Statistics and Computing (2025) 35:10
https://doi.org/10.1007/s11222-024-10535-0

ORIGINAL PAPER

Gaussian measures conditioned on nonlinear observations: consistency, MAP estimators, and simulation

Yifan Chen[1] · Bamdad Hosseini[2] · Houman Owhadi[3] · Andrew M. Stuart[3]

Received: 21 May 2024 / Accepted: 11 November 2024 / Published online: 9 December 2024

Abstract

The article presents a systematic study of the problem of conditioning a Gaussian random variable ξ on nonlinear observations of the form $F \circ \boldsymbol{\phi}(\xi)$ where $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^N$ is a bounded linear operator and F is nonlinear. Such problems arise in the context of Bayesian inference and recent machine learning-inspired PDE solvers. We give a representer theorem for the conditioned random variable $\xi \mid F \circ \boldsymbol{\phi}(\xi)$, stating that it decomposes as the sum of an infinite-dimensional Gaussian (which is identified analytically) as well as a finite-dimensional non-Gaussian measure. We also introduce a novel notion of the mode of a conditional measure by taking the limit of the natural relaxation of the problem, to which we can apply the existing notion of maximum a posteriori estimators of posterior measures. Finally, we introduce a variant of the Laplace approximation for the efficient simulation of the aforementioned conditioned Gaussian random variables towards uncertainty quantification.

Keywords Gaussian measures · Conditional probability · Bayesian inference

1 Introduction

We consider the problem of conditioning a Gaussian measure on a finite set of nonlinear observations in the form of a nonlinear transformation of bounded linear functionals. Let $\{\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}}, \|\cdot\|_{\mathcal{X}}\}$ be a separable Hilbert space with dual $\mathcal{X}^*$ and consider a Gaussian measure $\mu = N(0,\mathcal{K}) \in \mathcal{P}(\mathcal{X})$, where $\mathcal{P}(\mathcal{X})$ denotes the set of all Borel probability measures on $\mathcal{X}$. Let $\mathcal{K} : \mathcal{X} \to \mathcal{X}$ denote the covariance operator under μ. Fix a vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N) \in (\mathcal{X}^*)^{\otimes N}$ for $N \in \mathbb{N}$, along with a nonlinear map $F : \mathbb{R}^N \to \mathbb{R}^M$ for $M \in \mathbb{N}$. Let $\xi \sim \mu$ and $\beta > 0$ be a parameter; then our goal in this article is to characterize the family of measures

$$\mu_\beta^{\mathbf{y}} := \text{Law}\{\xi \mid \mathbf{y} \sim N(F(\boldsymbol{\phi}(\xi)), \beta^2 I)\}, \tag{1}$$

and their modes, in the limit of small β. The natural candidate for the $\beta = 0$ limit is

$$\mu_0^{\mathbf{y}} := \text{Law}\{\xi \mid F(\boldsymbol{\phi}(\xi)) = \mathbf{y}\}. \tag{2}$$

We refer to the measures $\mu_\beta^{\mathbf{y}}$ for $\beta > 0$ as *posteriors* and to their $\beta = 0$ limit $\mu_0^{\mathbf{y}}$ as *conditionals.* The modes of the posterior measures $\mu_\beta^{\mathbf{y}}$, which are often called the *maximum a posteriori (MAP)* estimator, can be characterized via optimization problems of the following form

$$u_\beta^{\mathbf{y}} := \underset{u \in \mathcal{K}^{1/2}\mathcal{X}}{\arg\min} \; \|\mathcal{K}^{-1/2}u\|_{\mathcal{X}}^2 + \frac{1}{\beta^2}|F(\boldsymbol{\phi}(u)) - \mathbf{y}|^2. \tag{3}$$

This variational characterization of the MAP estimator is a natural generalization of the definition in finite-dimensional Euclidean spaces (Kaipio and Somersalo 2006, Sec. 3.4) and appears in Dashti et al. (2013); it is one of several possible generalizations to infinite-dimensional spaces as we outline in Sect. 3.1; the key technical issue arising in infinite dimensions is the need to seek a minimizer in the Cameron–Martin

✉ Bamdad Hosseini
bamdadh@uw.edu

Yifan Chen
yifan.chen@nyu.edu

Houman Owhadi
owhadi@caltech.edu

Andrew M. Stuart
stuart@caltech.edu

[1] Courant Institute of Mathematical Science, New York University, New York, NY, USA

[2] Department of Applied Mathematics, University of Washington, Seattle, WA, USA

[3] Department of Computing and Mathematical Sciences, California Instutute of Technology, Pasadena, CA, USA

$$\begin{tikzcd}
u_\beta^{\mathbf{y}} \arrow[r, leftrightarrow, "\text{Theorem } 5", "\text{MAP}"'] \arrow[d, "\text{Theorem } 9"', "\beta\to 0"] & \mu_\beta^{\mathbf{y}} \arrow[d, "\beta\to 0"', "\text{Theorem } 2"] \\
u_0^{\mathbf{y}} \arrow[r, leftrightarrow, "\text{Conditional MAP}", "\text{Theorem } 7"'] & \mu_0^{\mathbf{y}}
\end{tikzcd}$$

Fig. 1 Diagram relating small-noise limits of posteriors $\mu_\beta^{\mathbf{y}}$ and their MAP estimators $u_\beta^{\mathbf{y}}$ to their conditional counterparts $\mu_0^{\mathbf{y}}$ and $u_0^{\mathbf{y}}$

space of the Gaussian measure, $\mathcal{K}^{1/2}\mathcal{X}$. Proceeding in analogy with the definition of $\mu_0^{\mathbf{y}}$, we may take the $\beta = 0$ limit of (3) to formally obtain a variational characterization of the mode of $\mu_0^{\mathbf{y}}$ via

$$\begin{aligned} u_0^{\mathbf{y}} := \operatorname*{arg\,min}_{u\in\mathcal{K}^{1/2}\mathcal{X}} \quad & \|\mathcal{K}^{-1/2}u\|_{\mathcal{X}} \quad \text{subject to (s.t.)} \\ F\big(\boldsymbol{\phi}(u)\big) = \mathbf{y}. & \end{aligned} \tag{4}$$

The rigorous characterization of the aforementioned posterior and conditional measures, along with their modes, is the primary focus of this article. To this end, we make the following contributions:

1. We establish the existence of appropriate limits of $\mu_\beta^{\mathbf{y}}$ and $u_\beta^{\mathbf{y}}$ as $\beta \to 0$, making precise the natural candidates for $\mu_0^{\mathbf{y}}$ and $u_0^{\mathbf{y}}$ defined above, and characterizing $u_0^{\mathbf{y}}$ as an approximate definition of the mode of the conditional measure $\mu_0^{\mathbf{y}}$. These relationships, and the theorems making them explicit, are summarized in Fig. 1.
2. We show that for $\beta \geq 0$, the posterior measures $\mu_\beta^{\mathbf{y}}$ can be decomposed as the convolution of a conditional Gaussian measure and a non-Gaussian measure that is finite-dimensional; this result is given in Theorem 4. This decomposition is analogous to representer theorems for the MAP estimator $u_\beta^{\mathbf{y}}$, stating that the minimizers of (3) are effectively finite-dimensional; see Theorem 6.
3. We introduce a technique for generating samples from the posteriors $\mu_\beta^{\mathbf{y}}$ by decomposing them into a finite-dimensional component, which is sampled by standard algorithms such as Markov chain Monte Carlo (MCMC) or variational inference, and an infinite-dimensional Gaussian component, which may be simulated exactly using analytical properties of Gaussian measures; see Sect. 4. In particular, we show that the non-Gaussian component is amenable to approximation using a Laplace or Gauss–Newton-type approximation in settings where lots of observations are available, leading to efficient numerical algorithms in applications such as PDE solvers.

1.1 Motivating examples

In what follows we give three motivating examples for the study of posterior measures of the form (1), their MAP estimators and conditional counterparts. In Sect. 1.1.1 we study a Bayesian inverse problem; Sect. 1.1.2 is devoted to the GP-PDE methodology for solving forward problems in PDEs; and Sect. 1.1.3 combines the two preceding subsections to study the GP-PDE methodology for inverse problems.

1.1.1 Inverse problems

Fix any $\beta > 0$. Then the posterior measures $\mu_\beta^{\mathbf{y}}$ solve the Bayesian inverse problem (BIP) of finding the conditional distribution of $u \mid \mathbf{y}$ when $u \sim \mu$, $\boldsymbol{\zeta} \sim N(0, \beta^2 I)$ independent of u, and $\mathbf{y}$ (the data) is given by the model

$$y = G(u) + \boldsymbol{\zeta}, \tag{5a}$$

$$G := F \circ \boldsymbol{\phi}. \tag{5b}$$

We can employ Bayes' rule (Stuart 2010) to characterize the $\mu_\beta^{\mathbf{y}}$ via their Radon-Nikodym derivatives with respect to μ :

$$\begin{aligned} \frac{\mathrm{d}\mu_\beta^{\mathbf{y}}}{\mathrm{d}\mu}(u) &= \frac{1}{\omega_\beta(\mathbf{y})} \exp\left(-\frac{1}{2\beta^2}|F(\boldsymbol{\phi}(u)) - \mathbf{y}|^2\right), \\ \omega_\beta(\mathbf{y}) &:= \mathbb{E}_{u\sim\mu} \exp\left(-\frac{1}{2\beta^2}|F(\boldsymbol{\phi}(u)) - \mathbf{y}|^2\right). \end{aligned} \tag{6}$$

Example 1 To illustrate this setting we consider the inverse problem of finding u from $y = (y_1, \ldots, y_M)$ where

$$y_m = p(\overline{\mathbf{x}}_m) + \zeta_m, \quad m = 1, \ldots, M$$

where $\{\overline{\mathbf{x}}_m\}_{m=1}^M \subset (0, 1)^2$ are a set of observation locations and p solves the elliptic PDE

$$\begin{cases} -\nabla \cdot \big(\exp(u(\mathbf{x})\nabla p(\mathbf{x})\big) = f(\mathbf{x}), & \mathbf{x} \in (0, 1)^2, \\ p(\mathbf{x}) = 0, & \mathbf{x} \in \partial(0, 1)^2. \end{cases} \tag{7}$$

Assume that the prior μ on u is a centred Gaussian random field and let $\boldsymbol{\phi}(u)$ denote the vector of the first N coefficients in an expansion of u in the (ordered according to decreasing eigenvalues) eigenfunctions of the covariance operator of μ. As the basis of a practical computational approach to the inverse problem, now consider solving the elliptic PDE with u replaced by its truncation u^N to N terms in this eigenbasis:

$$\begin{cases} -\nabla \cdot \big(\exp(u^N(\mathbf{x}))\nabla p^N(\mathbf{x})\big) = f(\mathbf{x}), & \mathbf{x} \in (0, 1)^2, \\ p^N(\mathbf{x}) = 0, & \mathbf{x} \in \partial(0, 1)^2. \end{cases}$$

The resulting solution p^N of this PDE, evaluated at the points $\{\overline{\mathbf{x}}_m\}_{m=1}^M$, defines a map $F : \mathbb{R}^N \to \mathbb{R}^M$ from $\boldsymbol{\phi}(u)$

to $\{p^N(\overline{\mathbf{x}}_m)\}_{m=1}^M$. We then consider the inverse problem of recovering u from $\mathbf{y}$ where

$$y_m = p^N(\overline{\mathbf{x}}_m) + \zeta_m, \quad m = 1, \ldots, M.$$

This may be written in the general form (5) for $G(u) = F \circ \boldsymbol{\phi}(u)$.

A common task in solving inverse problems and uncertainty quantification (UQ) is to estimate various statistics of the above posterior measures. The MAP $u_\beta^{\mathbf{y}}$ is a popular choice among practitioners, which highlights the importance of understanding its properties. Alternatively, one may choose to generate samples from $\mu_\beta^{\mathbf{y}}$ directly using MCMC and then compute empirical statistics such as posterior mean and variance. In either case, our finite-dimensional representations of $\mu_\beta^{\mathbf{y}}$ and its MAP $u_\beta^{\mathbf{y}}$ offer a path towards efficient calculations. Moreover, it is natural to characterize solutions of these problems in the small noise limit as $\beta \to 0$ to understand the consistency of the underlying inverse problems and their limit behavior.

1.1.2 Solving PDEs with Gaussian processes

One of the core problems of the field of scientific machine learning (ML) is the design of novel algorithms for the solution of PDEs based on ML techniques. An example of such a methodology was introduced by the authors in Chen et al. (2021) where a Gaussian Process (GP) solver was developed for the numerical solution of nonlinear PDEs; henceforth referred to as GP-PDE. We briefly recall this methodology in the context of a specific example from Chen et al. (2021).

Example 2 Consider the PDE

$$\begin{cases} -\Delta u(\mathbf{x}) + \tau(u(\mathbf{x})) = f(\mathbf{x}), & \mathbf{x} \in (0,1)^2, \\ u(\mathbf{x}) = 0, & \mathbf{x} \in \partial(0,1)^2, \end{cases} \tag{8}$$

for $\tau : \mathbb{R} \to \mathbb{R}$ and $f : (0,1)^2 \to \mathbb{R}$. We assume the existence of a unique solution $u^\star$ in the strong/classical sense. Then GP-PDE aims to find a numerical approximation $u_\beta^{\mathbf{y}}$ to $u^\star$ by the following recipe: First, choose a set of M collocation points $\mathbf{x}_1, \ldots, \mathbf{x}_M \in [0,1]^2$, with J in the interior and $M - J$ on the boundary, ordered so that $\mathbf{x}_1, \ldots, \mathbf{x}_J \in (0,1)^2$ while $\mathbf{x}_{J+1}, \ldots, \mathbf{x}_M \in \partial(0,1)^2$. Then define $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)$ for $N = J + M$ with the ϕ_m defined as

$$\begin{aligned} \phi_m(u) &= u(\mathbf{x}_m), && \text{for } m = 1, \ldots, M, \\ \phi_m(u) &= \Delta u(\mathbf{x}_{m-M}), && \text{for } m = M+1, \ldots, M+J. \end{aligned} \tag{9}$$

and the nonlinear function $F : \mathbb{R}^{M+J} \to \mathbb{R}^M$ defined row-wise as

$$F_m(\mathbf{z}) = \begin{cases} -z_{m+M} + \tau(z_j), & 1 \le m \le J, \\ z_m, & J+1 \le m \le M, \end{cases} \tag{10}$$

Furthermore, define the vector $\mathbf{y} \in \mathbb{R}^M$ defined element-wise as

$$y_m = \begin{cases} f(\mathbf{x}_m), & 1 \le m \le J, \\ 0, & J+1 \le m \le M. \end{cases} \tag{11}$$

With these definitions, we may now consider our PDE solver as an instance of (5).

Recalling the discussion earlier in Sect. 1 suggests that the resulting minimizer identifies the mode of an underlying posterior measure $\mu_\beta^{\mathbf{y}}$. This observation was discussed informally in Chen et al. (2021), in the setting of the GP-PDE methodology, and our Theorem 5 establishes this connection rigorously. The GP-PDE methodology relies on a representer theorem (see also Smola and Schölkopf 1998) that identifies the solution of (3), in the GP-PDE context, via a finite-dimensional optimization problem. In Chen et al. (2021) it is argued that the natural $\beta \to 0$ limit of (3), namely (4), can also be solved with a representer theorem. Theorem 4 and Theorem 6 can be viewed as establishing Bayesian analogs of these results from Chen et al. (2021), where the exposition is primarily focused on kernel methods.

1.1.3 Simultaneous solution of PDEs and inverse problems

For our final motivating example we consider a methodology introduced in Chen et al. (2021), for the solution of PDE based inverse problems, using the GP-PDE methodology, in which we seek both the solution of the PDE and an unknown coefficient at the same time. We outline how the formulation fromn that paper is also encompassed by the theoretical framework in this paper.

Example 3 Consider the elliptic PDE (7) in a 1D setting and with slightly modified notation

$$\begin{cases} -\partial_x \cdot \big(\exp(a(x))\partial_x p(x)\big) = f(x), & x \in (0,1), \\ p(0) = p(1) = 0, \end{cases}$$

and let $u = (a, p)$ be the parameter we wish to infer. Consider the inverse problem of finding u from data/observations $\mathbf{y}^{\text{obs}} \in \mathbb{R}^{M_{\text{obs}}}$ where

$$y_m^{\text{obs}} = p(\overline{x}_m) + \zeta_m^{\text{obs}}, \quad m = 1, \ldots, M_{\text{obs}}, \tag{12}$$

and the $\{\overline{x}_m\}_{m=1}^{M_{\text{obs}}} \subset (0,1)$ are once again our observation locations as in Example 1. Now consider a mesh

$\{x_m\}_{m=1}^{M_{\text{mesh}}} \subset (0,1)$ (ignoring boundary points for brevity) and discretize the PDE using the GP-PDE approach of Example 2. Consider the maps

$$\phi_m^{\text{obs}}(u) = p(\overline{x}_j), \qquad m = 1,\dots,M_{\text{obs}},$$
$$\phi_m^{p}(u) = \begin{cases} p(x_m) & m = 1,\dots,M_{\text{mesh}},\\ \partial_x p(x_m) & m = M_{\text{mesh}+1},\dots,2M_{\text{mesh}},\\ \partial_{xx} p(x_m) & m = 2M_{\text{mesh}+1},\dots,3M_{\text{mesh}},\end{cases}$$
$$\phi_m^{a}(u) = \begin{cases} a(x_m) & m = 1,\dots,M_{\text{mesh}},\\ \partial_x a(x_m) & m = M_{\text{mesh}+1},\dots,2M_{\text{mesh}},\end{cases}$$

along with the (artifically) noisy PDE data

$$y_m^{\text{mesh}} = f(x_m) + \zeta_m^{\text{mesh}}, \qquad m = 1,\dots,M_{\text{mesh}}.$$

Then the PDE constrained to the mesh points $\{x_m\}_{m=1}^{M_{\text{mesh}}}$ defines the relationship

$$y_m^{\text{mesh}} = -\exp(\phi_m^a(u)) \left[\phi^p_{m+2M_{\text{mesh}}}(u) + \phi^a_{m+M_{\text{mesh}}}(u)\cdot\phi^p_{m+M_{\text{mesh}}}(u)\right] + \zeta_m^{\text{mesh}}.$$

This model, together with (12), defines a nonlinear map $F : \mathbb{R}^N \to \mathbb{R}^M$ for $N = M_{\text{obs}} + 5M_{\text{mesh}}$ and $M = M_{\text{obs}} + M_{\text{mesh}}$ so that

$$\mathbf{y} = F(\boldsymbol{\phi}(u)) + \boldsymbol{\zeta}, \quad \text{where} \quad \mathbf{y} = \begin{pmatrix}\mathbf{y}^{\text{obs}}\\ \mathbf{y}^{\text{mesh}}\end{pmatrix},$$
$$\boldsymbol{\phi}(u) = \begin{pmatrix}\boldsymbol{\phi}^{\text{obs}}(u)\\ \boldsymbol{\phi}^{p}(u)\\ \boldsymbol{\phi}^{a}(u)\end{pmatrix}, \quad \boldsymbol{\zeta} = \begin{pmatrix}\boldsymbol{\zeta}^{\text{obs}}\\ \boldsymbol{\zeta}^{\text{mesh}}\end{pmatrix}.$$

Here we used bold letters to denote the vectorized versions of the maps and variables defined above. We place a GP prior on u, implying a joint GP prior on a and p. Furthermore we assume that this prior is chosen so that p satisfies the boundary conditions, justifying the use of collocation points only in the interior of $(0, 1)$. This allows us to cast the problem of simultaneous recovery of both p and a in the desired form (5).

1.2 Literature review

Below we give an overview of published literature that is relevant to our work. We do so with a particular focus on the theory of Bayesian inverse problems, GPs, and probabilistic methods in numerical analysis.

1.2.1 Bayesian inverse problems and MAP estimators

Bayesian inference (Gelman et al. 1995) is a cornerstone of modern statistics and data science. When applied in the context of infinite-dimensional or functional inference the methodology is best known under the term Bayesian inverse problems (Franklin 1970; Tarantola 2005; Kaipio and Somersalo 2006; Stuart 2010). Over the past decade, the algorithmic development and theoretical analysis for Bayesian inverse problems have become mature areas of research. Here, Bayesian inference with Gaussian prior measures is by far the most common setup for both algorithms and theoretical analysis. The overwhelming majority of function space MCMC algorithms (Tierney 1998; Beskos et al. 2011; Cotter et al. 2013; Cui et al. 2016; Beskos et al. 2017) are developed specifically for Gaussian priors; see Hosseini (2019) and references within for examples of algorithms for non-Gaussian priors. The well-posedness theory of Bayesian inverse problems was originally developed for the case of Gaussian (or sub-Gaussian) priors (Cotter et al. 2009; Stuart 2010) and was later extended to the non-Gaussian setting (Dashti et al. 2012; Hosseini and Nigam 2017; Hosseini 2017; Sullivan 2017; Sprungk 2020; Latz 2020) but the case of Gaussian priors remains most applicable as it allows for the widest range of nonlinear forward maps. From this perspective, this article makes important theoretical contributions towards the understanding and characterization of Bayesian posteriors under nonlinear observation models with Gaussian priors. Most importantly, our second contribution enables the use of finite-dimensional MCMC algorithms for nonlinear observation models without the need for direct discretization of the inverse problem.

Variational methods are also an important family of algorithms for the solution of Bayesian inverse problems. Perhaps the most common task here is computing a MAP estimator. Defining a MAP estimator in the function space setting is highly non-trivial. Several definitions, and resulting analyses, of modes of measures on infinite-dimensional spaces exist (Agapiou et al. 2018; Ayanbayev et al. 2021a, b; Clason et al. 2019; Dashti et al. 2013; Helin and Burger 2015) where the choice of the notion of the mode is closely tied to the choice of the prior measure. Once again, the Gaussian priors lead to the most natural definition of a MAP estimator (Dashti et al. 2013; Ikeda and Watanabe 2014) which is the same one we shall use to define $u_\beta^{\mathbf{y}}$ in (3). However, to our knowledge, a notion of a conditional mode, i.e., a precise definition of $u_0^{\mathbf{y}}$ as in (4) has not been studied before and constitutes one of our main contributions.

1.2.2 Gaussian measures and processes

The general theory of Gaussian measures in infinite-dimensional settings is a classic subject in probability theory and the theory of stochastic differential equations. We refer the reader to the work of Bogachev (1998) for the detailed treatment of this subject on topological vector spaces and Maniglia and Rhandi (2004) and Janson (1997) for the case of Hilbert spaces.

GPs, as a special instance of Gaussian measures, and, by extension, reproducing kernel Hilbert space (RKHS) methods (Kanagawa et al. 2018; van der Vaart et al. 2008) and support vector machines (Smola and Schölkopf 1998), have a long history in approximation theory (Wendland 2004), statistical modeling and inference (Giné and Nickl 2021), inverse problems (Cressie 1990), and machine learning (Rasmussen and Williams 2007; Smola and Schölkopf 1998). While in this article we mainly focus on solving differential equations with GPs as an application of our theory (Särkkä 2011; Owhadi 2015; Chkrebtii et al. 2016; Cockayne et al. 2017; Raissi et al. 2018; Swiler et al. 2020; Chen et al. 2021; Wang et al. 2021) (see also Sect. 1.2.3 below), GPs have wide applications in many modern areas of scientific computing and machine learning such as deep GPs (Damianou and Lawrence 2013; Dunlop et al. 2018; Jakkala 2021; Dutordoir et al. 2021; Owhadi 2023) as a model for deep learning, vector-valued GPs for operator learning (Batlle et al. 2024) and generative modeling (Murray et al. 2008; Casale et al. 2018; Fortuin et al. 2020; Pandey et al. 2024), and graphical models for semi-supervised learning (Bertozzi et al. 2018).

The reasons for this widespread use of GPs are their many desirable theoretical properties that lead to efficient algorithms. Perhaps the most useful are the facts that (1) GPs are completely identified by their mean and covariance operators; (2) GPs are closed under affine transformations; and (3) GPs conditioned on affine observations are also GPs that can be identified analytically; see Lemma 2. However, GPs conditioned on nonlinear observations are in general no longer GPs and cannot be identified analytically. Due to this fact, such conditional measures are often characterized computationally using MCMC (Cotter et al. 2013; Beskos et al. 2017; Robert and Casella 1999), or variational inference (Blei et al. 2017; Pinski et al. 2015). Such conditional measures are readily common in the field of inverse problems but they are increasingly common in modern machine learning applications mentioned in the previous paragraph as well. To this end, one of the main contributions of this article is to reveal the additional structure of conditioned GPs in the nonlinear setting that can be further leveraged by both MCMC and variational algorithms to further improve the accuracy and complexity of algorithms.

1.2.3 The intersection of numerical analysis and probability

As discussed in Owhadi et al. (2019), the fields of numerical approximation and statistical inference, traditionally viewed as distinct, are in fact deeply connected through their common purpose of making estimations with partial information (Owhadi and Scovel 2019, Chap. 20). This shared purpose has recently stimulated a growing interest in learning approaches to solving PDEs (Owhadi 2015; Raissi et al. 2017) and in the merging of numerical errors with modeling errors and UQ (Hennig et al. 2015). Although this trend may seem novel, the synergy between numerical approximation and statistical inference has historical roots, dating back to Poincaré's lectures on Probability Theory (Poincaré 1896), and extending through the pioneering work of Sul'din (1959), Palasti and Renyi (1956), Sard (1963), Kimeldorf and Wahba (1970), and Larkin (1972). While these studies initially "attracted little attention among numerical analysts" (Larkin 1972), they were revived in the fields of Information Based Complexity (Traub et al. 1988), Bayesian Numerical Analysis (Diaconis 1988), and more recently in Probabilistic Numerics (Hennig et al. 2015; Cockayne et al. 2019). This connection between inference and numerical approximation is also central to Bayesian/decision-theoretic approaches to solving ODEs (Skilling 1992) and PDEs (Owhadi 2015), in identifying operator adapted wavelets (Owhadi and Scovel 2019) and designing fast solvers for kernel matrices (Schäfer et al. 2021a, b; Chen et al. 2024), and in parameter estimation (Chen et al. 2021).

Another connection between numerical approximation and statistical inference arises in the framework of optimal recovery introduced by Micchelli and Rivlin (1977), Owhadi and Scovel (2019) and its connection to Bayesian inference and GP regression through decision and game theory (Wald 1945; von Neumann 1928). Optimal recovery was initially used for solving linear PDEs (Harder and Desmarais 1972; Duchon 1977; Owhadi 2015), but was extended to nonlinear PDEs in Chen et al. (2021) and to general computational graph completion/discovery problems in Owhadi (2022) and Bourdais et al. (2024) where the connection between optimal recovery and the GP perspective on solving PDEs is made explicit. Finally, we mention the recent papers (Long et al. 2022) and Vadeboncoeur et al. (2023) where numerical errors are analyzed as Bayesian posterior measures. Further details about the connection between optimal recovery, decision theory, and GPs can be found in Section A.

1.3 Notation and preliminaries

We use $|\cdot|$ to denote the finite-dimensional Euclidean norm. Since $\mathcal{X}$ is Hilbertian, all elements of the dual space $\mathcal{X}^*$ may be Reisz-represented by elements of $\mathcal{X}$ itself; if $\psi \in \mathcal{X}^*$ then we write $\psi^* \in \mathcal{X}$ for its Reisz-representer. Likewise, if $\theta \in \mathcal{X}$ then we write θ^* for the dual element it Reisz-represents. Throughout we will write $B_r(u) \subset \mathcal{X}$ to denote the ball of radius $r \geq 0$ centered at u.

We give a brief summary of the notation from Gaussian measure theory needed for this paper; we follow Hairer (2009, Section 3) and the reader seeking more details may consult (Bogachev 1998). We say that a measure $\mu \in \mathbb{P}(\mathcal{X})$ is a Gaussian measure (process) on $\mathcal{X}$ if and only if for any

$\psi \in \mathcal{X}^*$, the pushforward measure $\mu \circ \psi^{-1} =: \psi_\sharp\mu \in \mathbb{P}(\mathbb{R})$ is a Gaussian measure. Henceforth we write $\mu = N(m, \mathcal{K})$ to denote a Gaussian measure in $\mathbb{P}(\mathcal{X})$ with mean $m \in \mathcal{X}$ and covariance operator $\mathcal{K} : \mathcal{X} \to \mathcal{X}$. Whenever $m = 0$ we say μ is a centered Gaussian measure. Note that $\mathcal{K}$ is necessarily compact, indeed it is trace-class, and we may define the symmetric operator $\mathcal{K}^{\frac{1}{2}}$ by spectral calculus; operator $\mathcal{K}^{-\frac{1}{2}}$ can also be densely defined on $\mathcal{K}^{1/2}\mathcal{X}$. Indeed, associated to a centered Gaussian measure $\mu = N(0, \mathcal{K})$, we identify its Cameron–Martin space $\mathcal{H}(\mu) := \mathcal{K}^{1/2}\mathcal{X}$ which is Hilbertian with corresponding inner product

$$\langle u, v\rangle_{\mathcal{H}(\mu)} := \langle \mathcal{K}^{-1/2}u, \mathcal{K}^{-1/2}v\rangle_{\mathcal{X}}, \qquad \forall u, v \in \mathcal{H}(\mu);$$

we write $\|\cdot\|_{\mathcal{H}(\mu)}$ for the induced norm. The Cameron–Martin space is a Reproducing Kernel Hilbert Space (RKHS) if pointwise evaluation is defined in $\mathcal{H}(\mu)$; the kernel of the RKHS is the covariance function associated with μ; see van der Vaart et al. (2008, Sec 2.3). For any infinite-dimensional Gaussian measure, it is always true that $\mu\big(\mathcal{H}(\mu)\big) = 0$; in contrast, by construction, $\mu(\mathcal{X}) = 1$. Furthermore $\mathcal{H}(\mu)$ is compactly embedded into $\mathcal{X}$.

We will also review some preliminary definitions and results for conditional measures identified via a mapping as these ideas are central to our study. Our reference for this material is Bogachev (2007, Sec. 10.4). Let $\mathcal{X}, \mathcal{Y}$ be separable Hilbert spaces with $\mathcal{B}(\mathcal{X}), \mathcal{B}(\mathcal{Y})$ denoting their respective Borel σ-algebras together with a measure $\nu \in \mathbb{P}(\mathcal{X})$. Consider a $(\mathcal{B}(\mathcal{X}), \mathcal{B}(\mathcal{Y}))$-measurable map $T : \mathcal{X} \to \mathcal{Y}$. We then have the following definition of a system of conditional measures of ν generated by the mapping T:

Definition 1 A function $(A, y) \mapsto \nu^y(A)$ is a *system of conditional measures* for ν with respect to the map T if:

(a) for every fixed $y \in \mathcal{Y}$ the function $\nu^y \in \mathbb{P}(\mathcal{X})$;
(b) for every fixed $A \in \mathcal{B}(\mathcal{X})$ the function $y \mapsto \nu^y(A)$ is measurable with respect to $\mathcal{B}(\mathcal{Y})$ and $T_\sharp\nu$-integrable;
(c) for all $A \in \mathcal{B}(\mathcal{X})$ and $E \in \mathcal{B}(\mathcal{Y})$ it holds that

$$\nu(A \cap T^{-1}(E)) = \int_E \nu^y(A) T_\sharp\nu(\mathrm{d}y).$$

We also use the alternative notation $\nu(\mathrm{d}\xi \mid T(\xi) = y)$ to denote the system of conditional measures in the above definition; this notation succintly captures what is behind the definition. The next result is a consequence of Bogachev (2007, Lem. 10.4.3 and Cor. 10.4.10):

Proposition 1 *Consider the above setting and suppose $T : \mathcal{X} \to \mathcal{Y}$ is ν-measurable. Then it holds that:*

(a) there exists a system of conditional measures ν^y for ν with respect to the map T;
(b) the conditional measures ν^y are essentially unique, i.e., there exists a set $Z \in \mathcal{B}(\mathcal{Y})$ so that $T_\sharp\nu(Z) = 0$ and the ν^y are unique for all $y \in \mathcal{Y} \setminus Z$ (i.e., essentially unique);
(c) for $T_\sharp\nu$-a.e. y the measures ν^y concentrate on $T^{-1}(y)$, i.e., $\nu^y(\mathcal{X} \setminus T^{-1}(y)) = 0$.

Remark 1 In most of this paper we consider $T = G$ where G is defined in (5b); thus $\mathcal{Y}$ is finite-dimensional. However we do make some theoretical observations and remarks about the more general setting, which includes infinite-dimensional $\mathcal{Y}$.

1.4 Outline

In Sect. 2 we analyze the posterior measure, and limits as $\beta \to 0$. Section 3 is devoted to the modes, or MAP estimators, associated with the family of posterior measures, and their $\beta \to 0$ limit. In Sect. 4 we discuss algorithms to sample the posterior measures, exploiting the special structure of the observations and the decomposition of posterior measures. Finally, we give our conclusions in Sect. 5. Proofs of various technical results are collected in the appendix.

2 Analysis of posterior and conditional measures

In this section we study the posterior measures $\mu^{\mathbf{y}}_\beta$, and the conditionals $\mu^{\mathbf{y}}_0$. In Sect. 2.1 we prove a form of convergence, suitably defined, of $\mu^{\mathbf{y}}_\beta$ to $\mu^{\mathbf{y}}_0$. Section 2.2 studies decompositions of the conditionals and posteriors respectively into the convolution of finite-dimensional non-Gaussians with an infinite-dimensional Gaussian part.

2.1 Convergence of posterior measures to conditionals

In this subsection we show that in the limit $\beta \to 0$ the posterior measures $\mu^{\mathbf{y}}_\beta$ converge to the conditional measures $\mu^{\mathbf{y}}_0$ in an appropriate sense. We start by identifying conditions that ensure that the family of posterior measures $\mu^{\mathbf{y}}_\beta$ are well-defined for $\beta > 0$. To this end consider the set-up of Sect. 1.1.1. We formulate the BIP of determining $u|\mathbf{y}$, from (5), under the following assumptions:

Assumption 1 Assume that $u \sim \mu$, $\boldsymbol{\zeta} \sim \pi_\beta := N(0, \beta^2 I)$ and $u, \boldsymbol{\zeta}$ are independent. Assume further that the map $F : \mathbb{R}^N \to \mathbb{R}^M$ is finite at some point $\mathbf{z}' \in \mathbb{R}^N$ and that F is locally Lipschitz, i.e., for every $r > 0$ there exists $L(r) > 0$ such that

$$\|F(\mathbf{z}_1) - F(\mathbf{z}_2)\|_2 \le L(r)\|\mathbf{z}_1 - \mathbf{z}_2\|_2 \qquad \forall \mathbf{z}_1, \mathbf{z}_2 \in B_r(0).$$

Note that, since F is finite at one point $\mathbf{z}' \in \mathbb{R}^N$ then this assumption implies that F is locally bounded from above, i.e., for every $r > 0$ there exists $M(r) > 0$ such that

$$\|F(\mathbf{z})\|_2 \le M(r) \qquad \forall \mathbf{z} \in B_r(0).$$

Recalling definition (5b), we have:

Lemma 1 *Let Assumption 1 hold and consider the BIP for $u|\mathbf{y}$ defined via (5). Then, for every $\beta > 0$, the posterior distribution $\mu^{\mathbf{y}}_{\beta}$ is given by (6). Furthermore, the map G defines a unique (up to equivalence) system of conditional measures of μ, denoted $\mu^{\mathbf{y}}_0 := \mu(\mathrm{d}u \mid G(u) = \mathbf{y})$.*

Proof (Dashti and Stuart 2017, Thm. 10) establishes the result for $\beta > 0$. The result for $\beta = 0$ follows from Proposition 1, using the fact that, under the stated assumptions on F, G is continuous and hence μ-measurable as a map from $\mathcal{X}$ into $\mathbb{R}^M$. □

We now consider the limit of the measures $\mu^{\mathbf{y}}_{\beta}$ as $\beta \to 0$. It is convenient to express the result in terms of the joint measure $\mathbb{P}(\mathrm{d}u, \mathrm{d}\mathbf{y})$. This joint measure may be factored as $\mathbb{P}(\mathrm{d}u|\mathbf{y})\mathbb{P}(\mathrm{d}\mathbf{y})$ or as $\mathbb{P}(\mathrm{d}\mathbf{y}|u)\mathbb{P}(\mathrm{d}u)$. The latter necessarily involves a Dirac mass when $\beta = 0$ and is not convenient to work with; we hence use the former factorization.

Theorem 2 *Let Assumption 1 hold. Then the measures $\mu^{\mathbf{y}}_{\beta}(\mathrm{d}u)G_\sharp\mu * \pi_\beta(\mathrm{d}\mathbf{y})$ converge weakly to $\mu^{\mathbf{y}}_0(\mathrm{d}u)G_\sharp\mu(\mathrm{d}\mathbf{y})$ as $\beta \to 0$. That is, $\forall f \in C_b(\mathcal{X} \times \mathcal{Y})$*

$$\lim_{\beta\to 0}\int_{\mathcal{Y}}\int_{\mathcal{X}} f(u,\mathbf{y})\mu^{\mathbf{y}}_{\beta}(\mathrm{d}u)G_\sharp\mu * \pi_\beta(\mathrm{d}\mathbf{y})$$
$$= \int_{\mathcal{Y}}\int_{\mathcal{X}} f(u,\mathbf{y})\mu^{\mathbf{y}}_0(\mathrm{d}u)G_\sharp\mu(\mathrm{d}\mathbf{y}).$$

Proof It will be helpful to extend π_β to a measure on $\mathcal{X} \times \mathbb{R}^M$ by defining $\pi'_\beta := \delta_0 \times N(0, \beta^2 I)$. With this notation we note that

$$\mu^{\mathbf{y}}_{\beta}(\mathrm{d}u)G_\sharp\mu * \pi_\beta(\mathrm{d}\mathbf{y}) = (Id \times G)_\sharp\mu * \pi'_\beta(\mathrm{d}u, \mathrm{d}\mathbf{y}),$$

and that

$$\mu^{\mathbf{y}}_0(\mathrm{d}u)G_\sharp\mu(\mathrm{d}\mathbf{y}) = (Id \times G)_\sharp\mu(\mathrm{d}u, \mathrm{d}\mathbf{y}).$$

The desired result thus reduces to proving that, $\forall f \in C_b(\mathcal{X} \times \mathcal{Y})$,

$$\lim_{\beta\to 0}\int_{\mathcal{Y}}\int_{\mathcal{X}} f(u,\mathbf{y})(Id \times G)_\sharp\mu * \pi'_\beta(\mathrm{d}u, \mathrm{d}\mathbf{y})$$
$$= \int_{\mathcal{Y}}\int_{\mathcal{X}} f(u,\mathbf{y})(Id \times G)_\sharp\mu(\mathrm{d}u, \mathrm{d}\mathbf{y}).$$

Noting that π'_β converges weakly to a Dirac at the origin in $\mathcal{X} \times \mathbb{R}^M$ as $\beta \to 0$ gives the desired result. □

Remark 2 We note that the above result can be interpreted as an "almost" weak convergence result for the posterior measures $\mu^{\mathbf{y}}_{\beta}$. More precisely, take $f(u, \mathbf{y}) = g(u)h(\mathbf{y})$ where $g \in C_b(\mathcal{X})$ and $h \in C_b(\mathcal{X})$ is a continuous approximation to $\frac{1}{\mu(B_\epsilon(\mathbf{y}'))}\mathbf{1}_{B_\epsilon(\mathbf{y}')}$ for some fixed $\mathbf{y}' \in \mathbb{R}^M$ and $\epsilon > 0$. Then Theorem 2 tells us that the expectation of g with respect to $\mu^{\mathbf{y}'}_{\beta}$ converges to the conditional expectation with respect to $\mu^{\mathbf{y}'}_0$ so long as we average $\mathbf{y}$ in a ball with arbitrarily small but positive radius ϵ around $\mathbf{y}'$.

2.2 Finite-dimensional representation of conditional and posterior measures

The finite-dimensional representation of the conditionals is analogous to the family of representer theorems for kernel methods (Smola and Schölkopf 1998, Sec. 4.2), generalized to the probabilistic setting, and is stated as Theorem 3 below. To understand this proposition, we first recall a classic lemma pertaining to conditioning Gaussian measures on direct sums of Hilbert spaces, and a corollary thereof.

Let $\mathcal{X} = \mathcal{X}_1 \oplus \mathcal{X}_2$ where $\mathcal{X}_1, \mathcal{X}_2$ are separable Hilbert spaces and let μ be a Gaussian measure on $\mathcal{X}$ and $\Pi_i : \mathcal{X} \to \mathcal{X}_i$ denote the natural projection onto $\mathcal{X}_i$. Then by Stuart (2010, Lem. 4.3) and Owhadi and Scovel (2018) (see also Owhadi and Scovel 2019, Chap. 17.8) we have that the conditional measure of μ with respect to the maps Π_i is also Gaussian and can be characterized explicitly:

Lemma 2 *Let $\mu = N(m, \mathcal{K}) \in \mathbb{P}(\mathcal{X})$ where $\mathcal{X} = \mathcal{X}_1 \oplus \mathcal{X}_2$ as above. Write $m = (m_1, m_2)$ for the mean and let $\mathcal{K}$ be the positive definite covariance operator and define $\mathcal{K}_{ij} = \Pi_i\mathcal{K}\Pi_j^*$. Write μ^{x_1} for the system of conditional measures of μ with respect to Π_1. Then for $(\Pi_1)_\sharp\mu$-a.e. $x_1 \in \mathcal{X}_1$ it holds that $\mu^{x_1} = \delta_{x_1} \otimes N(m^{x_1}, \mathcal{K}_{2|1})$ where $N(m^{x_1}, \mathcal{K}_{2|1})$ is a Gaussian measure on $\mathcal{X}_2$ with mean $m^{x_1} = m_2 + \mathcal{K}_{21}\mathcal{K}_{11}^{-1}(x_1 - m_1)$ and covariance operator $\mathcal{K}_{2|1} = \mathcal{K}_{22} - \mathcal{K}_{21}\mathcal{K}_{11}^{-1}\mathcal{K}_{12}$.*

We will now present a corollary of the above lemma that characterizes the conditionals of Gaussian measures under general linear observations. Before presenting that result we introduce some notation which is used extensively throughout the rest of the article. Given a trace class covariance operator $\mathcal{K}$ on $\mathcal{X}$ and a vector of dual elements $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N) \in (\mathcal{X}^\star)^N$ we define the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N) \in \mathcal{X}^N$ and the symmetric matrix $\Theta \in \mathbb{R}^{N\times N}$ defined as

$$\theta_i := \mathcal{K}\phi_i^*, \qquad \text{and} \qquad \Theta_{ij} := \phi_i(\mathcal{K}\phi_j^*). \tag{13}$$

In the parlance of kernel methods the θ_i are referred to as features/ representers of the ϕ_i while Θ is the kernel matrix. Assuming Θ is invertible, and given any vector $\mathbf{z} \in \mathbb{R}^N$, we further define the conditional mean $u^{\mathbf{z}} \in \mathcal{X}$ and covariance operator $\mathcal{K}^{\boldsymbol{\phi}}$ as

$$u^{\mathbf{z}} = \boldsymbol{\theta}^\top \Theta^{-1}\mathbf{z} := \sum_{i,j=1}^{N} (\Theta^{-1})_{ij}\theta_i z_j, \tag{14a}$$

$$\mathcal{K}^{\boldsymbol{\phi}} = \mathcal{K} - \boldsymbol{\theta}^\top \Theta^{-1}\boldsymbol{\theta}^* := \mathcal{K} - \sum_{i,j=1}^{N} (\Theta^{-1})_{ij}\theta_i\theta_j^*. \tag{14b}$$

With this notation at hand we then obtain the following corollary that can be proven by applying Lemma 2 to the measure $\mu \otimes \boldsymbol{\phi}_\sharp\mu$ on the product space $\mathcal{X} \times \mathbb{R}^N$, using the fact that $\boldsymbol{\phi}_\sharp\mu = N(0, \Theta)$ and that the tensor product of two Gaussian measures is also Gaussian:

Corollary 1 *Suppose $\mu = N(0, \mathcal{K})$ with $\mathcal{K}$ a trace-class covariance operator on $\mathcal{X}$. Consider the map $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N) \in (\mathcal{X}^*)^N$ and the system of conditional measures $\mu^{\mathbf{z}} \equiv \mu\big(\mathrm{d}u \mid \boldsymbol{\phi}(u) = \mathbf{z}\big)$. If Θ is invertible then $\mu^{\mathbf{z}} = N(u^{\mathbf{z}}, \mathcal{K}^{\boldsymbol{\phi}})$.*

Remark 3 We often consider the vector of functions $\boldsymbol{\varphi} := \Theta^{-1}\boldsymbol{\theta} \in \mathcal{X}^N$; the entries φ_i of $\boldsymbol{\varphi}$ are referred to as the *Gamblets* in the parlance of Owhadi and Scovel (2019). We can then write $u^{\mathbf{z}} = \boldsymbol{\varphi}^\top\mathbf{z}$ and refer to $\boldsymbol{\varphi}^\top : \mathbb{R}^N \to \mathcal{X}$ as the Gamblet reconstruction map. In the rest of the article it is also useful to define the measures

$$\mu^{\boldsymbol{\phi}} := N(0, \mathcal{K}^{\boldsymbol{\phi}}), \quad \text{and} \quad \eta := N(0, \Theta),$$

where the former is conditional measure $\mu\big(\mathrm{d}u \mid \boldsymbol{\phi}(u) = 0\big)$ and the latter is the distribution of $\phi_\sharp\mu$.

Now notice that the measure μ can be reconstructed as the convolution $\mu = \mu^{\boldsymbol{\phi}} * \boldsymbol{\varphi}_\sharp^\top\eta$. Crucial to this fact is that the $u^{\mathbf{z}}$ depends on $\mathbf{z}$ whereas $\mathcal{K}^{\boldsymbol{\phi}}$ does not, it only depends on the linear map $\boldsymbol{\phi}$ and not the vector $\mathbf{z}$, and that $\mathbf{z} \sim \eta$ under μ.

Building on this remark we have the following useful factorization of the conditional $\mu_0^{\mathbf{y}}$ which is one of our main theoretical contributions.

Theorem 3 *Suppose Assumption 1 holds and that Corollary 1 is satisfied. Then $\mu = \mu^{\boldsymbol{\phi}} * \boldsymbol{\varphi}_\sharp^\top\eta$ and $\mu_0^{\mathbf{y}} = \mu^{\boldsymbol{\phi}} * \boldsymbol{\varphi}_\sharp^\top\eta_0^{\mathbf{y}}$, where $\eta_0^{\mathbf{y}} := \eta(\mathrm{d}\mathbf{z} \mid F(\mathbf{z}) = \mathbf{y})$ is the system of conditionals of $\eta = N(0, \Theta)$ with respect to the map F.*

Proof Let $u \sim \mu$. Conditional on $\boldsymbol{\phi}(u) = \mathbf{z}$ the distribution of u is $N(u^{\mathbf{z}}, \mathcal{K}^{\boldsymbol{\phi}})$, by Corollary 1. In the absence of observations, $\mathbf{z} \sim \eta$ and then $u^{\mathbf{z}} \sim \boldsymbol{\varphi}_\sharp^\top\eta$. When conditioned on $F(\mathbf{z}) = \mathbf{y}$, however, we obtain $\mathbf{z} \sim \eta_0^{\mathbf{y}}$ and $u^{\mathbf{z}} \sim \boldsymbol{\varphi}_\sharp^\top\eta_0^{\mathbf{y}}$. Because $\mu^{\boldsymbol{\phi}}$ is independent of $\mathbf{z}$ the two results follow by the properties of convolutions of measures. □

We may now generalize Theorem 3 to the setting $\beta > 0$; we show that the posterior measures in (6) can be decomposed as the convolution of a finite-dimensional (in general) non-Gaussian measure with an independent centered Gaussian measure. The result may also be viewed as a generalization of Corollary 1 to nonlinear measurements. This theorem is the second major theoretical contribution of our work.

Theorem 4 *Suppose Assumption 1 holds and that Corollary 1 is satisfied. Let $\mu_\beta^{\mathbf{y}}$ be as in (6) and let Λ denote the Lebesgue measure. Then $\mu_\beta^{\mathbf{y}} = \mu^{\boldsymbol{\phi}} * \boldsymbol{\varphi}_\sharp^\top\eta_\beta^{\mathbf{y}}$ where $\eta_\beta^{\mathbf{y}} \in \mathbb{P}(\mathbb{R}^N)$ has Lebesgue density*

$$\frac{\mathrm{d}\eta_\beta^{\mathbf{y}}}{\mathrm{d}\Lambda}(\mathbf{z}) = \frac{1}{\varpi_\beta(\mathbf{y})}\exp\left(-\frac{1}{2\beta^2}|F(\mathbf{z}) - \mathbf{y}|^2 - \frac{1}{2}\mathbf{z}^\top\Theta^{-1}\mathbf{z}\right),$$
$$\varpi_\beta(\mathbf{y}) := \int_{\mathbb{R}^N}\exp\left(-\frac{1}{2\beta^2}|F(\mathbf{z}) - \mathbf{y}|^2 - \frac{1}{2}\mathbf{z}^\top\Theta^{-1}\mathbf{z}\right)\Lambda(\mathrm{d}\mathbf{z}).$$

Proof Recall π_β from Assumption 1 and, for any measure π on a vector space, let $\pi(\mathrm{d}\mathbf{y} + \mathbf{m})$ denote the shift of π by a vector $\mathbf{m}$. Consider the measure $\nu(\mathrm{d}u, \mathrm{d}\mathbf{y}) := \mu(\mathrm{d}u)\pi_\beta(\mathrm{d}\mathbf{y} + F(\boldsymbol{\phi}(u)))$. By Bayes' rule the posterior measures $\mu_\beta^{\mathbf{y}}(\mathrm{d}u) \otimes \delta_{\mathbf{y}}(\mathrm{d}y)$ are precisely the conditionals of ν with respect to the projection $\Pi : \mathcal{X} \times \mathbb{R}^M \to \mathbb{R}^M$, i.e.,

$$\nu(A \cap \Pi^{-1}(E)) = \int_E \big(\mu_\beta^{\mathbf{y}} \otimes \delta_{\mathbf{y}}\big)(A)\big(F_\sharp(\boldsymbol{\phi}_\sharp\mu) * \pi_\beta\big)(\mathrm{d}\mathbf{y}),$$
$$A \in \mathcal{B}(\mathcal{X} \times \mathbb{R}^M),\ E \in \mathcal{B}(\mathbb{R}^M).$$

Further consider the measure $\tilde{\eta}(\mathrm{d}\mathbf{z}, \mathrm{d}\mathbf{y}) := \boldsymbol{\phi}_\sharp\mu(\mathrm{d}\mathbf{z})\pi_\beta(\mathrm{d}\mathbf{y} + F(\mathbf{z}))$. Applying Bayes' rule once again we identify $\eta_\beta^{\mathbf{y}}$ as the conditionals of $\tilde{\eta}$ with respect to the projection $\tilde{\Pi} : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}^M$,

$$\tilde{\eta}(B \cap \tilde{\Pi}^{-1}(E)) = \int_E \big(\eta_\beta^{\mathbf{y}} \otimes \delta_{\mathbf{y}}\big)(B)\big(F_\sharp(\boldsymbol{\phi}_\sharp\mu) * \pi_\beta\big)(\mathrm{d}\mathbf{y}),$$
$$B \in \mathcal{B}(\mathbb{R}^N \times \mathbb{R}^M),\ E \in \mathcal{B}(\mathbb{R}^M).$$

By Corollary 1 we have that $(I \times \boldsymbol{\phi})_\sharp\mu(\mathrm{d}u, \mathrm{d}\mathbf{z}) = \mu^{\boldsymbol{\phi}}(\mathrm{d}u + \boldsymbol{\varphi}^\top\mathbf{z})\boldsymbol{\phi}_\sharp\mu(\mathrm{d}\mathbf{z})$. Now define the measure $\tilde{\nu} := (I \times \boldsymbol{\phi})_\sharp\mu(\mathrm{d}u, \mathrm{d}\mathbf{z})\pi_\beta(\mathrm{d}\mathbf{y} + F(\mathbf{z})) \in \mathbb{P}(\mathcal{X} \times \mathbb{R}^N \times \mathbb{R}^M)$. We then have, by the above arguments and Remark 3,

$$\begin{aligned}\tilde{\nu}(\mathrm{d}u, \mathrm{d}\mathbf{z}, \mathrm{d}\mathbf{y}) &= \mu^{\boldsymbol{\phi}}(\mathrm{d}u + \boldsymbol{\varphi}^\top\mathbf{z})\boldsymbol{\phi}_\sharp\mu(\mathrm{d}\mathbf{z})\pi_\beta(\mathrm{d}\mathbf{y} + F(\mathbf{z}))\\ &= \mu^{\boldsymbol{\phi}}(\mathrm{d}u + \boldsymbol{\varphi}^\top\mathbf{z})\tilde{\eta}(\mathrm{d}\mathbf{z}, \mathrm{d}\mathbf{y})\\ &= \mu^{\boldsymbol{\phi}}(\mathrm{d}u + \boldsymbol{\varphi}^\top\mathbf{z})\eta_\beta^{\mathbf{y}}(\mathrm{d}\mathbf{z})\big(F_\sharp(\boldsymbol{\phi}_\sharp\mu) * \pi_\beta\big)(\mathrm{d}\mathbf{y}).\end{aligned}$$

Now observe that $\nu = T_\sharp\tilde{\nu}$ where $T : (u, \mathbf{z}, \mathbf{y}) \mapsto (u, \mathbf{y})$ so that we have the desired identity

$$\nu(\mathrm{d}u, \mathrm{d}\mathbf{y}) = \big(\mu^{\boldsymbol{\phi}} * \boldsymbol{\varphi}_\sharp^\top\eta_\beta^{\mathbf{y}}\big)(\mathrm{d}u)\big(F_\sharp(\boldsymbol{\phi}_\sharp\mu) * \pi\big)(\mathrm{d}\mathbf{y}).$$

□

3 Modes of posterior and conditional measures

In this section we analyze the modes of the posteriors $\mu_\beta^{\mathbf{y}}$ (i.e., the MAP estimators) and the conditionals $\mu_0^{\mathbf{y}}$. Section 3.1 defines the mode of the posterior; subsection Sect. 3.2 defines the mode of the conditional; and Sect. 3.3 considers the $\beta \to 0$ limit of the posterior modes.

3.1 Modes of measures

For a recent overview of definitions of mode in problems defined on infinite dimensional spaces, see (Lambley and Sullivan 2023). In this paper we employ the following notion of the mode of a measure, from Dashti et al. (2013):

Definition 2 Consider a measure $\nu \in \mathbb{P}(\mathcal{X})$. Any point $u^\dagger \in \mathcal{X}$ is a *mode* of ν if it satisfies

$$\lim_{r\to 0} \frac{\nu(B_r(u^\dagger))}{\sup_{u\in\mathcal{X}} \nu(B_r(u))} = 1.$$

This formalizes the idea of defining the mode as the centre of a small ball of maximal probability, in the limit of vanishing radius. The modes of the posterior measures $\mu_\beta^{\mathbf{y}} \in \mathbb{P}(\mathcal{X})$ defined in (6) are referred to as MAP estimators. The next proposition follows directly from Dashti et al. (2013, Cor. 3.10) which allows us to characterize the MAP estimators of $\mu_\beta^{\mathbf{y}}$ via the optimization problem (3). We emphasize that local minimizers of (3) may not be unique, but that a global minimizer exists provided that $F \circ \boldsymbol{\phi}$ is continuous on $\mathcal{X}$ (Dashti et al. 2013).

Theorem 5 *Suppose $\mu = N(0, \mathcal{K})$, $\mu_\beta^{\mathbf{y}}$ is defined as in (6) with $\beta > 0$ and $\mathbf{y} \in \mathbb{R}^M$, and the map $F : \mathbb{R}^N \to \mathbb{R}^M$ satisfies Assumption 1. Define the Onsager–Machlup (OM) functional $J_\beta^{\mathbf{y}} : \mathcal{X} \to [0, \infty]$ by*

$$J_\beta^{\mathbf{y}}(u) := \begin{cases} \dfrac{1}{2\beta^2}|F(\boldsymbol{\phi}(u)) - \mathbf{y}|^2 + \dfrac{1}{2}\|u\|^2_{\mathcal{H}(\mu)}, & \textit{if } u \in \mathcal{H}(\mu), \\ +\infty, & \textit{if } u \in \mathcal{X} \setminus \mathcal{H}(\mu). \end{cases}$$

Then a point $u_\beta^{\mathbf{y}} \in \mathcal{X}$ is a MAP estimator for $\mu_\beta^{\mathbf{y}}$, according to Definition 2, if and only if it is a minimizer of $J_\beta^{\mathbf{y}}$ over $\mathcal{X}$.

Proof To apply the stated corollary define $\Phi(u) := \frac{1}{2\beta^2} |F(\boldsymbol{\phi}(u)) - \mathbf{y}|^2$. Notice that Φ is bounded below uniformly on $\mathcal{X}$, is bounded above on bounded sets in $\mathcal{X}$ and is Lipschitz on bounded sets in $\mathcal{X}$. Then the result follows by a direct application of Dashti et al. (2013, Cor. 3.10). □

We now further characterize MAP estimators of $\mu_\beta^{\mathbf{y}}$ via a representer theorem for the minimizers of OM functionals. This theorem constitutes our main result towards the finite-dimensional characterization of MAP estimators.

Theorem 6 *Suppose that the conditions of Theorem 5 are satisfied. Then $u_\beta^{\mathbf{y}}$ is a MAP estimator for $\mu_\beta^{\mathbf{y}}$ if $u_\beta^{\mathbf{y}} = \boldsymbol{\varphi}^\top \mathbf{z}_\beta^{\mathbf{y}}$ and $\mathbf{z}_\beta^{\mathbf{y}} \in \mathbb{R}^N$ solves*

$$\operatorname*{minimize}_{\mathbf{z}\in\mathbb{R}^N} \frac{1}{2\beta^2}|F(\mathbf{z}) - \mathbf{y}|^2 + \frac{1}{2}\mathbf{z}^\top \Theta^{-1}\mathbf{z}. \tag{15}$$

Proof By Theorem 5 $u_\beta^{\mathbf{y}}$ is a MAP estimator for $\mu_\beta^{\mathbf{y}}$ if it is a minimizer of the OM functional. Applying Chen et al. (2021, Prop. 2.3) to characterize the minimizers of the OM functional yields the desired result. □

Remark 4 Let $\beta > 0$. Note that solutions of the optimization problem (3) (i.e., minimizers of the OM functional) are necessarily in $\mathcal{H}(\mu)$; samples from the posterior $\mu_\beta^{\mathbf{y}}$ given by (1), however, are almost surely not in $\mathcal{H}(\mu)$ because the posterior is absolutely continuous with respect to the prior μ and $\mu\big(\mathcal{H}(\mu)\big) = 0$. Simply put, we need $\mathcal{X}$ to be sufficiently regular so that $\phi_i \in \mathcal{X}^*$ for the probabilistic formulation to make sense, however, the optimization problems (3) and (4) require the ϕ_i to be bounded and linear functionals on both $\mathcal{H}(\mu)$ and $\mathcal{X}$.

This observation has important implications in the context of the GP-PDE solver of Sect. 1.1.2. In order to apply the optimization approaches (3) or (4) to solving PDEs as in Chen et al. (2021), it is necessary that pointwise evaluation of all derivatives appearing in the PDE is possible in $\mathcal{H}(\mu)$. To apply the probabilistic (Bayesian) approach (1) or (2) to the same problem, pointwise evaluation of all derivatives appearing in the PDE is needed over the support of μ, i.e., the space $\mathcal{X}$. Thus the probabilistic approach places a more stringent requirement on the Gaussian prior measure μ than does the optimization approach.

3.2 Modes of conditional measures

Here we define a novel notion of a mode for a conditional measure. We develop a theorem applicable for general maps T with respect to which conditional measures are defined and specified to the case $T = G$, with G given by (5b), in a corollary.

Definition 3 Consider separable Hilbert spaces $\mathcal{X}, \mathcal{Y}$, a measure $\nu \in \mathbb{P}(\mathcal{X})$, and a map $T : \mathcal{X} \to \mathcal{Y}$. Fix a point $y \in \operatorname{supp} T_\sharp \nu$. Then any point $u^\dagger \in T^{-1}(y)$ that satisfies

$$\lim_{r\to 0} \frac{\nu(B_r(u^\dagger))}{\sup_{u\in T^{-1}(y)} \nu(B_r(u))} = 1,$$

is a *conditional mode* of $\nu(\mathrm{d}u \mid T(u) = y)$.

The above definition of the conditional mode is a natural extension of Definition 2 and modifies that definition by restricting the feasible set of $u^\dagger$ to the subset $T^{-1}(y) \subseteq \mathcal{X}$.

Below we show that this definition leads to a natural characterization of conditional modes of Gaussian measures via constrained optimization problems, this is the conditional analog of Theorem 5 and constitutes one of our main theoretical contributions in the paper.

Theorem 7 *Let $\mathcal{X}, \mathcal{Y}$ be separable Hilbert spaces and suppose $T : \mathcal{X} \to \mathcal{Y}$ is continuous. Consider $\mu = N(0, \mathcal{K}) \in \mathbb{P}(\mathcal{X})$ with Cameron–Martin space $\mathcal{H}(\mu)$. Fix a point $y \in T(\mathcal{H}(\mu)) \cap \operatorname{supp} T_\sharp\mu$, assuming the intersection is nonempty. Then u^y is a conditional mode of $\mu(\mathrm{d}u \mid T(u) = y)$ if and only if it solves the optimization problem*

$$\underset{u \in \mathcal{X}}{\text{minimize}} \quad \|u\|_{\mathcal{H}(\mu)} \quad \text{s.t.} \quad T(u) = y. \tag{16}$$

The proof follows by adapting the proof techniques of Dashti et al. (2013, Cor. 3.10) to our definition of a conditional mode. The details are summarized in Appendix B for brevity.

Remark 5 The preceding theorem requires both that $y \in \operatorname{supp} T_\sharp\mu$ and that $y \in T(\mathcal{H}(\mu))$. The first condition is natural: we want the data to have arisen, in principle, from a map T applied to the realization of the measure μ. The second condition, however, says that it must also be realized as an application of the map T to a point in the Cameron–Martin space $\mathcal{H}(\mu)$. Recall that $\mu(\mathcal{H}(\mu)) = 0$. Requiring both of these conditions to hold leads to restrictions on the map T.

Consider the following example of a Gaussian measure from Dashti and Stuart (2017). Assume a centered Gaussian measure μ with a covariance operator which is the inverse of $-\frac{d^2}{dx^2}$ on $I := (0, 1)$, with homogeneous Dirichlet boundary conditions; this is a compact operator from $L^2(I)$ into itself. Thus μ is the Brownian bridge and we may take $\mathcal{X} = H^s(I)$, for any $s < \frac{1}{2}$ since all such Sobolev spaces are in the support of μ. Furthermore any draw from μ is almost surely *not* an element of $H^s(I)$ for any $s \geq \frac{1}{2}$. In particular the Cameron–Martin space is $H_0^1(I)$ and $\mu(H_0^1(I)) = 0$. Now define $t : \mathbb{R} \to \mathbb{R}$ by $t(u) = \min(1, u)$ and $T : \mathcal{X} \to \mathcal{X}$ by $T(u)(x) := t\big(u(x)\big)$. Applying such a function $t(\cdot)$ pointwise to any draw from μ results, almost surely, in a function with no more than $s < \frac{1}{2}$ weak derivatives in $L^2(I)$. Such a function cannot simultaneously be the image under a globally Lipschitz $T(\cdot)$ of an element of $H_0^1(I)$. Thus the preceding theorem cannot be applied.

On the other hand, working with the same measure μ, taking $\mathcal{Y} = \mathbb{R}$ and $T(u) = u(\frac{1}{2})$ it follows from the previous regularity discussions, and the properties of Brownian bridge at any point in the open interval I, that any $y \in \mathbb{R}$ is also in $T(\mathcal{H}(\mu)) \cap \operatorname{supp} T_\sharp\mu$. Thus the theorem can be applied.

Noting the ideas underlying the preceding remark, the following corollary of Theorem 7 is immediate, noting the finite-dimensionality of the image of $T := F \circ \boldsymbol{\phi}$.

Corollary 2 *Consider $\mu = N(0, \mathcal{K}) \in \mathbb{P}(\mathcal{X})$ with Cameron–Martin space $\mathcal{H}(\mu)$, and map $F : \mathbb{R}^N \to \mathbb{R}^M$ satisfying Assumption 1. Suppose $\mu_0^{\mathbf{y}}$ is defined as in Lemma 1 for some $\mathbf{y} \in F\Big(\boldsymbol{\phi}\big(\mathcal{H}(\mu)\big)\Big) \subseteq \mathbb{R}^M$ and with $\phi_i \in \mathcal{X}^\star$. Then a point $u_0^{\mathbf{y}} \in \mathcal{X}$ is a conditional mode for $\mu_0^{\mathbf{y}}$, according to Definition 3, if and only if it is a minimizer of the constrained optimization problem*

$$\underset{u \in \mathcal{X}}{\text{minimize}} \quad \|u\|_{\mathcal{H}(\mu)} \quad \text{s.t.} \quad F\big(\boldsymbol{\phi}(u)\big) = \mathbf{y}. \tag{17}$$

Using the representer theorem (Chen et al. 2021, Prop. 2.3), we can further characterize the conditional modes $u_0^{\mathbf{y}}$ via a finite-dimensional optimization problem. We recall this result for convenience.

Proposition 8 *Suppose Corollary 2 is satisfied. Then every conditional mode $u_0^{\mathbf{y}}$ of $\mu_0^{\mathbf{y}}$ can be written as $u_0^{\mathbf{y}} = \boldsymbol{\varphi}^\top \mathbf{z}_0^{\mathbf{y}}$ where $\mathbf{z}_0^{\mathbf{y}}$ is a solution of*

$$\underset{\mathbf{z} \in \mathbb{R}^N}{\text{minimize}} \quad \mathbf{z}^\top \Theta^{-1} \mathbf{z} \quad \text{s.t.} \quad F(\mathbf{z}) = \mathbf{y}. \tag{18}$$

3.3 Convergence of MAP estimators to conditional modes

Finally, we establish the convergence of the MAP estimators $u_\beta^{\mathbf{y}}$ to the conditional modes $u_0^{\mathbf{y}}$ in the setting where $T = G$, with G given by (5b).

Theorem 9 *Consider $\mu = N(0, \mathcal{K}) \in \mathbb{P}(\mathcal{X})$ with Cameron–Martin space $\mathcal{H}(\mu)$, and a map $F : \mathbb{R}^N \to \mathbb{R}^M$ satisfying Assumption 1. Fix a point $\mathbf{y} \in G\big(\mathcal{H}(\mu)\big)$ and consider the posteriors $\mu_\beta^{\mathbf{y}}$ and their MAP estimators $u_\beta^{\mathbf{y}}$, along with the conditional measures $\mu_0^{\mathbf{y}}$ and their conditional modes $u_0^{\mathbf{y}}$. Then for any sequence of $\beta \to 0$ there exists a subsequence $\beta_n \to 0$ so that $u_{\beta_n}^{\mathbf{y}}$ converges to a conditional mode $u_0^{\mathbf{y}}$.*

Proof First define $J_\beta(u) := \|u\|_{\mathcal{H}(\mu)}^2 + \frac{1}{\beta^2}|G(u) - \mathbf{y}|^2$, recalling that $\mathcal{H}(\mu) = \mathcal{K}^{\frac{1}{2}}\mathcal{X}$ is compactly embedded into $\mathcal{X}$. Note that $u_\beta^{\mathbf{y}}$ is a minimizer of J_β over $\mathcal{X}$ and that $u_0^{\mathbf{y}}$ minimizes $\|u\|_{\mathcal{H}(\mu)}$ in $G^{-1}(\mathbf{y}) \subseteq \mathcal{X}$. Hence, it holds that

$$\|u_\beta^{\mathbf{y}}\|_{\mathcal{H}(\mu)}^2 \leq J_\beta(u_\beta^{\mathbf{y}}) \leq J_\beta(u_0^{\mathbf{y}}) = \|u_0^{\mathbf{y}}\|_{\mathcal{H}(\mu)}^2. \tag{19}$$

Thus we have that $\|u_\beta^{\mathbf{y}}\|_{\mathcal{H}(\mu)} \leq \|u_0^{\mathbf{y}}\|_{\mathcal{H}(\mu)}$ for all $\beta > 0$. Since $\mathcal{H}(\mu)$ is a compact subset of $\mathcal{X}$ we have convergence of $u_\beta^{\mathbf{y}}$ in $\mathcal{X}$ to a limit $u_* \in \mathcal{H}(\mu)$, as well as weak convergence in $\mathcal{H}(\mu)$, along a subsequence β_n. It is immediate that $u_* \in G^{-1}(\mathbf{y})$ as otherwise (along a further relabelled subsequence) there is $\epsilon > 0$ and $N \in \mathbb{N}$ such that $J_{\beta_n}(u_{\beta_n}^{\mathbf{y}}) \geq \epsilon/\beta_n^2$ for all $n \geq N$, which contradicts (19) for all n such that β_n is sufficiently small. To show that u_* is equal to a minimizer of $\|u\|_{\mathcal{H}(\mu)}$ in $G^{-1}(\mathbf{y})$ we assume for contradiction that $\|u_*\|_{\mathcal{H}(\mu)} > \|u_0^{\mathbf{y}}\|_{\mathcal{H}(\mu)}$. By (19) we have

$$\|u^{\mathbf{y}}_{\beta_n}\|^2_{\mathcal{H}(\mu)} \le J_{\beta_n}(u^{\mathbf{y}}_{\beta_n}) \le J_{\beta_n}(u_0^{\mathbf{y}}) = \|u_0^{\mathbf{y}}\|^2_{\mathcal{H}(\mu)} < \|u_*\|^2_{\mathcal{H}(\mu)}.$$

However, by lower semi-continuity of Hilbert space norms, we also have

$$\liminf_{\beta_n \to 0} \|u^{\mathbf{y}}_{\beta_n}\|^2_{\mathcal{H}(\mu)} \ge \|u_*\|^2_{\mathcal{H}(\mu)},$$

giving the desired contradiction. □

4 Algorithms

In this section, we discuss algorithms to sample the posterior and conditional measures of Gaussian priors. According to Theorems 3 and 4, both measures can be represented by a convolution of a finite-dimensional measure that is possibly non-Gaussian, and an infinite-dimensional Gaussian measure that can be identified analytically. Our goal here is to exploit this structure to design efficient algorithms for simulation of the aforementioned posterior and conditional measures as summarized in Sects. 4.1 and 4.2. In Sect. 4.4 we present more concrete examples where posterior measures arising within the GP-PDE methodology are simulated.

4.1 Sampling strategies for posterior measures ($\beta^2 > 0$)

The key idea behind our proposed numerical algorithms is the observation that Theorem 4 enables the decomposition $\mu^{\mathbf{y}}_\beta = \mu^{\boldsymbol{\phi}} * \boldsymbol{\varphi}^\top_\sharp \eta^{\mathbf{y}}_\beta$ where $\mu^{\boldsymbol{\phi}} = N(0, \mathcal{K}^{\boldsymbol{\phi}})$ is a Gaussian whose covariance operator is given by (14), in terms of the measurement operator $\boldsymbol{\phi}$ and the prior covariance matrix $\mathcal{K}$. Thus, the measure $\mu^{\boldsymbol{\phi}}$ can be simulated via standard techniques for discretization and sampling of Gaussian processes and measures (Rasmussen and Williams 2007; Betz et al. 2014; Snelson and Ghahramani 2007). Furthermore, the map $\boldsymbol{\varphi}$ (recall Remark 3) is also defined using $\boldsymbol{\phi}$ and $\mathcal{K}$ and so can be approximated via appropriate discretization. It remains to simulate $\eta^{\mathbf{y}}_\beta$ which is, in general, non-Gaussian. We recall that Theorem 4 identifies $\eta^{\mathbf{y}}_\beta \in \mathbb{P}(\mathbb{R}^N)$ via its Lebesgue density

$$\frac{\mathrm{d}\eta^{\mathbf{y}}_\beta}{\mathrm{d}\Lambda}(\mathbf{z}) \propto \exp\left(-\frac{1}{2\beta^2}|F(\mathbf{z}) - \mathbf{y}|^2 - \frac{1}{2}\mathbf{z}^\top \Theta^{-1}\mathbf{z} \right).$$

At this level any sampling algorithm of choice such as MCMC (Robert and Casella 1999), sequential Monte Carlo (Doucet et al. 2001), or variational inference (Blei et al. 2017) can be used to simulate samples from $\eta^{\mathbf{y}}_\beta$, leading to an algorithm for simulating posterior samples as summarized in Algorithm 1. While this approach is accurate up to the discretization errors of $\mathcal{K}^{\boldsymbol{\phi}}$ and $\boldsymbol{\varphi}$ and the convergence of the utilized sampling algorithms for $\eta^{\mathbf{y}}_\beta$, it has limited utility in the limit $\beta \to 0$ which is particularly important in the context of the GP-PDE solver of Sect. 1.1.2. This is due to the well-understood phenomenon that as $\beta \to 0$ the measure $\eta^{\mathbf{y}}_\beta$ concentrates on the set $F^{-1}(\mathbf{y})$ which may have very small prior measure, leading to poor convergence rates for sampling algorithms such as MCMC.

Algorithm 1 Recipe for generating samples from $\mu^{\mathbf{y}}_\beta$ using MCMC on $\eta^{\mathbf{y}}_\beta$

```
1: Input: Prior covariance $\mathcal{K}$, maps $F$, $\boldsymbol{\phi}$, and $\beta > 0$
2: Output: Samples $u_j \sim \mu^{\mathbf{y}}_\beta$
3: Discretize the operators $\mathcal{K}^{\boldsymbol{\phi}}$ and $\boldsymbol{\varphi}$ as $\widehat{\mathcal{K}}^{\boldsymbol{\phi}}$ and $\widehat{\boldsymbol{\varphi}}$
4: for $j = 1, \dots,$ Number of samples do
5:     Simulate $w_j \sim \mu^{\boldsymbol{\phi}}$ by setting $w_j = (\widehat{\mathcal{K}}^{\boldsymbol{\phi}})^{1/2}\xi_j$ where $\xi_j \sim N(0, I)$
6:     Simulate $v_j \sim \eta^{\mathbf{y}}_\beta$ using MCMC (or similar algorithm)
7:     Set $u_j = w_j + \widehat{\boldsymbol{\varphi}}^\top v_j$
8: end for
```

Under the conjecture that $\eta^{\mathbf{y}}_\beta$ approaches a Gaussian measure in the limit of large data and small noise, we propose to replace Step 6 of Algorithm 1 with a Gaussian approximation step at the mode; this is sometimes referred to as the Laplace approximation to $\eta^{\mathbf{y}}_\beta$ (Kass et al. 1991). More precisely, letting $\mathbf{z}^{\mathbf{y}}_\beta$ be a mode of $\eta^{\mathbf{y}}_\beta$ obtained by solving (15), we define the Gaussian measure

$$\frac{\mathrm{d}\overline{\eta}^{\mathbf{y}}_\beta}{\mathrm{d}\Lambda}(\mathbf{z}) \propto \exp\left(-\frac{1}{2\beta^2}(\mathbf{z} - \mathbf{z}^{\mathbf{y}}_\beta)^\top \left(\nabla F(\mathbf{z}^{\mathbf{y}}_\beta)^\top \nabla F(\mathbf{z}^{\mathbf{y}}_\beta) + D^2 F(\mathbf{z}^{\mathbf{y}}_\beta)(F(\mathbf{z}^{\mathbf{y}}_\beta) - \mathbf{y}) \right) (\mathbf{z} - \mathbf{z}^{\mathbf{y}}_\beta) \right). \tag{20}$$

The above Laplace approximation leads to an efficient sampling algorithm for the posterior since $\overline{\eta}^{\mathbf{y}}_\beta$ is Gaussian and can be simulated exactly given access to the second variation D^2F. In situations where this second variation is expensive to compute we propose an alternative approximation to $\eta^{\mathbf{y}}_\beta$ as follows:

$$\frac{\mathrm{d}\widetilde{\eta}^{\mathbf{y}}_\beta}{\mathrm{d}\Lambda}(\mathbf{z}) \propto \exp\left(-\frac{1}{2\beta^2}|F(\mathbf{z}^{\mathbf{y}}_\beta) + \nabla F(\mathbf{z}^{\mathbf{y}}_\beta)^\top(\mathbf{z} - \mathbf{z}^{\mathbf{y}}_\beta) - \mathbf{y}|^2 - \frac{1}{2}\mathbf{z}^\top \Theta^{-1}\mathbf{z} \right). \tag{21}$$

We refer to this measure as the Gauss–Newton approximation to $\eta^{\mathbf{y}}_\beta$ as it arises from the probabilistic interpretation of the Gauss–Newton algorithm of Chen et al. (2021) that was proposed for finding the mode $\mathbf{z}^{\mathbf{y}}_\beta$. The advantage of the Gauss–Newton approximation over the regular Laplace approximation is that it only uses ∇F and not its second variation,

The Laplace and Gauss–Newton approximations are related to each other, indeed we have

$$\frac{\mathrm{d}\overline{\eta}_\beta^{\mathbf{y}}}{\mathrm{d}\Lambda}(\mathbf{z}) \propto \frac{\mathrm{d}\tilde{\eta}_\beta^{\mathbf{y}}}{\mathrm{d}\Lambda}(\mathbf{z}) \exp\left(-\frac{1}{2\beta^2}(\mathbf{z}-\mathbf{z}_\beta^{\mathbf{y}})^\top \left[D^2F(\mathbf{z}_\beta^{\mathbf{y}})(F(\mathbf{z}_\beta^{\mathbf{y}})-\mathbf{y})\right](\mathbf{z}-\mathbf{z}_\beta^{\mathbf{y}})\right),$$

implying that the Gauss–Newton approximation is close to Laplace whenever $F(\mathbf{z}_\beta^{\mathbf{y}}) - \mathbf{y}$ is small. We anticipate that this approximation is accurate in the regimes where density $\eta_\beta^{\mathbf{y}}$ would concentrate around the set $F^{-1}(\mathbf{y})$. Our numerical experiments indicate that this happens in the GP-PDE setting when we have a lot of observation points and $\beta \to 0$, however, we do not expect this approximation to be good in the setting where $\beta \to 0$, but only a few observations are available.

4.2 Sampling strategies for conditional measures ($\beta^2 = 0$)

The conditional measure $\mu_0^{\mathbf{y}}$ can be simulated using similar ideas from the previous section. By Theorem 3, we can write $\mu_0^{\mathbf{y}} = \mu^{\boldsymbol{\phi}} * \boldsymbol{\varphi}_\sharp^\top \eta_0^{\mathbf{y}}$. Once again, the measure $\mu^{\boldsymbol{\phi}}$ can be simulated (up to discretization errors) exactly and so it remains to generate samples from $\eta_0^{\mathbf{y}} = \eta(\mathrm{d}\mathbf{z}|F(\mathbf{z}) = \mathbf{y})$, the conditional of $\eta = N(0, \Theta)$ with respect to the map F. To do so, we will identify an explicit expression for the Lebesgue density of this conditional. For simplicity we assume that there is a decomposition $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$ such that $F(\mathbf{z}) - \mathbf{y} = 0$ is equivalent to $\mathbf{z}_2 - G(\mathbf{z}_1; \mathbf{y}) = 0$ for some mapping G depending on $\mathbf{y}$ (and implicitely F). Here $\mathbf{z}_1 \in \mathbb{R}^{N_1}$, $\mathbf{z}_2 \in \mathbb{R}^{N_2}$ such that $N = N_1 + N_2$. Such a decomposition is often easy to obtain in many practical applications including the GP-PDE example of Sect. 1.1.2 and can generally be guaranteed by the implicit function theorem under mild conditions on F.

With this decomposition, and a slight abuse of notation, we have

$$\eta_0^{\mathbf{y}} = \eta(\mathrm{d}\mathbf{z}|\mathbf{z}_2 = G(\mathbf{z}_1; \mathbf{y})) = \eta(\mathrm{d}\mathbf{z}_1|\mathbf{z}_2 = G(\mathbf{z}_1; \mathbf{y}))\delta_{G(\mathbf{z}_1;\mathbf{y})}(\mathbf{z}_2).$$

distribution with identity covariance in $\mathbb{R}^N$.

Proposition 10 *Let $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2) \sim N(0, \Theta)$, where $\mathbf{z}_1 \in \mathbb{R}^{N_1}$, $\mathbf{z}_2 \in \mathbb{R}^{N_2}$, and Θ is non-singular. Consider the measure $\breve{\eta}_\beta(\mathrm{d}\mathbf{z}_1) := \eta(\mathrm{d}\mathbf{z}_1|\mathbf{z}_2 = G(\mathbf{z}_1)+\beta\xi)$ where $\xi \sim N(0, I)$ and G is a measurable function[1] in $\mathbb{R}^{N_1}$. Then the density of $\breve{\eta}_\beta$ converges uniformly as $\beta \to 0$ to a density $\breve{\eta}_0$, where*

$$\frac{\mathrm{d}\breve{\eta}_0}{\mathrm{d}\Lambda}(\mathbf{z}_1) \propto \exp\left(-\frac{1}{2}(\mathbf{z}_1, G(\mathbf{z}_1))\,\Theta^{-1}\begin{pmatrix}\mathbf{z}_1\\ G(\mathbf{z}_1)\end{pmatrix}\right). \tag{22}$$

Proof We can write down the density of $\breve{\eta}_\beta$ using Bayes' formula:

$$\begin{aligned}\breve{\eta}_\beta(\mathrm{d}\mathbf{z}_1) &\propto \Lambda(\mathrm{d}\mathbf{z}_1)\int \exp\left(-\frac{1}{2}(\mathbf{z}_1, \mathbf{z}_2)\,\Theta^{-1}\begin{pmatrix}\mathbf{z}_1\\ \mathbf{z}_2\end{pmatrix}\right)\\ &\quad \exp\left(-\frac{|\mathbf{z}_2 - G(\mathbf{z}_1)|^2}{2\beta^2}\right)\Lambda(\mathrm{d}\mathbf{z}_2)\\ &\propto \Lambda(\mathrm{d}\mathbf{z}_1)\int \exp\left(-\frac{1}{2}(\mathbf{z}_1, G(\mathbf{z}_1)+\mathbf{z}_3)\,\Theta^{-1}\begin{pmatrix}\mathbf{z}_1\\ G(\mathbf{z}_1)+\mathbf{z}_3\end{pmatrix}\right)\\ &\quad \exp\left(-\frac{|\mathbf{z}_3|^2}{2\beta^2}\right)\Lambda(\mathrm{d}\mathbf{z}_3),\end{aligned} \tag{23}$$

where we have used the change of variables $\mathbf{z}_2 = G(\mathbf{z}_1)+\mathbf{z}_3$. Let us define

$$g(\mathbf{z}_1, \mathbf{z}_3) := \exp\left(-\frac{1}{2}(\mathbf{z}_1, G(\mathbf{z}_1)+\mathbf{z}_3)\,\Theta^{-1}\begin{pmatrix}\mathbf{z}_1\\ G(\mathbf{z}_1)+\mathbf{z}_3\end{pmatrix}\right),$$

so that we can write $\breve{\eta}_\beta(\mathrm{d}\mathbf{z}_1) \propto \Lambda(\mathrm{d}\mathbf{z}_1)\int g(\mathbf{z}_1, \mathbf{z}_3)\rho_\beta(\mathbf{z}_3)\Lambda(\mathrm{d}\mathbf{z}_3)$ where ρ_β is the density of a Gaussian random variable with mean 0 and covariance $\beta^2 I$. As ρ_β is a mollifier, it holds that $\lim_{\beta\to 0}\int g(\mathbf{z}_1, \mathbf{z}_3)\rho_\beta(\mathbf{z}_3)\Lambda(\mathrm{d}\mathbf{z}_3) = g(\mathbf{z}_1, 0)$ for any $\mathbf{z}_1$; here such convergence is also uniform for all $\mathbf{z}_1$ which yields the uniform convergence of the density of $\breve{\eta}_\beta$ to that of $\breve{\eta}_0$, as $\beta \to 0$ as desired.

To verify the claimed uniform convergence above, consider

$$\begin{aligned}&\left|\int g(\mathbf{z}_1, \mathbf{z}_3)\rho_\beta(\mathbf{z}_3)\Lambda(\mathrm{d}\mathbf{z}_3) - g(\mathbf{z}_1, 0)\right|\\ &\quad\leq \int \left|g(\mathbf{z}_1, \mathbf{z}_3) - g(\mathbf{z}_1, 0)\right|\rho_\beta(\mathbf{z}_3)\Lambda(\mathrm{d}\mathbf{z}_3)\\ &\quad\leq \sup_{\mathbf{z}_1,\mathbf{z}_3}\left|\nabla_{\mathbf{z}_3} g(\mathbf{z}_1, \mathbf{z}_3)\right| \int |\mathbf{z}_3|\rho_\beta(\mathbf{z}_3)\Lambda(\mathrm{d}\mathbf{z}_3) \leq C\beta,\end{aligned}$$

where C is a universal constant that depends only on the dimension N and the eigenvalues of Θ, but independent of $\mathbf{z}_1$, $\mathbf{z}_3$ and β since

$$\begin{aligned}\sup_{\mathbf{z}_1,\mathbf{z}_3}|\nabla_{\mathbf{z}_3} g(\mathbf{z}_1, \mathbf{z}_3)| &\leq \sup_{\mathbf{z}_1,\mathbf{z}_3}\left|\Theta^{-1}\begin{pmatrix}\mathbf{z}_1\\ G(\mathbf{z}_1)+\mathbf{z}_3\end{pmatrix}\right| \cdot |g(\mathbf{z}_1, \mathbf{z}_3)|\\ &\leq \sup_{\mathbf{x}}|\mathbf{x}|\exp\left(-\frac{1}{2}\mathbf{x}^\top\Theta\mathbf{x}\right) \leq C_1.\end{aligned}$$

Here $C_1 \geq 0$ is a constant that also depends on N and the spectrum of Θ. Moreover, by the standard moment formula for Gaussian distributions, it holds that $\int |z_3|\rho_\beta(\mathbf{z}_3)\Lambda(\mathrm{d}\mathbf{z}_3) \leq C_2\beta$. Taking $C = C_1C_2 \geq 0$ leads to the desired result. □

[1] Note that the dependence of G on $\mathbf{y}$ is suppressed here since the theorem holds for arbitrary measurable maps G.

Since Proposition 10 gives a closed form expression for the Lebesgue density of the conditional measure η_0, we can use standard algorithms, such as those discussed in Sect. 4.1, to (approximately) sample this measure. Notably, letting $\mathbf{z}_1^\dagger$ denote the mode of η_0, the Gauss–Newton approximation to (22) will now correspond to the measure

$$\frac{\mathrm{d}\tilde{\eta}_0}{\mathrm{d}\Lambda}(\mathbf{z}_1) \propto \exp\left(-\frac{1}{2}\left(\mathbf{z}_1, G(\mathbf{z}_1^\dagger) + \nabla G(\mathbf{z}_1^\dagger)(\mathbf{z}_1 - \mathbf{z}_1^\dagger)\right)\Theta^{-1}\begin{pmatrix}\mathbf{z}_1 \\ G(\mathbf{z}_1^\dagger) + \nabla G(\mathbf{z}_1^\dagger)(\mathbf{z}_1 - \mathbf{z}_1^\dagger)\end{pmatrix}\right). \tag{24}$$

4.3 On computational complexity and accuracy

Our proposed Laplace and Gauss–Newton methods can lead to significant gains in terms of computational complexity and algorithmic wall-clock times. Indeed, the main computational bottleneck for these algorithms is the inversion of the dense kernel matrix Θ. While this cost is also shared by MCMC algorithms, such as those used in our experiments in Sect. 4.4, the number of linear solves involving this matrix is drastically smaller for our algorithms. For example, in our experiments in Sect. 4.4 we routinely used $O(10^7)$ MCMC steps while Laplace and Gauss–Newton typically converge using $O(10)$ iterations. Naturally, Gauss–Newton is the most efficient option due to skipping Hessian calculations. We emphasize that the efficiency of Laplace and Gauss–Newton comes at the price of non-zero asymptotic accuracy due to the Gaussian approximation which is noticeable if $\eta_\beta^{\mathbf{y}}$ is highly non-Gaussian.

We also note that while Θ is in general dense, and so costs $O(N^3)$ operations to invert, in many practical settings it can be inverted much more efficiently. For example, when $\mathcal{K}$ is a covariance operator of a GP defined on some compact set $\Omega \subset \mathbb{R}^d$ then we can employ the sparse Cholesky algorithm developed in Chen et al. (2024) to obtain a near-linear complexity solver for Θ; more precisely, the sparse Cholesky solver achieves an ϵ-accurate approximation of Θ^{-1} at the cost of $\mathcal{O}(N \log^{2d} \frac{N}{\epsilon})$.

Finally, we emphasize that the choice of the operators ϕ_i has a nontrivial impact on the cost of constructing Θ. For examples such as the GP-PDE methodology, these operators are obtained as compositions of pointwise evaluation functions with differential operators that are often implemented efficiently using automatic differentiation or analytic formulae. However, if the ϕ_i are integral operators then construction of the θ_i and in turn the entries of Θ according to (13) may require additional quadrature steps and numerical approximation.

4.4 Numerical experiments

Our numerical experiments contain two parts: The first part investigates the Laplace and Gauss–Newton approximations introduced in Sects. 4.1 and 4.2, for (approximately) sampling the posterior and conditional distributions. The second part applies our methodology to GP-PDE solvers for example nonlinear PDEs. In Sect. 4.4.1, we compare, through numerical experiments, MCMC, the Laplace approximation and its Gauss-variant; we show that, on the examples considered, the Laplace and Gauss–Newton approximations are good approximations to MCMC in certain regimes as the posteriors concentrate around the true values of the parameter, making Gauss–Newton a good approximation to Laplace.

We apply our methodology to perform UQ as a proxy for error estimation for GP-PDE solvers in Sect. 4.4.2. In Sect. 4.4.3, we use UQ estimates for adaptive selection of collocation points for the solver.

4.4.1 Laplace vs Gauss–Newton

In this subsection, we numerically demonstrate, in a nonlinear elliptic PDE example, the accuracy of Laplace and Gauss–Newton approximations when compared to (the viewed as gold standard) MCMC algorithms. We consider the PDE (8) with $d = 2$ and $\tau(u) = 10u^3$ and choose the ground truth solution $u^\dagger(\mathbf{x}) = \sin(\pi\mathbf{x}_1)\sin(\pi\mathbf{x}_2) + \sin(3\pi\mathbf{x}_1)\sin(3\pi\mathbf{x}_2)$ and determine the right-hand side f which gives this solution, noticing that the Dirichlet boundary conditions are readily satisfied by the prescribed solution. We take J collocation points on a uniform grid in the interior of the domain and M uniform points on the boundary. We denote the interior points by $\mathbf{x}_1, ..., \mathbf{x}_J$ and the boundary points by $\mathbf{x}_{J+1}, ..., \mathbf{x}_M$. For our experiments we took $(J, M) = \{(16, 25), (49, 64), (81, 100)\}$. Following Sect. 1.1.2, we then define $F(\boldsymbol{\phi}(u))$ and $\mathbf{y}$, based on these collocation points and on f, such that identity $F(\boldsymbol{\phi}(u)) = \mathbf{y}$ encodes the PDE constraint at the collocation points.

Suppose u is a priori distributed according to the GP $\mu = N(0, \mathcal{K})$ where $\mathcal{K}$ is the integral operator corresponding to the Matérn kernel with regularity parameter $\nu = 7/2$ (Rasmussen and Williams 2007, Sec. 4.2.1) , with lengthscale parameter $l = 0.3$. The Matérn kernel is widely used for modeling of spatial fields and is particularly well suited to PDE applications since its Cameron–Martin space coincides with classic Sobolev spaces (Kanagawa et al. 2018, Ex. 2.6). This kernel was also used extensively in Chen et al. (2021) in the context of the GP-PDE solver. We further observed that our results are not very sensitive to the choice of ν, l as long as these values are not unusually small or large. Then the conditional $\mu_0^{\mathbf{y}}$ encodes information about the solution to the PDE. We compute the conditional mode $u_0^{\mathbf{y}}$ using the Gauss–Newton optimization algorithm of Chen et al. (2021)

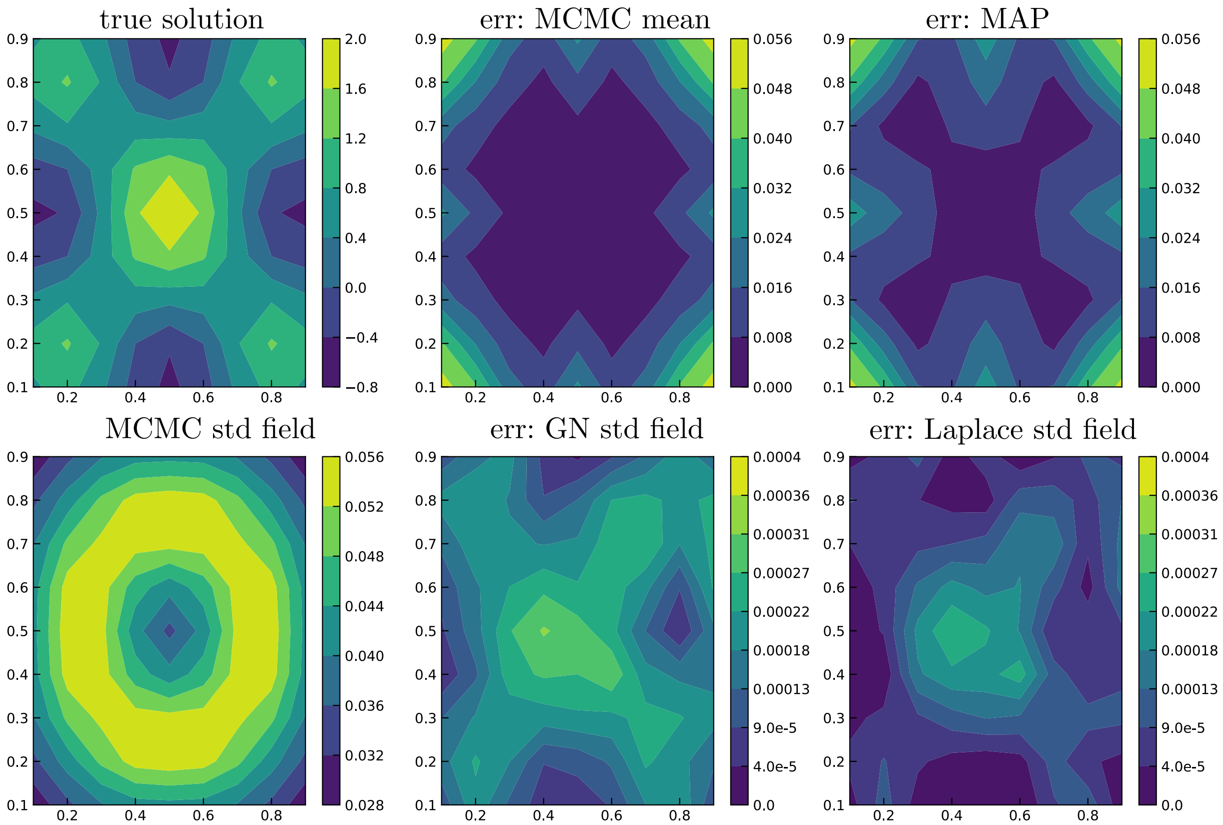

Fig. 2 Numerical results for nonlinear elliptic (8) as described in Sect. 4.4.1 with $(J, M) = (81, 100)$ collocation points. Top row: True solution, error of MCMC mean, and error of the MAP estimator obtained by the GP-PDE methodology. Bottom row: standard deviation field of MCMC samples followed by its difference from the standard deviation fields obtained using the Gauss–Newton and Laplace approximations. Boundary points are excluded from the plots for clarity, since we observe the exact values of the solution on the boundary

that gives a numerical solution to the optimization problem (18) that characterizes $u_0^{\mathbf{y}}$. Using this mode we further compute the Laplace and Gauss–Newton approximations to $\mu_0^{\mathbf{y}}$ following the approach of Sect. 4.2.

In Fig. 2 (top row), we compare the true solution of the PDE to the MAP estimator $u_0^{\mathbf{y}}$ and the posterior mean of the MCMC samples with $(J, M) = (81, 100)$. We observe that the MAP and the MCMC mean are comparable approximations to the true solution, indicating that the posterior measure is concentrated around the truth. This claim is further supported by Fig. 2 (bottom row) where we compare the pointwise standard deviations computed by MCMC, Gauss–Newton, and Laplace. We see good agreement between all three methods, suggesting that (a) the posterior is close to being Gaussian and (b) the Gauss–Newton approximation is as good as Laplace. In Table 1 we further compare the relative L^2 error between the MCMC mean and standard deviations with those of Laplace and Gauss–Newton approximations. We observe that not only does the MAP converge to the MCMC mean but that Laplace and Gauss–Newton approximations to the standard deviation fields converge to that of the MCMC samples. Moreover, the Gauss–Newton and Laplace errors are comparable, with Gauss–Newton achieving higher errors when collocation points are scarce.

In Fig. 3, we evaluate the posterior fields at the location $x = [0.6, 0.4]$ for different mesh sizes and compare the kernel density estimator of the MCMC samples to that of Laplace and Gauss–Newton approximations. Here we observe that (a) the Laplace and Gauss–Newton approximations are very close to each other and (b) as we refine the mesh, these two approximations converge to the MCMC posterior. We observed this behavior consistently at other locations as well, supporting the claim that the posterior is nearly Gaussian around the MAP.

4.4.2 UQ for GP-PDE

One of the advantages of the GP-PDE perspective is that the conditional/posterior uncertainties can be readily computed as a priori indicators of the performance of the algorithm. Here we will investigate the usefulness of such uncertainties in the context of our nonlinear elliptic PDE (8) as well as Burgers' equation.

Table 1 Relative L^2 error for the mean and stanfard deviation of the Laplace approximation and its Gauss–Newton variant at sampled points compared with MCMC for the nonlinear elliptic PDE (8) as described in Sect. 4.4.1. MCMC results were obtained using 10^7 time steps

Relative L^2 error	$(J, M) = (16, 25)$	$(J, M) = (49, 64)$	$(J, M) = (81, 100)$
MAP vs MCMC mean	1.086e−1	1.682e−2	6.320e−3
(std) Laplace vs MCMC	6.360e−2	5.557e−3	2.136e−3
(std) Gauss–Newton vs MCMC	7.934e−2	1.038e−2	4.086e−3

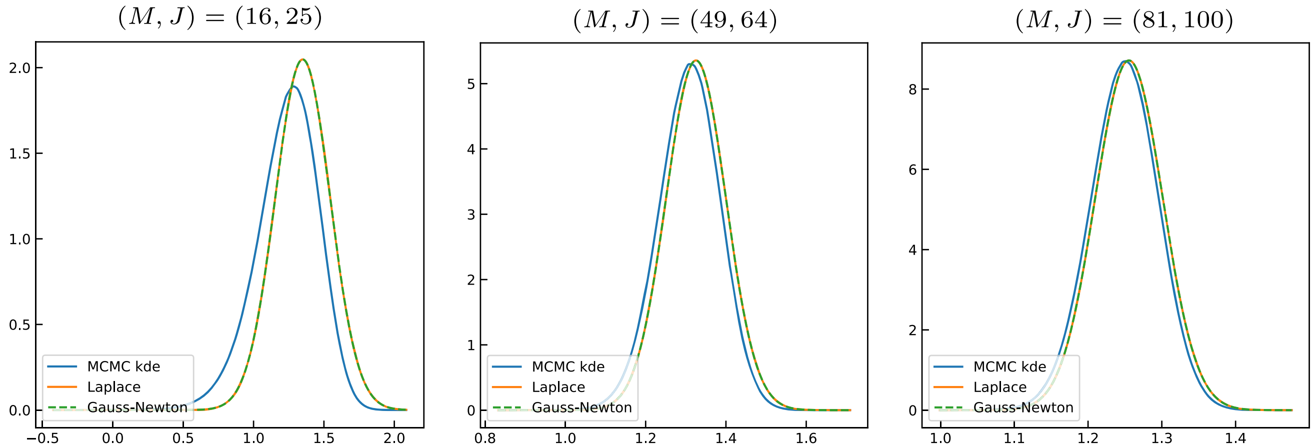

Fig. 3 Pointwise numerical results for the nonlinear elliptic PDE (8) as described in Sect. 4.4.1. Here we compared the conditional distribution of the solution to its various approximations at a single point [0.6, 0.4] with (Left) $(M, J) = (16, 25)$, (middle) $(M, J) = (49, 64)$, and (right) $(M, J) = (81, 100)$ collocation points

Nonlinear elliptic PDE

We start by considering the nonlinear elliptic PDE (8) once more with $\tau(u) = \alpha u^3$ along with prescribed solution $u^\dagger(\mathbf{x}) = \sin(\mathbf{x}_1)\sin(\mathbf{x}_2) + \sin(10\mathbf{x}_1)\sin(a\mathbf{x}_2)$ with scalar parameters $\alpha, a > 0$ to be chosen later. We solve the PDE using $(J, M) = (16, 25)$ with the prior $\mu = N(0, \mathcal{K})$ with $\mathcal{K}$ being the 7/2-Matérn kernel. To estimate the conditional mode and standard deviations we ran three steps of the Gauss–Newton algorithm for different choices of (α, a) as shown in Fig. 4. We observe that in the linear PDE setting where $\alpha = 0$, the resulting posterior standard deviation field is very smooth and is known to be independent of the PDE solution and only dependent on the collocation points. As expected, maximum standard deviation occurs in the middle of the domain as is often expected in GP regression. Interestingly, the posterior standard deviation fields appear to change noticeably with stronger nonlinearities. In particular, the maximum uncertainty no longer occurs in the middle of the domain but rather over a non-trivial set.

It is well-known, in the context of GP regression (Owhadi 2015, Thm.5.1) that if $u^\dagger$ is the ground truth and $u_0^{\mathbf{y}}$ is its GP interpolant, that the following error bound holds

$$|u^\dagger(\mathbf{x}) - u_0^{\mathbf{y}}(\mathbf{x})| \le \|u^\dagger\|_{\mathcal{H}(\mu)}\sigma(\mathbf{x}) \qquad \forall \mathbf{x} \in \Omega, \tag{25}$$

where $\sigma(\mathbf{x})$ is the standard deviation field of the conditioned GP and $\|\cdot\|_{\mathcal{H}(\mu)}$ denotes the Cameron–Martin/RKHS norm of $u^\dagger$ corresponding to the GP prior μ. It is therefore natural to investigate, numerically, whether this error bound remains valid in the case of the GP-PDE solver. Since in practice we do not have access to $\|u^\dagger\|_{\mathcal{H}(\mu)}$, we replace it with the Cameron–Martin norm of the MAP, i.e., $\|u_0^{\mathbf{y}}\|_{\mathcal{H}(\mu)}$.

In Fig. 5 we show a slice of the PDE solution $u^\dagger$ along with the GP-PDE solution and the requisite error bounds computed using the standard deviation fields for our nonlinear elliptic PDE example. We observe that in all three cases, the conditional mode $u_0^{\mathbf{y}}$ is a good approximation to $u^\dagger$ while the upper and lower bounds computed via (25) always contain both the numerical and true solutions. However, we note that the computed error bands appear to be too large compared to the actual error of the numerical solution.

Burgers' equation

Next we consider the viscous Burgers equation:

$$\begin{aligned}
&\partial_t u + u\partial_x u - 0.01\partial_x^2 u = 0, \quad \forall (x, t) \in (-1, 1) \times (0, 1]\,,\\
&u(x, 0) = -\sin(\pi x)\,,\\
&u(-1, t) = u(1, t) = 0\,.
\end{aligned} \tag{26}$$

We solved this equation using the space-time GP-PDE approach of Chen et al. (2021). Collocation points were

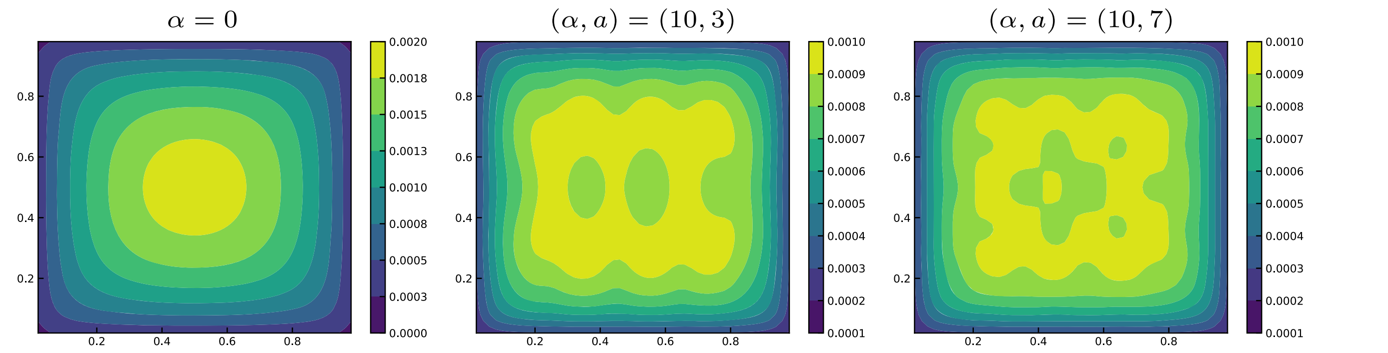

Fig. 4 Comparing posterior standard deviation fields for the nonlinear elliptic PDE (8) as described in Sect. 4.4.2. From left to right the panels show the standard deviation fields for increasingly stronger nonlinearities

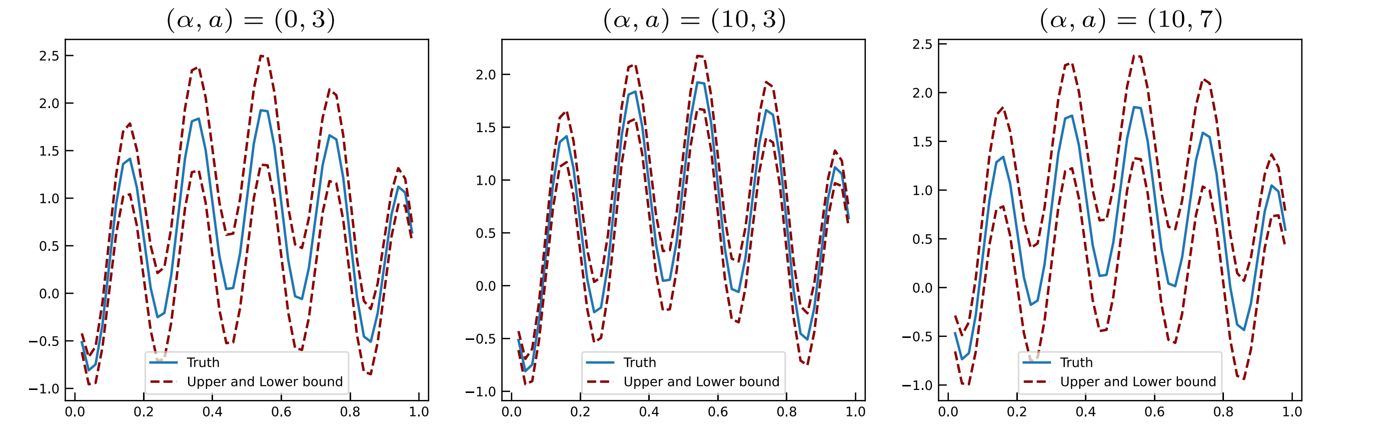

Fig. 5 Truth and the upper and lower error bound obtained by the GP-PDE method, for the slice $\mathbf{x}_2 = 0.5$, in the nonlinear elliptic PDE (8) as described in Sect. 4.4.1. From left to right the panels show the posterior mean with uncertainty bands for increasingly stronger nonlinearities

uniformly distributed on a regular grid with time step size $\mathrm{d}t = 0.05$ and spatial step size $\mathrm{d}x = 0.0125$. The kernel of the covariance function of the GP is chosen as the anisotropic Gaussian kernel, same as Chen et al. (2021):

$$K\Big((x,t),(x',t');\sigma\Big) = \exp\Big(-\sigma_1^{-2}(x-x')^2 - \sigma_2^{-2}(t-t')^2\Big) \qquad (27)$$

with $\sigma = (1/20, 1/3)$; this anisotropic kernel was also used in Chen et al. (2021) to account for different regularity of the solution of Burgers' equation in the temporal and spatial domains. The spatial and temporal lengthscales were tuned by hand. We ran 15 steps of Gauss–Newton to obtain the conditional mode and the corresponding approximation to the conditional covariance matrix. In Fig. 6 (left and middle) we show the GP-PDE solution to the Burgers' equation as well as the posterior standard deviation estimated using Gauss–Newton. We clearly observe that the standard deviation is peaked around the location of the (near) discontinuity in the solution, indicating that the standard deviation field is a good proxy for the adaptive placement of collocation points.

4.4.3 Adapting collocation points

Based on our observation in the previous section (e.g. Fig. 6) it is of interest to investigate whether the UQ estimates from the posterior/conditional measure can be used for the adaptation of collocation points for PDE solvers. For example, we may add more collocation points in areas of maximum variance under the posterior/conditional on the solution of the PDE.

For our first experiment we considered the Burgers equation (26) which was originally solved on a uniform grid and added 30 new collocation points in the region of maximum posterior variance which happens to surround the (smoothed) shock. This produces a non-uniform grid of collocation points as shown on the right panel of Fig. 6. In our experiments we observed that adding these new points leads to a factor 2 improvement in the L^∞ error of the solution at time $t = 1$. This demonstrates the effectiveness of using UQ estimates for adaptive selection of collocation points. We observed that when we continued to select points based on this greedy approach, the improvement in accuracy was less significant and sometimes even numerical instability occurs. We attribute this phenomenon to the use of a

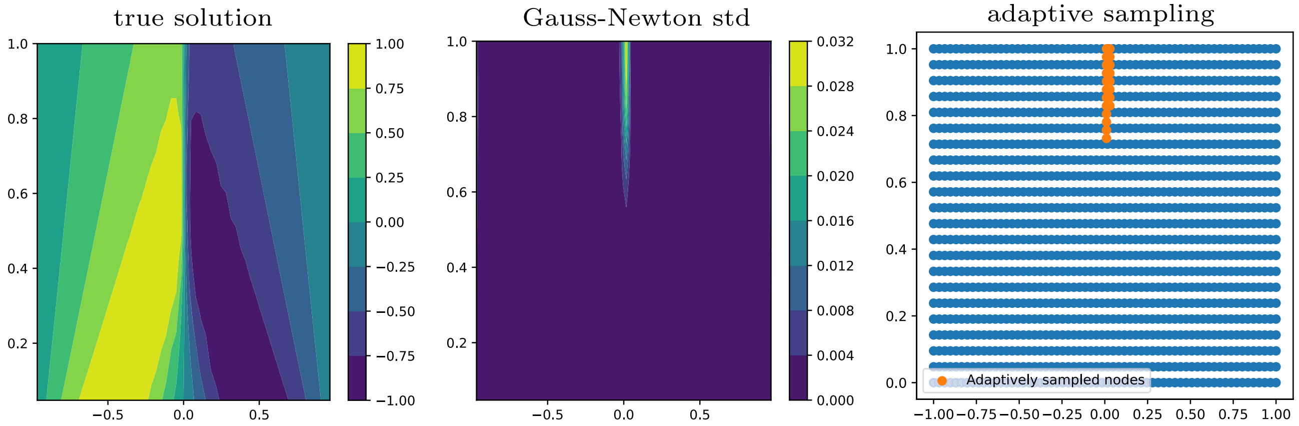

Fig. 6 Numerical experiments for the Burgers' PDE (26). Left: Contour plot of the MAP estimator of the solution in space-time; Middle: Contour plot of the conditional standard deviation field; Right: Adaptively sampled collocation points guided by areas of concentrated uncertainty

global space-time formulation, which overlooks the causality of time dependent PDEs and could lead to numerical challenges. This could also be attributed to the ill-conditioning of the involved kernel matrices associated to a large number of points packed in a small region of the domain which further warrants the use of a nugget term.

For our second experiment we return to the nonlinear elliptic PDE (8) with $\tau(x) = 10x^3$. We prescribe the exact solution $u(\mathbf{x}) = 2^{4p}\mathbf{x}_1^{2p}(1-\mathbf{x}_1)^p\mathbf{x}_2^{2p}(1-\mathbf{x}_2)^p$ with $p = 10$ as shown in Fig. 7; this example is designed to have a highly localized feature around the location $(2/3, 2/3)$. We then solve the PDE and adaptively add collocation points as follows: (1) Start with 100 uniformly sampled collocation points in the interior and on the boundary of the unit box; (2) compute the Gauss–Newton approximation to the posterior of the solution and sample 50 new collocation points in areas of largest posterior variance; (3) repeat step (2) for 10 iterations to get a total of 600 collocation points in the interior.

In the bottom left panel of Fig. 7 we show an instance of the collocation points obtained by the above procedure which may be compared with the top right panel, depicting a uniform set of collocation points. We see that the posterior adapted points are blind to the concentrated features of the solution to the PDE, contrary to our early example for Burgers' equation. We further modified our adaptive sampling of the collocation points to place new points in regions of large equation residual which produced the bottom right panel of Fig. 7. We observe that this new strategy leads to collocation points that are clustered around the main feature of the solution. We present L^2 and L^∞ errors of the solutions obtained by the three sampling strategies in Table 2, showing that the conditional variance adaptation scheme leads to an order of magnitude improvement in the error over uniform points while residual adaptation leads to yet another order of magnitude improvement.

These experiments show the advantages and potential limitations of using the posterior/conditional variance for adapting collocation points. Interestingly, in the case of Burgers' equation the conditional variance captures the interesting structures in the solution while this property is not prominent in the case of our nonlinear elliptic PDE.

5 Conclusions

Our focus in this article was the characterization of Gaussian measures conditioned on finite nonlinear observations that are obtained as the composition of a nonlinear map with a bounded and linear operator. We showed that (1) such conditionals can be characterized as the limit of posterior measures with noisy observations with vanishing small noise standard deviation. We showed that this small-noise limiting argument also applied to the MAP estimators of the resulting conditionals leading to the novel definition of a conditional MAP of a Gaussian measure; (2) We showed that the resulting posteriors/conditional measures can be decomposed as the convolution of a Gaussian measure that can be identified analytically with a finite-dimensional non-Gaussian measure. This decomposition mirrored well-known representer theorems from RKHS theory. Item (2) further led us to the design of novel algorithms for the simulation of Gaussians conditioned on nonlinear observations by focusing computational effort on the non-Gaussian component.

We applied our results to the particular case of the GP-PDE methodology, a collocation method for solving nonlinear PDEs that models the solution of the PDE as a GP conditioned on the PDE constraint at the collocation points. We developed two variational inference techniques for simulation of the non-Gaussian component in this case under the conjecture that, if the collocation points are sufficiently dense then the non-Gaussian component of the posterior should

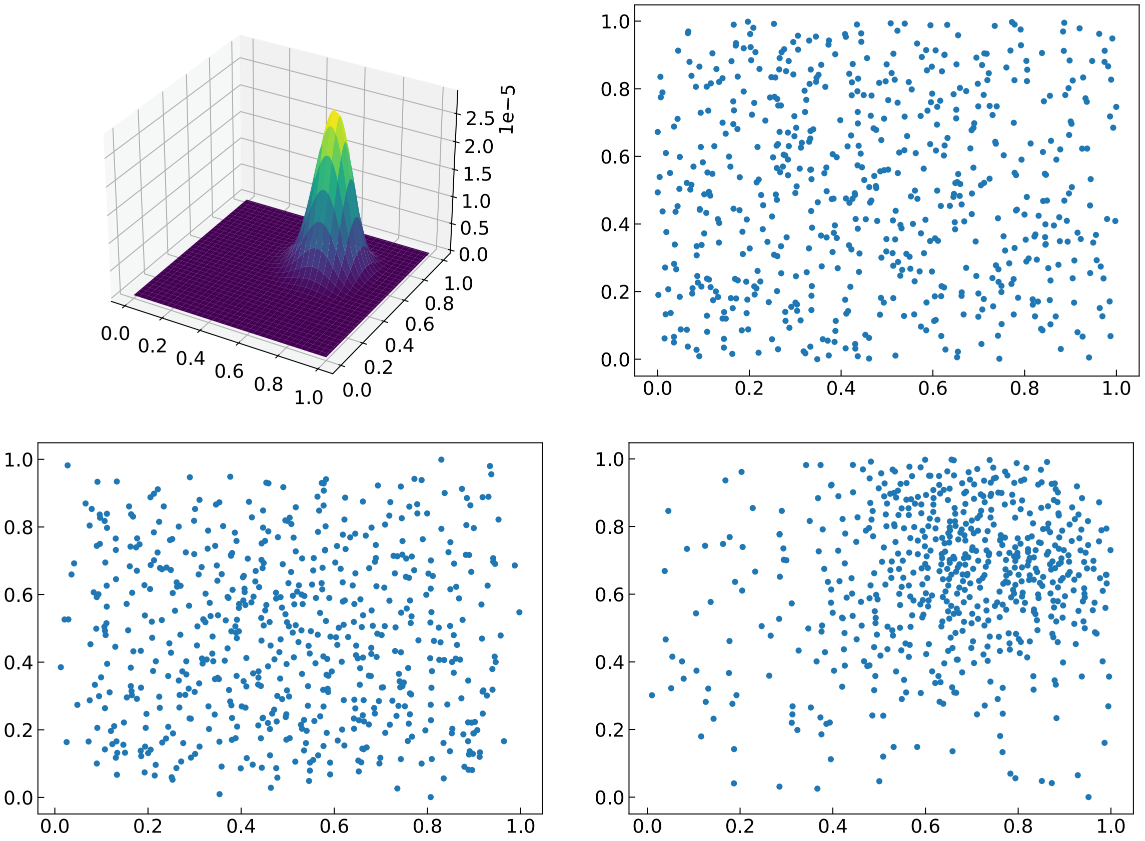

Fig. 7 An instance of adaptively selected collocation points for the nonlinear elliptic PDE (8) as described in Sect. 4.4.3. Top left: true solution; top right: uniform sampling; bottom left: greedy sampling based on the conditional standard deviation; bottom right: greedy sampling based on equation residues

Table 2 Relative L^2 and L^∞ errors of the numerical solutions of the nonlinear elliptic PDE (8) as described in Sect. 4.4.3

Sampling strategy	Uniform	Conditional variance	Equation residue
Relative L^2 error	2.337e−2	3.345e−3	1.365e−4
Relative L^∞ error	1.565e−2	2.554e−3	1.046e−4

Various strategies for adaptive sampling of collocation points were compared. The errors were averaged over 20 trials

be approximately Gaussian around its MAP. Our numerical experiments confirmed this claim. We also investigated the usefulness of the resulting uncertainty estimates for improving the accuracy of the PDE solver by adaptive selection of collocation points.

While the GP-PDE setting was the main motivation for our work, our results have wide application in the field of inverse problems where Gaussian priors are widely used in a function space setting. Here one often discretizes the problem and samples the posterior using a function space MCMC algorithm. However, our results here suggest that significant speed up may be achieved by performing MCMC only on the non-Gaussian component and directly simulating the Gaussian component, for example by computing the underlying precision matrix of the prior. Our experiments also suggest that this non-Gaussian component may be well approximated by a variational technique such as a Laplace approximation. We also observed that our Gauss–Newton approximation (which is only first order) appears to work well in practice, a topic that warrants more detailed theoretical analysis.

Appendix A On optimal recovery, game theory, and probabilistic numerics

As presented in Owhadi and Scovel (2019), the framework of optimal recovery of Micchelli and Rivlin (1977) provides a natural setting for understanding the correspondence between numerical approximation and Bayesian inference, which involves the counter-intuitive modeling of a perfectly known function as a sample from a random process.

To describe this consider a Banach space $(\mathcal{B}, \|\cdot\|)$ and write $[\cdot,\cdot]$ for the duality product between $\mathcal{B}$ and its dual space $(\mathcal{B}^*, \|\cdot\|_*)$. When $\mathcal{B}$ is infinite (or high) dimensional, as conceptualized in Information Based Complexity (Traub et al. 1988) (the branch of computational complexity founded on the observation that numerical implementation requires computation with partial information and limited resources), one cannot directly compute with $u \in \mathcal{B}$ but only with a finite number of *features* of u. The type of features we consider here are represented as a vector $\Phi(u) := \big([\phi_1, u], \ldots, [\phi_m, u]\big)$ corresponding to m linearly independent measurements $\phi_1, \ldots, \phi_m \in \mathcal{B}^*$. The objective is to recover/approximate u from the partial information contained in the feature vector $\Phi(u)$. Then, using the relative error in $\|\cdot\|$-norm as a loss, the classical numerical analysis approach is to approximate u with the minimizer $v^\dagger$ of

$$\min_v \max_u \frac{\|u - v(\Phi(u))\|}{\|u\|}, \tag{A1}$$

where the maximum is taken over all $u \in \mathcal{B}$ and the minimum is taken over all possible functions v of the m linear measurements. The minimax approximant is Micchelli and Rivlin (1977) and Owhadi and Scovel (2019, Chap. 18) then

$$v^\dagger(y) = \operatorname{argmin} \begin{cases} \text{Minimize } \|v\| \\ \text{Subject to } v \in \mathcal{B} \text{ and } \Phi(v) = y. \end{cases} \tag{A2}$$

Furthermore, the minmax problem (A1) can be viewed as the adversarial zero sum game in which Player I chooses an element u of the linear space $\mathcal{B}$ and Player II (who does not see u) must approximate Player I's choice based on seeing the finite number of linear measurements $\Phi(u)$ of u. The function $(u, v) \mapsto \frac{\|u - v(\Phi(u))\|}{\|u\|}$ has no saddle points, so to identify a minmax solution as a saddle point one can proceed, as in Wald's decision theory (Wald 1945), evidently influenced by von Neumann's theory of games (von Neumann 1928), by introducing mixed/randomized strategies and lift the problem to probability measures over all possible choices for players I and II. For the lifted version of the game, the optimal mixed strategy of Player I is a cylinder measure defined by the norm $\|\cdot\|$ and the optimal strategy of Player II is a pure strategy because $\|\cdot\|$ is convex. Furthermore if the norm $\|\cdot\|$ is quadratic, then the optimal strategy of Player I is a centered Gaussian field whose covariance operator $Q : \mathcal{B}^* \to \mathcal{B}$ is defined by the norm $\|\cdot\|$ and the identity $\|\phi\|_*^2 = [\phi, Q\phi]$. For further references on Gaussian measures on infinite-dimensional spaces, we refer to Bogachev (1998) and Maniglia and Rhandi (2004) (for Hilbert spaces). See also Janson (1997) for Gaussian fields on Hilbert spaces. The application of optimal recovery, initially focused on solving linear PDEs (Harder and Desmarais 1972; Duchon 1977; Owhadi 2015), has been extended to nonlinear PDEs in Chen et al. (2021) and to general computational graph completion problems in Owhadi (2022).

Appendix B Proof of Theorem 7

The main ideas required for the proof of Theorem 7 are contained in Proposition 12. The proposition and theorem themselves rest on several lemmas which we collect together in a preliminary subsection.

First we recall three technical results, concerning small ball probabilities, from Dashti et al. (2013).

Lemma 3 (Dashti et al. 2013, Lem. 3.6) *Let $\mu = N(0, \mathcal{K})$, $r > 0$ and $u \in \mathcal{X}$. Then there exists a constant $\alpha > 0$ independent of u, r so that*

$$\frac{\mu(B_r(u))}{\mu(B_r(0))} \le \exp\left(\frac{\alpha}{2} r^2\right) \exp\left(-\frac{\alpha}{2}\left(\|u\|_{\mathcal{X}} - r\right)^2\right).$$

Lemma 4 *(Dashti et al. 2013, Lem. 3.7) Suppose $u_0 \notin \mathcal{H}(\mu)$, $\{u_r\}_{r\ge 0} \subset \mathcal{X}$ and u_r converges weakly to u_0 in $\mathcal{X}$ as $r \to 0$. Then for any $\epsilon > 0$ there exists $r > 0$ small enough so that*

$$\frac{\mu(B_r(u_r))}{\mu(B_r(0))} < \epsilon.$$

Lemma 5 (Dashti et al. 2013, Lem. 3.9) *Consider a sequence $\{u_r\}_{r\ge 0} \subset \mathcal{X}$ and suppose u_r converges weakly and not strongly to 0 in $\mathcal{X}$ as $r \to 0$. Then for any $\epsilon > 0$, there exists r small enough such that*

$$\frac{\mu(B_r(u_r))}{\mu(B_r(0))} < \epsilon.$$

A fourth useful lemma concerning small ball probabilities is:

Lemma 6 (Bogachev 1998, Lem. 4.7.1) *For all $u \in \mathcal{H}(\mu)$ it holds that*

$$1 \le \frac{1}{\mu(B_r(0))} \int_{B_r(0)} \exp\left(\langle u, x\rangle_{\mathcal{H}(\mu)}\right) \mathrm{d}\mu(x).$$

For our final lemma we recall the following classic result [see for example Bogachev (1998, Cor. 4.7.8)] which is integral to the analysis in the following subsection.

Lemma 7 *Let $\mu = N(0, \mathcal{K}) \in \mathbb{P}(\mathcal{X})$. Then*

$$\lim_{r\to 0} \frac{\mu(B_r(u_1))}{\mu(B_r(u_2))} = \exp\left(\frac{1}{2}\|u_2\|^2_{\mathcal{H}(\mu)} - \frac{1}{2}\|u_1\|^2_{\mathcal{H}(\mu)}\right), \quad \forall u_1, u_2 \in \mathcal{H}(\mu).$$

Now recall Definition 3 of the conditional mode. Our goal is to show that such a point is equivalent to a minimizer of (16). We start by establishing the existence of such minimizers.

Proposition 11 *Let $\mu = N(0, \mathcal{K})$ and fix $y \in T(\mathcal{H}(\mu))$ for a continuous map $T : \mathcal{X} \to \mathcal{Y}$. Then there exists a minimizer u^y of (16).*

Proof Since $y \in T(\mathcal{H}(\mu))$ by assumption, then the feasible set $T^{-1}(y) \cap \mathcal{H}(\mu)$ is non-empty. Define $I := \inf\{\|u\|_{\mathcal{H}(\mu)} : u \in T^{-1}(y)\}$ and let $\{u_n\} \in T^{-1}(y)$ be a minimizing sequence. Then for any $\delta > 0$ there exists $N = N(\delta)$ so that

$$0 \le I \le \|u_n\|_{\mathcal{H}(\mu)} \le I + \delta, \qquad \forall n \ge N.$$

Since $\mathcal{H}(\mu)$ is a Hilbert space and $\{u_n\}$ is bounded we infer the existence of a limit point $u^y \in \mathcal{H}(\mu)$ (possibly along a subsequence) so that u_n converges to u^y weakly in $\mathcal{H}(\mu)$. The weak lower semicontinuity of the $\mathcal{H}(\mu)$-norm now yields, $I \le \|u^y\|_{\mathcal{H}(\mu)} \le I + \delta$ and the result follows since δ is arbitrary. □

Proposition 12 *Consider $\mu = N(0, \mathcal{K})$, a continuous map $T : \mathcal{X} \to \mathcal{Y}$ and a point $y \in T(\mathcal{X})$. Define*

$$u_r := \operatorname*{arg\,max}_{u \in T^{-1}(y)} \mu(B_r(u)). \tag{B3}$$

Then:

(i) the maximizer $u_r \in \mathcal{X}$ exists for every $r > 0$;
(ii) if y belongs to $T(\mathcal{H}(\mu))$ then there exists $u^y \in \mathcal{H}(\mu) \cap T^{-1}(y)$ and a subsequence of $\{u_r\}_{r \ge 0}$ which converges to u^y strongly in $\mathcal{X}$ as $r \to 0$;
(iii) if y belongs to $T(\mathcal{H}(\mu)) \cap \operatorname{supp} T_\sharp\mu$, and the intersection is not empty, then the limit u^y is both a conditional mode of $\mu(\mathrm{d}u | T(u) = y)$ and a minimizer of (16).

Proof (i) First observe that by assumption $T^{-1}(y)$ is not empty. By Lemma 3 we deduce that any maximizing sequence is bounded in $\mathcal{X}$. Extract a weakly convergent subsequence $\{u_r^{(n)}\}_{n \in \mathbb{N}}$ with limit u_r. Since $\mathcal{X}$ is a Hilbert space and $T^{-1}(y)$ is closed we conclude that $u_r \in T^{-1}(y)$. The Gaussian measures $\mu(\cdot + u_r^{(n)})$ then converge weakly as $n \to \infty$ to Gaussian measures $\mu(\cdot + u_r)$ (Bogachev 1998). Thus $\mu(B_r(u_r^{(n)})) \to \mu(B_r(u_r))$ since the indicator function of a ball is a bounded measurable function. Hence, since the subsequence is a maximizing subsequence, the result is proved.

(ii) Now consider the sequence $u_r = \arg\max_{u \in T^{-1}(y)} \mu(B_r(u))$, indexed over $r \ge 0$. Our first task is to show that $\{u_r\}_{r \ge 0}$ is bounded in $\mathcal{X}$. By the hypothesis that $y \in T(\mathcal{H}(\mu))$ we can pick a point $u^\star \in \mathcal{H}(\mu) \cap T^{-1}(y)$, which we will fix for the remainder of the proof of (ii). Since u_r is, by definition, the maximizer of $\mu(B_r(u))$ over $T^{-1}(y)$ then we have that

$$\frac{\mu\big(B_r(u_r)\big)}{\mu\big(B_r(u^\star)\big)} \ge 1. \tag{B4}$$

By the Cameron–Martin formula we can further write

$$\begin{aligned} 1 \le \frac{\mu\big(B_r(u_r)\big)}{\mu\big(B_r(u^\star)\big)} &= \frac{\mu\big(B_r(u_r)\big)}{\mu\big(B_r(0)\big)} \frac{\mu\big(B_r(0)\big)}{\mu\big(B_r(u^\star)\big)} \\ &= \frac{\mu\big(B_r(u_r)\big)}{\mu\big(B_r(0)\big)} \exp\left(\frac{1}{2}\|u^\star\|^2_{\mathcal{H}(\mu)}\right) \\ &\quad \frac{\mu(B_r(0))}{\int_{B_r(0)} \exp(-\langle u^\star, x\rangle_{\mathcal{H}(\mu)}) \mathrm{d}\mu(x)}. \end{aligned}$$

An application of Lemma 6 yields the lower bound

$$\frac{\mu(B_r(u_r))}{\mu(B_r(0))} \ge \exp\left(-\frac{1}{2}\|u^\star\|^2_{\mathcal{H}(\mu)}\right). \tag{B5}$$

Now suppose, to obtain a contradiction, that $\{u_r\}_{r \ge 0}$ is not bounded in $\mathcal{X}$, so that for any $R > 0$ there exists r_R so that $\|u_{r_R}\|_{\mathcal{X}} > R$ with $r_R \to 0$ and $R \to \infty$. Then the lower bound (B5) contradicts Lemma 3 for large R and sufficiently small r_R leading to the conclusion that $\{u_r\}_{r \ge 0}$ is bounded. Since $\mathcal{X}$ is a Hilbert space and $T^{-1}(y)$ is closed we infer there exists a point $u^y \in T^{-1}(y)$ and a subsequence $\{u_r\}$ which converges weakly to u^y in $\mathcal{X}$ as $r \to 0$.

Now suppose, again for contradiction, that either: (a) there is no strongly convergent subsequence of $\{u_r\}$ in $\mathcal{X}$; or (b) if there is such a subsequence its limit u_0 does not belong to $\mathcal{H}(\mu)$. We start with the case (b). Consider (B5) and apply Lemma 4 with $\epsilon = \frac{1}{2}\exp(-\frac{1}{2}\|u^\star\|^2_{\mathcal{H}(\mu)})$ to obtain

$$\begin{aligned} \exp\left(-\frac{1}{2}\|u^\star\|^2_{\mathcal{H}(\mu)}\right) &\le \frac{\mu(B_r(u_r))}{\mu(B_r(0))} \\ &< \frac{1}{2}\exp\left(-\frac{1}{2}\|u^\star\|^2_{\mathcal{H}(\mu)}\right), \end{aligned} \tag{B6}$$

which is a contradiction and so the limit point $u_0 \in \mathcal{H}(\mu)$. Now consider case (a) where there exists no strongly convergent subsequence that converges to u_0. Then the (sub)sequence $u_r - u_0$ satisfies the conditions of Lemma 5. We can then repeat the above argument with the same choice of ϵ to obtain (B6) once again which is a contradiction. This concludes the proof of part (ii).

(iii) In what follows we let $u_0 \in T^{-1}(y) \cap \mathcal{H}(\mu)$ denote the limit of the relabelled subsequence $\{u_s\}$ of $\{u_r\}$ as in part (ii). Now suppose either $\{u_s\}$ is not bounded in $\mathcal{H}(\mu)$ or if it is, it only converges weakly to u_0 and not strongly in $\mathcal{H}(\mu)$. This implies that $\|u_0\|_{\mathcal{H}(\mu)} \le \liminf_{s \to 0} \|u_s\|_{\mathcal{H}(\mu)}$ which in turn implies the existence of a sufficiently small s for which $\|u_0\|_{\mathcal{H}(\mu)} \le \|u_s\|_{\mathcal{H}(\mu)}$. Therefore Lemma 7 implies that $\limsup_{s\to 0}\frac{\mu(B_s(u_s))}{\mu(B_s(u_0))} \le 1$. On the other hand, by the definition of u_s we have that $\mu(B_s(u_s)) \ge \mu(B_s(u_0))$ and so $\liminf_{s\to 0}\frac{\mu(B_s(u_s))}{\mu(B_s(u_0))} \ge 1$, from which we conclude that

$$\lim_{s\to 0}\frac{\mu(B_s(u_s))}{\mu(B_s(u_0))} = 1. \tag{B7}$$

By Definition 3 it follows that u_0 is a conditional mode. It remains to consider the setting where $\{u_s\}$ converges strongly to u_0 in $\mathcal{H}(\mu)$. Then by the Cameron–Martin formula we have

$$\frac{\mu(B_s(u_s))}{\mu(B_s(u_0))} = \exp\left(\frac{1}{2}\|u_0\|^2_{\mathcal{H}(\mu)} - \frac{1}{2}\|u_s\|^2_{\mathcal{H}(\mu)}\right)$$
$$\frac{\int_{B_s(0)}\exp\left(-\langle u_s, v\rangle_{\mathcal{H}(\mu)}\right)\mu(\mathrm{d}v)}{\int_{B_s(0)}\exp\left(-\langle u_0, v\rangle_{\mathcal{H}(\mu)}\right)\mu(\mathrm{d}v)}.$$

It follows, from Bogachev (1998, Lem. 4.7.1; see also proof of Lem. 4.7.2), that the maps

$$u \mapsto \mu(B_s(0))^{-1}\int_{B_s(0)}\exp\left(-\langle u, v\rangle_{\mathcal{H}(\mu)}\right)\mu(\mathrm{d}v),$$

are locally Lipschitz on $\mathcal{H}(\mu)$ from which we infer (B7) once again.

We now show that u_0 solves (16). Suppose otherwise, so that $\|u_0\|_{\mathcal{H}(\mu)} - \|u^y\|_{\mathcal{H}(\mu)} > 0$. By Lemma 7 we have that

$$\frac{\mu(B_s(u_0))}{\mu(B_s(u^y))} \le K(s)\exp\left(\frac{1}{2}\|u^y\|^2_{\mathcal{H}(\mu)} - \frac{1}{2}\|u_0\|^2_{\mathcal{H}(\mu)}\right),$$

with $K(s) \to 1$ as $s \to 0$. Now choose $\tilde{s}$ sufficiently small so that

$$1 \le K(s) < \exp\left(\frac{1}{2}\|u_0\|^2_{\mathcal{H}(\mu)} - \frac{1}{2}\|u^y\|^2_{\mathcal{H}(\mu)}\right)$$

for any $s < \tilde{s}$. Then by the above display we have

$$\frac{\mu(B_s(u_0))}{\mu(B_s(u^y))} < 1.$$

Using this bound and (B7) we can then write

$$\limsup_{s\to 0}\frac{\mu(B_s(u_s))}{\mu(B_s(u^y))} = \limsup_{s\to 0}\frac{\mu(B_s(u_s))}{\mu(B_s(u_0))}\frac{\mu(B_s(u_0))}{\mu(B_s(u^y))}$$
$$< \limsup_{s\to 0}\frac{\mu(B_s(u_s))}{\mu(B_s(u_0))} \le 1,$$

which is a contradiction since by the definition of u_s we have $\mu(B_s(u_s)) \ge \mu(B_s(u^y))$ for any $s > 0$. Thus u_0 solves (16). □

Proof of Theorem 7 First let u^y be a conditional mode and take the sequence $\{u_r\}_{r\ge 0}$ as in (B3). By Proposition 12 there exists a relabelled subsequence $\{u_s\}$ which converges strongly in $\mathcal{X}$ to $u_0 \in \mathcal{H}(\mu) \cap T^{-1}(y)$ and u_0 is also a conditional mode and so by Definition 3

it holds that $\lim_{r\to 0}\frac{\mu(B_r(u_r))}{\mu(B_r(u_0))} = 1$. Since u^y is also a conditional mode we have

$$\lim_{r\to 0}\frac{\mu(B_r(u^y))}{\mu(B_r(u_0))} = \lim_{r\to 0}\frac{\mu(B_r(u^y))}{\mu(B_r(u_r))}\lim_{r\to 0}\frac{\mu(B_r(u_r))}{\mu(B_r(u_0))} = 1.$$

We infer from Lemma 4 that $u^y \in \mathcal{H}(\mu) \cap T^{-1}(y)$ since otherwise the limit $\lim_{r\to 0}\frac{\mu(B_r(u^y))}{\mu(B_r(u_r))}$ would vanish. Now suppose u^y does not solve (16). We can obtain a contradiction by repeating the last step of the proof of Proposition 12.

To prove the converse statement let u^y be a solution of (16) with u_0 defined as before. Then Lemma 7 implies $\lim_{r\to 0}\frac{\mu(B_r(u_0))}{\mu(B_r(u^y))} = 1$, and so we have

$$\lim_{r\to 0}\frac{\mu(B_r(u_r))}{\mu(B_r(u^y))} = \lim_{r\to 0}\frac{\mu(B_r(u_r))}{\mu(B_r(u_0))}\lim_{r\to 0}\frac{\mu(B_r(u_0))}{\mu(B_r(u^y))} = 1.$$

The result follows from Definition 3. □

Acknowledgements All four authors acknowledge support from the Air Force Office of Scientific Research under MURI award number FA9550-20-1-0358 (Machine Learning and Physics-Based Modeling and Simulation). Additionally HO acknowledges support by the Department of Energy under award number DE-SC0023163 (SEA-CROGS: Scalable, Efficient and Accelerated Causal Reasoning Operators, Graphs and Spikes for Earth and Embedded Systems) and a Department of Defense Vannevar Bush Faculty Fellowship. The work of AMS is also supported by a Department of Defense Vannevar Bush Faculty Fellowship. BH is also supported by the National Science Foundation grant DMS-2208535 (Machine Learning for Bayesian Inverse Problems). YC is also supported by a Courant Instructorship.

Author Contributions All authors contributed equally to the manuscript.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

Agapiou, S., Burger, M., Dashti, M., Helin, T.: Sparsity-promoting and edge-preserving maximum a posteriori estimators in non-parametric Bayesian inverse problems. Inverse Probl. **34**(4), 045002 (2018)

Ayanbayev, B., Klebanov, I., Lie, H.C., Sullivan, T.: Γ-convergence of Onsager–Machlup functionals: I. With applications to maximum a posteriori estimation in Bayesian inverse problems. Inverse Probl. **38**(2), 025005 (2021a)

Ayanbayev, B., Klebanov, I., Lie, H.C., Sullivan, T.J.: Γ-convergence of Onsager–Machlup functionals: II. Infinite product measures on Banach spaces. Inverse Probl. **38**(2), 025006 (2021b)

Batlle, P., Darcy, M., Hosseini, B., Owhadi, H.: Kernel methods are competitive for operator learning. J. Comput. Phys. **496**, 112549 (2024)

Bertozzi, A.L., Luo, X., Stuart, A.M., Zygalakis, K.C.: Uncertainty quantification in graph-based classification of high dimensional data. SIAM/ASA J. Uncertain. Quantif. **6**(2), 568–595 (2018)

Beskos, A., Pinski, F.J., Sanz-Serna, J.M., Stuart, A.M.: Hybrid Monte Carlo on Hilbert spaces. Stoch. Process. Appl. **121**(10), 2201–2230 (2011)

Beskos, A., Girolami, M., Lan, S., Farrell, P.E., Stuart, A.M.: Geometric mcmc for infinite-dimensional inverse problems. J. Comput. Phys. **335**, 327–351 (2017)

Betz, W., Papaioannou, I., Straub, D.: Numerical methods for the discretization of random fields by means of the Karhunen–Loève expansion. Comput. Methods Appl. Mech. Eng. **271**, 109–129 (2014)

Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: a review for statisticians. J. Am. Stat. Assoc. **112**(518), 859–877 (2017)

Bogachev, V.I.: Gaussian Measures. Amer. Math. Soc. Volume 62 of Mathematical Surveys and Monographs (1998)

Bogachev, V.I.: Measure Theory vol. 2, Springer (2007)

Bourdais, T., Batlle, P., Yang, X., Baptista, R., Rouquette, N., Owhadi, H.: Codiscovering graphical structure and functional relationships within data: a Gaussian process framework for connecting the dots. Proc. Natl. Acad. Sci. **121**(32), 2403449121 (2024)

Casale, F.P., Dalca, A., Saglietti, L., Listgarten, J., Fusi, N.: Gaussian process prior variational autoencoders. Adv. Neural Inf. Process. Syst. **31** (2018)

Chen, Y., Hosseini, B., Owhadi, H., Stuart, A.M.: Solving and learning nonlinear PDEs with Gaussian processes. J. Comput. Phys. **447**, 110668 (2021a)

Chen, Y., Owhadi, H., Stuart, A.: Consistency of empirical Bayes and kernel flow for hierarchical parameter estimation. Math. Comput. **90**(332), 2527–2578 (2021b)

Chen, Y., Owhadi, H., Schäfer, F.: Sparse Cholesky factorization for solving nonlinear pdes via Gaussian processes. Math. Comput. (2024)

Chkrebtii, O.A., Campbell, D.A., Calderhead, B., Girolami, M.A.: Bayesian solution uncertainty quantification for differential equations. Bayesian Anal. **11**(4), 1239–1267 (2016)

Clason, C., Helin, T., Kretschmann, R., Piiroinen, P.: Generalized modes in Bayesian inverse problems. SIAM/ASA J. Uncertain. Quantif. **7**(2), 652–684 (2019)

Cockayne, J., Oates, C., Sullivan, T., Girolami, M.: Probabilistic numerical methods for pde-constrained bayesian inverse problems. In: AIP Conference Proceedings, vol. 1853. AIP Publishing (2017)

Cockayne, J., Oates, C.J., Sullivan, T.J., Girolami, M.: Bayesian probabilistic numerical methods. SIAM Rev. **61**(4), 756–789 (2019)

Cotter, S.L., Dashti, M., Robinson, J.C., Stuart, A.M.: Bayesian inverse problems for functions and applications to fluid mechanics. Inverse Probl. **25**(11), 115008 (2009)

Cotter, S., Roberts, G., Stuart, A., White, D.: Mcmc methods for functions: modifying old algorithms to make them faster. Stat. Sci. **28**(3), 424 (2013)

Cressie, N.: The origins of kriging. Math. Geol. **22**, 239–252 (1990)

Cui, T., Law, K.J., Marzouk, Y.M.: Dimension-independent likelihood-informed mcmc. J. Comput. Phys. **304**, 109–137 (2016)

Damianou, A., Lawrence, N.D.: Deep Gaussian processes. In: Carlos M. Carvalho and Pradeep Ravikumar (eds.) Artificial Intelligence and Statistics, pp. 207–215 (2013)

Dashti, M., Stuart, A.M.: The Bayesian approach to inverse problems. In: Roger Ghanem, David Higdon, and Houman Owhadi (eds.) Handbook of Uncertainty Quantification, pp. 311–428. Springer (2017)

Dashti, M., Harris, S., Stuart, A.: Besov priors for Bayesian inverse problems. Inverse Probl. Imaging **6**(2), 183–200 (2012)

Dashti, M., Law, K.J., Stuart, A.M., Voss, J.: MAP estimators and their consistency in Bayesian nonparametric inverse problems. Inverse Probl. **29**(9), 095017 (2013)

Diaconis, P.: Bayesian numerical analysis. In: Statistical Decision Theory and Related Topics, IV, Vol. 1 (West Lafayette, Ind., 1986), pp. 163–175 (1988)

Doucet, A., De Freitas, N., Gordon, N.: An introduction to sequential Monte Carlo methods. Sequential Monte Carlo methods in practice, pp. 3–14 (2001)

Duchon, J.: Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In: Constructive Theory of Functions of Several Variables (Proceedings of Conference, Mathematical Research Institute, Oberwolfach, 1976), pp. 85–100571. Springer, Berlin (1977)

Dunlop, M.M., Girolami, M.A., Stuart, A.M., Teckentrup, A.L.: How deep are deep Gaussian processes? J. Mach. Learn. Res. **19**(54), 1–46 (2018)

Dutordoir, V., Hensman, J., Wilk, M., Ek, C.H., Ghahramani, Z., Durrande, N.: Deep neural networks as point estimates for deep Gaussian processes. Adv. Neural. Inf. Process. Syst. **34**, 9443–9455 (2021)

Fortuin, V., Baranchuk, D., Rätsch, G., Mandt, S.: Gp-vae: deep probabilistic time series imputation. In: International Conference on Artificial Intelligence and Statistics, pp. 1651–1661. PMLR (2020)

Franklin, J.N.: Well-posed stochastic extensions of ill-posed linear problems. J. Math. Anal. Appl. **31**(3), 682–716 (1970)

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis, third edition, CRC Press (2013)

Giné, E., Nickl, R.: Mathematical Foundations of Infinite-dimensional Statistical Models Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press (2021)

Hairer, M.: An introduction to stochastic PDEs (2009). arXiv:0907.4178

Harder, R.L., Desmarais, R.N.: Interpolation using surface splines. J. Aircr. **9**, 189–191 (1972)

Helin, T., Burger, M.: Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems. Inverse Probl. **31**(8), 085009 (2015)

Hennig, P., Osborne, M.A., Girolami, M.: Probabilistic numerics and uncertainty in computations. Proc. R. Soc. A: Math. Phys. Eng. Sci. **471**(2179), 20150142 (2015)

Hosseini, B.: Well-posed bayesian inverse problems with infinitely divisible and heavy-tailed prior measures. SIAM/ASA J. Uncertain. Quantif. **5**(1), 1024–1060 (2017)

Hosseini, B.: Two Metropolis-Hastings algorithms for posterior measures with non-Gaussian priors in infinite dimensions. SIAM/ASA J. Uncertain. Quantif. **7**(4), 1185–1223 (2019)

Hosseini, B., Nigam, N.: Well-posed bayesian inverse problems: priors with exponential tails. SIAM/ASA J. Uncertain. Quantif. **5**(1), 436–465 (2017)

Ikeda, N., Watanabe, S.: Stochastic Differential Equations and Diffusion Processes, North-Holland Mathematical Library, North-Holland Publishing Company (2014)

Jakkala, K.: Deep Gaussian processes: a survey (2021). arXiv:2106.12135

Janson, S.: Gaussian Hilbert Spaces Cambridge Tracts in Mathematics, Cambridge University Press, (1997)

Kaipio, J., Somersalo, E.: Statistical and Computational Inverse Problems. Springer, Berlin (2006)

Kanagawa, M., Hennig, P., Sejdinovic, D., Sriperumbudur, B.K.: Gaussian processes and kernel methods: a review on connections and equivalences (2018). arXiv:1807.02582

Kass, R.E., Tierney, L., Kadane, J.B.: Laplace's method in bayesian analysis. Contemp. Math. **115**, 89–99 (1991)

Kimeldorf, G.S., Wahba, G.: A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. Ann. Math. Stat. **41**, 495–502 (1970)

Lambley, H., Sullivan, T.J.: An order-theoretic perspective on modes and maximum a posteriori estimation in bayesian inverse problems. SIAM/ASA J. Uncertain. Quantif. **11**(4), 1195–1224 (2023)

Larkin, F.M.: Gaussian measure in Hilbert space and applications in numerical analysis. J. Math. **2**(3), 379–421 (1972)

Latz, J.: On the well-posedness of bayesian inverse problems. SIAM/ASA J. Uncertain. Quantif. **8**(1), 451–482 (2020)

Long, D., Wang, Z., Krishnapriyan, A., Kirby, R., Zhe, S., Mahoney, M.: Autoip: a united framework to integrate physics into Gaussian processes. In: International Conference on Machine Learning, pp. 14210–14222. PMLR (2022)

Maniglia, S., Rhandi, A.: Gaussian measures on separable Hilbert spaces and applications. Quaderni di Matematica **2004**(1) (2004)

Micchelli, C.A., Rivlin, T.J.: A survey of optimal recovery. In: Charles A Micchelli, Theodore J Rivlin (eds.) Optimal Estimation in Approximation Theory, pp. 1–54. Springer (1977)

Murray, I., MacKay, D., Adams, R.P.: The Gaussian process density sampler. Adv. Neural Inf. Process. Syst. **21** (2008)

Neumann, J.: Zur Theorie der Gesellschaftsspiele. Math. Ann. **100**(1), 295–320 (1928)

Owhadi, H.: Bayesian numerical homogenization. Multiscale Model. Simul. **13**(3), 812–828 (2015)

Owhadi, H.: Computational graph completion. Res. Math. Sci. **9**(2), 27 (2022)

Owhadi, H.: Do ideas have shape? Idea registration as the continuous limit of artificial neural networks. Phys. D **444**, 133592 (2023)

Owhadi, H., Scovel, C.: Conditioning Gaussian measure on Hilbert space. J. Math. Stat. Anal. **1**(1) (2018). arXiv:1506.04208

Owhadi, H., Scovel, C.: "Operator-Adapted Wavelets, Fast Solvers, and Numerical Homogenization". Cambridge Monographs on Applied and computational Mathematics No 35, Cambridge University Press (2019)

Owhadi, H., Scovel, C., Schäfer, F.: Statistical numerical approximation. Notices AMS **66**, 1 (2019)

Palasti, I., Renyi, A.: On interpolation theory and the theory of games. MTA Mat. Kat. Int. Kozl **1**, 529–540 (1956)

Pandey, B., Hosseini, B., Batlle, P., Owhadi, H.: Diffeomorphic measure matching with kernels for generative modeling (2024). arXiv:2402.08077

Pinski, F.J., Simpson, G., Stuart, A.M., Weber, H.: Kullback–Leibler approximation for probability measures on infinite dimensional spaces. SIAM J. Math. Anal. **47**(6), 4091–4122 (2015)

Poincaré, H.: Calcul des Probabilités (1896)

Raissi, M., Perdikaris, P., Karniadakis, G.E.: Inferring solutions of differential equations using noisy multi-fidelity data. J. Comput. Phys. **335**, 736–746 (2017)

Raissi, M., Perdikaris, P., Karniadakis, G.E.: Numerical Gaussian processes for time-dependent and nonlinear partial differential equations. SIAM J. Sci. Comput. **40**(1), 172–198 (2018)

Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning, Adaptive Computation and Machine Learning Series, MIT Press Direct (2007)

Robert, C.P., Casella, G.: Monte Carlo Statistical Methods, vol. 2. Springer, Berlin (1999)

Sard, A.: Linear Approximation, Mathematical Surveys and Monographs. Amer. Math. Soc. 9, (1963)

Särkkä, S.: Linear operators and stochastic partial differential equations in gaussian process regression. In: Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14–17, 2011, Proceedings, Part II 21, pp. 151–158. Springer (2011)

Schäfer, F., Sullivan, T.J., Owhadi, H.: Compression, inversion, and approximate pca of dense kernel matrices at near-linear computational complexity. Multiscale Model. Simul. **19**(2), 688–730 (2021a)

Schäfer, F., Katzfuss, M., Owhadi, H.: Sparse Cholesky factorization by Kullback-Leibler minimization. SIAM J. Sci. Comput. **43**(3), 2019–2046 (2021b)

Skilling, J.: Bayesian solution of ordinary differential equations. In: C. Ray Smith, Gary J. Erickson, Paul (eds.) Maximum Entropy and Bayesian Methods, pp. 23–37. Springer (1992)

Smola, A.J., Schölkopf, B.: "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond", Adaptive Computation and Machine Learning Series, MIT Press Direct (1998)

Snelson, E., Ghahramani, Z.: Local and global sparse gaussian process approximations. In: Marina Meila and Xiaotong Shen (eds.) Artificial Intelligence and Statistics, pp. 524–531. PMLR (2007)

Sprungk, B.: On the local lipschitz stability of bayesian inverse problems. Inverse Probl. **36**(5), 055015 (2020)

Stuart, A.M.: Inverse problems: a Bayesian perspective. Acta Numer. **19**, 451–559 (2010)

Sul'din, A.V.: Wiener measure and its applications to approximation methods. I. Izvestiya Vysshikh Uchebnykh Zavedenii. Matematika **6**, 145–158 (1959)

Sullivan, T.: Well-posed bayesian inverse problems and heavy-tailed stable quasi-banach space priors. Inverse Probl. Imaging **11**(5), 857–874 (2017)

Swiler, L.P., Gulian, M., Frankel, A.L., Safta, C., Jakeman, J.D.: A survey of constrained Gaussian process regression: approaches and implementation challenges. J. Mach. Learn. Model. Comput. **1**(2), 119–156 (2020)

Tarantola, A.: Inverse Problem Theory and Methods for Model Parameter Estimation. Soc Indus Appl Math (2005)

Tierney, L.: A note on Metropolis-Hastings kernels for general state spaces. Ann. Appl. Probab. **8**, 1–9 (1998)

Traub, J.F., Wasilkowski, G.W., Woźniakowski, H.: Information-Based Complexity. Academic Press (1988)

Vaart, A.W., Zanten, J.H., et al.: Reproducing kernel Hilbert spaces of gaussian priors. IMS Collect. **3**, 200–222 (2008)

Vadeboncoeur, A., Akyildiz, Ö.D., Kazlauskaite, I., Girolami, M., Cirak, F.: Fully probabilistic deep models for forward and inverse problems in parametric pdes. J. Comput. Phys. **491**, 112369 (2023)

Wald, A.: Statistical decision functions which minimize the maximum risk. Ann. Math. **2**(46), 265–280 (1945)

Wang, J., Cockayne, J., Chkrebtii, O., Sullivan, T.J., Oates, C.J.: Bayesian numerical methods for nonlinear partial differential equations. Stat. Comput. **31**, 1–20 (2021)

Wendland, H.: Scattered Data Approximation, Cambridge Monographs on Applied and Computational Mathematics No 17, Cambridge University Press (2004)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.