Introduction To Gaussian Process Regression In Bayesian Inverse Problems, With New Results On Experimental Design For Weighted Error Measures

T. Helin, A.M. Stuart, A.L. Teckentrup, K.C. Zygalakis

Abstract Bayesian posterior distributions arising in modern applications are often computationally intractable due to the large computational cost of evaluating the data likelihood. Examples include inverse problems in partial differential equation models arising in climate modeling and in subsurface fluid flow. To alleviate the problem of expensive likelihood evaluation, a natural approach is to use Gaussian process regression to build a surrogate model for the likelihood, resulting in an approximate posterior distribution that is amenable to computations in practice. This paper serves as an introduction to Gaussian process regression, in particular in the context of building surrogate models for inverse problems; we also present new insights into a suitable choice of training points, motivated by the use of Gaussian processes in approximate Bayesian inversion. We show that the error between the true and approximate posterior distribution can be bounded by the error between the true and approximate likelihood, measured in the L^2-norm weighted by the true posterior; furthermore we show that minimizing the error between the true and approximate likelihood in this norm suggests choosing the training points in the Gaussian process surrogate model based on the true posterior.

Tapio Helin

LUT University, School of Engineering, Yliopistokatu 4, P.O. Box 20, FI-53851 Lappeenranta, Finland , e-mail: tapio.helin@lut.fi

Andrew M Stuart

California Institute of Technology, Computing and Mathematical Sciences, 1200 E. California Blvd., MC 305-16, Pasadena, CA 91125, USA e-mail: astuart@caltech.edu

Aretha L Teckentrup

University of Edinburgh, School of Mathematics, James Clerk Maxwell Building, Peter Guthrie Tait road, EH9 3FD, Scotland e-mail: a.teckentrup@ed.ac.uk

Konstantinos C Zygalakis

University of Edinburgh, School of Mathemats, James Clerk Maxwell Building, Peter Guthrie Tait road, EH9 3FD, Scotland e-mail: k.zygalakis@ed.ac.uk

1 Introduction

Parameters in mathematical models in science and engineering are often not fully known and have to be estimated from observed data. Accurate reconstruction of the parameters, as well as an estimate of the uncertainty in the reconstruction, are crucial for reliable predictions and risk assessments. The recent explosion of available data, driven by the increase in large-scale scientific experiments and the development of sensor technology, means that there is a pressing need to develop new algorithms for the seamless integration of observed data with sophisticated mathematical models.

The complexity of models involved in modern applications, such as those based on partial differential equations, result in a huge computational cost and render many methods for solving inverse problems practically infeasible. A widespread solution to this problem is to approximate the model by a computationally cheaper surrogate model to obtain an approximate likelihood that is much faster to compute, feasible to use for simulations in practice and ideally is accurate where it needs to be accurate for solution of the parameter learning problem of interest [19, 25, 4, 24, 11, 42].

This paper discusses the use of surrogate models in the Bayesian approach to inverse problems, in which we find the *posterior* distribution of the unknown parameters conditioned on the observed data. A particular focus is on random surrogate models and Gaussian process regression. We provide new results that show that the context of the surrogate model, i.e. the Bayesian posterior distribution, should be taken into account when designing the surrogate model, by choosing training points in the area of parameter space where the posterior places significant mass. This poses a somewhat circular problem, since the reason we are using Gaussian process regression is to be able to compute the posterior. However, various computational approaches have been suggested to circumvent this problem in practice, including a sequential design strategy that only requires access to the approximate posterior [40], and the updating of training points while exploring the posterior with sampling methods such as Markov chain Monte Carlo [12, 6].

The structure of the remainder of the paper is the following. In section 2, we introduce Bayesian posterior distributions in inverse problems, discuss computational approximations using surrogate models, and prove new error bounds between the true and approximate posterior distributions in a norm weighted by the posterior. The main novel results are Theorems 1 and 2. In section 3, we introduce Gaussian process regression as surrogate models, and prove new results on the accuracy of Gaussian process regression and suitable choices of training points in the context of inverse problems. The main new results are Corollary 1, Corollary 2 and Theorem 3. Section 4 gives some simple numerical examples illustrating the theoretical findings.

2 Bayesian Inverse Problems and Their Approximation

In this section, we set up the Bayesian inverse problem and describe its approximation using random surrogate models. In subsection 2.1 we introduce the framework for

Bayesian inversion, and discuss MCMC methods for them. Subsection 2.2 contains two large-scale examples that motivate the need for surrogate modeling. In subsection 2.3 we explain surrogate modeling in detail leading, in subsection 2.4, to error estimates summarizing the effect of errors in the surrogate model on the solution of the Bayesian inverse problem.

2.1 Bayesian Inverse Problems

We are interested in solving the inverse problem of determining an unknown parameter $u \in U$ from noisy, indirect data $y \in \mathbb{R}^{d_y}$ given by

$$y = \mathcal{G}(u) + \eta, \tag{1}$$

for some observation operator $\mathcal{G} : U \to \mathbb{R}^{d_y}$. For ease of presentation we assume the noise η is a realization of the $\mathbb{R}^{d_y}$-valued Gaussian random variable $\mathcal{N}(0, \Gamma)$, for a known positive definite covariance matrix Γ, but other distributions on η can be dealt with similarly [23].

We adopt a Bayesian statistical perspective, in which the pair (u, y) is treated as a random variable (u in finite dimensions) or a random process (u in infinite dimensions). The aim is to find the distribution of the conditional random variable $u|y$. This approach leads to a well-posed problem in the sense of Hadamard: there exists a unique conditional distribution $u|y$ that depends continuously on y [41, 20].

In the absence of data, we assume u is distributed according to a prior measure μ_0. Equation (1) then defines the conditional distribution of $y|u$, assuming that η is independent of u. Suitable choices of prior measure will depend on the application. We are then interested in the posterior distribution μ^y on the conditioned random variable $u|y$, which can be characterized as follows through Bayes' Theorem. This delivers the the Radon-Nikodym derivative of the posterior with respect to the prior distribution (see e.g. [41]).

Proposition 1 *Suppose U is a separable Banach space, $\mathcal{G} : U \to \mathbb{R}^{d_y}$ is continuous and $\mu_0(U) = 1$. Then the posterior distribution μ^y on the conditioned random variable $u|y$ is absolutely continuous with respect to μ_0 and given by Bayes' Theorem:*

$$\frac{d\mu^y}{d\mu_0}(u) = \frac{1}{Z} \exp\big(-\Phi(u)\big),$$

where

$$\Phi(u) = \frac{1}{2} \|y - \mathcal{G}(u)\|_\Gamma^2, \qquad and \qquad Z = \mathbb{E}_{\mu_0}\Big(\exp\big(-\Phi(u)\big)\Big).$$

In the preceding we adopt the notational convention $\|\cdot\|_A = \|A^{-\frac{1}{2}} \cdot\|_2$ as in [41], where $\|\cdot\|_2$ is the Euclidean norm and A is any symmetric positive matrix. In a finite-dimensional setting, where $u \in U \subseteq \mathbb{R}^{d_u}$ and we are inferring a finite number

of unknown parameters, Bayes' Theorem can be written in terms of the probability density function (pdf) of the prior and posterior, denoted by π_0 and π^y, respectively [17]. This takes the form

$$\pi^y(u) = \frac{1}{Z} \exp\left(-\Phi(u)\right)\pi_0(u),$$

with Φ and Z as defined in Proposition 1.

The term $\exp\left(-\Phi(u)\right)$ is referred to as the *data likelihood*, and comes from the distribution of $y|u$. In other words, it characterizes how likely it is to observe the data y given a particular choice of the parameter u. Since $y = \mathcal{G}(u) + \eta$ and $\eta \sim \mathrm{N}(0, \Gamma)$, we have $y|u \sim \mathrm{N}(\mathcal{G}(u), \Gamma)$, and the pdf of $y|u$ is hence proportional to $\exp\left(-\Phi(u)\right)$. The normalization constant Z ensures that the posterior μ^y is a probability distribution, with $\mu^y(U) = 1$. By Bayes' Theorem, $Z = Z(y)$ is the marginal pdf of the data y, and hence characterizes how likely it is to observe the data y given the observational model $\mathcal{G}$. Z is therefore often referred to as the *model evidence*. The analytical value of Z is usually not known, and computing Z numerically is notoriously difficult (see e.g. [47]).

In applications, the goal is usually to compute a quantity of interest related to the posterior distribution μ^y. This could for example be an expected value $\mathbb{E}_{\mu^y}[g(u)]$, where $g : U \to U$ is chosen as the identity for the conditional mean $\mathbb{E}[u|y] = \mathbb{E}_{\mu^y}[u]$ or $g : U \to \{0, 1\}$ is chosen as the indicator function $\mathrm{I}_{u \in A}$ for computing event probabilities $\mathbb{P}[u|y \in A] = \mathbb{E}_{\mu^y}[\mathrm{I}_{u \in A}]$. The method of choice for sampling from the posterior distribution, enabling the computation of expected values and other quantities of interest, is often Markov chain Monte Carlo (MCMC) [14, 27, 35, 9, 13, 10]. A prototypical example is the Metropolis-Hastings algorithm, which consists of the following steps for sampling from a target density π on a finite-dimensional parameter space $U \subseteq \mathbb{R}^{d_u}$:

1. Choose $u^{(1)}$ with $\pi(u^{(1)}) > 0$.
2. At state $u^{(i)}$, sample a proposal u' from density $q(u' \mid u^{(i)})$.
3. Accept sample u' with probability

$$\alpha(u' \mid u^{(i)}) = \min\left(1, \frac{\pi(u')\, q(u^{(i)} \mid u')}{\pi(u^{(i)})\, q(u' \mid u^{(i)})}\right),$$

i.e. $u^{(i+1)} = u'$ with probability $\alpha(u' \mid u^{(i)})$; otherwise stay at $u^{(i+1)} = u^{(i)}$.

Steps 2 and 3 are repeated until the required number of samples have been generated. In the context of inverse problems and Bayesian posterior distributions, we note in particular that knowledge of the normalization constant of the target density π is not required, since this cancels in the ratio in α. The crucial ingredient in the algorithm is the *proposal density q* in step 2, and a wide range of options exists, from simple random walks to methods using (higher-order) derivatives of the target and the geometry of the parameter space (see e.g. [35, 10, 13]). There is a rich theory underpinning the Metropolis-Hastings algorithm, and in particular, it is guaranteed under mild assumptions that the distribution of $u^{(i)}$ converges to the target density π as $i \to \infty$ (see e.g. [35]).

MCMC methods typically require repeated evaluation of the data likelihood. In the Metropolis-Hastings algorithm above; this can be seen in step 3, where $\exp\left(-\Phi(u')\right)$ needs to be computed for every proposal u'. This quickly becomes infeasible in modern applications where the computation of the likelihood is very costly. This includes for example inverse problems in geophysics (subsurface flow model) and climate (general circulation model), where $\mathcal{G}$ involves the solution of one or more coupled partial differential equations (PDEs). Two such examples are given in the next subsection.

2.2 Examples of Large-Scale Complex Bayesian Inverse Problems

Subsurface Flow Model

A simple model for stationary subsurface fluid flow is given by

$$-\nabla \cdot (k(x)\nabla p(x)) = g(x), \qquad x \in D, \tag{2}$$

where k represents the permeability (or hydraulic conductivity) of the subsurface, p denotes the pressure head of the fluid, g incorporates any sources or sinks, and suitable boundary conditions are imposed on the computational domain D [46, 36]. This model comes from a combination of Darcy's law for single-phase fluid flow in a porous medium, $q(x) = -k(x)\nabla p(x)$, and conservation of mass, $\nabla \cdot q = g$.

A typical inverse problem in this context is to reconstruct the permeability k given noisy measurements of the pressure head $y = \{p(x_i) + \eta_i\}_{i=1}^{d_y}$ or the Darcy flux $y = \{k(x_i)\nabla p(x_i) + \eta_i\}_{i=1}^{d_y}$ at fixed locations $x_i \in D$. Although we in general wish to reconstruct the function $k \in L^2(D)$, we often choose a parametrization of k in computations. The inverse problem then becomes that of learning the coefficients $u \in \mathbb{R}^{d_u}$ in the parametrization. For example, we can choose a piecewise constant model

$$k(x; u) = \sum_{j=1}^{d_u} u_j I_{x \in D_j}(x),$$

where D_j is a partitioning of the domain D representing layers of different types of rock, and u_j is the value of the permeability k in the layer D_j. Since the permeability is always positive, and can vary over orders of magnitude between different types of rock, the prior distribution μ_0 on $u = \{u_j\}_{j=1}^{d_u}$ should reflect these properties, and log-normal distributions are often used.

Evaluation of the parameter-to-observation map $\mathcal{G} : \mathbb{R}_{>0}^{d_u} \to \mathbb{R}^{d_y}$, defined by $\mathcal{G}(u) = \{p(x_i; u)\}_{i=1}^{d_y}$ or $\mathcal{G}(u) = \{k(x_i; u)\nabla p(x_i; u)\}_{i=1}^{d_y}$, requires the solution of the PDE (2) with a different coefficient k for every u and is thus typically very costly.

General Circulation Model

General Circulation Models (GCMs) typically comprise discretization of a system of three space-dimensional time-dependent PDEs, with spatial dimensions representing Earth's longitude, latitude and vertical dimensions. Unknown parameters appear in sub-grid scale models which attempt to capture unresolvable physics on the scale of clouds. These parameters can in principle be learned from data in the form of time-averaged satellite measurements of the Earth's atmosphere. Uncertainty quantification, and hence the Bayesian approach, is crucial as estimation (and reduction) of uncertainty in climate models is a central goal of modern climate modeling research [38]. In this paper we will consider a specific GCM from [31]. The conservation laws, which are discretized to form the computational model, include two equations which take the form

$$\text{Moisture Conservation:} \quad \frac{\partial q}{\partial t} + v \cdot \nabla q = -\frac{q - q_{\text{ref}}(T; \theta)}{\tau_q(q, T; \theta)}$$

$$\text{Energy Conservation:} \quad \frac{\partial T}{\partial t} + v \cdot \nabla T = \frac{T - T_{\text{ref}}(q, T; \theta)}{\tau_T(q, T; \theta)} + \cdots,$$

coupled to further conservation laws for mass and momentum. The model includes two unknown parameters: θ_{RH}, the reference relative humidity; and θ_τ a relaxation timescale; the functional forms of q_{ref}, T_{ref}, τ_q and τ_T are known. The unknown parameter is $\theta = (\theta_{RH}, \theta_\tau)$. The inverse problem is to learn about θ, and uncertainty in θ, from climate statistics; the paper [12] employs 30-day averages of the free-tropospheric relative humidity, of the precipitation rate, and of a measure of the frequency of extreme precipitation. The parameter-to-data map thus requires simulation of the discretized conservation laws. This leads to $O(10^2)$ noisy and indirect observations from which to learn about the 2 unknowns. Similar, related inverse problems may be found in [15][subsection 5.10], in which the number of unknown parameters is 4.

2.3 Surrogate Models

To make computations feasible in practice, it is common to approximate the log-likelihood Φ, or the forward model $\mathcal{G}$, by a surrogate model, also known as an emulator, meta-model or reduced order model. A wide range of suitable surrogate models exist, see e.g. [42, 23] and the references therein. Recently, a particular focus has been on the use of *random* surrogate models. These have been shown to lead to better approximations of the posterior distribution when the error in the surrogate model is large compared to the noise in the observations [8, 2], and to improve the statistical efficiency of MCMC methods [11]. Examples of random surrogate models are methods from the field of probabilistic numerics [8, 22], randomized projection methods [21, 29], and Gaussian process emulators [19, 42].

Given a random surrogate model for the log-likelihood Φ or the forward model $\mathcal{G}$, there are then different ways of using this to approximate the posterior distribution μ^y. We will henceforth fix our notation to using a surrogate model for Φ, since using a surrogate model for $\mathcal{G}$ can be translated into a surrogate model for Φ using the specific form of the likelihood. Let us denote the random surrogate model by Φ_N, where N denotes a discretization parameter such as the number of training points in a Gaussian process emulator or the number of grid points used in a probabilistic PDE solver. Using the mean $\mathbb{E}[\Phi_N]$ of the surrogate model to approximate Φ, we obtain the *mean-based approximation*

$$\frac{d\mu_{\text{mean}}^{y,N}}{d\mu_0}(u) = \frac{1}{Z_N^{\text{mean}}} \exp\left(-\mathbb{E}[\Phi_N(u)]\right), \tag{3}$$

$$Z_N^{\text{mean}} = \mathbb{E}_{\mu_0}\left(\exp\left(-\mathbb{E}[\Phi_N(u)]\right)\right).$$

Alternatively, we can use the random surrogate Φ_N directly to approximate Φ, and take the expected value of the resulting random approximation of the likelihood. This gives the *marginal approximation*

$$\frac{d\mu_{\text{marginal}}^{y,N}}{d\mu_0}(u) = \frac{1}{Z_N^{\text{marginal}}} \mathbb{E}\left(\exp\left(-\Phi_N(u)\right)\right), \tag{4}$$

$$Z_N^{\text{marginal}} = \mathbb{E}_{\mu_0}\left(\mathbb{E}\left(\exp\left(-\Phi_N(u)\right)\right)\right).$$

Intuitively, the marginal approximation introduces additional uncertainty in the approximate posterior distribution, coming from the randomness of the surrogate model. One interpretation of the randomness of our surrogate model is that since we have used a finite amount of information/resources in the construction, there is still some uncertainty (or *error*) associated to it. Including this uncertainty in the approximate posterior distribution then allows us to take into account the error in the surrogate model when inferring the parameter u. In practical applications, where the accuracy of the surrogate model might be limited due to computational resources, the uncertainty (or error) in the surrogate model can be large (or comparable) to the uncertainty present in the observations y, and it is crucial to take this into account to avoid over-confident and biased predictions [8, 2].

In special cases, it can be shown analytically that the marginal approximation results in a form of variance inflation in the likelihood. One such instance is when the forward model $\mathcal{G}$ is approximated by a Gaussian process $\mathcal{G}_N \sim \text{GP}(m_N^{\mathcal{G}}, K_N)$ with mean $m_N^{\mathcal{G}}$ and covariance kernel K_N (see section 3 for more details). The mean-based approximation assumes the observational model $y = m_N^{\mathcal{G}}(u) + \eta$, whereas the marginal approximation uses

$$y = \mathcal{G}_N(u) + \eta = m_N^{\mathcal{G}}(u) + \xi(u) + \eta,$$

with $\xi(u) \sim \mathrm{N}(0, K_N(u, u))$. The resulting approximate likelihoods are proportional to

$$\frac{1}{\sqrt{\det(\Gamma)}} \exp\left(-\frac{||y - m_N^{\mathcal{G}}(u)||_\Gamma^2}{2}\right)$$

for the mean-based approximation, and

$$\frac{1}{\sqrt{\det(\Gamma + K_N(u, u))}} \exp\left(-\frac{||y - m_N^{\mathcal{G}}(u)||_{(\Gamma + K_N(u, u))}^2}{2}\right)$$

for the marginal approximation, see e.g. [5, 7, 2]. The difference between the two likelihoods hence depends on the relationship between Γ, the uncertainty in the observations y, and K_N, the uncertainty in the surrogate model Φ_N. If K_N is small compared to Γ, the two approximate posteriors will be similar.

The approximate posterior distributions can also be motivated from a decision theoretic point of view [40, 16]. Suppose U is a compact subset of $\mathbb{R}^{d_u}$, and we approximate the negative log-likelihood Φ by a Gaussian process $\Phi_N \sim \mathrm{GP}(m_N^\Phi, k_N)$. Then the (un-normalized) marginal approximation minimizes the expected L^2-loss:

$$\mathbb{E}\left(\exp\left(-\Phi_N(u)\right)\right)\pi_0(u) = \underset{f \in L^2(U)}{\arg\min} \; \mathbb{E} \int_U \left(\exp\left(-\Phi_N(u)\right)\pi_0(u) - f(u)\right)^2 du.$$

The (un-normalized) mean-based approximation instead minimizes the expected L^1-loss:

$$\exp\left(-m_N^\Phi(u)\right)\pi_0(u) = \underset{f \in L^1(U)}{\arg\min} \; \mathbb{E} \int_U \left|\exp\left(-\Phi_N(u)\right)\pi_0(u) - f(u)\right| du.$$

Remark 1 (Alternative approximate posteriors) In addition to the approximations described above, we can also consider the *sample-based* approximation

$$\frac{d\mu_{\mathrm{sample}}^{y,N}}{d\mu_0}(u) = \frac{1}{Z_N^{\mathrm{sample}}} \exp\left(-\Phi_N(u)\right),$$

$$Z_N^{\mathrm{sample}} = \mathbb{E}_{\mu_0}\left(\exp\left(-\Phi_N(u)\right)\right),$$

which results in a random approximation of the posterior [42, 23]. If there is considerable uncertainty in the random surrogate model Φ_N, the posterior approximations corresponding to different samples of Φ_N may look very dissimilar.

An alternative marginal approximation can be defined by taking the expected value of the normalized random likelihood defined above,

$$\frac{d\mu_{\mathrm{marginal}'}^{y,N}}{d\mu_0}(u) = \mathbb{E}\left(\frac{1}{Z_N^{\mathrm{sample}}} \exp\left(-\Phi_N(u)\right)\right).$$

However, this approximation is not easily amenable to sampling methods such as (pseudo-marginal) MCMC, since the computation of Z_N^{sample} is intractable. ∎

2.4 Error in the Surrogate-accelerated Posterior Distribution

To justify the use of the approximate posterior distributions $\mu_{\text{mean}}^{y,N}$ and $\mu_{\text{marginal}}^{y,N}$ in practice, we want to show that they converge to the true posterior μ^y as $N \to \infty$. We will measure the error in the Hellinger distance

$$d_{\text{Hell}}(\mu^y, \mu^{y,N}) = \left(\frac{1}{2} \int_U \left(\sqrt{\frac{d\mu^y}{d\nu}} - \sqrt{\frac{d\mu^{y,N}}{d\nu}} \right)^2 d\nu \right)^{1/2},$$

where ν is any measure such that μ^y and $\mu^{y,N}$ are both absolutely continuous with respect to ν (and the value of the distance is independent of the choice of ν.) In [23, 42, 43] bounds were obtained on $d_{\text{Hell}}(\mu^y, \mu_{\text{mean}}^{y,N})$ and $d_{\text{Hell}}(\mu^y, \mu_{\text{marginal}}^{y,N})$ in terms of $\|\Phi - \mathbb{E}[\Phi_N]\|_{L_{\mu_0}^p(U)}$ and $\left\| \mathbb{E}\left[|\Phi - \Phi_N|^{q_1} \right]^{1/q_1} \right\|_{L_{\mu_0}^{q_2}(U)}$, respectively, for p, q_1 and q_2 depending on the regularity of the problem. These results suggest that choosing Φ_N to be a good approximation to Φ in all regions where the prior μ_0 places significant mass is sufficient for accurate surrogate modeling in the context of Bayesian inference. However we would expect that, in fact, Φ_N only needs to be a good approximation to Φ in regions where the posterior μ^y places significant mass. This is captured in the following results, which follow by analyzing the Hellinger distance with the choice $\nu = \mu^y$.

Theorem 1 (Convergence of mean-based approximation)

Suppose that there exist scalars $C_1, C_2 \geq 0$, independent of N, such that for the Hölder-conjugate exponent pair (p_1, p_1'), we have[1]

(i) $\left\| \exp\left((\Phi - \mathbb{E}[\Phi_N]) \right) \right\|_{L_{\mu^y}^{p_1}(U)} \leq C_1(p_1)$;

(ii) $C_2^{-1} \leq Z_N^{\text{mean}} \leq C_2$.

Then there exists a constant $C_{\text{Thm1}} = C_{\text{Thm1}}(C_1, C_2, Z)$, independent of N, such that

$$d_{\text{Hell}}(\mu^y, \mu_{\text{mean}}^{y,N}) \leq C_{\text{Thm1}} \left\| \Phi - \mathbb{E}[\Phi_N] \right\|_{L_{\mu^y}^{2p_1'}(U)}.$$

Theorem 2 (Convergence of marginal approximation)

Suppose that there exist scalars $C_1, C_2 \geq 0$, independent of N, such that, for the Hölder-conjugate exponent pairs (p_1, p_1') and (p_2, p_2'), we have[1]

---

[1] For $p_1 = \infty$, assumption (i) requires bounding the standard $L^\infty(U)$-norm, due to the corresponding Hölder inequality $\int_U fg\mu^y(du) \leq \sup_{u \in U} |f(u)| \int_U |g|\mu^y(du) = \|f\|_{L^\infty(U)} \|g\|_{L_{\mu^y}^1(U)}$.

(i) $\left\| \mathbb{E}\left[\left(\exp\left(\Phi - \Phi_N \right) \right)^{p_2} \right]^{1/p_2} \right\|_{L^{2p_1}_{\mu^y}(U)} \leq C_1(p_1, p_2)$

(ii) $C_2^{-1} \leq Z_N^{\mathrm{marginal}} \leq C_2.$

Then there exists a constant $C_{\mathrm{Thm2}} = C_{\mathrm{Thm2}}(C_1, C_2, Z)$, independent of N, such that

$$d_{\mathrm{Hell}}\left(\mu, \mu^{y,N}_{\mathrm{marginal}} \right) \leq C_{\mathrm{Thm2}} \left\| \mathbb{E}\left[\left(\Phi - \Phi_N \right)^{p'_2} \right]^{1/p'_2} \right\|_{L^{2p'_1}_{\mu^y}(U)}.$$

Note that these results hold in the general setting where the parameter space U is a Banach space, and do not require U to be finite-dimensional. The proofs of these results are similar to those in [42, 23], and can be found in the appendix. The assumptions in the above theorems are generally applicable to any random surrogate model Φ_N, and in fact also to general distributions on the noise η. They need to be checked on a case-by-case basis. This is done for Gaussian noise and surrogate models based on Gaussian process regression in section 3.

Assumption *(i)* in Theorems 1 and 2 is related to integrability properties of our surrogate model Φ_N. The exponent p_1, present in both theorems, is related to the decay rate of the tails in u. The exponent p_2, related to the decay rate of the tails of the distribution of Φ_N, is only needed in Theorem 2, since the mean-based approximation $\mu^{y,N}_{\mathrm{mean}}$ does not make use of the randomness in Φ_N. Assumption *(ii)* in Theorems 1 and 2 ensures that the normalization constants Z_N^{mean} and Z_N^{marginal}, and hence the approximate posteriors, are well-defined.

3 Gaussian Process Regression

We now want to use Gaussian process regression to build a surrogate model for the data likelihood. Gaussian process regression (a.k.a Gaussian process emulation, or kriging) is a way of building an approximation to a function f, based on a finite number of evaluations of f at a chosen set of training points (a.k.a. design points).

We will here consider emulation of either the parameter-to-observation map $\mathcal{G} : U \to \mathbb{R}^{d_y}$ or the negative log-likelihood $\Phi : U \to \mathbb{R}$. We will for simplicity focus on the emulation of scalar valued functions, and assume that an emulator of $\mathcal{G}$ in the case $d_y > 1$ is constructed by emulating each entry independently. However, we remark here that in practice including correlation between different entries in $\mathcal{G}$ can significantly improve the quality of the approximate posterior [2].

In subsection 3.1 we set-up the framework of Gaussian process regression, followed in subsection 3.2 by discussion of the use of this regression technique in the context of surrogate modeling. Subsection 3.3 returns to the two examples from subsection 2.2, in order to illustrate the potential for reduction in computational cost achieved by the surrogate model f_N compared to the original model f. Subsection 3.4 is devoted to estimates of the error in the approximate posterior in terms of the error in a GP surrogate model used to approximate the likelihood contribution to the

posterior. In subsection 3.5 error estimates for Gaussian process regression, which may be used in the analysis of the previous subsection, are provided.

3.1 Set-up

Let $U \subseteq \mathbb{R}^{d_u}$ be finite-dimensional, and let $f : U \to \mathbb{R}$ be an arbitrary function. Gaussian process regression is a Bayesian procedure, and the starting point is to put a Gaussian process prior on the function f. In other words, we model f as a sample of the Gaussian process

$$f_0 \sim \mathrm{GP}(m(u), k(u, u')), \tag{7}$$

with $m : U \to \mathbb{R}$ a chosen mean function, giving $\mathbb{E}[f_0(u)] = m(u)$, and $k : U \times U \to \mathbb{R}$ a chosen symmetric, positive-definite covariance kernel, giving $\mathbb{C}\mathrm{ov}(f_0(u), f_0(u')) = k(u, u')$. This means that for any set of points $\{\tilde{u}^m\}_{m=1}^M \subseteq U$, the vector $[f_0(\tilde{u}^1); \dots; f_0(\tilde{u}^M)] \in \mathbb{R}^M$ follows a multivariate Gaussian distribution, with mean $[m(\tilde{u}^1); \dots; m(\tilde{u}^M)] \in \mathbb{R}^M$ and covariance matrix $K \in \mathbb{R}^{M \times M}$ with entries $K_{i,j} = k(\tilde{u}^i, \tilde{u}^j)$.

Typical choices of the mean function m include the zero function and polynomials [33]. A family of covariance functions k frequently used in applications are the Matèrn covariance functions [26], given by

$$k_{\nu, \lambda, \sigma_k^2}(u, u') = \sigma_k^2 \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(\frac{\|u - u'\|_2}{\lambda} \right)^\nu B_\nu \left(\frac{\|u - u'\|_2}{\lambda} \right), \tag{8}$$

where Γ denotes the Gamma function, B_ν denotes the modified Bessel function of the second kind and ν, λ and σ_k^2 are positive parameters. The parameter λ is referred to as the *correlation length*, and governs the length scale at which $f_0(u)$ and $f_0(u')$ are correlated. The parameter σ_k^2 is the *marginal variance* $\sigma_k^2 = k_{\nu, \lambda, \sigma_k^2}(u, u) = \mathbb{V}[f_0(u)]$, and governs the typical magnitude of $f_0(u)$. Finally, the parameter ν is referred to as the *smoothness parameter*, and governs the regularity of sample paths of f_0 as a function of u. Sample paths of f_0 are in the Sobolev space $H^{\nu-\epsilon}(U)$ almost surely, for any $\epsilon > 0$, see e.g. [18].

In the limit $\nu \to \infty$, we obtain the Gaussian covariance function

$$k_{\infty, \lambda, \sigma_k^2}(u, u') = \sigma_k^2 \exp\left(-\frac{\|u - u'\|_2^2}{2\lambda^2} \right),$$

also known as the squared exponential or radial basis function (RBF) covariance function. The formula for the Matèrn covariance function furthermore simplifies when ν is a half integer, and popular choices include

$$k_{\frac{1}{2},\lambda,\sigma_k^2}(u,u') = \sigma_k^2 \exp\left(-\frac{\|u-u'\|_2}{\lambda}\right),$$

$$k_{\frac{3}{2},\lambda,\sigma_k^2}(u,u') = \sigma_k^2 \left(1 + \sqrt{3}\frac{\|u-u'\|_2}{\lambda}\right) \exp\left(-\sqrt{3}\frac{\|u-u'\|_2}{\lambda}\right).$$

The choice $\nu = \frac{1}{2}$ gives the exponential covariance function, also known as the Laplace covariance function; in the setting of one-dimensional input variable u it is the covariance function of the Ornstein-Uhlenbeck process.

Figure 1 shows 5 independent sample paths of the Gaussian process f_0 with $m = 0$ and 4 different choices of the parameters in the Matérn covariance function. We also show the mean $\mathbb{E}[f(u)] = m(u) = 0$ as the solid black line, and the marginal standard deviation $\sqrt{k(u,u)} = \sigma = 1$ as the light grey interval around the mean. The two plots in the top row show $\nu = \frac{1}{2}$, leading to sample paths that are continuous but not differentiable, and the two plots in the bottom row show $\nu = \infty$, in which case the sample paths are infinitely smooth. The left column shows a long correlation length λ, leading to mostly large scale variations, whereas the right column shows a shorter correlation length, resulting in more small scale fluctuations.

Now suppose we are given data in the form of a set of distinct training points $D_N := \{u^n\}_{n=1}^N \subseteq U$, together with corresponding function values

$$f(D_N) := [f(u^1); \ldots; f(u^N)] \in \mathbb{R}^N. \tag{9}$$

Since f_0 is a Gaussian process, the vector $[f_0(u^1); \ldots; f_0(u^N); f_0(\tilde{u}^1); \ldots, f_0(\tilde{u}^M)] \in \mathbb{R}^{N+M}$, for any set of test points $\{\tilde{u}^m\}_{m=1}^M \subseteq U$, follows a multivariate Gaussian distribution. The conditional distribution of $f_0(\tilde{u}^1), \ldots, f_0(\tilde{u}^M)$, given the values $f_0(u^1) = f(u^1), \ldots, f_0(u^N) = f(u^N)$, is then again Gaussian, with mean and co-variance given by the standard formulas for the conditioning of Gaussian random variables [33]. Conditioning the Gaussian process (7) on the known values $f(D_N)$, we hence obtain another Gaussian process f_N, known as the *predictive process*. We have

$$f_N \sim \mathrm{GP}(m_N^f(u), k_N(u,u')), \tag{10}$$

where the predictive mean $m_N^f : U \to \mathbb{R}$ and predictive covariance $k_N : U \times U \to \mathbb{R}$ are known explicitly, and depend on the modeling choices made in (7). We will from now on focus on the popular choice $m \equiv 0$; the case of a non-zero mean is discussed in Remark 2. When $m \equiv 0$, we have

$$m_N^f(u) = k(u, D_N)^T K(D_N, D_N)^{-1} f(D_N), \tag{11}$$

$$k_N(u,u') = k(u,u') - k(u, D_N)^T K(D_N, D_N)^{-1} k(u', D_N), \tag{12}$$

where $k(u, D_N) = [k(u, u^1); \ldots; k(u, u^N)] \in \mathbb{R}^N$ and $K(D_N, D_N) \in \mathbb{R}^{N \times N}$ is the matrix with ij^{th} entry equal to $k(u^i, u^j)$ [33].

Figure 2 is similar to Figure 1, and shows samples from the predictive process f_N (with $N = 3$) for different choices of hyper-parameters in the Matèrn covariance kernel. We also show the updated mean $\mathbb{E}[f_N(u)] = m_N^f(u)$ and the updated marginal
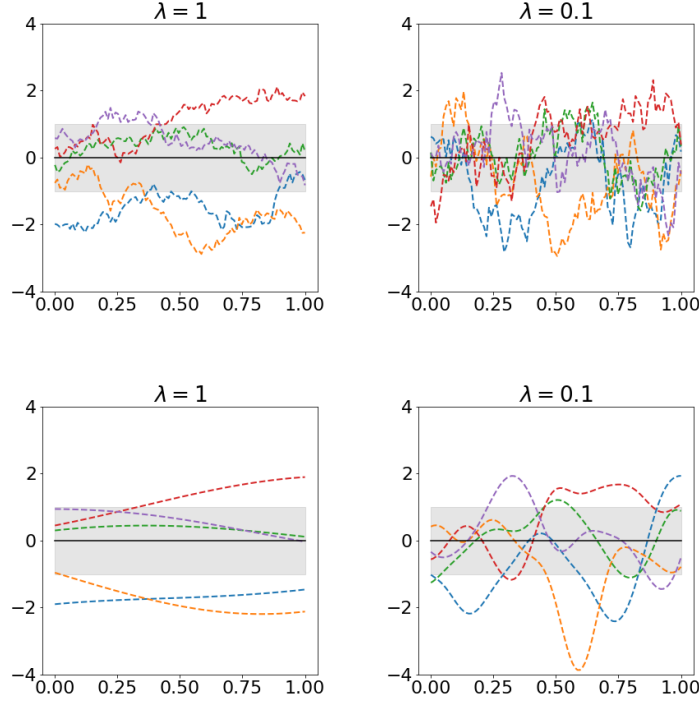
Fig. 1: Sample paths of the Gaussian process f_0 with Matérn covariance function $k_{\nu,\lambda,\sigma_k^2}$. Top left: $\{\nu, \lambda, \sigma_k^2\} = \{\frac{1}{2}, 1, 1\}$. Top right: $\{\nu, \lambda, \sigma_k^2\} = \{\frac{1}{2}, 0.1, 1\}$. Bottom left: $\{\nu, \lambda, \sigma_k^2\} = \{\infty, 1, 1\}$. Bottom right: $\{\nu, \lambda, \sigma_k^2\} = \{\infty, 0.1, 1\}$.

standard deviation $\sqrt{k_N(u, u)}$. We see that the choice of hyper-parameters can have a significant influence on the behaviour of the predictive process. Good values of hyper-parameters are often not known a priori, and have to be estimated from the observed values $f(D_N)$ in (9) (see e.g. [33] and the references therein). This is done for λ and σ^2 in Figure 2 using `scikit-learn` [32].

From (11), we note that m_N^f interpolates the function f at the training points D_N, since the vector $k(u^n, D_N)$ is the nth row of the matrix $K(D_N, D_N)$. In other words, we have $m_N^f(u^n) = f(u^n)$, for all $n = 1, \ldots, N$. For the predictive covariance k_N, we note that $k_N(u, u) < k(u, u)$ for all $u \in U$, since $K(D_N, D_N)$ is positive-definite by assumption. Furthermore, we also note that $k_N(u^n, u^n) = 0$, for $n = 1, \ldots, N$, since $k(u^n, D_N)^T K(D_N, D_N)^{-1} k(u^n, D_N) = k(u^n, u^n)$.

Remark 2 (Prior with non-zero mean) If in (7) we use a non-zero mean $m(\cdot)$, the formula for the predictive mean m_N^f changes to

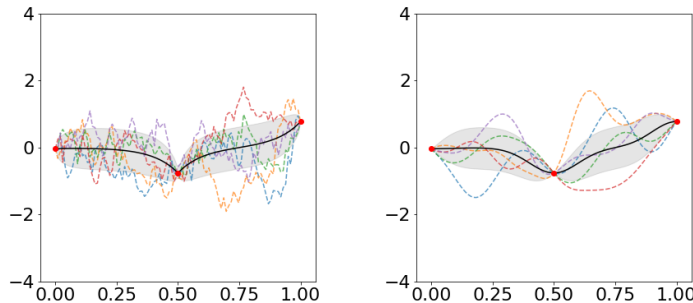$$m_N^{f,m}(u) = m(u) + k(u, D_N)^T K(D_N, D_N)^{-1}(f(D_N) - m(D_N)),$$

Fig. 2: Sample paths of the Gaussian process f_3 with Matérn covariance function $k_{\nu,\lambda,\sigma_k^2}$, for $f(x) = \sin(x - 2.5)^2$. Left: $\{\nu, \lambda, \sigma_k^2\} = \{\frac{1}{2}, 0.1, 0.63^2\}$ Right: $\{\nu, \lambda, \sigma_k^2\} = \{\infty, 0.1, 0.63^2\}$.

where $m(D_N) := [m(u^1); \ldots; m(u^N)] \in \mathbb{R}^N$. The predictive covariance $k_N(u, u')$ is as in (11). As in the case $m \equiv 0$, we have $m_N^f(u^n) = f(u^n)$, for $n = 1, \ldots, N$, and m_N^f is an interpolant of f. Under suitable assumptions on m, any error bounds derived in the case $m \equiv 0$ can be transferred to the general case, see e.g [43]. ∎

Remark 3 (Noisy function values) If instead of exact function values as in (9), we observe noisy function values

$$d := [f(u^1) + \varepsilon_1; \ldots; f(u^N) + \varepsilon_N] \in \mathbb{R}^N,$$

with $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ i.i.d., then the formulas for the predictive process f_N presented above hold with $K(D_N, D_N)$ replaced by $K(D_N, D_N) + \sigma^2 \mathrm{I}$. ∎

3.2 Gaussian Process Regression as Surrogate Model

There are two main use cases for Gaussian process regression: (i) the true function f generating the input-output pairs $\{u^n, f(u^n)\}_{n=1}^N$ is unknown, or (ii) the function f generating $\{u^n, f(u^n)\}_{n=1}^N$ is known, but computationally very expensive to evaluate. In scenario (i), we wish to learn what the function f is from the observed input-output pairs. This is often based on real data, and so the assumption that the observed function values contain noise is common. In the case of (ii), we wish to construct a surrogate model (a.k.a. reduced model, meta-model or emulator) that is cheaper to evaluate than the original f. This is often based on data $\{u^n, f(u^n)\}_{n=1}^N$ obtained from simulating a mathematical model on a computer, and so assuming no noise in the function values may be more appropriate.

To construct a surrogate model for f, we simulate N model runs at inputs $\{u^n\}_{n=1}^N$ to obtain outputs $\{f(u^n)\}_{n=1}^N$, and then follow the Bayesian procedure outlined in the previous section to obtain the predictive process $f_N \sim \text{GP}(m_N^f(u), k_N(u, u'))$. As such, it is similar to many other surrogate models based on model runs or *snapshots*.

The Gaussian process f_N is a random surrogate model for f. The uncertainty in f_N, encoded in the predictive variance k_N, is a model for the remaining uncertainty about f after observing the finite amount of information $\{u^n, f(u^n)\}_{n=1}^N$, or in other words a model for the error between f and m_N^f, based on the assumption that f is a sample of the Gaussian process prior (7). At the training points D_N, where we have observed f, there is no uncertainty about what value f takes and there is no error in m_N^f (i.e. $k_N(u^n, u^n) = 0$ and $m_N^f(u^n) = f(u^n)$, cf Figure 2). Away from the training points, where we do not know the value of f, there is still some uncertainty/error (i.e. $k_N(u, u) > 0$ and $m_N^f(u) \neq f(u)$ in general).

It remains to see that f_N is indeed a surrogate model in the sense that it is much cheaper to evaluate than the original model f. To this end, note that in the approximate posteriors in section 2.3, we typically need to evaluate the predictive mean m_N^f and the predictive variance k_N. As can be seen from (11), the predictive mean m_N^f is a linear combination of kernel evaluations:

$$m_N^f(u) = \sum_{n=1}^N \alpha_n k(u, u^n), \qquad \alpha = K(D_N, D_N)^{-1} f(D_N) \in \mathbb{R}^N.$$

To evaluate $m_N^f(u)$ at an unobserved input $u \in U \setminus D_N$, we hence only need to evaluate the sum on the RHS of the expression above. Firstly, we note that the number of summands N is typically small. This corresponds to the number of times we need to run our original model f, and in many practical applications, this will be limited by computational resources. Secondly, we note that the computation of the coefficients α can, in general, be somewhat costly, since it involves finding the Cholesky factorization (or similar) of $K(D_N, D_N)$ which incurs computational cost $O(N^3)$; however, as noted previously, N is typically small in the applications of surrogate modeling that we consider in this paper. Furthermore, the coefficients α only need to be computed *once*, rather than for every u. Thirdly, we note that the evaluation of the kernel $k(u, u^n)$ is cheap, since this typically involves the evaluation of standard functions such as polynomials and exponentials. Similar arguments apply to evaluations of the predictive variance $k_N(u, u)$.

3.3 Examples of Gaussian Process Regression as Surrogate Model

We return to the two examples introduced in subsection 2.2 in order to illustrate the potential speedups afforded by use of Gaussian process surrogate modeling.

Computational Timings in Subsurface Flow Example

Consider the subsurface flow example from subsection 2.2. The table below gives representative computational timings comparing cost of the evaluation of the solution of the PDE, $f(u)$, with the two primary costs incurred in the Gaussian process surrogate modeling, namely the determination of the coefficients of the kernel representation of the mean, α, and evaluation of the mean m_N^f. Further details, and more extensive numerical experiments, may be found in [2]. The numerical solution of the PDE for a given parameter u, representing the true model f, is implemented in Firedrake [34].

Set-up	Computing $f(u)$	Computing $m_N^f(u)$	Computing α
$d_u = 2, d_y = 10, D = (0, 1), N = 10$	2.6×10^{-1}s	3.6×10^{-5}s	3.2×10^{-3}s
$d_u = 2, d_y = 10, D = (0, 1), N = 20$	2.6×10^{-1}s	4.0×10^{-5}s	1.2×10^{-2}s
$d_u = 10, d_y = 10, D = (0, 1), N = 10$	2.6×10^{-1}s	5.6×10^{-5}s	3.2×10^{-3}s

Computational Motivation for GCM Example

In the paper [12] Gaussian process regression is used to accelerate Bayesian inference for parameters appearing in an idealized general circulation model (GCM) used in climate modeling and described in subsection 2.2. Evaluation of the parameter-to-data map is computationally expensive and use of Gaussian process surrogates, as overviewed in this paper, can be used to accelerate the computations. In particular the number of unknown parameters in the example considered in [12] is 2, the number of observations is on the order of $O(10^2)$, whilst the number of variables in the computational model is on the order of $O(10^5)$; by using Gaussian process surrogates, which sidestep the need to work in space of dimension $O(10^5)$, computational costs can be significantly reduced. The natural measure of computational cost is the number of evaluations of the GCM required to perform MCMC to obtain solution of the Bayesian inverse problem. Without emulation this is $O(10^5)$; with Gaussian process emulation it is possible to achieve the same accuracy in solution of the Bayesian inverse problem with $O(10^2)$ evaluations of the forward model [12]. These evaluations arise from providing training data for the Gaussian process and in particular from the use of the ensemble Kalman based calibrate-emulate-sample approach to the design of this training data [6]. The wish to put the calibrate-emulate-sample methodology on firm theoretical foundations motivates our work in this paper on the optimal choice of design points for Gaussian process surrogate modeling in the context of Bayesian inverse problems.

3.4 Error in GP-accelerated Posterior Distribution

Suppose now that we use Gaussian process regression as our surrogate model in the approximate posterior distributions (3) and (4). An application of Theorems 1 and 2 gives the following error bounds. We denote by $H_k(U)$ the reproducing kernel Hilbert space (RKHS) (see e.g. [44]) of the prior covariance kernel k in (7).

Corollary 1 *Suppose $U \subseteq \mathbb{R}^{d_u}$, $\Phi \in H_k(U)$, and the random surrogate model is constructed by applying Gaussian process regression to Φ, resulting in $\Phi_N \sim GP(m_N^\Phi(u), k_N(u, u'))$. Then there exist constants $C_{\mathrm{Cor1}}, C'_{\mathrm{Cor1}} > 0$, independent of N, such that*

$$d_{\mathrm{Hell}}(\mu^y, \mu_{\mathrm{mean}}^{y,N}) \leq C_{\mathrm{Cor1}} \left\| \Phi - m_N^\Phi \right\|_{L^2_{\mu^y}(U)},$$

and, under the additional assumptions that U is bounded with Lipschitz boundary and $\sup_{u \in U} k_N(u, u) \to 0$ as $N \to \infty$,

$$d_{\mathrm{Hell}}(\mu^y, \mu_{\mathrm{marginal}}^{y,N}) \leq C'_{\mathrm{Cor1}} \left(\left\| \Phi - m_N^\Phi \right\|_{L^2_{\mu^y}(U)} + \left\| k_N^{1/2}(\cdot, \cdot) \right\|_{L^2_{\mu^y}(U)} \right).$$

Corollary 2 *Suppose $U \subseteq \mathbb{R}^{d_u}$, $\mathcal{G}^j \in H_k(U)$ for $j = 1, \ldots, d_y$, and the random surrogate model is constructed by applying Gaussian process regression component-wise to $\mathcal{G}$, resulting in $\mathcal{G}_N^j \sim GP(m_N^{\mathcal{G}^j}(u), k_N(u, u'))$ and $\Phi_N(u) = \frac{1}{2} \| y - \mathcal{G}_N(u) \|_\Gamma^2$. Then there exist constants $C_{\mathrm{Cor2}}, C'_{\mathrm{Cor2}} > 0$, independent of N, such that*

$$d_{\mathrm{Hell}}(\mu^y, \mu_{\mathrm{mean}}^{y,N}) \leq C_{\mathrm{Cor2}} \sum_{j=1}^{d_y} \left\| \mathcal{G}^j - m_N^{\mathcal{G}^j} \right\|_{L^2_{\mu^y}(U)},$$

and, under the additional assumptions that U is bounded with Lipschitz boundary and $\sup_{u \in U} k_N(u, u) \to 0$ as $N \to \infty$,

$$d_{\mathrm{Hell}}(\mu^y, \mu_{\mathrm{marginal}}^{y,N}) \leq C'_{\mathrm{Cor2}} \left(\sum_{j=1}^{d_y} \left\| \mathcal{G}^j - m_N^{\mathcal{G}^j} \right\|_{L^2_{\mu^y}(U)} + \left\| k_N^{1/2}(\cdot, \cdot) \right\|_{L^2_{\mu^y}(U)} \right).$$

Proofs of the above results are similar to those in [42] developed using the L^2-norm weighted by the prior measure μ_0, $L^2_{\mu_0}(U)$. The proofs are outlined in the appendix. However, in contrast to [42], for the new results derived here, all error measures are computed in the L^2-norm weighted by the true posterior measure μ^y, $L^2_{\mu^y}(U)$. This enables us to deduce that the behaviour of the random surrogate model Φ_N in parts of the parameter space to which μ^y does not attach significant mass does not have a big influence on the accuracy of the approximate posterior distributions.

For the mean-based approximation (3) the results show that we only need to control the error in the predictive means m_N^Φ and $m_N^{\mathcal{G}^j}$, respectively. For the marginal approximations (4), we also need to control the predictive variance k_N.

3.5 Error Bounds for GP Regression in Weighted Spaces

We now study the behaviour of the predictive mean m_N^f and predictive variance k_N as $N \to \infty$, for a general function $f : U \to \mathbb{R}$ that we wish to emulate. Using a representer theorem ([33, 39]), the predictive mean m_N^f defined in (11) can be shown to be the minimum norm interpolant in the RKHS $H_k(U)$ of the prior covariance kernel k:

$$m_N^f = \underset{\substack{g \in H_k(U) \text{ s.t.} \\ g(u^{(n)}) = f(u^{(n)}), \, 1 \le n \le N}}{\arg\min} \|g\|_{H_k(U)}. \tag{13}$$

For the Matèrn kernels defined in (8), it is known that the RKHS is equal to the Sobolev space $H^{\nu + d_u/2}(U)$ as a vector space, with equivalent norms [44]. So there exist constants $C_{\text{low}}(U), C_{\text{up}}(U) > 0$ such that for all $g \in H_k(U)$, we have

$$C_{\text{low}}(U)\|g\|_{H^{\nu+d_u/2}(U)} \le \|g\|_{H_k(U)} \le C_{\text{up}}(U)\|g\|_{H^{\nu+d_u/2}(U)}. \tag{14}$$

Given the training points D_N we introduce the mapping $m_N^{(\cdot)} : H_k(U) \to H_k(U)$, which is built on the definition of the predictive mean m_N^f given in (11). With $g(D_N) := [g(u^1); \ldots ; g(u^N)] \in \mathbb{R}^N$ as in (9), we let

$$g(u) \mapsto m_N^g(u) = k(u, D_N)^T K(D_N, D_N)^{-1} g(D_N). \tag{15}$$

The predictive variance k_N satisfies the equality in Proposition 2, which follows, allowing us to transfer convergence results on m_N^f to convergence results on k_N [37, 42].

Proposition 2 *Suppose k_N and $m_N^{(\cdot)}$ are given by (12) and (15), respectively. Then for any $u \in U$ we have*

$$k_N(u, u)^{\frac{1}{2}} = \underset{\substack{g \in H_k(U) \text{ s.t.} \\ \|g\|_{H_k(U)} = 1}}{\sup} |g(u) - m_N^g(u)|.$$

For any bounded set $\tilde{U} \subseteq U$, we define the fill distance $h_{D_N, \tilde{U}}$ as

$$h_{D_N, \tilde{U}} := \underset{u \in \tilde{U}}{\sup} \, \underset{u^n \in D_N \cap \tilde{U}}{\inf} \|u - u^n\|_2.$$

The fill distance is the maximum distance any point in $\tilde{U}$ can be from a training point $u^n \in D_N \cap \tilde{U}$, and we require a space-filling set of training points D_N (in $\tilde{U}$) to ensure that the fill distance goes to zero as $N \to \infty$. The following result follows from standard results in scattered data approximation [44, 28, 1].

Proposition 3 *Suppose $\tilde{U} \subseteq U$ is a Lipschitz domain that satisfies an interior cone condition with angle θ, and $\tilde{U}$ is contained in the cube $B(u_c, R_c) = \{u \in \mathbb{R}^{d_u} : \|u - u_c\|_\infty \le R_c\}$, for some $u_c \in \mathbb{R}^{d_u}$ and $0 < R_c < \infty$. Suppose further that the*

RKHS $H_k(\tilde{U})$ is isomorphic to the Sobolev space $H^\tau(\tilde{U})$ and $f \in H^\tau(\tilde{U})$. Then there exist constants $C_{\text{Prop3}}, C'_{\text{Prop3}} > 0$, independent of f, D_N and N, such that

$$\|f - m_N^f\|_{H^\beta(\tilde{U})} \le C_{\text{Prop3}}(\theta) h_{D_N,\tilde{U}}^{\tau-\beta} \|f\|_{H^\tau(U)}, \qquad \text{for any } \beta \le \tau,$$

for all sets D_N with $h_{D_N,\tilde{U}} \le h_0(\tilde{U}) = C'_{\text{Prop3}} R_c$.

Proposition 3 allows us to bound the error $f - m_N^f$ locally in a subdomain $\tilde{U}$, in the sense that $\|f - m_N^f\|_{H^\beta(\tilde{U})}$ can be bounded in terms of the local fill distance $h_{D_N,\tilde{U}}$. The proof is given in the appendix.

The behaviour of the fill distance $h_{D_N,\tilde{U}}$ in terms of N can be characterized explicitly for many point sets D_N, see e.g. [43, 45] and the references therein. The fastest possible decay as $N \to \infty$ is $h_{D_N,\tilde{U}} = O(N^{-\frac{1}{d_u}})$, which is obtained for example by uniform tensor grids. As seen in Proposition 4 below, the fill distance of randomly sampled point sets also decays at (almost) the optimal rate.

Proposition 4 *Suppose*

(i) $\tilde{U} \subseteq \mathbb{R}^{d_u}$ is a bounded Lipschitz domain that satisfies an interior cone condition, and $\tilde{U}$ is contained in the cube $B(u_c, R_c) = \{u \in \mathbb{R}^{d_u} : \|u - u_c\|_\infty \le R_c\}$, for some $u_c \in \mathbb{R}^{d_u}$ and $0 < R_c < \infty$,

(ii) $g : [0, \infty) \to [0, \infty)$ is continuous, monotonically increasing, and satisfies $g(0) = 0$ and $\lim_{x\downarrow 0} g(x) \exp(x^{-3d_u}) = \infty$,

(iii) training points $\tilde{D}_N \subseteq \tilde{U}$ are sampled i.i.d. from a measure ν with density ρ satisfying $\rho(u) \ge \rho_{\min} > 0$ for all $u \in \overline{\tilde{U}}$.

Then there exists constants $C_{\text{Prop4}}, C'_{\text{Prop4}} > 0$ and $0 < C''_{\text{Prop4}} \le 1$, independent of N, u_c and R_c, such that for any $\varepsilon > 0$, we have

$$\mathbb{E}_\nu[g(h_{\tilde{D}_N,\tilde{U}})] \le C_{\text{Prop4}} R_c \, g(N^{-\frac{1}{d_u}+\varepsilon}),$$

$$\mathbb{P}_\nu[h_{\tilde{D}_N,\tilde{U}} > h] \le C'_{\text{Prop4}} \left(\frac{h}{R_c}\right)^{-d_u} \left(1 - C''_{\text{Prop4}} \left(\frac{h}{R_c}\right)^{d_u}\right)^N, \quad \text{for any } 0 \le h \le R_c.$$

The proof of Proposition 4 in the special case $\tilde{U} \subseteq [0,1]^{d_u}$ can be found in [30]. The general case follows from a simple transformation $u \mapsto R_c u + u_c$. The constants C_{Prop4} and C''_{Prop4} deteriorate, to ∞ and 0 respectively, as $\rho_{\min} \to 0$. The results of Proposition 4 further extend to the setting where the training points are not sampled i.i.d. from ν, but are instead generated from a uniformly ergodic Markov chain with ν as its stationary distribution [30].

Motivated by the results in Corollaries 1 and 2, we now study the quantitites $\mathbb{E}_\nu\left[\|f - m_N^f\|_{L^2_{\mu^y}(U)}\right]$ and $\mathbb{E}_\nu\left[\|k_N^{1/2}(\cdot,\cdot)\|_{L^2_{\mu^y}(U)}\right]$, where the design points D_N are assumed to be sampled from ν. We are particularly interested in the interplay between the sampling measure ν (with density ρ) and the posterior μ^y (with density π^y). Intuitively, we would expect $\nu \approx \mu^y$, since Gaussian process regression will be more accurate in regions with a higher density of training points.

Theorem 3 *Suppose U is a bounded Lipschitz domain, $H_k(U)$ is isomorphic to the Sobolev space $H^\tau(U)$, and $f \in H^\tau(U)$. Further suppose that for all $N \in \mathbb{N}$,*

(i) $U_N \subseteq \mathbb{R}^{d_u}$ is compact and $U_N \subseteq \left\{ u \in \mathbb{R}^{d_u} : \pi^y(u) \leq C_1^2 N^{-\frac{2\tau}{d_u}} \right\}$,
(ii) the training points D_N are sampled i.i.d. from a measure ν_N with density ρ_N satisfying $\rho_N(u) \geq \rho_{\min} > 0$ for all $u \in \overline{U \setminus U_N}$, and $\rho_N(u) = 0$ otherwise,
(iii) $U \setminus U_N$ is a Lipschitz domain that satisfies an interior cone condition with angle θ, and $U \setminus U_N$ is contained in the cube $B(u_c, R_c^N) = \{ u \in \mathbb{R}^{d_u} : \|u - u_c\|_\infty \leq R_c^N \}$, for some $u_c \in \mathbb{R}^{d_u}$ and $0 < R_c^N < C_2 \log N$.

Then there exists a constant $C_{\text{Thm3}} > 0$, independent of f and N, such that for all $0 \leq \beta \leq \tau$ and $\varepsilon > 0$ we have

$$\mathbb{E}_{\nu_N} \left[\|f - m_N^f\|_{H_{\mu^y}^\beta(U)} \right] \leq C_{\text{Thm3}} \, N^{-\frac{\tau - \beta}{d_u} + \varepsilon} \, \|f\|_{H^\tau(U)}.$$

Furthermore, for any partitioning $U \setminus U_N \subseteq \cup_{i=1}^r B_i$, where each B_i is a bounded Lipschitz domain that satisfies an interior cone condition with angle θ', there exists a constant C'_{Thm3} such that for all $0 \leq \beta \leq \tau$ we have

$$\mathbb{E}_{\nu_N} \left[\|f - m_N^f\|_{H_{\mu^y}^\beta(U)} \mathbb{I}_{\{h_{D_N, B_i} \leq h_0(B_i), 1 \leq i \leq n\}} \right]$$
$$\leq C'_{\text{Thm3}} \left(\left(\sup_{u \in U_N} \pi^y(u) \right)^{\frac{1}{2}} + \sum_{i=1}^r \left(\sup_{u \in B_i} \pi^y(u) \right)^{\frac{1}{2}} \mathbb{E}_{\nu_N} \left[h_{D_N, B_i}^{\tau - \beta} \right] \right).$$

The choice $\beta = 0$ in Theorem 3 gives a bound on $\|f - m_N^f\|_{L_{\mu^y}^2(U)}$, whereas $\beta = d_u/2 + \delta$, for any $\delta > 0$, gives a bound on $\|k_N^{1/2}(\cdot, \cdot)\|_{L_{\mu^y}^2(U)}$ via Proposition 2 and the Sobolev embedding theorem as in [42]. The assumption that U is bounded is required only since we need $\rho_{\min} > 0$ in assumption *(ii)*. The assumption that $R_c^N \leq C_2 \log N$ in *(iii)* can be dropped since U is bounded and hence R_c^N is uniformly bounded in N. However, R_c^N does appear as a factor in C_{Thm3} and so different choices of R_c^N could lead to pre-asymptotic effects slowing down convergence. In assumption *(iii)* we assume that the angle θ is independent of N, which ensures that the constant $C_{\text{Prop3}}(\theta)$ in Proposition 3 is also independent of N.

The results in Theorem 3 give us insight into the interplay between the choice of training points D_N and the posterior μ^y. We introduce the set U_N as the part of the parameter space U which does not carry significant posterior mass. We then place training points only in $U \setminus U_N$, according to a sampling measure ν_N. The first claim shows that even by ignoring the set U_N, and placing training points only where the posterior density is sufficiently large, we still obtain (almost) the optimal rate of convergence in N. Optimality here refers to the fastest obtainable rate N^{-1/d_u} of the fill distance in d_u dimensions, as well as the fastest obtainable rate $N^{-\tau/d_u}$ for the approximation of $f \in H^\tau(U)$ in $L^2(U)$ by N function values (see e.g. [43] and the references therein). The first claim holds for any choice of sampling measure ν_N on $U \setminus U_N$, but in general we expect C_{Thm3} to deteriorate as $\rho_{\min}$ gets close to zero.

The second claim shows that the sampling measure ν_N should be chosen such that the local fill distance h_{D_N, B_i} is small where the posterior mass $\sup_{u \in B_i} \pi^y(u)$ is large. In other words, we want to put training points in the regions with highest posterior density. We also note that the bound in the second claim suggests oversampling in the tails of the posterior density π^y, since the fill distance should be balanced with the *square-root of π^y*. This might be related to the fact that extrapolation is generally much harder than interpolation.

Finally, we remark that the above observations agree with previous studies on the subject. In [40], the authors develop a sequential strategy for optimally choosing the training points D_N, in the context of the marginal approximation (4). In practice this results in the training points being placed in regions of high posterior density. Similarly, the choice of design points attained by the calibrate-emulate-sample approach in [6, 12], gives training points in regions of high posterior density for the solution of the Bayesian inverse problem in the GCM example from subsection 2.2. The results in this paper are the first to provide a theoretical justification for this choice of training points in terms of accuracy of the approximate posterior distributions.

The related problem of optimal choice of training points D_N in Bayesian quadrature, i.e. when we are interested in bounding the error

$$\left| \int_U f(u)\pi(u)\mathrm{d}u - \int_U m_N^f(u)\pi(u)\mathrm{d}u \right|,$$

for some target measure π, is studied numerically in [3]. They observe that choosing the design points D_N according to a slightly inflated version of π, i.e. oversampling in the tails of π, seems to give the smallest error.

4 Numerical Examples

We finish this work with two simple illustrative examples. In the first example, our function is $f(u) = u$, and our posterior measure is $\mu \sim \mathcal{N}(1, 1)$. We then take our design measure to be $\nu \sim \mathcal{N}(1, \sigma^2)$. We are interested in analyzing the following error quantity

$$e(N, \nu) = \int_{-\infty}^{\infty} \mathbb{E}_\nu |m_N^f(u) - f(u)|^2 \mu(du) \tag{16}$$

both as a function of the value of the variance σ^2 of the design measure ν as well as the number of points N used in the Gaussian process regression. In all our calculations we have used 10^3 realizations for the design points in order to approximate the expectation in (16). Note that this example does not fulfill all assumptions of Theorem 3, since U is unbounded and there is no a-priori truncation of the parameter space by discarding U_N, but the behaviour is still as predicted by that theorem.

As we can in Figure 3, the results agree with the Theorem 3. In particular, as expected we observe that as the number of design points N increases $e(N, \nu)$

decreases, and at the same rate for different choices of σ. Furthermore, perhaps more interestingly we observe that as we increase the variance of the design measure the error overall decreases until it reaches a minimum value nearby the value of the variance of the true posterior and then slowly increases.
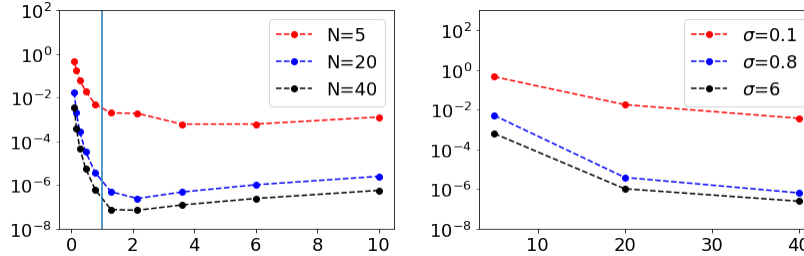

Fig. 3: $e(N, \nu)$ as a function of σ and N for the case of Gaussian posterior μ and design measure ν, with Gaussian covariance kernel k.

In the second numerical experiment we now repeat the first experiment, again with $f(u) = u$, but now with the true posterior measure $\mu \sim U[-1, 1]$ and the design measure $\nu \sim U[-\epsilon, \epsilon]$. As we can see in Figure 4 the results are similar to the Gaussian case studied before with the error reducing as a function of N for a fixed value of ϵ.
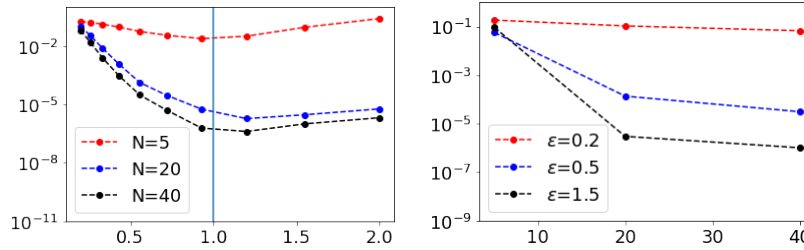

Fig. 4: $e(N, \nu)$ as a function of ϵ and N for the case of uniform posterior μ and design measure ν, with Gaussian covariance kernel k.

Acknowledgements

References

1. R. Arcangéli, M. C. López de Silanes, and J. J. Torrens, *Extension of sampling inequalities to sobolev semi-norms of fractional order and derivative data*, Numer. Math., 121 (2012), pp. 587–608.
2. T. Bai, A. L. Teckentrup, and K. C. Zygalakis, *Gaussian processes for Bayesian inverse problems associated with linear partial differential equations*. In preperation, 2023.
3. F.-X. Briol, C. J. Oates, J. Cockayne, W. Y. Chen, and M. Girolami, *On the sampling problem for kernel quadrature*, in ICML17, PMLR, 2017, pp. 586–595.
4. T. Bui-Thanh, K. Willcox, and O. Ghattas, *Model reduction for large-scale systems with high-dimensional parametric input space*, SIAM. J. Sci. Comput., 30 (2008), pp. 3270–3288.
5. D. Calvetti, M. Dunlop, E. Somersalo, and A. Stuart, *Iterative updating of model error for Bayesian inversion*, Inverse Probl., 34 (2018), p. 025008.
6. E. Cleary, A. Garbuno-Inigo, S. Lan, T. Schneider, and A. M. Stuart, *Calibrate, emulate, sample*, J. Comput. Phys., 424 (2021), p. 109716.
7. J. Cockayne, C. Oates, T. Sullivan, and M. Girolami, *Probabilistic numerical methods for PDE-constrained Bayesian inverse problems*, in AIP Conference Proceedings 1853, 2017, p. 060001.
8. P. R. Conrad, M. Girolami, S. Särkkä, A. Stuart, and K. Zygalakis, *Statistical analysis of differential equations: introducing probability measures on numerical solutions*, Stat. Comput., 27 (2017), pp. 1065–1082.
9. P. R. Conrad, Y. M. Marzouk, N. S. Pillai, and A. Smith, *Accelerating asymptotically exact mcmc for computationally intensive models via local approximations*, J. Am. Stat. Assoc., 111 (2016), pp. 1591–1607.
10. S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White, *MCMC methods for functions: modifying old algorithms to make them faster*, Stat. Sci., 28 (2013), pp. 424–446.
11. T. Cui, C. Fox, and M. O'Sullivan, *Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm*, Water Resour. Res., 47 (2011).
12. O. R. Dunbar, A. Garbuno-Inigo, T. Schneider, and A. M. Stuart, *Calibration and uncertainty quantification of convective parameters in an idealized gcm*, J. Adv. in Model. Earth Sy., 13 (2021), p. e2020MS002454.
13. M. Girolami and B. Calderhead, *Riemann manifold Langevin and Hamiltonian Monte Carlo methods*, J. R. Stat. Soc. B, 73 (2011), pp. 123–214.
14. W. Hastings, *Monte-Carlo sampling methods using Markov chains and their applications*, Biometrika, 57 (1970), pp. 97–109.
15. D. Z. Huang, T. Schneider, and A. M. Stuart, *Iterated Kalman methodology for inverse problems*, J. Comput. Phys., 463 (2022), p. 111262.
16. M. Järvenpää, M. U. Gutmann, A. Vehtari, and P. Marttinen, *Parallel Gaussian process surrogate Bayesian inference with noisy likelihood evaluations*, Bayesian Anal., 16 (2021), pp. 147–178.
17. J. P. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer, 2005.
18. M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur, *Gaussian processes and kernel methods: A review on connections and equivalences*, arXiv preprint arXiv:1807.02582, (2018).

19. M. C. Kennedy and A. O'Hagan, *Bayesian calibration of computer models*, J. R. Stat. Soc. B, 63 (2001), pp. 425–464.

20. J. Latz, *On the well-posedness of Bayesian inverse problems*, SIAM/ASA J. Uncert. Quant., 8 (2020), pp. 451–482.

21. E. B. Le, A. Myers, T. Bui-Thanh, and Q. P. Nguyen, *A data-scalable randomized misfit approach for solving large-scale pde-constrained inverse problems*, Inverse Probl., 33 (2017), p. 065003.

22. H. C. Lie, A. M. Stuart, and T. J. Sullivan, *Strong convergence rates of probabilistic integrators for ordinary differential equations*, Stat. Comput., 29 (2019), pp. 1265–1283.

23. H. C. Lie, T. J. Sullivan, and A. L. Teckentrup, *Random forward models and log-likelihoods in Bayesian inverse problems*, SIAM/ASA J. Uncert. Quant., 6 (2018), pp. 1600–1629.

24. Y. Marzouk and D. Xiu, *A stochastic collocation approach to Bayesian inference in inverse problems*, Commun. Comput. Phys., 6 (2009), pp. 826–847.

25. Y. M. Marzouk, H. N. Najm, and L. A. Rahn, *Stochastic spectral methods for efficient Bayesian solution of inverse problems*, J. Comput. Phys., 224 (2007), pp. 560–586.

26. B. Matérn, *Spatial Variation*, Springer, 2013.

27. N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, *Equation of state calculations by fast computing machines*, J. Chem. Phys., 21 (1953), p. 1087.

28. F. Narcowich, J. Ward, and H. Wendland, *Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting*, Math. Comput., 74 (2005), pp. 743–763.

29. A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optimiz., 19 (2009), pp. 1574–1609.

30. C. J. Oates, J. Cockayne, F.-X. Briol, and M. Girolami, *Convergence rates for a class of estimators based on Stein's method*, Bernoulli, 25 (2019), pp. 1141–1159.

31. P. A. O'Gorman and T. Schneider, *The hydrological cycle over a wide range of climates simulated with an idealized gcm*, J. Climate, 21 (2008), pp. 3815–3832.

32. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine learning in Python*, J. Mach. Learn. Res., 12 (2011), pp. 2825–2830.

33. C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*, MIT Press, 2006.

34. F. Rathgeber, D. A. Ham, L. Mitchell, M. Lange, F. Luporini, A. T. McRae, G.-T. Bercea, G. R. Markall, and P. H. Kelly, *Firedrake: automating the finite element method by composing abstractions*, ACM T. Math. Software, 43 (2016), pp. 1–27.

35. C. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, 1999.

36. Y. Rubin, *Applied stochastic hydrogeology*, Oxford University Press, 2003.

37. M. Scheuerer, R. Schaback, and M. Schlather, *Interpolation of spatial data–A stochastic or a deterministic problem?*, Eur. J. Appl. Math., 24 (2013), pp. 601–629.

38. T. Schneider, *Earth system modeling 2.0: Toward data-informed climate models with quantified uncertainties*, B. Am. Phys. Soc., 65 (2020).

39. B. Schoelkopf and A. Smola, *Learning with Kernels*, MIT Press, 2002.

40. M. Sinsbeck and W. Nowak, *Sequential design of computer experiments for the solution of Bayesian inverse problems*, SIAM/ASA J. Uncert. Quant., 5 (2017), pp. 640–664.

41. A. M. Stuart, *Inverse problems*, vol. 19 of Acta Numer., Cambridge University Press, 2010, pp. 451–559.

42. A. M. Stuart and A. L. Teckentrup, *Posterior Consistency for Gaussian Process Approximations of Bayesian Posterior Distributions*, Math. Comput., 87 (2018), pp. 721–753.

43. A. L. Teckentrup, *Convergence of Gaussian process regression with estimated hyperparameters and applications in Bayesian inverse problems*, SIAM/ASA J. Uncert. Quant., 8 (2020), pp. 1310–1337.

44. H. Wendland, *Scattered Data Approximation*, Cambridge University Press, 2004.

45. G. Wynne, F.-X. Briol, and M. Girolami, *Convergence guarantees for Gaussian process means with misspecified likelihoods and smoothness*, J. Mach. Learn. Res., 22 (2021).

46. D. ZHANG, *Stochastic methods for flow in porous media: coping with uncertainties*, Elsevier, 2001.
47. Y. ZHOU, A. M. JOHANSEN, AND J. A. ASTON, *Toward automatic model comparison: an adaptive sequential Monte Carlo approach*, J. Comput. Graph. Stat., 25 (2016), pp. 701–726.

Appendix: Proofs of Results

Proof (of Theorem 1) First, note that the mean-based posterior approximation $\mu_{\mathrm{mean}}^{y,N}$ is absolutely continuous with respect to the true posterior μ^y, since $\exp\left(-\mathbb{E}[\Phi_N(u)]\right)$ and $\exp\left(-\Phi(u)\right)$ are both positive and μ_0 is a probability measure:

$$\mu^y(A) = \int_A \exp\left(-\Phi(u)\right)\mu_0(\mathrm{d}u) = 0 \Rightarrow \mu_{\mathrm{mean}}^{y,N}(A) = \int_A \exp\left(-\mathbb{E}[\Phi_N(u)]\right)\mu_0(\mathrm{d}u) = 0.$$

For the same reasons, the prior μ_0 is absolutely continuous with respect to the true posterior μ^y. We then have the Radon-Nikodym derivative

$$\frac{d\mu_{\mathrm{mean}}^{y,N}}{d\mu^y}(u) = \frac{d\mu_{\mathrm{mean}}^{y,N}}{d\mu_0}(u)\frac{d\mu_0}{d\mu^y}(u) = \frac{d\mu_{\mathrm{mean}}^{y,N}}{d\mu_0}(u)\left(\frac{d\mu^y}{d\mu_0}(u)\right)^{-1} = \frac{\frac{1}{Z_N^{\mathrm{mean}}}\exp\left(-\mathbb{E}[\Phi_N(u)]\right)}{\frac{1}{Z}\exp\left(-\Phi(u)\right)}.$$

Using the definition of the Hellinger distance with $\nu = \mu^y$ and the inequality

$$\left(1-\frac{ab}{cd}\right)^2 = \frac{1}{c^2d^2}(cd-ab)^2 = \frac{1}{c^2d^2}(cd-cb+cb-ab)^2 \leq \frac{2}{c^2d^2}\left((cd-cb)^2+(cb-ab)^2\right),$$

for real numbers $a,b,c,d \in \mathbb{R}$, we then have

$$2\, d_{\mathrm{Hell}}^2(\mu^y, \mu_{\mathrm{mean}}^{y,N}) = \int_U \left(1-\sqrt{\frac{\frac{1}{Z_N^{\mathrm{mean}}}\exp\left(-\mathbb{E}[\Phi_N(u)]\right)}{\frac{1}{Z}\exp\left(-\Phi(u)\right)}}\right)^2 \mu^y(\mathrm{d}u)$$

$$\leq 2\int_U \left(1-\exp\left(\frac{1}{2}(\Phi(u)-\mathbb{E}[\Phi_N(u)])\right)\right)^2 \mu^y(\mathrm{d}u) +$$

$$2\,Z\left(Z^{-1/2}-(Z_N^{\mathrm{mean}})^{-1/2}\right)^2 \int_U \exp\left((\Phi(u)-\mathbb{E}[\Phi_N(u)])\right)\mu^y(\mathrm{d}u)$$

$$=: I + II.$$

Using the local Lipschitz continuity of the exponential function, Hölder's inequality with conjugate exponents (p_1, p_1'), the inequality $(a+b)^2 \leq 2(a^2+b^2)$ for $a, b \in \mathbb{R}$, the triangle equality in $L_{\mu^y}^{p_1}$, and Jensen's inequality for $x \to x^2$, we have

$$I \leq \frac{1}{2} \int_U \left(1 + \exp\left(\frac{1}{2}(\Phi(u) - \mathbb{E}[\Phi_N(u)])\right)\right)^2 (\Phi(u) - \mathbb{E}[\Phi_N(u)])^2 \, \mu^y(du)$$

$$\leq \frac{1}{2} \left\| \left(1 + \exp\left(\frac{1}{2}(\Phi - \mathbb{E}[\Phi_N])\right)\right)^2 \right\|_{L_{\mu^y}^{P_1}} \left\| (\Phi - \mathbb{E}[\Phi_N])^2 \right\|_{L_{\mu^y}^{P_1'}}$$

$$\leq (1 + C_1(p_1)) \left\| \Phi - \mathbb{E}[\Phi_N] \right\|^2_{L_{\mu^y}^{2P_1'}}$$

For II, using the inequality $(a^{-1/2} - b^{-1/2})^2 \leq (a-b)^2 \max\{a^{-3}, b^{-3}\}$, for $a, b > 0$, we have

$$II \leq 2 Z \max(Z^{-3}, (Z_N^{\mathrm{mean}})^{-3}) \left(Z - (Z_N^{\mathrm{mean}})\right)^2 \int_U \exp\left((\Phi(u) - \mathbb{E}[\Phi_N(u)])\right) \mu^y(du).$$

Then, as in the bound for I and using Jensen's inequality for $x \to x^2$, we have

$$\left(Z - (Z_N^{\mathrm{mean}})\right)^2 = \left(\int_U \exp\left(-\Phi(u)\right) - \exp\left(-\mathbb{E}[\Phi_N(u)]\right)) \mu_0(du)\right)^2$$

$$= \left(\int_U \left(1 - \exp\left(\frac{1}{2}(\Phi(u) - \mathbb{E}[\Phi_N(u)])\right)\right) \mu^y(du)\right)^2$$

$$\leq (1 + C_1(p_1)) \left\| \Phi - \mathbb{E}[\Phi_N] \right\|^2_{L_{\mu^y}^{2P_1'}}$$

and so

$$II \leq 2Z \max\left\{Z^{-3}, C_2^3\right\} (1 + C_1(p_1)) C_1(1) \left\| \Phi - \mathbb{E}[\Phi_N] \right\|^2_{L_{\mu^y}^{2P_1'}}.$$

This completes the proof, with

$$C_{\mathrm{Thm1}} = \sqrt{(1 + C_1(p_1)) \left(\frac{1}{2} + Z \max\left\{Z^{-3}, C_2^3\right\} C_1(1)\right)}.$$

Proof (of Theorem 2) Similar to the proof of Theorem 1, we compute

$$\frac{d\mu_{\mathrm{marginal}}^{y,N}}{d\mu^y}(u) = \frac{\frac{1}{Z_N^{\mathrm{marginal}}}\mathbb{E}\left(\exp\left(-\Phi_N(u)\right)\right)}{\frac{1}{Z}\exp\left(-\Phi(u)\right)},$$

and using the definition of the Hellinger distance with $\nu = \mu^y$, we have

$$2\, d_{\text{Hell}}^2(\mu^y, \mu_{\text{marginal}}^{y,N}) = = \int_U \left(1 - \sqrt{\frac{\frac{1}{Z_N^{\text{marginal}}}\mathbb{E}\left(\exp\left(-\Phi_N(u)\right)\right)}{\frac{1}{Z}\exp\left(-\Phi(u)\right)}}\,\right)^2 \mu^y(\mathrm{d}u)$$

$$\leq 2\int_U \left(1 - \sqrt{\mathbb{E}\left[\exp\left(\Phi(u) - \Phi_N(u)\right)\right]}\,\right)^2 \mu^y(\mathrm{d}u) +$$

$$2\,Z\left(Z^{-1/2} - (Z_N^{\text{marginal}})^{-1/2}\right)^2 \int_U \mathbb{E}\left[\exp\left(\Phi(u) - \Phi_N(u)\right)\right]\mu^y(\mathrm{d}u)$$

$$=: I + II.$$

Using the inequality $(a-b)^2 = \left(\frac{a^2-b^2}{a+b}\right)^2 \leq \frac{(a^2-b^2)^2}{a^2+b^2}$, for $a, b \in \mathbb{R}$, together with Hölder's inequality with conjugate exponents q_1 and q_1', we obtain

$$I \leq 2\left\|\left(1 - \mathbb{E}\left[\exp\left(\Phi - \Phi_N\right)\right]\right)^2\right\|_{L_{\mu^y}^{q_1'}} \left\|\left(1 + \mathbb{E}\left[\exp\left(\Phi - \Phi_N\right)\right]\right)^{-1}\right\|_{L_{\mu^y}^{q_1}}.$$

We estimate the second factor on the right-hand side above as in the proof of [23, Theorem 3.1]:

$$\left\|\left(1 + \mathbb{E}\left[\exp\left(\Phi - \Phi_N\right)\right]\right)^{-1}\right\|_{L_{\mu^y}^{q_1}} \leq \min\left\{\|1\|_{L_{\mu^y}^{q_1}},\left\|\mathbb{E}\left[\exp\left(\Phi - \Phi_N\right)\right]^{-1}\right\|_{L_{\mu^y}^{q_1}}\right\} = 1$$

This holds for any $q_1 \in [1, \infty]$, and so we can choose $q_1 = \infty$ and $q_1' = 1$.

For the first factor, the linearity of expectation, local Lipschitz continuity of the exponential function, Hölder's inequality with conjugate exponents p_2, p_2' with respect to $\mathbb{E}$ and p_1, p_1' with respect to μ^y, the inequality $|a + b|^{p_2} \leq 2^{p_2-1}(|a|^{p_2} + |b|^{p_2})$ and the triangle inequality in $L_{\mu^y}^{2p_1}$ give

$$I \leq 2\left\|\left(1 - \mathbb{E}\left[\exp\left(\Phi - \Phi_N\right)\right]\right)^2\right\|_{L_{\mu^y}^1}$$

$$\leq 2\int_U \left(\mathbb{E}\left[\left(1 + \exp\left(\Phi(u) - \Phi_N(u)\right)\right)\left(\Phi(u) - \Phi_N(u)\right)\right]\right)^2 \mu^y(\mathrm{d}u)$$

$$\leq 2\left\|\mathbb{E}\left[\left(1 + \exp\left(\Phi - \Phi_N\right)\right)^{p_2}\right]^{1/p_2}\right\|_{L_{\mu^y}^{2p_1}}^2 \left\|\mathbb{E}\left[\left(\Phi - \Phi_N\right)^{p_2'}\right]^{1/p_2'}\right\|_{L_{\mu^y}^{2p_1'}}^2$$

$$\leq 2^{p_2}(1 + C_1(p_1, p_2))\left\|\mathbb{E}\left[\left(\Phi - \Phi_N\right)^{p_2'}\right]^{1/p_2'}\right\|_{L_{\mu^y}^{2p_1'}}^2.$$

For II, using the inequality $(a^{-1/2} - b^{-1/2})^2 \leq (a-b)^2 \max\{a^{-3}, b^{-3}\}$, for $a, b > 0$, we have

$$II \leq 2\,Z \max(Z^{-3}, (Z_N^{\text{marginal}})^{-3})\left(Z - (Z_N^{\text{marginal}})\right)^2 \int_U \mathbb{E}\left[\exp\left(\Phi(u) - \Phi_N(u)\right)\right]\mu^y(\mathrm{d}u).$$

Then, as in the bound for I and using Jensen's inequality for $x \to x^2$, we have

$$
\begin{aligned}
\left(Z - (Z_N^{\mathrm{marginal}}) \right)^2 &= \left(\int_U \exp\left(-\Phi(u)\right) - \mathbb{E}\left[\exp\left(-\Phi_N(u)\right)\right] \mu_0(\mathrm{d}u) \right)^2 \\
&= \left(\int_U \left(\mathbb{E}\left[1 - \exp\left(\Phi(u) - \Phi_N(u)\right)\right] \right)^2 \mu^y(\mathrm{d}u) \right)^2 \\
&\leq 2^{p_2 - 1}(1 + C_1(p_1, p_2)) \left\| \mathbb{E}\left[(\Phi - \Phi_N)^{p_2'} \right]^{1/p_2'} \right\|_{L_{\mu^y}^{2p_1'}}^2
\end{aligned}
$$

and so

$$
II \leq Z \max\left\{ Z^{-3}, C_2^3 \right\} 2^{p_2}(1 + C_1(p_1, p_2))C_1(1,1) \left\| \mathbb{E}\left[(\Phi - \Phi_N)^{p_2'} \right]^{1/p_2'} \right\|_{L_{\mu^y}^{2p_1'}}^2.
$$

This completes the proof, with

$$
C_{\mathrm{Thm2}} = \sqrt{2^{p_2 - 1}(1 + C_1(p_1, p_2)) \left(1 + Z \max\left\{ Z^{-3}, C_2^3 \right\} C_1(1,1)\right)}.
$$

Proof (of Corollary 1) We check the assumptions in Theorems 1 and 2. For assumption (ii), we can follow the proof of Lemmas 4.1 and 4.7 in [42], respectively. Note that instead of assuming that $\sup_{u \in U} \left| \Phi(u) - m_N^\Phi(u) \right|$ converges to 0 as N tends to infinity and $\sup_{u \in U} \|\mathcal{G}(u)\| \leq C_{\mathcal{G}} < \infty$, we can directly bound $\sup_{u \in U} m_N^\Phi(u)$ independently of N using the definition of the RKHS $H_k(U)$ of the kernel k, the Cauchy-Schwarz inequality and the minimum norm interpolant property (13):

$$
\sup_{u \in U} \left| m_N^\Phi(u) \right| = \sup_{u \in U} \left| \langle m_N^\Phi(\cdot), k(\cdot, u) \rangle_{H_k(U)} \right| \leq \|\Phi\|_{H_k(U)} \sup_{u \in U} \sqrt{k(u, u)}.
$$

We can then bound Z_N^{mean} as in the proof of [42, Lemma 4.1], and Z_N^{marginal} as in the proof of [42, Lemma 4.7]. For assumption (i) in Theorem 1, this then immediately follows with $p_1 = \infty$, since as above

$$
\left\| \exp\left((\Phi - \mathbb{E}[\Phi_N])\right) \right\|_{L^\infty(U)} \leq \exp\left(\|\Phi\|_{H_k(U)} \left(1 + \sup_{u \in U} \sqrt{k(u, u)}\right) \right).
$$

For assumption (i) in Theorem 2, we similarly have that for $p_1 = \infty$, the quantity

$$
\sup_{u \in U} \mathbb{E}\left[(\exp\left(\Phi(u) - \Phi_N(u)\right))^{p_2} \right]^{1/p_2} \leq \mathbb{E}\left[\exp\left(p_2 \sup_{u \in U}(\Phi(u) + \Phi_N(u))\right) \right]^{1/p_2}
$$

can be bounded for any $p_2 < \infty$. The choice $p_2 = 2$ then gives the desired result. □

Proof (of Corollary 2) This is very similar to the proof of Corollary 1, with the same modification to bound $\sup_{u \in U} \left| m_N^{\mathcal{G}^j}(u) \right|$ independently of N.   □

Proof (of Proposition 3) It follows from [1, Theorem 3.2] that

$$\|f - m_N^f\|_{H^\beta(\tilde{U})} \le C(\theta) h_{D_N,\tilde{U}}^{\tau-\beta} \|f - m_N^f\|_{H^\tau(\tilde{U})},$$

for any $\beta \le \tau$ and for all sets D_N with $h_{D_N,\tilde{U}} \le h_0(\tilde{U}) = C'_{\text{Prop3}} R_c$. Using $\tilde{U} \subseteq U$, the triangle inequality in $H^\tau(U)$, the norm equivalence between $H_k(U)$ and $H^\tau(U)$ from (14), and $\|m_N^f\|_{H_k(U)} \le \|f\|_{H_k(U)}$ from (13) then gives

$$\begin{aligned}
\|f - m_N^f\|_{H^\tau(\tilde{U})} &\le \|f\|_{H^\tau(U)} + \|m_N^f\|_{H^\tau(U)} \\
&\le \left(1 + C_{\text{low}}(U)^{-1} C_{\text{up}}(U)\right) \|f\|_{H^\tau(U)}.
\end{aligned}$$

This finishes the proof, with $C_{\text{Prop3}}(\theta) = C(\theta)\left(1 + C_{\text{low}}(U)^{-1} C_{\text{up}}(U)\right)$.   □

Proof (of Theorem 3) Step 1: We split the error as

$$\mathbb{E}_{\nu_N}\left[\|f - m_N^f\|_{H_{\mu^y}^\beta(U)}\right] \le \mathbb{E}_{\nu_N}\left[\|f - m_N^f\|_{H_{\mu^y}^\beta(U\backslash U_N)}\right] + \mathbb{E}_{\nu_N}\left[\|f - m_N^f\|_{H_{\mu^y}^\beta(U_N)}\right].$$

Step 2: For $\mathbb{E}_{\nu_N}\left[\|f - m_N^f\|_{H_{\mu^y}^\beta(U_N)}\right]$, we have by assumption (*i*) and $\beta \le \tau$ that

$$\begin{aligned}
\mathbb{E}_{\nu_N}\left[\|f - m_N^f\|_{H_{\mu^y}^\beta(U_N)}\right] &\le \left(\sup_{u \in U_N} \pi^y(u)\right)^{\frac{1}{2}} \mathbb{E}_{\nu_N}\left[\|f - m_N^f\|_{H^\beta(U_N)}\right] \\
&\le C_1 N^{-\frac{\tau}{d_u}} \mathbb{E}_{\nu_N}\left[\|f - m_N^f\|_{H^\tau(U_N)}\right] \\
&\le C_1\left(1 + C_{\text{low}}(U)^{-1} C_{\text{up}}(U)\right) N^{-\frac{\tau}{d_u}} \|f\|_{H^\tau(U)},
\end{aligned}$$

where we have bounded $\|f - m_N^f\|_{H^\tau(U_N)}$ as in the proof of Proposition 3.

Step 3: For $\mathbb{E}_{\nu_N}\left[\|f - m_N^f\|_{H_{\mu^y}^\beta(U\backslash U_N)}\right]$, the linearity of expectation gives

$$\begin{aligned}
\mathbb{E}_{\nu_N}\left[\|f - m_N^f\|_{H^\beta(U\backslash U_N)}\right] = &\mathbb{E}_{\nu_N}\left[\|f - m_N^f\|_{H^\beta(U\backslash U_N)} \mathrm{I}_{\{h_{D_N,U\backslash U_N} \le h_0(U\backslash U_N)\}}\right] \\
&+ \mathbb{E}_{\nu_N}\left[\|f - m_N^f\|_{H^\beta(U\backslash U_N)} \mathrm{I}_{\{h_{D_N,U\backslash U_N} > h_0(U\backslash U_N)\}}\right].
\end{aligned}$$

For the first term, an application of Propositions 3 and 4, and Hölder's inequality with conjugate exponents $p = \infty$ and $q = 1$, gives

$$\begin{aligned}
&\mathbb{E}_{\nu_N}\left[\|f - m_N^f\|_{H^\beta(U\backslash U_N)} \mathrm{I}_{\{h_{D_N,U\backslash U_N} \le h_0(U\backslash U_N)\}}\right] \\
&\le C_{\text{Prop3}}(\theta) \mathbb{E}_{\nu_N}\left[h_{D_N,U\backslash U_N}^{\tau-\beta}\right] \sup_{D_N} \|f - m_N^f\|_{H^\tau(U\backslash U_N)} \\
&\le C_{\text{Prop3}}(\theta) C_{\text{Prop4}} C_2 \left(1 + C_{\text{low}}(U)^{-1} C_{\text{up}}(U)\right) N^{-\frac{\tau-\beta}{d_u}+\varepsilon} \|f\|_{H^\tau(U)},
\end{aligned}$$

for any $\varepsilon > 0$, where we have bounded $\|f - m_N^f\|_{H^\tau(U\backslash U_N)}$ as in step 2.

For the second term, we use Hölder's inequality with conjugate exponents $p = \infty$ and $q = 1$ to obtain

$$
\mathbb{E}_{\nu_N} \left[\|f - m_N^f\|_{H^\beta(U \setminus U_N)} \mathrm{I}_{\{h_{D_N, U \setminus U_N} > h_0(U \setminus U_N))\}} \right]
$$

$$
\leq \sup_{D_N} \|f - m_N^f\|_{H^\beta(U \setminus U_N)} \mathbb{P}_{\nu_N} \left[h_{D_N, U \setminus U_N} > h_0(U \setminus U_N) \right].
$$

The second factor can be bounded by Proposition 4. For the first factor, we again use the bound on $\|f - m_N^f\|_{H^\beta(U \setminus U_N))}$ as in step 2. Adding the two terms, we have

$$
\mathbb{E}_{\nu_N} \left[\|f - m_N^f\|_{H^\beta(U \setminus U_N)} \right] \leq \left(1 + C_{\mathrm{low}}(U)^{-1} C_{\mathrm{up}}(U) \right) \|f\|_{H^\tau(U)}
$$

$$
\left(C_{\mathrm{Prop3}}(\theta) \, C_{\mathrm{Prop4}} \, C_2 \, N^{-\frac{\tau - \beta}{d_u} + \varepsilon} + C_{\mathrm{Prop4}}' \, (C_{\mathrm{Prop3}}')^{-d_u} \left(1 - C_{\mathrm{Prop4}}''(C_{\mathrm{Prop3}}')^{d_u} \right)^N \right).
$$

Since the geometrically decaying term is bounded by the algebraically converging term for N sufficiently large, and both terms are monotonically decreasing in N, it follows that

$$
\mathbb{E}_{\nu_N} \left[\|f - m_N^f\|_{H^\beta(U \setminus U_N)} \right] \leq C N^{-\frac{\tau - \beta}{d_u} + \varepsilon} \|f\|_{H^\tau(U)},
$$

with $C = \left(1 + C_{\mathrm{low}}(U)^{-1} C_{\mathrm{up}}(U) \right) \max\{ C_{\mathrm{Prop3}}(\theta) C_{\mathrm{Prop4}} \, C_2, C_{\mathrm{Prop4}}' \, (C_{\mathrm{Prop3}}')^{-d_u} \}$. This proves the first claim, with $C_{\mathrm{Thm3}} = C + C_1 \left(1 + C_{\mathrm{low}}(U)^{-1} C_{\mathrm{up}}(U) \right)$.

Step 5: For the second claim, we split $\mathbb{E}_{\nu_N} \left[\|f - m_N^f\|_{H^\beta_\mu(U \setminus U_N)} \right]$ further over the subdomains $U \setminus U_N \subseteq \{B_i\}_{i=1}^r \subseteq U$ to obtain

$$
\mathbb{E}_{\nu_N} \left[\|f - m_N^f\|_{H^\beta_{\mu^y}(U \setminus U_N)} \mathrm{I}_{\{h_{D_N, B_i} \leq h_0(B_i), 1 \leq i \leq n\}} \right]
$$

$$
\leq \sum_{i=1}^r \left(\sup_{u \in B_i} \pi^y(u) \right)^{\frac{1}{2}} \mathbb{E}_{\nu_N} \left[\|f - m_N^f\|_{H^\beta(B_i)} \mathrm{I}_{\{h_{D_N, B_i} \leq h_0(B_i)\}} \right].
$$

Then, as in step 3, we have

$$
\mathbb{E}_{\nu_N} \left[\|f - m_N^f\|_{H^\beta(B_i)} \mathrm{I}_{\{h_{\tilde{D}_N, B_i} \leq h_0(B_i)\}} \right]
$$

$$
\leq C_{\mathrm{Prop3}}(\theta') \, \mathbb{E}_{\nu_N} \left[h_{D_N, B_i} \right]^{\tau - \beta} \left(1 + C_{\mathrm{low}}(U)^{-1} C_{\mathrm{up}}(U) \right) \|f\|_{H^\tau(U)}.
$$

This finishes the proof of the second claim, with

$$
C_{\mathrm{Thm3}}' = \left(1 + C_{\mathrm{low}}(U)^{-1} C_{\mathrm{up}}(U) \right) \|f\|_{H^\tau(U)} \left(1 + r C_{\mathrm{Prop3}}(\theta') \right).
$$