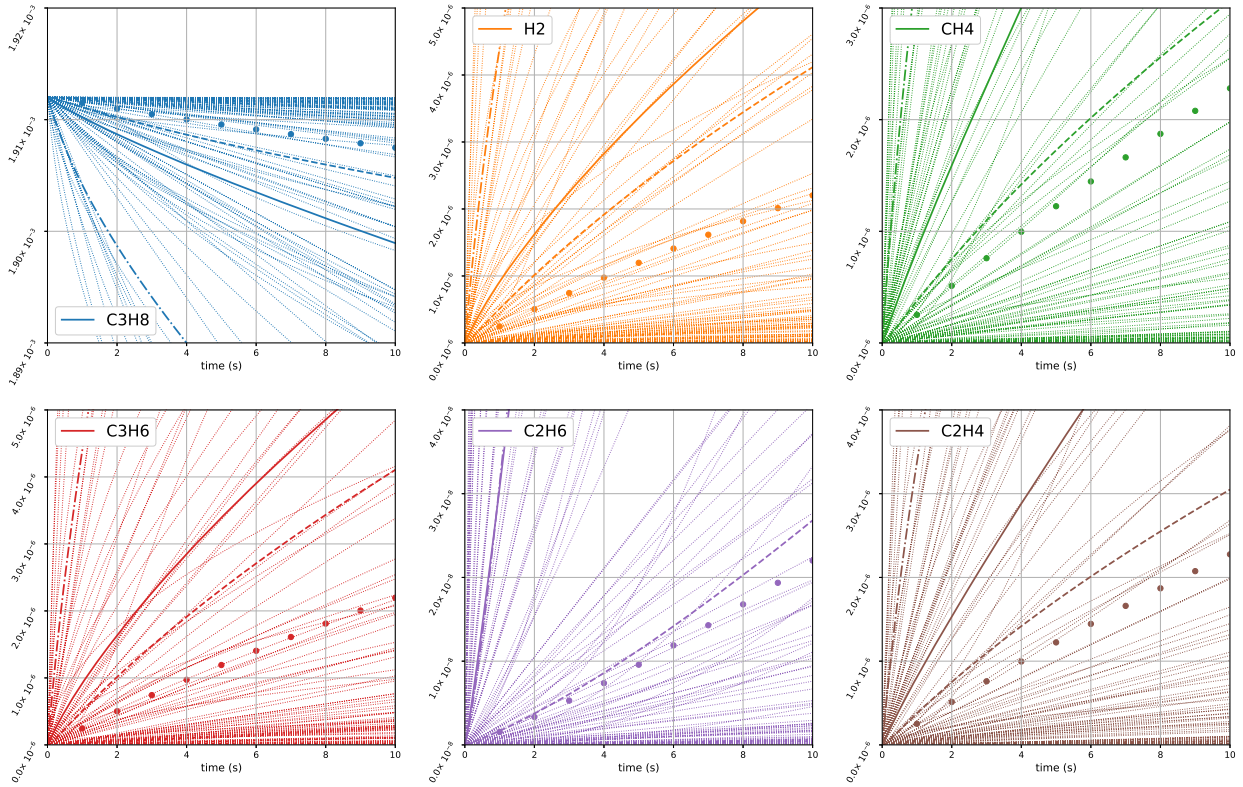


Fig. 11. Concentrations for the observed species C_3H_8, H_2, CH_4, C_3H_6, C_2H_6 and C_2H_4 for the optimal density for $\lambda = 10^{-2}$. The dotted lines represent sample trajectories from the optimal density. The solid line represents the average trajectory, while the circles represent the average value of the observations. The dashed line represents the trajectory associated to the reference values k^*. The dashed-dotted line represents the trajectory associated to the average rates $\mathbf{E}_{k\sim\Pi^*}[k]$.

When the density does not accumulate near the boundary, the interplay between Ω, and dynamics of the system and the measured quantities of interest will determine how the optimal density distributes mass in Ω. The experiments in Section 6.2 show that even though this may lead to a better estimate of certain parameters, e.g., the initial condition, the optimal density may assign probability to non-physical values. As mentioned before, this can be mitigated by choosing in (13) a suitable set C of constraints for the expected value. It is interesting to see that some choices of Ω may be better suited for some tasks. As an example, our results in Section 6.2 show that to estimate the initial condition with the expected value it is better to allow for wider intervals. In contrast, to capture the relation between the capacities it may be more appropriate to allow for narrower intervals.

In our experiments we have emphasized the role of the expected value of the initial conditions to assess the performance of the method. However, our results provide a closed-form expression for the density from which other statistics could be evaluated. As an example, in the experiments in Section 6.1 a much better estimate would be the mode. In fact, Fig. 4 suggests that when the system and the measurements are linear, the mode will be at the boundary of Ω. The experiments in Section 6.2 suggest the same. In general, the choice of a suitable statistic would depend on the properties of the system and the available measurements of quantities of interest.

Finally, our results show that the structure of the family of densities that can be found by the method depends strongly on how complex the dynamics of the system are, or how complex are the observations about the system. For linear ODEs and linear observations our results show that the probability *density* $\rho^\star$ must be the positive part of an affine function. As a consequence, to add expressiveness to the probability *measure* $\Pi^\star$ found by the method it may be necessary to use more complex reference measures μ_Ω on Ω or mixtures.

8. Conclusion

In this work, we propose a Bayesian model to quantify input uncertainty on random differential equations. Furthermore, we construct a family of priors for the probability density of the initial condition for which the *maximum a posteriori* estimate for the density can be computed using Tikhonov regularization on moment constraints for the density. This provides an underlying statistical model for the method proposed by Meyers et al., built on early work by Banks et al., to perform uncertainty quantification in aggregate data problems. Under a convexity assumption on the model, we leveraged duality to deduce an equivalent finite-dimensional formulation, which for typical cases becomes an unconstrained convex problem with a smooth objective function. This allowed us to provide a full theoretical characterization of the family of densities that can be obtained by the method, along with their dependence in the parameters of the problem. Although the trade-off is that the objective function involves a high-dimensional integral, we have shown that the problem can be solved efficiently using standard approximation methods. We believe that our results may lead to the development of more efficient implementations by using tailored numerical integration methods. Furthermore, our analysis yields insight about the flexibility of the approach when performing uncertainty quantification for high-dimensional systems by showing the interplay between the support of the density, the dynamics of the system, and the constraints being enforced.

Our work leaves some questions open. First, our method shows that there is a subtle interplay between the dynamics of the system, the observations about it, and the support of the probability measure that is to be found. Understanding this interplay would help practitioners make an informed choice of support or constraints for the expected values depending on the properties of the system. Second, we have shown that the structure of the probability densities found by the method is restricted by the system and the measured quantities of interest. A question of interest is to determine the size of the class of densities with this structure. Finally, it is an open question whether the optimal probability density can be *refined* to improve the performance of the method according to some criterion. These questions will be explored as future lines of research.

CRediT authorship contribution statement

Elena Villalón: Writing – original draft, Software, Methodology, Conceptualization. **Qian Yang:** Software, Resources, Methodology, Data curation. **Carlos A. Sing Long:** Writing – review & editing, Writing – original draft, Supervision, Software, Methodology, Conceptualization.

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.

Data availability

Data will be made available on request.

Acknowledgements

We thank the reviewers for their comments and suggestions. C. A. Sing Long thanks A. Jara and M. Castro for their insightful comments and discussions on Bayesian nonparametric methods. C. A. Sing Long and E. Villalón were partially supported by the grant ANID – FONDECYT – 1211643. C. A. Sing Long was also partially supported by the grants ANID – Millennium Science Initiative

Program – NCN17_059 and ANID – Millennium Science Initiative Program – NCN17_129, and by the National Center for Artificial Intelligence CENIA FB210017, Basal ANID. Q. Yang acknowledges support from the National Science Foundation (NSF) Grant No. DMR-2102406.

Appendix A. Proof of main results in Section 2

A.1. Proof of Proposition 1

Define for $i \in \{1, \dots, n\}$ the averages

$$\bar{\varphi}_i = \int_\Omega \varphi_i(\boldsymbol{x}) d\mu_\Omega(\boldsymbol{x})$$

and the vector $\bar{\boldsymbol{\varphi}} \in \mathbb{R}^n$ accordingly. Let

$$C_n := \bigcap\nolimits_{x\in\Omega} \{\boldsymbol{a} \in \mathbb{R}^n : \ \langle \boldsymbol{a}, \bar{\boldsymbol{\varphi}}\rangle = 1, \langle a, \boldsymbol{\varphi}(x)\rangle \geq 0\}.$$

It is clear this is a closed and convex set in $\mathbb{R}^n$. If $\rho \in P_{\mu,n}$ then $\rho = \langle \boldsymbol{a}, \boldsymbol{\varphi}\rangle$ for some $\boldsymbol{a} \in \mathbb{R}^n$. Since $\rho \geq 0$ and its integral with respect to μ_Ω equals 1 we deduce that

$$\langle \boldsymbol{a}, \bar{\boldsymbol{\varphi}}\rangle = 1 \qquad \text{and} \qquad x \in \Omega : \ \langle \boldsymbol{a}, \boldsymbol{\varphi}(x)\rangle \geq 0$$

whence $\boldsymbol{a} \in C_n$. Conversely, if $\boldsymbol{a} \in C_n$ then $\rho := \langle \boldsymbol{a}, \boldsymbol{\varphi}\rangle$ is such that $\rho \geq 0$ then it is straightforward to see that $\rho \in P_{\mu,n}$.

From the hypotheses it follows that $\bar{\varphi}_1 > 0$ whence $\bar{\varphi}_1^{-1}\varphi_1 \in P_{\mu,n}$. Thus, both $P_{\mu,n}$ and C_n are non-empty sets. Let $\boldsymbol{a}_n := \bar{\boldsymbol{\varphi}}/\|\bar{\boldsymbol{\varphi}}\|_2^2$. Then $\langle \boldsymbol{a}_n, \bar{\boldsymbol{\varphi}}\rangle = 1$. Let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{n-1} \in \mathbb{R}^n$ be orthonormal vectors such that $\langle \boldsymbol{v}_i, \bar{\boldsymbol{\varphi}}\rangle = 0$ for $i \in \{1, \dots, n-1\}$, and let $\boldsymbol{V}$ be the matrix with $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{n-1}$ as its columns. By construction we have that $\boldsymbol{V}^*\boldsymbol{a}_n = 0$. We define the set

$$\Theta_{V_n} := \left\{\theta \in \mathbb{R}^{n-1} : \ \boldsymbol{a}_n + \sum\nolimits_{i=1}^{n-1} \theta_i \boldsymbol{v}_i \in C_n\right\}$$

which is closed and convex since it is the preimage of the affine map $\boldsymbol{\theta} \mapsto \boldsymbol{a}_n + \boldsymbol{V}\boldsymbol{\theta}$. If $\boldsymbol{\theta} \in \Theta_{V_n}$ then, by construction, $\boldsymbol{a}_n + \boldsymbol{V}\boldsymbol{\theta} \in C_n$. Conversely, if $\boldsymbol{a} \in C_n$ we can always write

$$\boldsymbol{a} = \alpha_0 \boldsymbol{a}_n + \sum\nolimits_{i=1}^{n-1} \theta_i \boldsymbol{v}_i,$$

for some $\boldsymbol{\theta} \in \mathbb{R}^{n-1}$. It suffices to show that $\alpha_0 = 1$. Note that

$$1 = \langle \bar{\boldsymbol{\varphi}}, \boldsymbol{a}\rangle = \alpha_0 \langle \bar{\boldsymbol{\varphi}}, \boldsymbol{a}_n\rangle = \alpha_0$$

whence $\boldsymbol{\theta} \in \Theta_{V_n}$. It remains to show that Θ_{V_n} has non-empty interior. Define $\boldsymbol{a}_1 = \bar{\varphi}_1^{-1}\boldsymbol{e}_1$. Then $\rho = \langle \boldsymbol{a}_1, \boldsymbol{\varphi}\rangle = \bar{\varphi}_1^{-1}\varphi_1$ belongs to $P_{\mu,n}$ whence $\boldsymbol{a}_1 \in C_n$. Therefore, let

$$\boldsymbol{a} = \boldsymbol{a}_n + \boldsymbol{V}(\boldsymbol{V}^*\boldsymbol{a}_1 + \boldsymbol{\theta}) = \boldsymbol{a}_1 + \boldsymbol{V}\boldsymbol{\theta}$$

for some $\boldsymbol{\theta}$ to determine. Define

$$\rho = \langle \boldsymbol{a}, \boldsymbol{\varphi}\rangle = \langle \boldsymbol{a}_1, \boldsymbol{\varphi}\rangle + \langle \boldsymbol{V}\boldsymbol{\theta}, \boldsymbol{\varphi}\rangle = \frac{1}{\bar{\varphi}_1}\varphi_1 + \langle \boldsymbol{V}\boldsymbol{\theta}, \boldsymbol{\varphi}\rangle.$$

Note that ρ has integral equal to one for any choice of $\boldsymbol{\theta}$. Therefore, we must show it is non-negative to conclude that $\rho \in P_{\mu,n}$. Since $\varphi_1, \dots, \varphi_n$ are bounded, there is $B > 0$ such that $\|\boldsymbol{\varphi}(x)\|_2 \leq B$ for $x \in \Omega$. Hence, it suffices to have

$$|\langle \boldsymbol{V}\theta, \boldsymbol{\varphi}(x)\rangle| \leq \|\boldsymbol{V}\boldsymbol{\theta}\|_2 \|\boldsymbol{\varphi}(x)\|_2 \leq B\|\boldsymbol{\theta}\|_2 \leq \bar{\varphi}_1^{-1} \inf\nolimits_{x\in\Omega} \varphi_1(x) \leq \bar{\varphi}_1^{-1}\varphi_1(x)$$

to conclude that $\rho \geq 0$. Since this is satisfied for any $\boldsymbol{\theta}$ such that

$$\|\boldsymbol{\theta}\|_2 \leq \frac{1}{B\bar{\varphi}_1} \inf\nolimits_{x\in\Omega} \varphi_1(x)$$

and the right-hand side is strictly positive, we conclude that Θ_{V_n} has non-empty interior.

A.2. Proof of Proposition 2

Let $\{\rho^{\star,n}\}_{n\in\mathbb{N}}$ be the sequence of solutions to the finite-dimensional problems (10), and let F denote the objective function, which is independent of n. As L is non-negative, F is non-positive. First, since $V_1 \subset V_n$ for any $n \in \mathbb{N}$ we have that $P_{\mu,1} \subset P_{\mu,n}$ whence $F(\rho^{\star,1}) \leq F(\rho^{\star,n})$. Since L is non-negative, this implies that $2F(\rho^{\star,1}) \geq \lambda\|\rho^{\star,n}\|_{L^2_\mu}^2$. Hence, using the fact that $P_{\mu,n} \subset P_\mu$, we conclude that the sequence $\{\rho^{\star,n}\}_{n\in\mathbb{N}}$ is a norm-bounded sequence in P_μ. By the Banach-Alaoglu theorem, by possibly passing to a

subsequence, we can assume without loss that it converges weakly to a density $\rho^{\star,\infty}$. Since P_μ is convex and closed, it is also weakly closed. Thus $\rho^{\star,\infty} \in P_\mu$.

Now, let $\rho^\star$ be the solution to the infinite-dimensional problem (12) and let $\rho_n^\star$ be the projection of $\rho^\star$ onto the closed convex set $P_{\mu,n}$. Since the closure of the union of the subspaces $\{V_n\}_{n\in\mathbb{N}}$ is all of $L^2_\mu(\Omega)$, we have that $\{\rho_n^\star\}_{n\in\mathbb{N}}$ converges in norm to $\rho^\star$. By construction, we have that $F(\rho_n^\star) \le F(\rho^{\star,n}) \le F(\rho^\star)$.

Now, remark that the map

$$\rho \mapsto \int_\Omega \rho(\boldsymbol{x}_0)\boldsymbol{U}(\boldsymbol{x}_0)\, d\mu_\Omega(\boldsymbol{x}_0)$$

is strongly (or norm) continuous *and* weakly continuous. Since the L^2_μ-norm also strongly continuous, we conclude that F is strongly continuous whence

$$F(\rho^\star) = \limsup_{n\to\infty} F(\rho_n^\star).$$

On the other hand, the L^2_μ-norm is also strongly lower semicontinuous and thus weakly sequentially lower semicontinuous [27, Theorem 9.1]. This implies that F is weakly sequentially upper semicontinuous whence

$$\limsup_{n\to\infty} F(\rho^{\star,n}) \le F(\rho^{\star,\infty}).$$

Combining these two facts, we conclude that

$$F(\rho^\star) = \limsup_{n\to\infty} F(\rho_n^\star) \le \limsup_{n\to\infty} F(\rho^{\star,n}) \le F(\rho^{\star,\infty}) \le F(\rho^\star)$$

whence $\rho^{\star,\infty}$ is a solution to the infinite-dimensional problem, proving the claim.

Appendix B. Proofs of statements in Section 3

B.1. Proof of Proposition 3

Define the linear map $\mathcal{U}^\curlyvee : L^2_\mu(\Omega) \to \mathbb{R}^q$ as

$$\mathcal{U}^\curlyvee(\rho) := \int_\Omega \boldsymbol{U}(\boldsymbol{x}_0)\rho(\boldsymbol{x}_0)\, d\mu_\Omega(\boldsymbol{x}_0)$$

which, by (6) and the Cauchy-Schwarz inequality, is bounded. As $\mathbb{R}^q$ is finite-dimensional, $\mathcal{U}^\curlyvee$ has closed range. Consider the auxiliary problem

$$\begin{aligned} \underset{\rho\in L^2_\mu(\Omega)}{\text{minimize}} \quad & \frac{1}{2}\|\rho\|^2_{L^2_\mu} + \frac{1}{\lambda} L(\boldsymbol{A}\mathcal{U}^\curlyvee(\rho), \boldsymbol{y}) \\ \text{subject to} \quad & \rho \ge 0, \quad \langle 1, \rho\rangle_{L^2_\mu} = 1, \quad \mathcal{U}^\curlyvee(\rho) \in C. \end{aligned}$$

where 1 denotes the function that is identically equal to one over Ω; since μ_Ω is finite, $1 \in L^2_\mu$. The objective function in the above is strongly convex [27, Definition 10.7] and the feasible set is closed and convex. Hence, there is a unique minimizer $\rho^\star$ for the above problem [27, Proposition 11.8]. Let $\bar{\boldsymbol{u}}^\star := \mathcal{U}^\curlyvee(\rho^\star)$. We claim $(\rho^\star, \bar{\boldsymbol{u}}^\star)$ is an optimal solution to (13). If $(\rho, \bar{\boldsymbol{u}})$ is feasible for (13) then ρ is feasible for the above auxiliary problem and thus

$$\begin{aligned} \frac{1}{2}\|\rho^\star\|^2_{L^2_\mu} + \frac{1}{\lambda} L(\boldsymbol{A}\bar{\boldsymbol{u}}^\star, \boldsymbol{y}) &= \frac{1}{2}\|\rho^\star\|^2_{L^2_\mu} + \frac{1}{\lambda} L(\boldsymbol{A}\mathcal{U}^\curlyvee(\rho^\star), \boldsymbol{y}) \\ &\le \frac{1}{2}\|\rho\|^2_{L^2_\mu} + \frac{1}{\lambda} L(\boldsymbol{A}\mathcal{U}^\curlyvee(\rho), \boldsymbol{y}) \\ &= \frac{1}{2}\|\rho\|^2_{L^2_\mu} + \frac{1}{\lambda} L(\boldsymbol{A}\bar{\boldsymbol{u}}, \boldsymbol{y}) \end{aligned}$$

where we used the fact that $\bar{\boldsymbol{u}} = \mathcal{U}^\curlyvee(\rho)$ by feasibility of $(\rho, \bar{\boldsymbol{u}})$. Hence, the objective value of $(\rho, \bar{\boldsymbol{u}})$ is at least that of $(\rho^\star, \bar{\boldsymbol{u}}^\star)$. We conclude $(\rho^\star, \bar{\boldsymbol{u}}^\star)$ is an optimal solution to (13). To show it is the only solution, if we let $(\rho, \bar{\boldsymbol{u}})$ be any other solution, then, from the strong convexity of the quadratic term on ρ in the objective, it follows that $\rho = \rho^\star$. By feasibility, we conclude that $\bar{\boldsymbol{u}} = \mathcal{U}^\curlyvee(\rho) = \mathcal{U}^\curlyvee(\rho^\star) = \bar{\boldsymbol{u}}^\star$.

B.2. Proof of Theorem 1

First, it is apparent that $\mathcal{B} : X_P \to X_D$ is bounded and that its adjoint $\mathcal{B}^* : X_D \to X_P$ is given by

$$\mathcal{B}^*(\eta, \boldsymbol{\xi}, \boldsymbol{\omega}) = \begin{bmatrix} -(\eta + \langle \boldsymbol{\omega}, \boldsymbol{U}(\boldsymbol{x}_0)\rangle_2) \\ \boldsymbol{\omega} - \boldsymbol{A}^*\boldsymbol{\xi} \\ \boldsymbol{\xi} \end{bmatrix}$$

The function $f : X_P \to (-\infty, +\infty]$ is f is proper, convex and lower semicontinuous, with domain

$$\mathbf{dom}(f) = \{(\rho, \bar{\boldsymbol{u}}, \boldsymbol{v}) \in X_P : \rho \geq 0, \bar{\boldsymbol{u}} \in C\},$$

whereas $g : X_D \to (-\infty, +\infty]$ is proper, convex and lower semicontinuous. Furthermore, it is polyhedral. Note that

$$\mathbf{dom}(g) = \{\boldsymbol{r}\}$$

and that $\mathbf{dom}(g) \cap \mathbf{relint}(\mathcal{B}(\mathbf{dom}(f))) \neq \emptyset$. In fact,

$$\mathcal{B}(\mathbf{dom}(f)) = \{(s, \boldsymbol{w}, \boldsymbol{z}) \in \mathbb{R} \times \mathbb{R}^q \times \mathbb{R}^m : s < 0\}.$$

This follows from the fact that we may choose

$$\rho \equiv \frac{|s|}{\mu_\Omega(\Omega)}, \quad \bar{\boldsymbol{u}} = \frac{|s|}{\mu_\Omega(\Omega)} \int_\Omega \boldsymbol{U}(\boldsymbol{x}_0) d\mu_\Omega(\boldsymbol{x}_0) + \boldsymbol{w} \quad \text{and} \quad \boldsymbol{v} = \frac{|s|}{\mu_\Omega(\Omega)} \int_\Omega \boldsymbol{A}\boldsymbol{U}(\boldsymbol{x}_0) d\mu_\Omega(\boldsymbol{x}_0) + \boldsymbol{A}\boldsymbol{w} + \boldsymbol{z}$$

to obtain $(s, \boldsymbol{w}, \boldsymbol{z}) = \mathcal{B}(\rho, \bar{\boldsymbol{u}}, \boldsymbol{v})$. By [27, Fact 15.25] we conclude that the origin belongs to the strong relative interior of $\mathbf{dom}(g) - \mathcal{B}(\mathbf{dom}(f))$. Therefore, by [27, Theorem 15.23] strong duality holds. On one hand, the convex dual of f is

$$\begin{aligned}
f^\star(\rho, \bar{\boldsymbol{u}}, \boldsymbol{v}) &= \sup_{(\rho', \bar{\boldsymbol{u}}', \boldsymbol{v}') \in X_P} \langle (\rho, \bar{\boldsymbol{u}}, \boldsymbol{v}), (\rho', \bar{\boldsymbol{u}}', \boldsymbol{v}') \rangle - f(\rho', \bar{\boldsymbol{u}}', \boldsymbol{v}') \\
&= \sup_{\rho' \in L^2_\mu} \left(\langle \rho, \rho' \rangle_{L^2_\mu} - \frac{1}{2} \|\rho'\|^2_{L^2_\mu} - \mathbb{I}_{L^{2+}_\mu}(\rho) \right) \\
&\quad + \sup_{\bar{\boldsymbol{u}}' \in \mathbb{R}^q} \left(\langle \bar{\boldsymbol{u}}, \bar{\boldsymbol{u}}' \rangle_2 - \mathbb{I}_C(\bar{\boldsymbol{u}}') \right) + \sup_{\boldsymbol{v}' \in \mathbb{R}^m} \left(\langle \boldsymbol{v}, \boldsymbol{v}' \rangle_2 - \frac{1}{\lambda} L(\boldsymbol{v}', \boldsymbol{y}) \right) \\
&= \frac{1}{2} \|\rho\|^2_{L^2_\mu} + \sup_{\rho' \in L^2_\mu} \left(-\frac{1}{2} \|\rho - \rho'\|^2_{L^2_\mu} - \mathbb{I}_{L^{2+}_\mu}(\rho) \right) + \mathbb{I}^\star_C(\bar{\boldsymbol{u}}) + \frac{1}{\lambda} L^\star(\lambda \boldsymbol{v}, \boldsymbol{y}) \\
&= \frac{1}{2} \|\rho\|^2_{L^2_\mu} - \frac{1}{2} \|\rho_-\|^2_{L^2_\mu} + \mathbb{I}^\star_C(\bar{\boldsymbol{u}}) + \frac{1}{\lambda} L^\star(\lambda \boldsymbol{v}, \boldsymbol{y}) \\
&= \frac{1}{2} \|\rho_+\|^2_{L^2_\mu} + \mathbb{I}^\star_C(\bar{\boldsymbol{u}}) + \frac{1}{\lambda} L^\star(\lambda \boldsymbol{v}, \boldsymbol{y})
\end{aligned}$$

where ρ_+, ρ_- denote the positive and negative parts of ρ. On the other, the convex dual of g is

$$g^\star(\eta, \boldsymbol{\omega}, \boldsymbol{\xi}) = \langle (\eta, \boldsymbol{\omega}, \boldsymbol{\xi}), \boldsymbol{r} \rangle = -\eta.$$

We conclude that the dual objective becomes

$$\frac{1}{2} \int_\Omega (\eta + \langle \boldsymbol{\omega}, \boldsymbol{U}(\boldsymbol{x}_0) \rangle_2)^2_+ d\mu_\Omega(\boldsymbol{x}_0) + \mathbb{I}^\star_C(\boldsymbol{A}^* \boldsymbol{\xi} - \boldsymbol{\omega}) + \frac{1}{\lambda} L^\star(-\lambda \boldsymbol{\xi}, \boldsymbol{y}) - \eta$$

as desired. Finally, by [27, Theorem 19.1], if $(\rho^\star, \bar{\boldsymbol{u}}^\star, \boldsymbol{v}^\star) \in X_P$ and $(\eta^\star, \boldsymbol{\omega}^\star, \boldsymbol{\xi}^\star) \in X_D$ are optimal primal and dual variables then

$$-\mathcal{B}^*(\eta^\star, \boldsymbol{\omega}^\star, \boldsymbol{\xi}^\star) \in \partial f(\rho^\star, \bar{\boldsymbol{u}}^\star, \boldsymbol{v}^\star) \quad \text{and} \quad (\eta^\star, \boldsymbol{\omega}^\star, \boldsymbol{\xi}^\star) \in \partial g(\mathcal{B}(\rho^\star, \bar{\boldsymbol{u}}^\star, \boldsymbol{v}^\star)).$$

From

$$\partial f(\rho, \bar{\boldsymbol{u}}, \boldsymbol{v}) = \{(\rho + \varphi, 0, \boldsymbol{g}) : \varphi \leq 0, \rho\varphi = 0, \boldsymbol{g} \in \lambda^{-1} \partial L(\boldsymbol{v}, \boldsymbol{y})\}$$

it follows that the first condition implies

$$\eta^\star + \langle \boldsymbol{\omega}^\star, \boldsymbol{U}(\boldsymbol{x}_0) \rangle = \rho^\star(\boldsymbol{x}_0) + \varphi^\star(\boldsymbol{x}_0).$$

This is possible only if

$$\rho^\star(\boldsymbol{x}_0) = (\eta^\star + \langle \boldsymbol{\omega}^\star, \boldsymbol{U}(\boldsymbol{x}_0) \rangle)_+.$$

B.3. Proof of Theorem 2

This result is proven by slightly modifying the arguments used in the proof of Theorem 1. We use the same notation as in that proof. Define the space $X^n_P := V_n \times \mathbb{R}^q \times \mathbb{R}^m$. Consider the restriction $\mathcal{B} : X^n_P \to X_D$. The critical difference is to observe that the adjoint $\mathcal{B}^* : X_D \to X^n_P$ becomes

$$\mathcal{B}^*(\eta,\boldsymbol{\xi},\boldsymbol{\omega}) = \begin{bmatrix} -(\eta 1_{V_n} + \langle \boldsymbol{\omega}, \boldsymbol{U}_{V_n}(\boldsymbol{x}_0)\rangle_2) \\ \boldsymbol{\omega} - \boldsymbol{A}^*\boldsymbol{\xi} \\ \boldsymbol{\xi} \end{bmatrix}.$$

Define the convex function $f_n : X_P^n \to \mathbb{R}$

$$f_n(\rho,\bar{\boldsymbol{u}},\boldsymbol{v}) = \frac{1}{2}\|\rho\|^2_{L^2_\mu} + \mathbb{I}_{V_n^+}(\rho) + \frac{1}{\lambda}L(\boldsymbol{v},\boldsymbol{y}) + \mathbb{I}_C(\bar{\boldsymbol{u}})$$

Since there exists a strictly positive function $\varphi_1 \in V_n$ we have that V_n^+ is non-empty. Therefore, the primal problem associated to the finite-dimensional problem is

$$\underset{(\rho,\bar{\boldsymbol{u}},\boldsymbol{v})\in X_P^n}{\text{minimize}} \quad f_n(\rho,\bar{\boldsymbol{u}},\boldsymbol{v}) + g(\mathcal{B}(\rho,\bar{\boldsymbol{u}},\boldsymbol{v}) - \boldsymbol{r}) \tag{B.1}$$

with dual

$$\underset{(\eta,\boldsymbol{\omega},\boldsymbol{\xi})\in X_D}{\text{maximize}} \quad -f_n^\star(-\mathcal{B}^*(\eta,\boldsymbol{\omega},\boldsymbol{\xi})) - g^\star(\eta,\boldsymbol{\omega},\boldsymbol{\xi}). \tag{B.2}$$

To prove strong duality holds, note that f_n is proper, convex and lower semicontinuous, with domain

$$\mathbf{dom}(f_n) = \{(\rho,\bar{\boldsymbol{u}},\boldsymbol{v}) \in X_P^n : \ \rho \geq 0, \bar{\boldsymbol{u}} \in C\}.$$

As in the proof of Theorem 1, we have that $\mathbf{dom}(g) \cap \mathbf{relint}(\mathcal{B}(\mathbf{dom}(f_n))) \neq \emptyset$. In fact,

$$\mathcal{B}(\mathbf{dom}(f_n)) = \{(s,\boldsymbol{w},\boldsymbol{z}) \in \mathbb{R}\times\mathbb{R}^q\times\mathbb{R}^m : \ s<0\}$$

follows from the fact that we may choose

$$\rho \equiv \frac{|s|}{\bar{\varphi}_1}\varphi_1, \quad \bar{\boldsymbol{u}} = \frac{|s|}{\mu_\Omega(\Omega)}\int_\Omega \boldsymbol{U}(\boldsymbol{x}_0)d\mu_\Omega(\boldsymbol{x}_0) + \boldsymbol{w} \quad \text{and} \quad \boldsymbol{v} = \frac{|s|}{\mu_\Omega(\Omega)}\int_\Omega \boldsymbol{A}\boldsymbol{U}(\boldsymbol{x}_0)d\mu_\Omega(\boldsymbol{x}_0) + \boldsymbol{A}\boldsymbol{w} + \boldsymbol{z}$$

to obtain $(s,\boldsymbol{w},\boldsymbol{z}) = \mathcal{B}(\rho,\bar{\boldsymbol{u}},\boldsymbol{v})$. Hence, strong duality holds. It remains to compute $f_n^\star$. We have that

$$\begin{aligned}
f_n^\star(\rho,\bar{\boldsymbol{u}},\boldsymbol{v}) &= \sup_{(\rho',\bar{\boldsymbol{u}}',\boldsymbol{v}')\in X_P^n} \langle(\rho,\bar{\boldsymbol{u}},\boldsymbol{v}),(\rho',\bar{\boldsymbol{u}}',\boldsymbol{v}')\rangle - f_(\rho',\bar{\boldsymbol{u}}',\boldsymbol{v}') \\
&= \sup_{\rho'\in V_n}\left(\langle\rho,\rho'\rangle_{L^2_\mu} - \frac{1}{2}\|\rho'\|^2_{L^2_\mu} - \mathbb{I}_{V_n^+}(\rho)\right) \\
&\quad + \sup_{\bar{\boldsymbol{u}}'\in\mathbb{R}^q}\left(\langle\bar{\boldsymbol{u}},\bar{\boldsymbol{u}}'\rangle_2 - \mathbb{I}_C(\bar{\boldsymbol{u}}')\right) + \sup_{\boldsymbol{v}'\in\mathbb{R}^m}\left(\langle\boldsymbol{v},\boldsymbol{v}'\rangle_2 - \frac{1}{\lambda}L(\boldsymbol{v}',\boldsymbol{y})\right) \\
&= \frac{1}{2}\|\rho\|^2_{L^2_\mu} + \sup_{\rho'\in V_n}\left(-\frac{1}{2}\|\rho-\rho'\|^2_{L^2_\mu} - \mathbb{I}_{V_n^+}(\rho)\right) + \mathbb{I}_C^\star(\bar{\boldsymbol{u}}) + \frac{1}{\lambda}L^\star(\lambda\boldsymbol{v},\boldsymbol{y}) \\
&= \frac{1}{2}\|\rho\|^2_{L^2_\mu} - \frac{1}{2}\|\rho_-\|^2_{L^2_\mu} + \mathbb{I}_C^\star(\bar{\boldsymbol{u}}) + \frac{1}{\lambda}L^\star(\lambda\boldsymbol{v},\boldsymbol{y}) \\
&= \frac{1}{2}\|\rho_+\|^2_{L^2_\mu} + \mathbb{I}_C^\star(\bar{\boldsymbol{u}}) + \frac{1}{\lambda}L^\star(\lambda\boldsymbol{v},\boldsymbol{y}),
\end{aligned}$$

proving the theorem.

B.4. Proof of Proposition 4

Let $h(s) = s_+^2/2$. Then h is differentiable with Lipschitz derivative $h'(s) = s_+$; its Lipschitz constant is 1. Fix $\boldsymbol{x}_0$ and let

$$q(\alpha,\boldsymbol{\beta}) = \alpha + \langle\boldsymbol{\beta},\boldsymbol{U}(\boldsymbol{x}_0)\rangle_2.$$

Then

$$\begin{aligned}
&\frac{|h(q(\alpha,\boldsymbol{\beta})) - h(q(\alpha_0,\boldsymbol{\beta}_0)) - h'(q(\alpha_0,\boldsymbol{\beta}_0))(q(\alpha,\boldsymbol{\beta}) - q(\alpha_0,\boldsymbol{\beta}_0)|}{((\alpha-\alpha_0)^2 + \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2^2)^{1/2}} \\
&\qquad = \frac{|q(\alpha,\boldsymbol{\beta}) - q(\alpha_0,\boldsymbol{\beta}_0)|}{((\alpha-\alpha_0)^2 + \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2^2)^{1/2}}\left|\int_0^1 h'((1-s)q(\alpha_0,\boldsymbol{\beta}_0) + sq(\alpha,\boldsymbol{\beta})) - h'(q(\alpha_0,\boldsymbol{\beta}_0))\,ds\right| \\
&\qquad \leq \frac{1}{2}\frac{(q(\alpha,\boldsymbol{\beta}) - q(\alpha_0,\boldsymbol{\beta}_0))^2}{((\alpha-\alpha_0)^2 + \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2^2)^{1/2}}
\end{aligned}$$

$$\leq \frac{(\alpha-\alpha_0)^2+\|\boldsymbol{U}(\boldsymbol{x}_0)\|_2^2\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2^2}{((\alpha-\alpha_0)^2+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2^2)^{1/2}}$$

$$\leq \max\{1,\|\boldsymbol{U}(\boldsymbol{x}_0)\|_2^2\}((\alpha-\alpha_0)^2+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2^2)^{1/2}.$$

Hence,

$$\frac{|H(\alpha,\boldsymbol{\beta})-H(\alpha_0,\boldsymbol{\beta}_0)-\partial_\alpha H(\alpha_0,\boldsymbol{\beta}_0)(\alpha-\alpha_0)-\langle\nabla_{\boldsymbol{\beta}}H(\alpha_0,\boldsymbol{\beta}_0),\,\boldsymbol{\beta}-\boldsymbol{\beta}_0\rangle_2|}{((\alpha-\alpha_0)^2+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2^2)^{1/2}}$$

$$\leq\left(\int_\Omega \max\{1,\|\boldsymbol{U}(\boldsymbol{x}_0)\|_2^2\}\,d\mu_\Omega(\boldsymbol{x}_0)\right)((\alpha-\alpha_0)^2+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2^2)^{1/2}$$

from where the first statement follows. The second statement follows from similar arguments, and the fact that h' is Lipschitz continuous. We omit the details for brevity.

B.5. Proof of Theorem 3

The theorem is a consequence of a change of variables using T. Although it is invertible with continuous inverse, it is not differentiable and it is not Lipschitz over $\bar{B}_2$. For this reason, we cannot apply the change of variables formula directly.

We proceed as follows. Let $\delta>0$. Define

$$A^\delta := \bar{B}_2\setminus\delta B_2=\{\boldsymbol{x}\in\bar{B}_2:\ \delta\leq\|\boldsymbol{x}\|_2\leq 1\}$$

and

$$\Omega_0^\delta:=\{T(\boldsymbol{x}):\ \boldsymbol{x}\in A^\delta\}=\{\boldsymbol{x}\in\Omega_0:\ \delta\leq\gamma_0(\boldsymbol{x})\leq 1\}.$$

Both $\{A^\delta\}_{\delta>0}$ and $\{\Omega_0^\delta\}_{\delta>0}$ are decreasing sequences. Consider the approximation

$$H^\delta(\alpha,\boldsymbol{\beta})=\frac{1}{2}\int_{\Omega_0^\delta}(\alpha+\langle\boldsymbol{\beta},\boldsymbol{U}(\boldsymbol{x}_0+\boldsymbol{x}_c)\rangle_2)_+^2\,d\boldsymbol{x}_0.$$

The integrand is non-negative and $H^\delta\to H$ pointwise as $\delta\to 0$ by the monotone convergence theorem [51, Theorem 2.4.1]. We will use the change of variables formula for each H^δ and then conclude by applying a limit argument.

For simplicity, we write $q(\boldsymbol{x})=\|\boldsymbol{x}\|_2/\gamma_0(\boldsymbol{x})$. Note q is well-defined for $\boldsymbol{x}\neq 0$ and there exists $C_q>0$ such that

$$\forall\,\boldsymbol{x}\neq 0:\ \frac{1}{C_q}\leq q(\boldsymbol{x})\leq C_q.$$

We first prove the following auxiliary result.

Lemma 1. *The map $T:A^\delta\to\Omega_0^\delta$ defined in (20) is Lipschitz with Lipschitz inverse $T^{-1}:\Omega_0^\delta\to A^\delta$ defined in (21).*

Proof of Lemma 1. Let $\boldsymbol{x},\boldsymbol{x}'\in A^\delta$. Then

$$T(\boldsymbol{x})-T(\boldsymbol{x}')=q(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{x}')+(q(\boldsymbol{x})-q(\boldsymbol{x}'))\boldsymbol{x}'.$$

The first term can be bounded by $\|\boldsymbol{x}-\boldsymbol{x}'\|_2$. For the second term, note that

$$|q(\boldsymbol{x})-q(\boldsymbol{x}')|\leq\frac{|\|\boldsymbol{x}\|_2-\|\boldsymbol{x}'\|_2|}{\gamma_0(\boldsymbol{x})}+\|\boldsymbol{x}'\|_2\left|\frac{1}{\gamma_0(\boldsymbol{x})}-\frac{1}{\gamma_0(\boldsymbol{x}')}\right|$$

$$\leq\frac{1}{\delta}\|\boldsymbol{x}-\boldsymbol{x}'\|_2+\left|\frac{1}{\gamma_0(\boldsymbol{x})}-\frac{1}{\gamma_0(\boldsymbol{x}')}\right|.$$

Finally, note that for $t,s\in[\delta,1]$ with $s\leq t$

$$\left|\frac{1}{t}-\frac{1}{s}\right|=\left|\int_s^t\frac{d\alpha}{\alpha^2}\right|\leq\frac{1}{\delta^2}|t-s|$$

from where

$$\|T(\boldsymbol{x})-T(\boldsymbol{x}')\|_2\leq C_q\|\boldsymbol{x}-\boldsymbol{x}'\|_2+\frac{1}{\delta}\|\boldsymbol{x}-\boldsymbol{x}'\|_2+\frac{1}{\delta^2}|\gamma_0(\boldsymbol{x})-\gamma_0(\boldsymbol{x}')|$$

and we conclude from the fact that γ_0 is Lipschitz continuous. To prove this claim, remark that since γ_0 is continuous, it is uniformly continuous on $\bar{B}_2(0,1)$. Let $\varepsilon > 0$ and let $\delta > 0$ be such that

$$\|\boldsymbol{x}' - \boldsymbol{x}\|_2 < \delta \;\Rightarrow\; |\gamma_0(\boldsymbol{x}') - \gamma_0(\boldsymbol{x})| < \varepsilon$$

for $\boldsymbol{x}', \boldsymbol{x} \in \bar{B}_2(0,1)$. Then, for any $\boldsymbol{x}', \boldsymbol{x} \in \mathbb{R}^d$ with $\boldsymbol{x}' \neq \boldsymbol{x}$ we can choose $\alpha = \delta/2\|\boldsymbol{x}' - \boldsymbol{x}\|_2$ so that

$$|\gamma_0(\boldsymbol{x}') - \gamma_0(\boldsymbol{x})| = \frac{1}{\lambda}|\gamma_0(\lambda\boldsymbol{x}') - \gamma_0(\lambda\boldsymbol{x})| < \frac{\varepsilon}{\lambda} = \frac{\varepsilon}{\delta}\|\boldsymbol{x}' - \boldsymbol{x}\|_2.$$

The same arguments can be applied to prove T^{-1} is Lipschitz continuous. We omit the details for brevity. □

Hence, T is differentiable Lebesgue a.e. by Rademacher's theorem [52, Theorem 2, Section 3.1.2]. Let DT denote its differential matrix. Then, we can apply the change of variables formula [52, Lemma 1, Section 3.3.1 and Theorem 2, Section 3.3.3], to obtain

$$H^\delta(\alpha, \boldsymbol{\beta}) = \frac{1}{2}\int_{A^\delta} (\alpha + \langle \boldsymbol{\beta}, \boldsymbol{U}(T(\boldsymbol{x}_0) + \boldsymbol{x}_c)\rangle_2)_+^2 |\det(DT(\boldsymbol{x}_0))|\, d\boldsymbol{x}_0.$$

We can further simplify the term involving the Jacobian using the following lemma.

Lemma 2. *The function $q : A^\delta \to \mathbb{R}^p$ is differentiable Lebesgue a.e. Furthermore, if $\boldsymbol{x}$ is a point of differentiability, then $\langle \nabla q(\boldsymbol{x}), \boldsymbol{x}\rangle_2 = 0$.*

Proof of Lemma 2. The arguments in the proof of Lemma 1 show q is Lipschitz continuous. Hence, by Rademacher's theorem it is differentiable Lebesgue a.e. If $\boldsymbol{x}$ is a point of differentiability, then

$$\lim_{\boldsymbol{x}' \to \boldsymbol{x}} \frac{|q(\boldsymbol{x}') - q(\boldsymbol{x}) - \langle \nabla q(\boldsymbol{x}), \boldsymbol{x}' - \boldsymbol{x}\rangle_2|}{\|\boldsymbol{x}' - \boldsymbol{x}\|_2} = 0.$$

However, for any $\alpha > 0$ and $\boldsymbol{x} \neq 0$

$$q(\alpha\boldsymbol{x}) = \frac{\|\alpha\boldsymbol{x}\|_2}{\gamma_0(\alpha\boldsymbol{x})} = \frac{|\alpha|\|\boldsymbol{x}\|_2}{|\alpha|\gamma_0(\boldsymbol{x})} = q(\boldsymbol{x})$$

whence q is homogeneous of degree 0. If we take $\boldsymbol{x}' = (1+\alpha)\boldsymbol{x}$ for α sufficiently small, it follows that $\boldsymbol{x}' - \boldsymbol{x} = \alpha\boldsymbol{x}$ and

$$\lim_{\alpha \to 0} \frac{|\alpha||\langle \nabla q(\boldsymbol{x}), \boldsymbol{x}\rangle_2|}{|\alpha|\|\boldsymbol{x}\|_2} = \frac{|\langle \nabla q(\boldsymbol{x}), \boldsymbol{x}\rangle_2|}{\|\boldsymbol{x}\|_2} = 0$$

whence the lemma follows. □

Therefore, there is a subset of A^δ of full Lebesgue measure where both T and q are differentiable. If $\boldsymbol{x}$ is such a point, then

$$DT(\boldsymbol{x}) = q(\boldsymbol{x})\boldsymbol{I}_d + \boldsymbol{x}\nabla q(\boldsymbol{x})^t \;\Rightarrow\; \det(DT(\boldsymbol{x})) = q(\boldsymbol{x})^d(1 + q(\boldsymbol{x})^{-d}\langle \nabla q(\boldsymbol{x}), \boldsymbol{x}\rangle_2) = q(\boldsymbol{x})^d$$

where we used the identity $\det(\boldsymbol{I} + \bar{\boldsymbol{u}}\boldsymbol{v}^t) = 1 + \langle \bar{\boldsymbol{u}}, \boldsymbol{v}\rangle_2$. We conclude that

$$H^\delta(\alpha, \boldsymbol{\beta}) = \frac{1}{2}\int_{A^\delta} \left(\alpha + \langle \boldsymbol{\beta}, \boldsymbol{U}(q(\boldsymbol{x})\boldsymbol{x} + \boldsymbol{x}_c)\rangle_2\right)_+^2 q(\boldsymbol{x})^d d\boldsymbol{x}.$$

As $H^\delta \to H$ pointwise, the theorem follows.

Appendix C. Proof of main results in Section 4

C.1. Proof of Proposition 5

We use the same notation as in the proof of Proposition 5. Note that

$$R_U(\bar{\boldsymbol{u}}) = \inf\left\{\frac{1}{2}\|\rho\|_{L^2_\mu}^2 \,:\, \rho \in L^2_\mu(\Omega),\, \rho \geq 0,\, \langle 1, \rho\rangle_{L^2_\mu} = 1,\, \mathcal{U}(\rho) = \bar{\boldsymbol{u}}\right\}. \tag{C.1}$$

Hence, $R_U(\bar{\boldsymbol{u}}) = +\infty$ if and only if the set

$$\{\rho \in L^2_\mu(\Omega) \,:\, \rho \geq 0,\, \langle 1, \rho\rangle_{L^2_\mu} = 1,\, \mathcal{U}(\rho) = \bar{\boldsymbol{u}}\}$$

is empty. Hence, the domain of R_U is given by (24). It is direct to verify it is convex. To show it is non-empty, it suffices to choose $\rho \equiv 1/\mu_\Omega(\Omega)$ to conclude that

$$\frac{1}{\mu_\Omega(\Omega)}\int_\Omega \boldsymbol{U}(\boldsymbol{x}_0) d\mu_\Omega(\boldsymbol{x}_0) \in \mathbf{dom}(R_U)$$

whence R_U is proper.

To prove the convexity of R_U, observe that the objective in (C.1) is strongly convex [27, Definition 10.7] and that the feasible set is closed. Then, there exists a unique ρ_u such that [27, Proposition 11.8]

$$R_U(\bar{\boldsymbol{u}}) = \frac{1}{2}\|\rho_u\|^2_{L^2_\mu}.$$

Let $\bar{\boldsymbol{u}}_1, \bar{\boldsymbol{u}}_2 \in \mathbf{dom}(R_U)$, let $\rho_1 = \rho_{u_1}$ and $\rho_2 = \rho_{u_2}$, and let $\theta \in [0,1]$. Since

$$\theta\rho_1 + (1-\theta)\rho_2 \geq 0, \quad \langle 1, \theta\rho_1 + (1-\theta)\rho_2\rangle_{L^2_\mu} = 1 \quad \text{and} \quad \mathcal{U}(\theta\rho_1 + (1-\theta)\rho_2) = \theta\bar{\boldsymbol{u}}_1 + (1-\theta)\bar{\boldsymbol{u}}_2$$

we see that

$$\begin{aligned} R_U(\theta\bar{\boldsymbol{u}}_1 + (1-\theta)\bar{\boldsymbol{u}}_2) &\leq \frac{1}{2}\|\theta\rho_1 + (1-\theta)\rho_2\|^2_{L^2_\mu} \\ &\leq \frac{1}{2}\theta\|\rho_1\|^2_{L^2_\mu} + \frac{1}{2}(1-\theta)\|\rho_2\|^2_{L^2_\mu} \\ &= \theta R_U(\bar{\boldsymbol{u}}_1) + (1-\theta)R_U(\bar{\boldsymbol{u}}_2). \end{aligned}$$

Therefore, we conclude R_U is convex.

C.2. Proof of Proposition 6

By the choice of $\boldsymbol{U}$ we have $\mathbf{dom}(R_U) \subset \Omega$. To show $\mathbf{int}(\Omega) \subset \mathbf{dom}(R_U)$ let $\boldsymbol{\mu}_0 \in \mathbf{int}(\Omega)$ and let $r > 0$ be such that $B(\boldsymbol{\mu}_0, r) \subset \mathbf{int}(\Omega)$. Then, it suffices to choose

$$\rho(\boldsymbol{x}_0) := \frac{1}{\mu(B(\boldsymbol{\mu}_0, r))}\chi_{B(\boldsymbol{\mu}_0, r)}$$

to conclude that $\boldsymbol{\mu}_0 \in \mathbf{dom}(R_U)$. Now, it remains to show that $\boldsymbol{\mu}_0 \in \mathbf{bd}(\Omega) \notin \mathbf{dom}(R_U)$. Suppose there is $\rho \in L^2_\mu(\Omega)$ with $\rho \geq 0$ and $\langle 1, \rho\rangle_{L^2_\mu} = 1$ such that

$$\boldsymbol{\mu}_0 = \int_\Omega \rho(\boldsymbol{x}_0)\boldsymbol{x}_0\, d\boldsymbol{x}_0 \;\Rightarrow\; \int_\Omega \rho(\boldsymbol{x}_0)(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)\, d\boldsymbol{x}_0 = 0.$$

Let $\boldsymbol{z} \in N_\Omega(\boldsymbol{\mu}_0)$ where N_Ω denotes the normal cone. Define for $s \geq 0$

$$\Omega_s := \{\boldsymbol{x}_0 \in \Omega : \langle \boldsymbol{z}, \boldsymbol{x}_0 - \boldsymbol{\mu}_0\rangle \leq s\}$$

and observe that $\Omega = \Omega_0$. If $\rho \geq 0$ on a measurable set $A \subset \Omega_s$ for some $s < 0$ then

$$\int_A \rho(\boldsymbol{x}_0)\langle \boldsymbol{z}, \boldsymbol{x}_0 - \boldsymbol{\mu}_0\rangle_2\, d\boldsymbol{x}_0 \leq s\int_A \rho(\boldsymbol{x}_0)d\boldsymbol{x}_0 < 0.$$

However, this implies

$$\begin{aligned} 0 &= \int_{\Omega\cap A} \rho(\boldsymbol{x}_0)\langle \boldsymbol{z}, \boldsymbol{x}_0 - \boldsymbol{\mu}_0\rangle_2\, d\boldsymbol{x}_0 + \int_{\Omega\setminus A} \rho(\boldsymbol{x}_0)\langle \boldsymbol{z}, \boldsymbol{x}_0 - \boldsymbol{\mu}_0\rangle_2\, d\boldsymbol{x}_0 \\ &< s\int_A \rho(\boldsymbol{x}_0)d\boldsymbol{x}_0 + \int_{\Omega\setminus A} \rho(\boldsymbol{x}_0)\langle \boldsymbol{z}, \boldsymbol{x}_0 - \boldsymbol{\mu}_0\rangle_2\, d\boldsymbol{x}_0. \end{aligned}$$

We conclude that

$$0 < -s\int_A \rho(\boldsymbol{x}_0)d\boldsymbol{x}_0 < \int_{\Omega\setminus A} \rho(\boldsymbol{x}_0)\langle \boldsymbol{z}, \boldsymbol{x}_0 - \boldsymbol{\mu}_0\rangle_2\, d\boldsymbol{x}_0 \leq 0$$

which is a contradiction unless $s = 0$. Therefore, ρ must be supported on

$$\Omega \cap \{\boldsymbol{x}_0 \in \mathbb{R}^d : \langle \boldsymbol{z}, \boldsymbol{x}_0 - \boldsymbol{\mu}_0\rangle_2 = 0\} \subset \{\boldsymbol{x}_0 \in \mathbb{R}^d : \langle \boldsymbol{z}, \boldsymbol{x}_0 - \boldsymbol{\mu}_0\rangle_2 = 0\}.$$

Since the set in the right-hand side has Lebesgue measure zero, we conclude $\rho = 0$. This contradiction proves the first statement of the proposition. To prove the second, let $\boldsymbol{\mu}_0 \in \mathbf{bd}(\Omega)$ and let $\{\boldsymbol{\mu}_n\}_{n\in\mathbb{N}}$ be a sequence in $\mathbf{int}(\Omega)$ with $\boldsymbol{\mu}_n \to \boldsymbol{\mu}_0$. Let $\{\rho_n\}_{n\in\mathbb{N}}$ be the sequence of optimal solutions to (14) for each $\boldsymbol{\mu}_n$. Suppose that $\{R_U(\boldsymbol{\mu}_n)\}_{n\in\mathbb{N}}$ remains bounded by a constant $C > 0$. This implies that $\|\rho_n\|_{n\in\mathbb{N}}$ is norm bounded in $L^2(\Omega)$. Since norm bounded sequences are weakly convergent the sequence has a weak limit $\rho_0 \in L^2(\Omega)$ [27, Theorem 2.33]. Since the set $\{\rho \in L^2(\Omega) : \rho \geq 0, \langle 1, \rho\rangle_{L^2} = 1\}$ is norm closed, it is weakly closed and ρ_0 also belongs to this set [27, Theorem 3.34]. However, the function $f(\boldsymbol{x}_0) = x_{0,i}$ is in $L^2(\Omega)$ for each $i \in \{1, \ldots, d\}$. Therefore,

$$\int_\Omega \rho_n(\boldsymbol{x}_0) x_{0,i} d\boldsymbol{x}_0 = \boldsymbol{\mu}_n \to \boldsymbol{\mu}_0 = \int_\Omega \rho_0(\boldsymbol{x}_0) x_{0,i} d\boldsymbol{x}_0.$$

However, $\boldsymbol{\mu}_0$ lies in the boundary. Using the same arguments as before, we reach a contradiction. This proves the claim.

C.3. Proof of Proposition 7

We show that R_U° is invariant under transformations of the form $T(\boldsymbol{\mu}) = \boldsymbol{Q}(\boldsymbol{\mu} - \boldsymbol{\mu}_0) + \boldsymbol{\mu}_0$ for any $d \times d$ orthogonal matrix $\boldsymbol{Q}$. The value of $R_U^\circ(T(\boldsymbol{x}_0))$ is the optimal value of the problem

$$\begin{aligned} &\underset{\rho \in L^2_\mu(\Omega):\, \rho \geq 0}{\text{minimize}} \quad \frac{1}{2}\int_\Omega \rho(\boldsymbol{x}_0)^2 d\boldsymbol{x}_0 \\ &\text{subject to} \quad \int_\Omega \rho(\boldsymbol{x}_0) d\mu_\Omega(\boldsymbol{x}_0) = 1, \quad \int_\Omega \rho(\boldsymbol{x}_0)\boldsymbol{x}_0 d\mu_\Omega(\boldsymbol{x}_0) = T(\boldsymbol{\mu}). \end{aligned}$$

However, T induces an isometric isomorphism $\mathcal{T} : L^2_\mu(\Omega) \mapsto L^2_\mu(\Omega)$ given by $\mathcal{T}(\rho)(\boldsymbol{x}_0) = \rho(T(\boldsymbol{x}_0))$. Therefore, the optimal value to the above problem is equivalent to

$$\begin{aligned} &\underset{\rho \in L^2_\mu(\Omega):\, \rho \geq 0}{\text{minimize}} \quad \frac{1}{2}\int_\Omega \rho(\boldsymbol{x}_0)^2 d\boldsymbol{x}_0 \\ &\text{subject to} \quad \int_\Omega \rho(\boldsymbol{x}_0) d\boldsymbol{x}_0 = 1, \quad \int_\Omega \rho(\boldsymbol{x}_0) T(\boldsymbol{x}_0) d\boldsymbol{x}_0 = T(\boldsymbol{\mu}), \end{aligned}$$

whence $R_U^\circ(T(\boldsymbol{\mu})) = R_U^\circ(\boldsymbol{\mu})$. It follows that the function $f : \bar{B}_2(0,1) \to \mathbb{R}$ given by $f(\boldsymbol{z}) = R_U^\circ(r\boldsymbol{z} + \boldsymbol{\mu}_0)$ is invariant under any orthogonal transformation. We conclude that $f(\boldsymbol{z}) = \varphi(\|\boldsymbol{z}\|_2)$ for some non-negative function φ and, by a suitable change of variable, we arrive to the first statement. Note that $\varphi \geq 0$ as $R_U^\circ \geq 0$. Furthermore, since $f(\boldsymbol{z}) = \varphi(\|\boldsymbol{x}\|_2)$ is convex on $\bar{B}_2(0,1)$ then φ must be convex and non-decreasing. Finally, from Proposition 6 we deduce that $\varphi(s) \to \infty$ as $s \to 1$ as this implies $\boldsymbol{\mu} \to \mathbf{bd}(\Omega)$. The second statement follows by standard arguments and the fact that φ is non-decreasing. We omit the proof for brevity.

Appendix D. Proof of main results in Section 5

D.1. Proof of Proposition 8

We have the bound

$$\begin{aligned} \sigma^2(\alpha, \boldsymbol{\beta}) &= \int_\Omega (h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0) - H(\alpha, \boldsymbol{\beta}))^2 d\Pi_\Omega(\boldsymbol{x}_0) \\ &= \int_\Omega \left(\int_\Omega (h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0) - h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0'))) d\Pi_\Omega(\boldsymbol{x}_0') \right)^2 d\Pi_\Omega(\boldsymbol{x}_0) \\ &\leq \iint_\Omega (h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0) - h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0'))^2 d\Pi_\Omega(\boldsymbol{x}_0') d\Pi_\Omega(\boldsymbol{x}_0). \end{aligned}$$

From the identity

$$\frac{1}{2}(c + x_2)_+^2 - \frac{1}{2}(c + x_1)_+^2 = (x_2 - x_1) \int_0^1 (c + \theta x_2 + (1-\theta)x_1)_+ \, d\theta$$

for any $c, x_1, x_2 \in \mathbb{R}$ we deduce

$$\begin{aligned} |h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0) - h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0')| &\leq \left(\int_0^1 (\alpha + \langle \boldsymbol{\beta}, (1-\theta)\boldsymbol{U}(\boldsymbol{x}_0') + \theta \boldsymbol{U}(\boldsymbol{x}_0) \rangle)_+ d\theta \right) |\langle \boldsymbol{\beta}, \boldsymbol{U}(\boldsymbol{x}_0) - \boldsymbol{U}(\boldsymbol{x}_0') \rangle_2| \\ &\leq \left(\int_0^1 ((1-\theta)(\alpha + \langle \boldsymbol{\beta}, \boldsymbol{U}(\boldsymbol{x}_0') \rangle_2)_+ + \theta(\alpha + \langle \boldsymbol{\beta}, \boldsymbol{U}(\boldsymbol{x}_0) \rangle_2)_+ d\theta \right) |\langle \boldsymbol{\beta}, \boldsymbol{U}(\boldsymbol{x}_0) - \boldsymbol{U}(\boldsymbol{x}_0') \rangle_2| \\ &= \left(\frac{1}{2}(\alpha + \langle \boldsymbol{\beta}, \boldsymbol{U}(\boldsymbol{x}_0') \rangle_2)_+ + \frac{1}{2}(\alpha + \langle \boldsymbol{\beta}, \boldsymbol{U}(\boldsymbol{x}_0) \rangle_2)_+ \right) |\langle \boldsymbol{\beta}, \boldsymbol{U}(\boldsymbol{x}_0) - \boldsymbol{U}(\boldsymbol{x}_0') \rangle_2| \end{aligned}$$

$$\leq \|\boldsymbol{\beta}\|_2 \left(|\alpha| + \frac{1}{2}\|\boldsymbol{\beta}\|_2 (\|\boldsymbol{U}(\boldsymbol{x}_0)\|_2 + \|\boldsymbol{U}(\boldsymbol{x}_0')\|_2) \right) \|\boldsymbol{U}(\boldsymbol{x}_0) - \boldsymbol{U}(\boldsymbol{x}_0')\|_2$$

the proposition follows.

D.2. Proof of Proposition 9

Since

$$\boldsymbol{g}_n(\alpha, \boldsymbol{\beta}) - \nabla H(\alpha, \boldsymbol{\beta}) = \frac{1}{M} \sum_{m=1}^{M} \iint_{\Omega} (\nabla h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0^{(n,m)}) - \nabla h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0)) d\Pi_\Omega(\boldsymbol{x}_0) d\Pi_\Omega(\boldsymbol{x}_0^{(n,m)})$$

we have

$$\begin{aligned}
\mathbf{E}_{\boldsymbol{x}_0 \sim \Pi_\Omega} \|\boldsymbol{g}_n(\alpha, \boldsymbol{\beta}) - \nabla H(\alpha, \boldsymbol{\beta})\|_2^2 &= \frac{1}{M^2} \sum_{m=1}^{M} \iint_{\Omega} \|\nabla h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0^{(n,m)}) - \nabla h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0)\|_2^2 d\Pi_\Omega(\boldsymbol{x}_0) d\Pi_\Omega(\boldsymbol{x}_0^{(n,m)}) \\
&= \frac{1}{M} \iint_{\Omega} \|\nabla h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0') - \nabla h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0)\|_2^2 d\Pi_\Omega(\boldsymbol{x}_0) d\Pi_\Omega(\boldsymbol{x}_0')
\end{aligned}$$

From

$$\begin{aligned}
|\partial_\alpha h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0') - \partial_\alpha h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0)| &= |(\alpha + \langle \boldsymbol{\beta}, \boldsymbol{U}(\boldsymbol{x}_0') \rangle_2)_+ - (\alpha + \langle \boldsymbol{\beta}, \boldsymbol{U}(\boldsymbol{x}_0) \rangle_2)_+| \\
&\leq \|\boldsymbol{\beta}\|_2 \|\boldsymbol{U}(\boldsymbol{x}_0') - \boldsymbol{U}(\boldsymbol{x}_0)\|_2
\end{aligned}$$

and

$$\begin{aligned}
\|\nabla_{\boldsymbol{\beta}} h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0') - \nabla_{\boldsymbol{\beta}} h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0)\|_2 &\leq |(\alpha + \langle \boldsymbol{\beta}, \boldsymbol{U}(\boldsymbol{x}_0') \rangle_2)_+ - (\alpha + \langle \boldsymbol{\beta}, \boldsymbol{U}(\boldsymbol{x}_0) \rangle_2)_+| \|\boldsymbol{U}(\boldsymbol{x}_0')\|_2 \\
&\quad + (\alpha + \langle \boldsymbol{\beta}, \boldsymbol{U}(\boldsymbol{x}_0) \rangle_2)_+ \|\boldsymbol{U}(\boldsymbol{x}_0') - \boldsymbol{U}(\boldsymbol{x}_0)\|_2 \\
&\leq (\|\boldsymbol{\beta}\|_2 \|\boldsymbol{U}(\boldsymbol{x}_0')\|_2 + |\alpha| + \|\boldsymbol{\beta}\|_2) \|\boldsymbol{U}(\boldsymbol{x}_0') - \boldsymbol{U}(\boldsymbol{x}_0)\|_2
\end{aligned}$$

we deduce

$$\|\nabla h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0') - \nabla h(\alpha, \boldsymbol{\beta}, \boldsymbol{x}_0)\|_2^2 \leq 2\|\boldsymbol{U}(\boldsymbol{x}_0') - \boldsymbol{U}(\boldsymbol{x}_0)\|_2^2 (\|\boldsymbol{\beta}\|_2^2 + (\|\boldsymbol{\beta}\|_2 \|\boldsymbol{U}(\boldsymbol{x}_0')\|_2 + |\alpha| + \|\boldsymbol{\beta}\|_2)^2)$$

from where the inequality follows. We omit the details for brevity.

References

[1] M. Braun, Differential Equations and Their Applications: An Introduction to Applied Mathematics, Texts in Applied Mathematics, vol. 11, Springer, New York, New York, NY, 1993, http://link.springer.com/10.1007/978-1-4612-4360-1.

[2] S. Wiggins, Introduction to Applied Nonlinear Dynamical Systems and Chaos, Texts in Applied Mathematics, vol. 2, Springer-Verlag, New York, 2003, http://link.springer.com/10.1007/b97481.

[3] E. Hairer, G. Wanner, C. Lubich, Geometric Numerical Integration, Springer Series in Computational Mathematics, vol. 31, Springer-Verlag, Berlin/Heidelberg, 2006, http://link.springer.com/10.1007/3-540-30666-8.

[4] A. Iserles, A First Course in the Numerical Analysis of Differential Equations, 2nd edition, Cambridge University Press, 2008, https://www.cambridge.org/core/product/identifier/9780511995569/type/book.

[5] M. Arnold, Numerical methods for simulation in applied dynamics, in: G. Maier, J. Salençon, W. Schneider, B. Schrefler, P. Serafini, M. Arnold, W. Schiehlen (Eds.), Simulation Techniques for Applied Dynamics, in: CISM International Centre for Mechanical Sciences., vol. 507, Springer, Vienna, Vienna, 2008, pp. 191–246, http://link.springer.com/10.1007/978-3-211-89548-1_5.

[6] H.-G. Beyer, B. Sendhoff, Robust optimization – a comprehensive survey, Comput. Methods Appl. Mech. Eng. 196 (33–34) (2007) 3190–3218, https://doi.org/10.1016/j.cma.2007.03.003, https://linkinghub.elsevier.com/retrieve/pii/S0045782507001259.

[7] C.J. Roy, W.L. Oberkampf, A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing, Comput. Methods Appl. Mech. Eng. 200 (25–28) (2011) 2131–2144, https://doi.org/10.1016/j.cma.2011.03.016, https://linkinghub.elsevier.com/retrieve/pii/S0045782511001290.

[8] R.C. Smith, Uncertainty Quantification: Theory, Implementation, and Applications, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013, https://epubs.siam.org/doi/book/10.1137/1.9781611973228.

[9] D. Kumar, F. Ahmed, S. Usman, A. Alajo, S.B. Alam, Recent advances in uncertainty quantification methods for engineering problems, in: AI Assurance, Elsevier, 2023, pp. 453–472, https://linkinghub.elsevier.com/retrieve/pii/B9780323919197000275.

[10] H.T. Banks, K.B. Flores, I.G. Rosen, E.M. Rutter, M. Sirlanci, W.C. Thompson, The Prohorov metric framework and aggregate data inverse problems for random PDEs, Commun. Appl. Anal. 22 (3) (2018) 415–446.

[11] C. Schacht, A. Meade, H. Banks, H. Enderling, D. Abate-Daga, Estimation of probability distributions of parameters using aggregate population data: analysis of a CAR T-cell cancer model, Math. Biosci. Eng. 16 (6) (2019) 7299–7326, https://doi.org/10.3934/mbe.2019365, http://www.aimspress.com/article/10.3934/mbe.2019365.

[12] H. Banks, A.E. Meade, C. Schacht, J. Catenacci, W.C. Thompson, D. Abate-Daga, H. Enderling, Parameter estimation using aggregate data, Appl. Math. Lett. 100 (2020) 105999, https://doi.org/10.1016/j.aml.2019.105999, https://linkinghub.elsevier.com/retrieve/pii/S0893965919303234.

[13] H.T. Banks, S. Hu, W.C. Thompson, H.T. Banks, Modeling and Inverse Problems in the Presence of Uncertainty, Monographs and Research Notes in Mathematics, CRC Press/Chapman & Hall, Boca Raton, Fla., 2014.

[14] H.T. Banks, W. Botsford, F. Kappel, C. Wang, Modeling and estimation in size-structured population models, in: 2nd Course on Mathematical Ecology, Trieste, December 8–12, 1986, World Press, Singapore, 1988, pp. 521–541.
[15] H.T. Banks, H.T. Banks, A Functional Analysis Framework for Modeling, Estimation and Control in Science and Engineering, A Chapman & Hall book CRC Press, Chapman & Hall, Boca Raton, Fla., 2012.
[16] J. Meyers, J. Rogers, A. Gerlach, Koopman operator method for solution of generalized aggregate data inverse problems, J. Comput. Phys. 428 (2021) 110082, https://doi.org/10.1016/j.jcp.2020.110082, https://linkinghub.elsevier.com/retrieve/pii/S0021999120308561.
[17] A.R. Gerlach, A. Leonard, J. Rogers, C. Rackauckas, The Koopman expectation: an operator theoretic method for efficient analysis and optimization of uncertain hybrid dynamical systems, arXiv:2008.08737 [math], http://arxiv.org/abs/2008.08737, Aug. 2020.
[18] A.N. Tikhonov, V.Y. Arsenin, Solutions of Ill-Posed Problems, Halsted Press, New York, 1977.
[19] X. Jiang, S. Li, R. Furfaro, Z. Wang, Y. Ji, High-dimensional uncertainty quantification for Mars atmospheric entry using adaptive generalized polynomial chaos, Aerosp. Sci. Technol. 107 (2020) 106240, https://doi.org/10.1016/j.ast.2020.106240, https://linkinghub.elsevier.com/retrieve/pii/S1270963820309226.
[20] I. Abdallah, C. Lataniotis, B. Sudret, Parametric hierarchical Kriging for multi-fidelity aero-servo-elastic simulators — application to extreme loads on wind turbines, Probab. Eng. Mech. 55 (2019) 67–77, https://doi.org/10.1016/j.probengmech.2018.10.001, https://linkinghub.elsevier.com/retrieve/pii/S0266892017302126.
[21] Z. He, Z. Zhang, High-dimensional uncertainty quantification via tensor regression with rank determination and adaptive sampling, IEEE Trans. Compon. Packag. Manuf. Technol. 11 (9) (2021) 1317–1328, https://doi.org/10.1109/TCPMT.2021.3093432, https://ieeexplore.ieee.org/document/9467329/.
[22] P. Hall, On Kullback-Leibler loss and density estimation, Ann. Stat. 15 (4) (Dec. 1987), https://doi.org/10.1214/aos/1176350606, https://projecteuclid.org/journals/annals-of-statistics/volume-15/issue-4/On-Kullback-Leibler-Loss-and-Density-Estimation/10.1214/aos/1176350606.full.
[23] S.N. MacEachern, Nonparametric Bayesian methods: a gentle introduction and overview, Commun. Stat. Appl. Methods 23 (6) (2016) 445–466, https://doi.org/10.5351/CSAM.2016.23.6.445, http://www.csam.or.kr/journal/view.html?doi=10.5351/CSAM.2016.23.6.445.
[24] B.O. Koopman, Hamiltonian systems and transformation in Hilbert space, Proc. Natl. Acad. Sci. 17 (5) (1931) 315–318, https://doi.org/10.1073/pnas.17.5.315, https://pnas.org/doi/full/10.1073/pnas.17.5.315.
[25] M. Budišić, R. Mohr, I. Mezić, Applied koopmanism, chaos: an interdisciplinary, J. Nonlinear Sci. 22 (4) (2012) 047510, https://doi.org/10.1063/1.4772195, https://pubs.aip.org/cha/article/22/4/047510/341880/Applied-Koopmanisma.
[26] R.T. Rockafellar, R.J.B. Wets, Variational Analysis, Grundlehren der Mathematischen Wissenschaften, vol. 317, Springer, Berlin Heidelberg, 1998, http://link.springer.com/10.1007/978-3-642-02431-3.
[27] H.H. Bauschke, P.L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces, CMS Books in Mathematics, Springer, New York, NY, 2011, https://link.springer.com/10.1007/978-1-4419-9467-7.
[28] Y. Nesterov, A method for solving the convex programming problem with convergence rate $O(1/k^2)$, Dokl. Akad. Nauk SSSR 269 (3) (1983) 543–547.
[29] Y. Nesterov, Gradient methods for minimizing composite functions, Math. Program. 140 (1) (2013) 125–161, https://doi.org/10.1007/s10107-012-0629-5, http://link.springer.com/10.1007/s10107-012-0629-5.
[30] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Sci. 2 (1) (2009) 183–202, https://doi.org/10.1137/080716542, http://epubs.siam.org/doi/10.1137/080716542.
[31] A.M. Rubinov, A.A. Yagubov, The space of star-shaped sets and its applications in nonsmooth optimization, in: M.L. Balinski, R.W. Cottle, L.C.W. Dixon, B. Korte, M.J. Toddla, E.L. Allgower, W.H. Cunningham, J.E. Dennis, B.C. Eaves, R. Fletcher, D. Goldfarb, J.-B. Hiriart-Urruty, M. Iri, R.G. Jeroslow, D.S. Johnson, C. Lemarechal, L. Lovasz, L. McLinden, M.J.D. Powell, W.R. Pulleyblank, A.H.G. Rinnooy Kan, K. Ritter, R.W.H. Sargent, D.F. Shanno, L.E. Trotter, H. Tuy, R.J.B. Wets, E.M.L. Beale, G.B. Dantzig, L.V. Kantorovich, T.C. Koopmans, A.W. Tucker, P. Wolfe, V.F. Demyanov, L.C.W. Dixon (Eds.), Quasidifferential Calculus, in: Mathematical Programming Studies, vol. 29, Springer, Berlin Heidelberg, 1986, pp. 176–202, http://link.springer.com/10.1007/BFb0121146.
[32] A.S. Nemirovskii, D.B. Yudin, Problem Complexity and Method Efficiency in Optimization, Wiley-Interscience Series in Discrete Mathematics, Wiley Chichester, New York, 1983.
[33] A.H. Stroud, D. Secrest, Gaussian Quadrature Formulas, Prentice-Hall, Englewood Cliffs, NJ, 1966.
[34] J. Glaubitz, Constructing positive interpolatory cubature formulas, arXiv:2009.11981 (Sep. 2020), http://arxiv.org/abs/2009.11981.
[35] T. Gerstner, M. Griebel, Numerical integration with sparse grids, Numer. Algorithms 18 (3/4) (1998) 209–232, https://doi.org/10.1023/A:1019129717644, http://link.springer.com/10.1023/A:1019129717644.
[36] G. Petrova, Cubature formulae for spheres, simplices and balls, J. Comput. Appl. Math. 162 (2) (2004) 483–496, https://doi.org/10.1016/j.cam.2003.08.036, https://linkinghub.elsevier.com/retrieve/pii/S0377042703007611.
[37] D. Williams, L. Shunn, A. Jameson, Symmetric quadrature rules for simplexes based on sphere close packed lattice arrangements, J. Comput. Appl. Math. 266 (2014) 18–38, https://doi.org/10.1016/j.cam.2014.01.007, https://linkinghub.elsevier.com/retrieve/pii/S0377042714000211.
[38] S.A. Smolyak, Quadrature and interpolation formulas for tensor products of certain classes of functions, Dokl. Akad. Nauk SSSR 148 (5) (1963) 1042–1045.
[39] J. Lu, D.L. Darmofal, Higher-dimensional integration with Gaussian weight for applications in probabilistic design, SIAM J. Sci. Comput. 26 (2) (2004) 613–624, https://doi.org/10.1137/S1064827503426863, http://epubs.siam.org/doi/10.1137/S1064827503426863.
[40] I. Mysovskikh, The approximation of multiple integrals by using interpolatory cubature formulae, in: Quantitative Approximation, Elsevier, 1980, pp. 217–243, https://linkinghub.elsevier.com/retrieve/pii/B9780122136504500258.
[41] J.H. Halton, On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals, Numer. Math. 2 (1) (1960) 84–90, https://doi.org/10.1007/BF01386213, http://link.springer.com/10.1007/BF01386213.
[42] H. Niederreiter, Low-discrepancy and low-dispersion sequences, J. Number Theory 30 (1) (1988) 51–70, https://doi.org/10.1016/0022-314X(88)90025-X, https://linkinghub.elsevier.com/retrieve/pii/0022314X8890025X.
[43] N.A. Butler, A new class of equal-weight integration rules on the hypercube, Numer. Math. 99 (2) (2004) 349–363, https://doi.org/10.1007/s00211-004-0562-5, http://link.springer.com/10.1007/s00211-004-0562-5.
[44] A. Nemirovski, A. Juditsky, G. Lan, A. Shapiro, Robust stochastic approximation approach to stochastic programming, SIAM J. Optim. 19 (4) (2009) 1574–1609, https://doi.org/10.1137/070704277, http://epubs.siam.org/doi/10.1137/070704277.
[45] W.J. Morokoff, R.E. Caflisch, Quasi-Monte Carlo integration, J. Comput. Phys. 122 (2) (1995) 218–230, https://doi.org/10.1006/jcph.1995.1209, https://linkinghub.elsevier.com/retrieve/pii/S0021999185712090.
[46] J. Brauchart, E. Saff, I. Sloan, R. Womersley, QMC designs: optimal order quasi Monte Carlo integration schemes on the sphere, Math. Comput. 83 (290) (2014) 2821–2851, https://doi.org/10.1090/S0025-5718-2014-02839-1, https://www.ams.org/mcom/2014-83-290/S0025-5718-2014-02839-1/.
[47] R.L. Smith, Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions, Oper. Res. 32 (6) (1984) 1296–1308.
[48] L. Lovász, S. Vempala, Hit-and-run from a corner, SIAM J. Comput. 35 (4) (2006) 985–1005, https://doi.org/10.1137/S009753970544727X, http://epubs.siam.org/doi/10.1137/S009753970544727X.
[49] D.J. Higham, Modeling and simulating chemical reactions, SIAM Rev. 50 (2) (2008) 347–368, https://doi.org/10.1137/060666457, http://epubs.siam.org/doi/10.1137/060666457.
[50] Q. Yang, C.A. Sing-Long, E.J. Reed, Rapid data-driven model reduction of nonlinear dynamical systems including chemical reaction networks using ℓ^1-regularization, Chaos 30 (5) (2020) 053122, https://doi.org/10.1063/1.5139463, https://pubs.aip.org/aip/cha/article/341976.
[51] D.L. Cohn, Measure Theory, Birkhäuser Advanced Texts Basler Lehrbücher, second edition, Springer, New York, NY, 2013, https://link.springer.com/10.1007/978-1-4614-6956-8.
[52] L.C. Evans, R.F. Gariepy, Measure Theory and Fine Properties of Functions, 1st edition, Chapman and Hall/CRC, 2015, https://www.taylorfrancis.com/books/9781482242393.