Journal of Machine Learning Research 19 (2018) 1-46 Submitted 1/18; Revised 8/18; Published 9/18

How Deep Are Deep Gaussian Processes?

Matthew M. Dunlop MDUNLOP@CALTECH.EDU
Computing and Mathematical Sciences
Caltech
Pasadena, CA 91125, USA

Mark A. Girolami M.GIROLAMI@IMPERIAL.AC.UK
Department of Mathematics
Imperial College London
London, SW7 2AZ, UK
and
The Alan Turing Institute
96 Euston Road
London, NW1 2DB, UK

Andrew M. Stuart ASTUART@CALTECH.EDU
Computing and Mathematical Sciences
Caltech
Pasadena, CA 91125, USA

Aretha L. Teckentrup A.TECKENTRUP@ED.AC.UK
School of Mathematics
University of Edinburgh
Edinburgh, EH9 3FD, UK
and
The Alan Turing Institute
96 Euston Road
London, NW1 2DB, UK

Editor: Neil Lawrence

Abstract

Recent research has shown the potential utility of deep Gaussian processes. These deep structures are probability distributions, designed through hierarchical construction, which are conditionally Gaussian. In this paper, the current published body of work is placed in a common framework and, through recursion, several classes of deep Gaussian processes are defined. The resulting samples generated from a deep Gaussian process have a Markovian structure with respect to the depth parameter, and the effective depth of the resulting process is interpreted in terms of the ergodicity, or non-ergodicity, of the resulting Markov chain. For the classes of deep Gaussian processes introduced, we provide results concerning their ergodicity and hence their effective depth. We also demonstrate how these processes may be used for inference; in particular we show how a Metropolis-within-Gibbs construction across the levels of the hierarchy can be used to derive sampling tools which are robust to the level of resolution used to represent the functions on a computer. For illustration, we consider the effect of ergodicity in some simple numerical examples.

Keywords: deep learning, deep Gaussian processes, deep kernels

1. Introduction

This section provides background on the study and application of Gaussian and deep Gaussian processes, outlines the contribution and setup of the paper, and establishes notation to be used in subsequent sections.

1.1 Background

Gaussian processes have proved remarkably successful as a tool for various statistical inference and machine learning tasks (Rasmussen and Williams, 2006; Kennedy and O'Hagan, 2001; Higdon et al., 2004; Stein, 1999). This success relates in part to the ease with which computations may be performed in the Gaussian framework, and also to the flexible ways in which Gaussian processes may be used, for example when combined with thresholding to perform classification tasks via probit models (Neal, 1997; Rasmussen and Williams, 2006) or to find interfaces in Bayesian inversion (Iglesias et al., 2016). Nonetheless there are limits to the sort of phenomena that are readily expressible via direct use of Gaussian processes, such as in the sparse data scenario, where the constructed probability distribution is far from posterior contraction. Recognizing this fact, there have been a number of interesting research activities which seek to represent new phenomena via the hierarchical cascading of Gaussians. Early work of this type includes the PhD thesis of Paciorek (2003) (see also Paciorek and Schervish, 2004) in which the aim is to reproduce spatially non-stationary phenomena, and this is achieved by means of a Gaussian process whose covariance function itself depends on another Gaussian process. This idea was recently re-visited by Roininen et al. (2016), using the precision operator viewpoint, rather than covariance function, and building on the explicit link between Gaussian processes and stochastic partial differential equations (SPDEs) (Lindgren et al., 2011). A different approach was adopted by Damianou and Lawrence (2013) where a Gaussian process was directly composed with another Gaussian process; furthermore the idea was implemented recursively, leading to what is referred to as deep Gaussian processes (DGP). These ingenious constructions open up new possibilities for problems in non-parametric inference and machine learning and the purpose of this paper is to establish, and utilize, a common framework for their study. Relevant to our analysis is the early work of Diaconis and Freedman (1999) which studied iterations of random Lipschitz functions and the conditions required for their convergence.

1.2 Our Contribution

In the paper we make three main contributions:

- We demonstrate a unifying perspective on the hierarchical Gaussian processes described in the previous subsection, leading to a wide class of deep Gaussian processes, with a common framework within which new deep Gaussian processes can be constructed.
- By exploiting the fact that this common framework has a Markovian structure, we interpret the depth of the process in terms of the ergodicity or non-ergodicity of this process; in simple terms ergodic constructions have effective depth given by the mixing time.

- We demonstrate how these processes may be used for inference; in particular we show how a Metropolis-within-Gibbs construction across the levels of the hierarchy can be used to derive sampling tools which are robust to the level of resolution used to represent the functions on a computer.

We also describe numerical experiments which illustrate the theory, and which demonstrate some of the limitations of the framework in the inference context, suggesting the need for further algorithmic innovation and theoretical understanding. We now summarize the results and contributions by direct reference to the main theorems in the paper.

- Theorem 4 shows that a composition-based deep Gaussian process will, with sufficiently many layers, produce samples that are approximately constant. This pathology can be avoided by, for example, increasing the width of each hidden layer, or allowing each layer to depend on the input layer.

- Theorem 8 shows the ergodicity of a class of discretized deep Gaussian processes, constructed using non-stationary covariance *functions*. As a consequence, there is little benefit in adding additional layers after a certain point. This observation elucidates the mechanism underlying the choices of DGPs with a small number of layers for inference in numerous papers (for example Cutajar et al., 2017; Salimbeni and Deisenroth, 2017; Dai et al., 2015).

- Theorem 14 establishes a similar result as Theorem 8 on function space, for a different class of deep Gaussian processes constructed using non-stationary covariance *operators*.

- Theorem 16 establishes the asymptotic properties of a deep Gaussian process formed by iterated convolution of fairly general classes of Gaussian random fields. Specifically it is shown that such processes will either converge weakly to zero or diverge as the number of layers is increased, and so they will provide little flexibility for inference in practice.

1.3 Overview

The general framework in which we place the existing literature, and which we employ to analyze deep Gaussian processes, and to construct algorithms for related inference tasks, is as follows. We consider sequences of functions $\{u_n\}$ which are conditionally Gaussian:

$$u_{n+1}|u_n \sim N\big(m(u_n), C(u_n)\big); \tag{CovOp}$$

here $m(u_n)$ denotes the mean function and $C(u_n)$ the covariance operator. We will also sometimes work with the covariance function representation, in which case we will write

$$u_{n+1}|u_n \sim \mathrm{GP}\big(m(x;u_n), c(x,x';u_n)\big). \tag{GP}$$

Note that the covariance function is the kernel of the covariance operator when the latter is represented as an integral operator over the approximate domain $D \subseteq \mathbb{R}^d$:

$$\big(C(u_n)\phi\big)(x) = \int c(x,x';u_n)\phi(x')dx'.$$

In most of the paper we consider the centred case where $m \equiv 0$, although the flexibility of allowing for non-zero mean will be important in some applications, as discussed in the conclusions. When the mean is zero, the iterations (CovOp) and (GP) can be written in the form

$$u_{n+1} = L(u_n)\xi_{n+1}, \tag{ZeroMean}$$

where $\{\xi_n\}$ form an i.i.d. Gaussian sequence and, for each u, $L(u)$ is a linear operator. For example if the ξ_n are white then the covariance operator is $C(u) = L(u)L(u)^\top$ with $\top$ denoting the adjoint operation and $L(u)$ is a Cholesky factor of $C(u)$. The formulation (ZeroMean) is useful in much of our analysis. For the purpose of this paper, we will refer to any sequence of functions constructed as in (ZeroMean) as a deep Gaussian process.

In section 2 we discuss the hierarchical Gaussian constructions referenced above, and place them in the setting of equations (CovOp), (GP) and (ZeroMean). Section 3 studies the ergodicity of the resulting deep Gaussian processes, using the Markov chain which defines them. In section 4 we provide supporting numerical experiments; we give illustrations of draws from deep Gaussian process priors, and we discuss inference. In the context of inference we describe a methodology for MCMC, using deep Gaussian priors, which is defined in the function space limit and is hence independent of the level of resolution used to represent the functions u_n; numerical illustrations are given. We conclude in section 5 in which we describe generalizations of the settings considered in this paper, and highlight future directions.

1.4 Notation

The structure of the deep Gaussian processes above means that they can be interpreted as Markov chains on a Hilbert space $\mathcal{H}$ of functions. Let $\mathcal{B}(\mathcal{H})$ denote the Borel σ-algebra on $\mathcal{H}$. We denote by $\mathsf{P} : \mathcal{H} \times \mathcal{B}(\mathcal{H}) \to \mathbb{R}$ the one-step transition probability distribution,

$$\mathsf{P}(u, A) = \mathbb{P}(u_n \in A \,|\, u_{n-1} = u), \tag{1}$$

and denote by $\mathsf{P}^n : \mathcal{H} \times \mathcal{B}(\mathcal{H}) \to \mathbb{R}$ the n-step transition probability distribution,

$$\mathsf{P}^n(u, A) = \mathbb{P}(u_n \in A \,|\, u_0 = u). \tag{2}$$

Thus, for example, in the case of the covariance operator construction (CovOp) we have

$$\mathsf{P}(u, \cdot) = N\big(0, C(u)\big),$$

when the mean is zero. This Markovian structure will be exploited when showing ergodicity, or lack of ergodicity, of the chains.

2. Four Constructions

This section provides examples of four constructions of deep Gaussian processes, all of which fall into our general framework. The reader will readily design others.

2.1 Composition

Let $D \subseteq \mathbb{R}^d$, $D' \subseteq \mathbb{R}^l$, $u_n : D \to \mathbb{R}^m$ and $F : \mathbb{R}^m \to D'$. If $\{\xi_n\}$ is a collection of i.i.d. centred Gaussian processes taking values in the space of continuous functions $C(D';\mathbb{R}^m)$ then we define the Markov chain

$$u_{n+1}(x) = \xi_{n+1}\Big(F\big(u_n(x)\big)\Big). \tag{3}$$

The case $m = l$, $F = \mathrm{id}$ and $D = D' = \mathbb{R}^m$ was introduced by Damianou and Lawrence (2013) and the generalization here is inspired by the formulation of Duvenaud et al. (2014). The case where two layers are employed could be interpreted as a form of warped Gaussian process: a generalization of Gaussian processes that have been used successfully in a number of inference problems (Snelson et al., 2004; Schmidt and O'Hagan, 2003).

We note that the mapping $\xi \mapsto \xi \circ F \circ u$ is linear, and we may thus define $L(u)$ by $L(u)\xi = \xi \circ F \circ u$; hence the Markov chain may be written in the form (ZeroMean). If $\xi_1 \sim N(0,\Sigma)$ then the Markov chain has the form (CovOp), with mean zero and $C(u) = L(u)\Sigma L(u)^*$; if $\xi_1 \sim \mathrm{GP}\big(0, k(z,z')\big)$ then the Markov chain has the form (GP) with mean zero and $c(x,x';u) = k\Big(F\big(u(x)\big), F\big(u(x')\big)\Big)$.

2.2 Covariance Function

Paciorek (2003) gives a general strategy to construct anisotropic versions of isotropic covariance functions. Let $\Sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ be such that $\Sigma(z)$ is symmetric positive definite for all $z \in \mathbb{R}^d$, and define the quadratic form

$$Q(x,x') = (x-x')^T \left(\frac{\Sigma(x)+\Sigma(x')}{2}\right)^{-1}(x-x'), \qquad x,x' \in \mathbb{R}^d.$$

If the isotropic correlation function $\rho_S(\cdot)$ is positive definite on $\mathbb{R}^d$, for all $d \in \mathbb{N}$, then the function

$$c(x,x') = \sigma^2 \frac{2^{\frac{d}{2}} \det(\Sigma(x))^{\frac{1}{4}} \det(\Sigma(x'))^{\frac{1}{4}}}{\det(\Sigma(x)+\Sigma(x'))^{\frac{1}{2}}} \rho_S(\sqrt{Q(x,x')})$$

is positive definite on $\mathbb{R}^d \times \mathbb{R}^d$ and may thus be used as a covariance function. We make these statements precise below. If we choose Σ to depend on u_n then this may be used as the basis of a deep Gaussian process. To be concrete we choose

$$\Sigma(x) = F\big(u(x)\big) I_d$$

where $F : \mathbb{R} \to \mathbb{R}_{\geq 0}$ for $u : D \subseteq \mathbb{R}^d \to \mathbb{R}$. We then write $c(x,x';u)$. Now let $u_n : D \to \mathbb{R}$ and consider the Markov chain (GP) in the mean zero case. Paciorek (2003) considered this iteration over one-step with $u_0 \sim \mathrm{GP}(0, \sigma^2\rho_S(\|x-x'\|))$ and u_1 was shown to exhibit interesting non-stationary effects. Here we generalize and consider the deep process that results from this construction for arbitrary $n \in \mathbb{N}$. By considering the covariance operator

$$\big(C(u)\varphi\big)(x) = \int_{\mathbb{R}^d} c(x,x';u)\varphi(x')\,\mathrm{d}x'$$

we may write the iteration in the form (CovOp). The form (ZeroMean) follows with $L(u) = C(u)^{\frac{1}{2}}$ and ξ_{n+1} being white noise.

Various generalizations of this construction are possible, for example allowing the pointwise variance of the process σ^2 to be spatially varying (Heinonen et al., 2016) and to depend on $u_n(x)$. These may be useful in applications, but we confine our analysis to the simpler setting for expository purposes; however in Remark 12 we discuss this generalization.

In order to make the statements made above precise, let $\rho_S : [0, \infty) \to \mathbb{R}$ be a stationary covariance kernel, where the covariance between locations x and y depends only on the Euclidean distance $\|x - y\|_2$. We make the following assumption on ρ_S.

Assumptions 1 *(i) The covariance kernel $\rho_S(\|x-y\|_2)$ is positive definite*[1] *on $\mathbb{R}^d \times \mathbb{R}^d$: for any $N \in \mathbb{N}$, $b \in \mathbb{R}^N \backslash \{0\}$ and pairwise distinct $\{x_i\}_{i=1}^N \subseteq \mathbb{R}^d$, we have*

$$\sum_{i=1}^{N} \sum_{j=1}^{N} b_i b_j \rho_S\big(\|x_i - x_j\|_2\big) > 0.$$

(ii) ρ_S is normalized to be a correlation kernel, i.e. $\rho_S(0) = 1$.

Using (Wendland, 2004, Theorem 6.11), sufficient conditions for ρ_S to fulfill Assumptions 1(i) are that ρ_S, as a function of $x - y$, is continuous, bounded and in $L^1(\mathbb{R}^d)$, with a Fourier transform that is non-negative and non-vanishing. These sufficient conditions are satisfied, for example, for the family of Matèrn covariance functions and the Gaussian covariance. To satisfy Assumptions 1(ii), any positive definite kernel $\tilde{\rho}_S$ can simply be rescaled by $\tilde{\rho}_S(0)$.

We now have the following proposition, a slightly weaker version of which is proved by Paciorek (2003), where it is shown that $\rho(\cdot, \cdot)$ is positive semi-definite if ρ_S is positive semi-definite. Our proof, which is in the Appendix, follows closely that of (Paciorek, 2003, Theorem 1), but sharpens the result using a characterization of positive definite kernels proved in (Wendland, 2004, Theorem 7.14).

Proposition 1 *Let Assumptions 1 hold. Suppose $\Sigma : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ is such that $\Sigma(z)$ is symmetric positive definite for all $z \in \mathbb{R}^d$, and define the quadratic form*

$$Q(x, x') = (x - x')^T \left(\frac{\Sigma(x) + \Sigma(x')}{2} \right)^{-1} (x - x'), \qquad x, x' \in \mathbb{R}^d.$$

Then the function $\rho(\cdot, \cdot)$, defined by

$$\rho(x, x') = \frac{2^{\frac{d}{2}} |\Sigma(x)|^{\frac{1}{4}} |\Sigma(x')|^{\frac{1}{4}}}{|\Sigma(x) + \Sigma(x')|^{\frac{1}{2}}} \rho_S\Big(\sqrt{Q(x, x')}\Big),$$

is positive definite on $\mathbb{R}^d \times \mathbb{R}^d$, for any $d \in \mathbb{N}$, and is a non-stationary correlation function.

1. If the double sum in this definition is only non-negative, we say that the kernel ρ_S is positive semi-definite. We are thus adopting the terminology used by Wendland (2004), where the kernel ρ_S is called positive definite if the double sum in Assumptions 1(i) is positive, and positive semi-definite if the sum is non-negative. For historical reasons, there is an alternative terminology, used by for example Paciorek (2003), where our notion of positive definite is referred to as strictly positive definite, and our notion of positive semi-definite is referred to as positive definite.

Non-stationary covariance functions $c(x, y)$, for which $c(x, x) \neq 1$, can be obtained from the non-stationary correlation function $\rho(x, y)$ through multiplication by a standard deviation function $\sigma : \mathbb{R}^d \to \mathbb{R}$, in which case we have $c(x, y) = \sigma(x)\sigma(y)\rho(x, y)$. Since the product of two positive definite kernels is also positive definite by (Wendland, 2004, Theorem 6.2), the kernel $c(x, y)$ can be ensured to be positive definite by a proper choice of σ. We discuss generalizations such as this in the conclusions section 5.

We are interested in studying the behaviour of Gaussian processes with non-stationary correlation functions $\rho(x, y)$ of the form derived in Proposition 1, in the particular case where the matrices $\Sigma(z)$ are derived from another Gaussian process. Specifically, we consider the following hierarchy of conditionally Gaussian processes on a bounded domain $D \subseteq \mathbb{R}^d$ defined as follows:

$$u_0 \sim \mathrm{GP}(0, \rho_{\mathrm{S}}(\cdot)), \tag{4a}$$

$$u_{n+1}|u_n \sim \mathrm{GP}(0, \rho(\cdot, \cdot; u_n)), \quad \text{for } n \in \mathbb{N}. \tag{4b}$$

Here, $\rho(\cdot, \cdot; u_n)$ denotes a non-stationary correlation function constructed from $\rho_{\mathrm{S}}(\cdot)$ as in Proposition 1, with the map Σ defined through u_n. Typical choices for Σ are $\Sigma(z) = (u_n(z))^2 \, \mathrm{I}_d$ and $\Sigma(z) = \exp(u_n(z)) \, \mathrm{I}_d$. Choices such as the first of these lead to the possibility of positive semi-definite Σ and, in the worst case, $\Sigma \equiv 0$. If $\Sigma \equiv 0$ the resulting correlation function is given by

$$\rho_{\mathrm{S}}(0) = 1, \quad \text{and} \quad \rho_{\mathrm{S}}(r) = 0 \quad \text{for any } r > 0.$$

This does not correspond to any (function valued) Gaussian process on $\mathbb{R}^d$ (Kallianpur, 2013): heuristically the resulting process would be a white noise process, but normalized to zero. However, it is possible to sample from any set of finite dimensional distributions when $\Sigma \equiv 0$: the correlation matrix is then the identity. To allow for the possibility of $F(\cdot)$ taking the value zero, we therefore only study the finite dimensional process defined as follows:

$$\mathbf{u}_0 \sim N(0, \mathbf{R}_{\mathrm{S}}), \tag{5a}$$

$$\mathbf{u}_{n+1} \mid \mathbf{u}_n \sim N(0, \mathbf{R}(\mathbf{u}_n)), \quad \text{for } n \in \mathbb{N}. \tag{5b}$$

The vector $\mathbf{u}_n$ has entries $(\mathbf{u}_n)_i = u_n(x_i)$. Here, $\mathbf{R}_{\mathrm{S}}$ is the covariance matrix with entries $(\mathbf{R}_{\mathrm{S}})_{ij} = \rho_{\mathrm{S}}(\|x_i - x_j\|_2)$, and $\mathbf{R}(\mathbf{u}_n)$ is the covariance matrix with entries $(\mathbf{R}(\mathbf{u}_n))_{ij} = \rho(x_i, x_j; u_n)$. The set $\{x_j\}$ comprises a finite set of points in $\mathbb{R}^d$.

We may now generalize Proposition 1 to allow for Σ becoming zero. In order to do this we make the following assumptions:

Assumptions 2 *(i) We have $\Sigma(z) = G(z)\mathrm{I}_d$, for some non-negative, bounded function $G : \mathbb{R} \to \mathbb{R}_{\geq 0}$.*

(ii) The correlation function ρ_{S} is continuous, with $\lim_{r\to\infty} \rho_{\mathrm{S}}(r) = 0$.

We then have the following result on the positive-definiteness of $\rho(\cdot, \cdot)$,

Proposition 2 *Let Assumptions 1 and 2 hold. Then the kernel $\rho(\cdot, \cdot)$ defined in Proposition 1 is positive definite on $\mathbb{R}^d \times \mathbb{R}^d$.*

Remark 3 *This proposition applies to the process* (5) *with* $\Sigma(z) = F\big(u_n(z)\big)\mathrm{I}_d$ *and* $F : \mathbb{R} \to \mathbb{R}_{\geq 0}$ *locally bounded, by taking* $G = F \circ u_n$*, proving that* $\rho(\cdot, \cdot; u_n)$ *is positive definite on* $D \times D$ *for all bounded functions* u_n *on* D*. Here we generalize the notion of positive-definite in the obvious way to apply on* $D \subseteq \mathbb{R}^d$ *rather than on the whole of* $\mathbb{R}^d$*.*

2.3 Covariance Operator

Here we demonstrate how precision (inverse covariance) operators may be used to make deep Gaussian processes. Because precision operators encode conditional independence and sparsity this can be a very attractive basis for fast computations (Lindgren et al., 2011). Our approach is inspired by the hierarchical Gaussian process introduced by Roininen et al. (2016), where one-step of the Markov chain which we introduce here was considered. Let $D \subseteq \mathbb{R}^d$, $u_n : D \to \mathbb{R}$ and $X := C(D; \mathbb{R})$. Assume that $F : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a bounded function. Let C_- be a covariance operator associated to a Gaussian process taking values in X and let P be the associated precision operator. Define the multiplication operator $\Gamma(u)$ by $\big(\Gamma(u)v\big)(x) = F\big(u(x)\big)v(x)$ and the covariance operator $C(u)$ by

$$C(u)^{-1} = P + \Gamma(u)$$

and consider the Markov chain (CovOp) with mean zero; this defines our deep Gaussian process. We note that formulation (GP) can be obtained by observing that the covariance function $c(u) := c(x, x'; u)$ is the Green's function associated with the precision operator for $C(u)$:

$$C(u)^{-1} c(\cdot, x'; u) = \delta_{x'}(\cdot)$$

where $\delta_{x'}$ is a Dirac delta function centred at point x'. Computationally we will typically choose P to be a differential operator, noting that then fast methods may be employed to sample the Gaussian process $u_{n+1}|u_n$ by means of SPDEs (Lindgren et al., 2011; Dashti and Stuart, 2017). If P is chosen as a differential operator, then the order of this operator will be related to the order of regularity of samples, and F will be related to the length scale of the samples. These relations are made explicit in the case of certain Whittle-Matérn distributions when F is constant (Lindgren et al., 2011); some boundary effects may be present when $D \neq \mathbb{R}^d$, though methodology is available to ameliorate these (Daon and Stadler, 2018). As in the previous subsection, the form (ZeroMean) follows with $L(u) = C(u)^{\frac{1}{2}}$ and ξ_{n+1} being white noise.

Generalizations of the construction in this subsection are possible, and we highlight these in subsection 5; however for expository purposes we confine our analysis to the setting described in this subsection. For theoretical investigation of the equivalence, as measures, of Gaussians defined by addition of an operator to a given precision operator (see Pinski et al., 2015).

2.4 Convolution

We consider the case (ZeroMean) where $L(u)\xi := u * \xi$ is a convolution. To be concrete we let $D = [0, 1]^d$ and construct a sequence of functions $u_n : D \to \mathbb{R}$ (or $u_n : D \to \mathbb{C}$) defined via the iteration

$$u_{n+1}(x) = (u_n * \xi_{n+1})(x) := \int_{[0,1]^d} u_n(x - y)\xi_{n+1}(y)\, dy,$$

where $\{\xi_n\}$ are a sequence of i.i.d. centred real-valued Gaussian random functions on D. Here we implicitly work with periodic extension of u_n from D to the whole of $\mathbb{R}^d$ in order to define the convolution.

3. The Role of Ergodicity

The purpose of this section is to demonstrate that the iteration (ZeroMean) is, in many situations, ergodic. This has the practical implication that the effective depth of the deep Gaussian process is limited by the mixing time of the Markov chain. In some cases the ergodic behaviour may be trivial (convergence to a constant). Furthermore, even if the chain is not ergodic, the large iteration number dynamics may blow-up, prohibiting use of the iteration at significant depth. The take home message is that in many cases the effective depth is not that great. Great care will be needed to design deep Gaussian processes whose depth, and hence approximation power, is substantial. This issue was first identified by Duvenaud et al. (2014), and we here provide a more general analysis of the phenomenon within the broad framework we have introduced for deep Gaussian processes.

3.1 Composition

We first consider the case where the iteration is defined by (3), which includes examples considered by Damianou and Lawrence (2013); Duvenaud et al. (2014). Duvenaud et al. (2014) observed that after a number of iterations, sample paths are approximately piecewise constant. We investigate this effect in the context of ergodicity. We first make two observations:

(i) if u_0 is piecewise constant, then u_n is piecewise constant for all $n \in \mathbb{N}$;

(ii) if u_0 has discontinuity set Z_0, and Z_n denotes the discontinuity set of the nth iterate, then $\mathsf{Z}_{n+1} \subseteq \mathsf{Z}_n$ for all $n \in \mathbb{N}$.

Due to point (ii) above, if the sequence $\{u_n\}$ is to be ergodic, then necessarily it must be the case that $\mathsf{Z}_n \to \varnothing$, or else the process will have retained knowledge of the initial condition. In particular, if the initial condition is piecewise constant, then ergodicity would force the limit to be constant in space.

In what follows we assume that the iteration is given by

$$u_{n+1}(x) = \xi_{n+1}\big(u_n(x)\big), \quad \xi^j_{n+1} \sim \mathrm{GP}\big(0, h(\|x - x'\|_2)\big) \text{ i.i.d.}$$

where h is a stationary covariance function. We therefore make the choice $m = l$ and $F = \mathrm{id}$ in (3) so that we are in the same setup as Damianou and Lawrence (2013); Duvenaud et al. (2014); the inclusion of more general maps F is discussed in Remark 5. Then for any $x, x' \in \mathbb{R}$ we have

$$\begin{pmatrix} u^j_{n+1}(x) \\ u^j_{n+1}(x') \end{pmatrix} \Big| u_n \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} h(0) & h\big(\|u_n(x) - u_n(x')\|_2\big) \\ h\big(\|u_n(x) - u_n(x')\|_2\big) & h(0) \end{pmatrix} \right).$$

A common choice of covariance function is the squared exponential kernel:

$$h(z) = \sigma^2 e^{-z^2/2w^2} \tag{6}$$

where $\sigma^2, w^2 > 0$ are scalar parameters. In Duvenaud et al. (2014), in the case $m = d = 1$, the choice $\sigma^2/w^2 = \pi/2$ is made above to ensure that the expected magnitude of the derivative remains constant through iterations. We show in the next proposition that if σ^2, w^2 are chosen such that $\sigma^2 < w^2/m$, then the limiting process is trivial in a sense to be made precise.

Theorem 4 *Assume that $h(\cdot)$ is given by the squared exponential kernel (6) and that u_0 is bounded on bounded sets almost-surely. Then if $\sigma^2 < w^2/m$,*

$$\mathbb{P}\big(\|u_n(x) - u_n(x')\|_2 \to 0 \text{ for all } x, x' \in D\big) = 1$$

where $\mathbb{P}$ denotes the law of the process $\{u_n\}$ over the probability space Ω.

Proof Since $1 - e^{-x} \leq x$ for $x \geq 0$ it follows that, for all $z \in \mathbb{R}$,

$$2h(0) - 2h(z) \leq \frac{\sigma^2}{w^2} z^2,$$

with equality when $z = 0$. Then we have

$$\begin{aligned}
\mathbb{E}\big(\|u_n(x) - u_n(x')\|_2^2 \big| u_{n-1}\big) &= \sum_{j=1}^{m} \mathbb{E}\big(|u_n^j(x) - u_n^j(x')|^2 \big| u_{n-1}\big) \\
&= \sum_{j=1}^{m} \big(2h(0) - 2h\big(\|u_{n-1}(x) - u_{n-1}(x')\|_2\big)\big) \\
&\leq m\frac{\sigma^2}{w^2}\|u_{n-1}(x) - u_{n-1}(x')\|_2^2
\end{aligned}$$

and so using induction and the tower property of conditional expectations,

$$\begin{aligned}
\mathbb{E}\|u_n(x) - u_n(x')\|_2^2 &\leq \left(\frac{m\sigma^2}{w^2}\right)\mathbb{E}\|u_{n-1}(x) - u_{n-1}(x')\|_2^2 \\
&\leq \left(\frac{m\sigma^2}{w^2}\right)^n \mathbb{E}\|u_0(x) - u_0(x')\|_2^2 \\
&\leq \left(\frac{m\sigma^2}{w^2}\right)^n \kappa(x, x')
\end{aligned}$$

for some constant $\kappa(x, x')$. By the Markov inequality, we see that for any $\varepsilon > 0$,

$$\mathbb{P}\big(\|u_n(x) - u_n(x')\|_2 \geq \varepsilon\big) \leq \frac{1}{\varepsilon^2}\left(\frac{m\sigma^2}{w^2}\right)^n \kappa(x, x'), \tag{7}$$

and so, since $\sigma^2 < w^2/m$, applying the first Borel-Cantelli lemma we deduce that

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty}\{\|u_m(x) - u_m(x')\|_2 \geq \varepsilon\}\right) = 0.$$

We therefore have that

$$\begin{aligned}\mathbb{P}(\|u_n(x)-u_n(x')\|_2\to 0) &= \mathbb{P}\left(\bigcap_{k=1}^{\infty}\bigcup_{n=1}^{\infty}\bigcap_{m=n}^{\infty}\{\|u_m(x)-u_m(x')\|_2<1/k\}\right)\\ &= 1-\mathbb{P}\left(\bigcup_{k=1}^{\infty}\left(\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty}\{\|u_m(x)-u_m(x')\|_2\geq 1/k\}\right)\right)\\ &\geq 1-\sum_{k=1}^{\infty}\mathbb{P}\left(\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty}\{\|u_m(x)-u_m(x')\|_2\geq 1/k\}\right)=1.\end{aligned}$$

Hence given any $x, x' \in D$, we can find $\Omega(x,x') \subseteq \Omega$ with $\mathbb{P}(\Omega(x,x'))=1$ such that for any $\omega \in \Omega(x,x')$, $\|u_n(x;\omega)-u_n(x';\omega)\|_2 \to 0$. Let $\{q_j\}$ be a countable dense subset of D, and define

$$\Omega_* = \bigcap_{i,j\in\mathbb{N}} \Omega(q_i,q_j),$$

noting that $\mathbb{P}(\Omega_*)=1$. Then for any $\omega\in\Omega_*$ and any $x,x'\in\{q_j\}$, $\|u_n(x;\omega)-u_n(x';\omega)\|_2\to 0$. Since sample paths are almost-surely continuous, the above can be extended to all $x,x'\in D$, so that

$$\mathbb{P}\big(\|u_n(x)-u_n(x')\|_2\to 0 \text{ for all } x,x'\in D\big)=1.$$

■

Remark 5 *1. If a more general transformation map $F:\mathbb{R}^m\to D'$ is included, then the above result still holds provided we take $\sigma^2 < w^2/(\|F'\|_\infty m)$. The convergence to a constant hence occurs when the length scale w is large or $\|F'\|_\infty$ is small (so each Gaussian random field doesn't change too rapidly across the domain), or when the amplitude σ is small (so inputs are not warped too far).*

2. *The condition of the above theorem is less likely to be satisfied as the width m of each layer is increased, and so this triviality pathology is unlikely to arise for large m; this may be observed in practice numerically.*

3. *Following Neal (1995); Duvenaud et al. (2014), recent works (such as Dai et al., 2015; Cutajar et al., 2017) connect all layers to the input layer in order to avoid certain pathologies. The Markovian structure of the process is maintained in this case: with the above notation, the process is then defined by*

$$u_{n+1}(x)=\xi_{n+1}(u_n(x),x), \quad \xi^j_{n+1}\sim \mathrm{GP}\big(0,h(\|x-x'\|_2)\big)\ i.i.d,$$

where now $\xi_n:\mathbb{R}^m\times\mathbb{R}^d\to\mathbb{R}^m$. Defining $\beta=m\sigma^2/w^2<1$, if $\sigma\geq 1$ we may use the same argument as the proof above to deduce that

$$\mathbb{E}\big(\|u_n(x)-u_n(x')\|_2^2\big|u_{n-1}\big)\leq \beta\|u_{n-1}(x)-u_{n-1}(x')\|_2^2+\beta\|x-x'\|_2^2,$$

which leads to

$$\mathbb{E}\|u_n(x) - u_n(x')\|_2^2 \leq \beta^n \mathbb{E}\|u_0(x) - u_0(x')\|_2^2 + \beta\left(\frac{1-\beta^n}{1-\beta}\right)\|x - x'\|_2^2.$$

The right hand side does not vanish as $n \to \infty$, and so we can no longer use the first Borel-Cantelli lemma to reach the same conclusion as the case where the layers are not connected to the input layer. This could provide some intuition as to why including the connection of each layer to the input layer provides greater stability than not doing so.

3.2 Covariance Function

In order to study ergodicity of the deep Gaussian process defined through covariance functions, we will restrict attention in the remainder of this subsection to hierarchies of finite-dimensional multivariate Gaussian random variables as in (5). Note that although we have here defined $\mathbf{u}_0 \sim N(0, \mathbf{R}_S)$, following e.g. Paciorek (2003), the ergodicity of the deep Gaussian process will be proved for fixed $\mathbf{u}_0 \in \mathbb{R}^N$ (cf. Theorem 8). The following result is immediate from Proposition 2.

Corollary 6 *Let Assumptions 1 and 2 hold. Then the covariance matrix $\mathbf{R}(\mathbf{u}_n)$ is positive definite for all $\mathbf{u}_n \in C$, for any compact subset of $C \subseteq \mathbb{R}^N$.*

Note that, because we have chosen to work with a correlation kernel, we have

$$\mathrm{Tr}\big(\mathbf{R}(\mathbf{u}_n)\big) = N. \tag{8}$$

We will use this fact explicitly in the ergodicity proof; however it may be relaxed as discussed in the Remark 12 below.

We view the sequence of random variables $\{\mathbf{u}_n\}_{n=0}^{\infty}$ as a Markov chain, with $\mathbf{u}_0 \in \mathbb{R}^N$ given, and we want to show the existence of a stationary distribution. Recall the one-step transition kernel P of the Markov chain given by (1), and its $n-$fold composition given by (2). In order to prove ergodicity of the Markov chain we will follow the proof technique of Mattingly et al. (2002); Meyn and Tweedie (2012), which establishes geometric ergodicity with the following proposition.

Proposition 7 *Suppose the Markov chain $\{\mathbf{u}_n\}_{n=0}^{\infty}$ satisfies, for some compact set $C \in \mathcal{B}(\mathbb{R}^N)$, the following:*

(i) For some $y^* \in int(C)$ and for any $\delta > 0$, we have

$$\mathsf{P}(u, \mathcal{B}_\delta(y^*)) > 0 \qquad \text{for all } u \in C.$$

(ii) The transition kernel $\mathsf{P}(u, \cdot)$ possesses a density $p(u, y)$ in C, precisely

$$\mathsf{P}(u, A) = \int_A p(u, y)\,\mathrm{d}y, \quad \text{for all } u \in C,\ A \in \mathcal{B}(\mathbb{R}^N) \cap \mathcal{B}(C),$$

and $p(u, y)$ is jointly continuous on $C \times C$.

(iii) There is a function $V : \mathbb{R}^N \to [1, \infty)$, *with* $\lim_{u\to\infty} V(u) = \infty$, *and real numbers* $\alpha \in (0, 1)$ *and* $\beta \in [0, \infty)$ *such that*

$$\mathbb{E}(V(\mathbf{u}_{n+1}) \mid \mathbf{u}_n) \leq \alpha V(\mathbf{u}_n) + \beta.$$

If we can choose the compact set C *such that*

$$C = \left\{ u : V(u) \leq \frac{2\beta}{\gamma - \alpha} \right\},$$

for some $\gamma \in (\sqrt{\alpha}, 1)$, *then there exists a unique invariant measure* π. *Furthermore, there is* $r(\gamma) \in (0, 1)$ *and* $\kappa(\gamma) \in (0, \infty)$ *such that for all* $\mathbf{u}_0 \in \mathbb{R}^N$ *and all measurable* g *with* $|g(u)| \leq V(u)$ *for all* $u \in \mathbb{R}^N$, *we have*

$$|\mathbb{E}^{\mathsf{P}^n(\mathbf{u}_0,\cdot)}(g) - \pi(g)| \leq \kappa r^n V(\mathbf{u}_0).$$

We may verify the assumptions of Proposition 7 leading to the following theorem concerning the ergodicity of deep Gaussian processes defined via the covariance function:

Theorem 8 *Suppose Assumptions 1 and 2 hold. Then the Markov chain* $\{\mathbf{u}_n\}_{n=0}^{\infty}$ *satisfies the assumptions of Proposition 7. As a consequence, there exists* $\varepsilon \in (0, 1)$ *such that for any* $\mathbf{u}_0 \in \mathbb{R}^N$, *there is a* $K(\mathbf{u}_0) > 0$ *with*

$$\|\mathsf{P}^n(\mathbf{u}_0, \cdot) - \pi\|_{TV} \leq K(1 - \varepsilon)^n \quad \textit{for all } n \in \mathbb{N},$$

and so the chain is ergodic.

The proof rests on the following three lemmas, and is given after stating and proving them. The first lemma shows that, on average, the norm of states of the chain remains constant as the length of the chain is increased. The second shows that, given any current state in $\mathbb{R}^N$ and any ball around the origin in $\mathbb{R}^N$, there is a positive probability that the next state will belong to that ball. The third lemma shows that the probability that the Markov chain moves to a set may be found via integration of a continuous function over that set.

Lemma 9 (Boundedness) *Suppose Assumptions 1 and 2 hold. For all* $n \in \mathbb{N}$, *we have*

$$\mathbb{E}\big(\|\mathbf{u}_{n+1}\|_2^2 \mid \mathbf{u}_n\big) = N.$$

Proof Let $n \geq 0$. Since the random variable $\mathbf{u}_{n+1}|\mathbf{u}_n$ has zero mean, the linearity of expectation implies (using (8)) that

$$\mathbb{E}\left(\|\mathbf{u}_{n+1}\|_2^2 \mid \mathbf{u}_n\right) = \mathbb{E}\left(\sum_{j=1}^{N} (\mathbf{u}_{n+1})_j^2 \,\middle|\, \mathbf{u}_n \right) = \mathrm{Tr}(\mathbf{R}(\mathbf{u}_n)) = N,$$

for all $n \in \mathbb{N}$. ■

Lemma 10 (Positive probability of a ball around zero) *Suppose Assumptions 1 and 2 hold. For all $u \in \mathbb{R}^N$ and $\delta > 0$, we have*

$$\mathsf{P}\big(u, \mathcal{B}_\delta(0)\big) > 0.$$

Proof We have the equality $\mathbf{u}_{n+1}|(\mathbf{u}_n = u) = \sqrt{\mathbf{R}(u)}\xi_{n+1}$ in distribution, where $\sqrt{\mathbf{R}(u)}$ denotes the Cholesky factor of the correlation matrix $\mathbf{R}(u)$ and $\xi_{n+1} \sim \mathrm{N}(0, \mathrm{I}_N)$. Then

$$\begin{aligned}
\mathsf{P}\big(u, \mathcal{B}_\delta(0)\big) &= \mathbb{P}\big(\|\mathbf{u}_n\|_2 \le \delta \,|\, \mathbf{u}_{n-1} = u\big) \\
&= \mathbb{P}\Big(\big\|\sqrt{\mathbf{R}(u)}\xi_{n+1}\big\|_2 \le \delta\Big) \\
&\ge \mathbb{P}\Big(\big\|\sqrt{\mathbf{R}(u)}\big\|_2 \|\xi_{n+1}\|_2 \le \delta\Big) \\
&= \mathbb{P}\Big(\|\xi_{n+1}\|_2 \le \delta \big\|\sqrt{\mathbf{R}(u)}\big\|_2^{-1}\Big).
\end{aligned}$$

To show that the latter probability is positive, we need to show that $\delta\big\|\sqrt{\mathbf{R}(u)}\big\|_2^{-1} > 0$. Since $\delta > 0$ is fixed, we only need to show $\big\|\sqrt{\mathbf{R}(u)}\big\|_2 < \infty$. Since $\big\|\sqrt{\mathbf{R}(u)}\big\|_2^2 = \rho(\mathbf{R}(u))$, the spectral radius of $\mathbf{R}(u)$, we have

$$\big\|\sqrt{\mathbf{R}(u)}\big\|_2^2 = \rho(\mathbf{R}(u)) \le \mathrm{Tr}(\mathbf{R}(u)) = N.$$

The claim then follows. ■

Lemma 11 (Transition probability has a density) *Suppose Assumptions 1 and 2 hold. Then the transition probability $\mathsf{P}(u, \cdot)$ has a jointly continuous density $p(u, y)$ for all $u \in C$, for any compact set $C \subseteq \mathbb{R}^N$.*

Proof We have $\mathbf{u}_{n+1}|(\mathbf{u}_n = u) \sim N(0, \mathbf{R}(u))$, and the existence of a jointly continuous density of the transition probability in C follows if $\mathbf{R}(u)$ is positive definite for all $u \in C$. The claim then follows by Proposition 2. ■

We may now use the three preceding lemmas to prove the main ergodic theorem for deep Gaussian processes defined through the covariance function.

Proof of Theorem 8 Lemma 10 shows that assumption (i) is satisfied, for any C containing $y^* = 0$, and Lemma 11 shows that assumption (ii) is satisfied, for any compact set C. It follows from Lemma 9 that assumption (iii) is satisfied, with $V(u) = \|u\|_2^2 + 1$, any $\alpha \in (0, 1)$ and $\beta = N + 1$. Now choose $\alpha = 1/4$ and $\gamma = 3/4 \in (\sqrt{\alpha}, 1)$, so that the set

$$C = \left\{u : V(u) \le \frac{2\beta}{\gamma - \alpha}\right\} = \big\{u : \|u\|_2^2 \le 4N + 4\big\}$$

is compact. Then there is a unique invariant measure π, and there is $r(\gamma) \in (0, 1)$ and $\kappa(\gamma) \in (0, \infty)$ such that for $\mathbf{u}_0 \in \mathbb{R}^N$ and all measurable g with $|g(u)| \le V(u)$ for all $u \in \mathbb{R}^N$, we have

$$|\mathbb{E}^{\mathsf{P}^n(\mathbf{u}_0, \cdot)}(g) - \pi(g)| \le \kappa r^n V(\mathbf{u}_0). \tag{9}$$

Since $V(u) \geq 1$ for all $u \in \mathbb{R}^N$, the above holds in particular for all measurable g with $\|g\|_\infty \leq 1$. Taking the supremum over all such g in (9) yields the given total variation bound, with $K = \kappa V(\mathbf{u}_0)$ and $\varepsilon = 1 - r$. ■

Remark 12 *(Covariance vs correlation kernels) In this subsection we have restricted our attention to correlation kernels $\rho_S(\|x_i - x_j\|_2)$ and $\rho(x_i, x_j; u_n)$, rather than more general covariance kernels*

$$c_S(\|x_i - x_j\|_2) = \sigma_S^2 \rho_S(\|x_i - x_j\|_2),$$
$$c(x_i, x_j; u_n) = \sigma(x_i; u_n)\sigma(x_j; u_n)\rho(x_i, x_j; u_n),$$

for stationary and non-stationary marginal standard deviation functions $\sigma_S \in (0, \infty)$ and $\sigma : \mathbb{R}^d \to (0, \infty)$ respectively. This restriction is solely for ease of presentation; the analysis presented readily extends to $c(x_i, x_j; u_n)$, under suitable assumptions on σ. In particular the analysis may be adapted to the case of general covariance kernels $c_S(\|x_i - x_j\|_2)$ and $c(x_i, x_j; u_n)$ under the assumption that there exist positive constants σ^-, σ^+ such that $\sigma^- \leq \sigma(z) \leq \sigma^+$, for all $z \in \mathbb{R}^d$. When general covariances are used then it is possible to ensure that every multivariate Gaussian random variable in the hierarchy is of the same amplitude by scaling the corresponding covariance matrix $\mathbf{C}(\mathbf{u}_n)$ to have constant trace N at each iteration n; the average variance over all points $\{x_i\}_{i=1}^N$ is then 1 for every n.

3.3 Covariance Operator

We consider the class of covariance operators introduced in section 2.3 and show that, under precise assumptions detailed below, the iteration (ZeroMean) produces an ergodic Markov chain. Unlike the previous subsection, where we worked on $\mathbb{R}^N$, here we will work on the separable Hilbert space $\mathcal{H} = L^2(D; \mathbb{R})$. To begin with, define the precision operators (densely defined on $\mathcal{H}$: Hairer et al., 2005; Pinski et al., 2015),

$$C_-^{-1} = P,$$
$$C_+^{-1} = P + F_+ I,$$
$$C(u)^{-1} = P + \Gamma(u), \quad u \in \mathcal{H},$$

and the probability measures

$$\mu_- = N(0, C_-),$$
$$\mu_+ = N(0, C_+),$$
$$\mu(\cdot\,; u) = N(0, C(u)), \quad u \in \mathcal{H}.$$

Throughout the rest of this section we make the following assumptions on C_- and F:

Assumptions 3

1. *The operator $C_- : \mathcal{H} \to \mathcal{H}$ is symmetric and positive, and its eigenvalues $\{\lambda_j^2\}$ have algebraic decay $\lambda_j^2 \asymp j^{-r}$ for some $r > 1$.*
2. *The function $F : \mathcal{H} \to \mathbb{R}$ is continuous, and there exists $F_+ \geq 0$ such that $0 \leq F(u) \leq F_+$ for all $u \in \mathcal{H}$.*

Remark 13 *1. The assumption on algebraic decay of the eigenvalues can be relaxed to the operator C_- being trace-class on $\mathcal{H}$; however the arguments that follow are cleaner when we assume this explicit decay which, of course, implies the trace condition. Note also that, under the stated assumption on algebraic decay, Gaussian measures on $L^2(D;\mathbb{R})$ will be supported on $X = C(D;\mathbb{R})$ under mild conditions on the eigenfunctions of C_- (Dashti and Stuart, 2017) so that $F\big(u(x)\big)$ will be defined for all $x \in D$ rather than x a.e. in D. Then $\Gamma(u)v$ makes sense pointwise when $v \in X$.*

2. The assumed form of the precision operator together with Assumptions 3 mean that the resulting family of measures $\{\mu(\cdot\,;u)\}_{u\in\mathcal{H}}$ will be mutually equivalent. This allows for the total variation metric between measures to be used, and a concise proof of ergodicity to be obtained. If the measures were singular, a different metric such as the Wasserstein metric would be required to quantify the convergence.

We now prove the following ergodic theorem for the deep Gaussian processes constructed through covariance operators.

Theorem 14 *Let Assumptions 3 hold, and let the Markov chain $\{u_n\}$ be given by (ZeroMean) with $L(u) = C(u)^{\frac{1}{2}}$ as defined above. Then there exists a unique invariant distribution π, and there exists $\varepsilon > 0$ such that for any $u_0 \in \mathcal{H}$,*

$$\|\mathsf{P}^n(u_0,\cdot) - \pi\|_{TV} \le (1-\varepsilon)^n \quad \text{for all } n \in \mathbb{N}.$$

In particular, the chain is ergodic.

The following lemma will be used to show a minorization condition, as well as establish further notation, key to the proof of Theorem 14 which follows it. It essentially shows a stronger form of equivalence of the family of measures $\{\mu(\cdot,u)\}_{u\in\mathcal{H}}$.

Lemma 15 *Let Assumptions 3 hold. Then there exists $\varepsilon > 0$ such that for any $u, v \in \mathcal{H}$,*

$$\frac{\mathrm{d}\mu(\cdot\,;u)}{\mathrm{d}\mu_+}(v) \ge \varepsilon.$$

Proof The assumptions on F mean that the measures $\mu(\cdot\,;u)$, μ_- and μ_+ are mutually absolutely continuous, with

$$\begin{aligned}
\frac{\mathrm{d}\mu(\cdot\,;u)}{\mathrm{d}\mu_-}(v) &= \frac{1}{Z(u)}\exp\Big(-\frac{1}{2}\langle v, F(u)v\rangle\Big),\\
Z(u) &= \mathbb{E}^{\mu_-}\Big[\exp\Big(-\frac{1}{2}\langle v, F(u)v\rangle\Big)\Big];\\
\frac{\mathrm{d}\mu_+}{\mathrm{d}\mu_-}(v) &= \frac{1}{Z_+}\exp\Big(-\frac{1}{2}\langle v, F_+v\rangle\Big),\\
Z_+ &= \mathbb{E}^{\mu_-}\Big[\exp\Big(-\frac{1}{2}\langle v, F_+v\rangle\Big)\Big].
\end{aligned}$$

Observe that we may bound $Z(u) \leq 1$ uniformly in $u \in H$ since $F \geq 0$. Additionally, we have that

$$Z_+ \geq \mathbb{E}^{\mu_-}\left[\exp\left(-\frac{1}{2}\langle v, F_+ v\rangle\right)\mathbb{1}_{\|v\|^2\leq 1}\right] \geq \exp\left(-\frac{1}{2}F_+\right)\mu_-\left(\|v\|^2 \leq 1\right) =: \varepsilon > 0.$$

Note that ε is positive since $\mathcal{H}$ is separable, and thus all balls have positive measure (Hairer, 2009). It follows that

$$\begin{aligned}\frac{\mathrm{d}\mu(\cdot\,; u)}{\mathrm{d}\mu_+}(v) &= \frac{\mathrm{d}\mu(\cdot\,; u)}{\mathrm{d}\mu_-}(v) \times \left(\frac{\mathrm{d}\mu_+}{\mathrm{d}\mu_-}(v)\right)^{-1} \\ &= \frac{1}{Z(u)}\exp\left(-\frac{1}{2}\langle v, F(u)v\rangle\right) \times Z_+ \exp\left(\frac{1}{2}\langle v, F_+ v\rangle\right) \\ &\geq \varepsilon \exp\left(\frac{1}{2}\langle v, (F_+ - F(u))v\rangle\right) \\ &\geq \varepsilon\end{aligned}$$

since F_+ bounds F above uniformly. ■

Proof of Theorem 14 We first establish existence of at least one invariant distribution by showing that chain $\{u_n\}$ is (strong) Feller, and that for each $u_0 \in \mathcal{H}$ the family $\{\mathsf{P}^n(u_0, \cdot)\}$ of transition kernels is tight. To see the former, let $f : \mathcal{H} \to \mathbb{R}$ be any bounded measurable function. We have that, for any $v \in \mathcal{H}$,

$$\begin{aligned}(\mathsf{P}f)(u) &:= \int_{\mathcal{H}} f(v)\mathsf{P}(u, \mathrm{d}v) \\ &= \int_{\mathcal{H}} f(v)\frac{1}{Z(u)}\exp\left(-\frac{1}{2}\langle v, F(u)v\rangle\right)\mu_-(\mathrm{d}v).\end{aligned}$$

Since $F(u) \leq F_+$ it follows that $Z(u)$ is bounded below by a positive constant, uniformly with respect to u. Additionally F is continuous and non-negative, and so the integrand is bounded and continuous with respect to u. Hence given any sequence $u^{(k)} \to u$ in $\mathcal{H}$, we may apply the dominated convergence theorem to see that $(\mathsf{P}f)(u^{(k)}) \to (\mathsf{P}f)(u)$. The function $\mathsf{P}f$ is therefore continuous, and so the chain $\{u_n\}$ is strong Feller.

We now show tightness. The assumptions on the operator $C_- : \mathcal{H} \to \mathcal{H}$ imply that it is trace-class, and so in particular compact. It is also positive and symmetric, and so by the spectral theorem, admits a complete orthonormal system of eigenvectors $\{\varphi_j\}$ with corresponding positive eigenvalues $\{\lambda_j^2\}$ such that $\lambda_j^2 \to 0$. Given $s > 0$, define the subspace $\mathcal{H}^s \subset \mathcal{H}$ by

$$\mathcal{H}^s = \left\{ v \in \mathcal{H} \;\middle|\; \|v\|_{\mathcal{H}^s}^2 := \sum_{j=1}^{\infty} j^{2s}|\langle \varphi_j, v\rangle|^2 < \infty \right\}.$$

It is standard to show that $\mathcal{H}^s$ is compactly embedded in $\mathcal{H}$ for any $s > 0$ (see for example Robinson, 2001, Appendix A.2). By the Karhunen-Loéve theorem, any $v \sim \mu_-$ may be represented as

$$v = \sum_{j=1}^{\infty} \lambda_j \xi_j \varphi_j, \quad \xi_j \sim N(0,1) \text{ i.i.d.}$$

Hence, by the orthonormality of the $\{\varphi_j\}$ and the assumed decay of the eigenvalues, we have that

$$\mathbb{E}^{\mu_-}\big(\|v\|_{\mathcal{H}^s}^2\big) = \sum_{j=1}^{\infty} j^{2s}\lambda_j^2 \asymp \sum_{j=1}^{\infty} j^{2s-r}$$

and so

$$\mathbb{E}^{\mu_-}\big(\|v\|_{\mathcal{H}^s}^2\big) < \infty \quad \text{if and only if} \quad s < \frac{r}{2} - \frac{1}{2}.$$

Since $r > 1$ by assumption, we can always choose $s > 0$ such that this holds; fix such an s in what follows. Observe that, for any $n \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\big(\|u_n\|_{\mathcal{H}^s}^2\big) &= \mathbb{E}\big(\mathbb{E}\big(\|u_n\|_{\mathcal{H}^s}^2 | u_{n-1}\big)\big) \\
&= \mathbb{E}\big(\mathbb{E}^{\mu(\cdot\,;u_{n-1})}\big(\|v\|_{\mathcal{H}^s}^2\big)\big) \\
&= \mathbb{E}\bigg(\int_{\mathcal{H}} \|v\|_{\mathcal{H}^s}^2 \frac{1}{Z(u_{n-1})} \exp\bigg(-\frac{1}{2}\big\langle v, F(u_{n-1})v\big\rangle\bigg)\,\mu_-(\mathrm{d}v)\bigg) \\
&\leq \frac{1}{Z_+}\mathbb{E}^{\mu_-}\big(\|v\|_{\mathcal{H}^s}^2\big) \\
&=: M < \infty.
\end{aligned}$$

We have bounded $Z(u_{n-1}) \geq Z_+$ using that $F(u_{n-1}) \leq F_+$. Applying the Chebychev inequality, we have for each $n \in N$ and $R > 0$

$$\mathbb{P}\big(\|u_n\|_{\mathcal{H}^s} > R\big) \leq \frac{\mathbb{E}\big(\|u_n\|_{\mathcal{H}^s}^2\big)}{R^2} \leq \frac{M}{R^2},$$

and so given any $\kappa > 0$,

$$\mathbb{P}\bigg(\|u_n\|_{\mathcal{H}^s} \leq \sqrt{\frac{M}{\kappa}}\bigg) \geq 1 - \kappa.$$

This can be rewritten as

$$\mathsf{P}^n(u_0, K_\kappa) \geq 1 - \kappa$$

where $K_\kappa = \big\{u \in \mathcal{H} \,\big|\, \|u\|_{\mathcal{H}^s} \leq \sqrt{M/\kappa}\big\}$ is compact in $\mathcal{H}$, since $\mathcal{H}^s$ is compactly embedded in $\mathcal{H}$; this shows tightness of the sequence of probability measures $\mathsf{P}^n(u_0, \cdot)$. Since tightness implies boundedness in probability on average, an application of (Meyn and Tweedie, 2012, Theorem 12.0.1) gives existence of an invariant distribution.

Lemma 15 shows that $\{u_n\}$ satisfies a global minorization condition for the one-step transition probabilities: for any $u_0 \in \mathcal{H}$ and any measurable $A \subseteq \mathcal{H}$,

$$\mathsf{P}\big(u_0, A\big) = \mathbb{E}^{\mu(\cdot\,;u_0)}\big(\mathbb{1}_A(v)\big) = \mathbb{E}^{\mu_+}\bigg(\frac{\mathrm{d}\mu(\cdot\,;u_0)}{\mathrm{d}\mu_+}(v)\mathbb{1}_A(v)\bigg) \geq \varepsilon\mu_+(A).$$

Combined with the existence of an invariant distribution above, a short coupling argument (Meyn and Tweedie, 2012, Theorem 16.2.4) gives the result with the same ε as above. ■

3.4 Convolution

The convolution iteration has the advantage that, through use of Fourier series and the law of large numbers, its long time behaviour can be completely characterized analytically. We consider the convolution as a random map on $\mathcal{H} = L^2(D;\mathbb{C})$, $D = (0,1)^d$. The iteration is given by

$$u_{n+1}(x) = (u_n * \xi_{n+1})(x) := \int_D u_n(x-y)\xi_{n+1}(y)\,dy, \quad \xi_{n+1} \sim N(0,C) \text{ i.i.d.} \tag{10}$$

where we implicitly work with periodic extensions to define the convolution. We assume that C is a negative fractional power of a differential operator so that it diagonalizes in Fourier space; such a form of covariance operator is common in applications, as it includes, for example, Whittle-Matérn distributions (Lindgren et al., 2011). For example, we may take

$$C = (I - \triangle)^{-\alpha}, \quad D(-\triangle) = H^2_{\text{per}}([0,1]^d) \subset \mathcal{H},$$

in which case the samples $\xi_{n+1} \sim N(0,C)$ will (almost surely) possess s fractional Sobolev and Hölder derivatives for any $s < \alpha - d/2$ (see Dashti and Stuart, 2017, for details).

We choose the orthonormal Fourier basis

$$\varphi_k(x) = e^{2\pi i k\cdot x}, \quad k \in \mathbb{Z}^d,$$

which are the eigenvectors of C; we denote the corresponding eigenvalues $\{\lambda_k^2\}$. Given $u \in \mathcal{H}$ and $k \in \mathbb{Z}^d$, define the Fourier coefficient $\hat{u}(k) \in \mathbb{C}$ by

$$\hat{u}(k) := \langle \varphi_k, u\rangle_{L^2} = \int_D \overline{\varphi_k(x)}u(x)\,\mathrm{d}x.$$

Then it can be readily checked that for any $u, v \in \mathcal{H}$ and $k \in \mathbb{Z}^d$,

$$\widehat{(u*v)}(k) = \hat{u}(k)\hat{v}(k). \tag{11}$$

We use this property to establish the following theorem.

Theorem 16 *Let $C : \mathcal{H} \to \mathcal{H}$ be a negative fractional power of a differential operator such that C is positive, symmetric and trace-class, with eigenvectors $\{\psi_k\}$ and eigenvalues $\{\lambda_k^2\}$. Define the Markov chain $\{u_n\}$ by (10). Then for any $u_0 \in \mathcal{H}$,*

$$\lim_{n\to\infty} |\hat{u}_n(k)|^2 = \begin{cases} 0 & |\lambda_k|^2 < 2e^\gamma \\ \infty & |\lambda_k|^2 > 2e^\gamma \end{cases} \quad \textit{almost surely}$$

where $\gamma \approx 0.577$ is the Euler-Mascheroni constant. In particular, if $|\lambda_k|^2 < 2e^\gamma$ for all $k \in \mathbb{Z}^d$, then every Fourier coefficient of u_n tends to zero almost surely and hence $u_n \rightharpoonup 0$ in $\mathcal{H}$ almost surely.

Proof First observe that by the Karhunen-Loéve theorem, we may express $\xi_{n+1} \sim N(0,C)$ as

$$\xi_{n+1} = \sum_{k\in\mathbb{Z}^d} \lambda_k \eta_{n,k}\varphi_k, \quad \eta_{n,k} \sim N(0,1) \text{ i.i.d.}$$

and so, since $\{\varphi_k\}$ is orthonormal,

$$\hat{\xi}_{n+1}(k) = \lambda_k \eta_{n,k}.$$

Then by the property (11), we see that for each $k \in \mathbb{Z}^d$ and $n \in \mathbb{N}$,

$$\hat{u}_{n+1}(k) = \hat{u}_n(k)\hat{\xi}_{n+1}(k) = \hat{u}_n(k)\lambda_k \eta_{n,k} \tag{12}$$

where the second equality is in distribution. The problem has now been reduced to an independent family of scalar problems. We can write $\hat{u}_n(k)$ explicitly as

$$\hat{u}_n(k) = \hat{u}_0(k) \prod_{j=1}^{n} \lambda_k \eta_{j,k}. \tag{13}$$

Now observe that

$$\begin{aligned} |\hat{u}_n(k)|^2 &= |\hat{u}_0(k)|^2 \prod_{j=1}^{n} |\lambda_k|^2 |\eta_{j,k}|^2 \\ &= |\hat{u}_0(k)|^2 \exp\left(n \cdot \frac{1}{n} \sum_{j=1}^{n} \log\left(|\lambda_k|^2 |\eta_{j,k}|^2 \right) \right) \\ &= |\hat{u}_0(k)|^2 \exp\left(n \cdot \left(\frac{1}{n} \sum_{j=1}^{n} \log |\eta_{j,k}|^2 + \log |\lambda_k|^2 \right) \right). \end{aligned} \tag{14}$$

By the strong law of large numbers, the scaled sum inside the exponential converges almost surely to $\mathbb{E}(\log |\eta_{1,k}|^2)$. This can be calculated as

$$\mathbb{E}(\log |\eta_{1,k}|^2) = -\gamma - \log 2.$$

If the bracketed term inside the exponential in (14) is eventually negative almost surely, then the limit of $|u_n(k)|^2$ will be zero almost surely. This is guaranteed when $-\gamma - \log 2 + \log |\lambda_k|^2 < 0$, i.e. $|\lambda_k|^2 < 2e^{\gamma}$. Similarly we get divergence if the bracketed term is eventually positive, which happens when $|\lambda_k|^2 > 2e^{\gamma}$. ■

Remark 17 *It is interesting to note that we may take expectations in (12) to establish that*

$$\mathbb{E}|\hat{u}_n(k)|^2 = |\hat{u}_0(k)|^2 |\lambda_k|^{2n}$$

and so

$$\lim_{n\to\infty} \mathbb{E}|\hat{u}_n(k)|^2 = \begin{cases} 0 & |\lambda_k|^2 < 1 \\ \infty & |\lambda_k|^2 > 1 \end{cases}.$$

In particular, if $|\lambda_k|^2 \in (1, 2e^{\gamma})$, then $|\hat{u}_n(k)|^2$ converges to zero almost surely, but diverges in mean square.

Via a slight modification of the above proof to account for different boundary conditions, we have the following result.

Corollary 18 *Let $D = (0,1)$ and let $\{u_n\}$ be defined by the iteration (10), where each ξ_{n+1} is a Brownian bridge. Then $u_n \rightharpoonup 0$ almost surely.*

Proof The Brownian bridge on $[0,1]$ has covariance operator $(-\triangle)^{-1}$, where

$$D(-\triangle) = \{u \in H^2_{\text{per}}([0,1]) \,|\, u(0) = u(1) = 0\}.$$

The result of Theorem 16 cannot be applied directly, since the basis functions $\{\varphi_k\}$ do not satisfy the boundary conditions. The eigenfunctions with the correct boundary conditions are given by

$$\psi_j(x) = \sqrt{2}\sin(j\pi x) = \frac{1}{\sqrt{2}i}\big(\varphi_j(x) - \varphi_{-j}(x)\big), \quad j \geq 1$$

with corresponding eigenvalues $\alpha_j^2 = (\pi^2 j^2)^{-1}$. A Brownian bridge $\xi_{n+1} \sim N(0, (-\triangle)^{-1})$ can then be expressed as

$$\xi_{n+1} = \sum_{j=1}^{\infty} \alpha_j \zeta_{n,j} \psi_j, \quad \zeta_{n,j} \sim N(0,1) \text{ i.i.d.}$$

by the Karhunen-Loéve theorem. We calculate

$$\begin{aligned}
\hat{u}_{n+1}(k) &= \hat{u}_n(k)\hat{\xi}_n(k) \\
&= \hat{u}_n(k) \sum_{j=1}^{\infty} \alpha_j \zeta_{n,j} \langle \varphi_k, \psi_j \rangle \\
&= \hat{u}_n(k) \sum_{j=1}^{\infty} \alpha_j \zeta_{n,j} \frac{1}{\sqrt{2}i} \big(\langle \varphi_k, \varphi_j \rangle - \langle \varphi_k, \varphi_{-j} \rangle \big) \\
&= \hat{u}_n(k) \frac{\operatorname{sgn}(k)\alpha_{|k|}}{\sqrt{2}i} \zeta_{n,|k|} \\
&= \hat{u}_n(k) \lambda_k \eta_{n,k}, \quad \eta_{n,k} \sim N(0,1).
\end{aligned}$$

We can now proceed as in Theorem 16 to deduce that $|u_n(k)|^2 \to 0$ whenever $|\lambda_k|^2 < 2e^{\gamma}$; note that the correlations between $\hat{u}_n(k)$ and $\hat{u}_n(-k)$ do not affect the argument. Now observe that $|\lambda_k|^2 = (2\pi^2 k^2)^{-1} < 1 < 2e^{\gamma}$ for all k, and the result follows. ■

Remark 19 *The preceding results also holds if we replace the Brownian bridge by a Gaussian process with precision operator the negative Laplacian subject to Neumann boundary conditions and spatial mean zero; the eigenfunctions are then*

$$\psi_j(x) = \sqrt{2}\cos(j\pi x) = \frac{1}{\sqrt{2}}\big(\varphi_j(x) + \varphi_{-j}(x)\big), \quad j \geq 1.$$

The argument is identical, except no $\operatorname{sgn}(k)$ *term appears in* λ_k.

4. Numerical Illustrations

We now study two of the constructions of deep Gaussian processes numerically. In subsection 4.1 we look at realizations of the deep Gaussian process constructed using the covariance function formulation, and in subsection 4.2 we perform similar experiments for the covariance operator formulations. Finally we consider Bayesian inverse problems, in which we choose deep Gaussian processes as our prior distributions; we introduce a function space MCMC algorithm, which scales well under mesh refinement of the functions to be inferred, for sampling.

For the composition construction, numerical experiments are given by, for example, Damianou and Lawrence (2013); Duvenaud et al. (2014). We do not provide numerical experiments for the convolution construction; Theorem 16 tells us that interesting behaviour cannot be expected in this case.

4.1 Covariance Function

We start by investigating typical realizations of a deep Gaussian process, constructed through anisotropic covariance kernels as in section 2.2. As the basis of our construction, we choose a stationary Gaussian correlation kernel, given by

$$\rho_S(r) = \exp(-r^2), \qquad r > 0.$$

The function F determining the length scale of the kernel $\rho(\cdot,\cdot;u_n)$ is chosen as $F(x) = x^2$, such that $\Sigma(z) = (u_n(z))^2\ \mathrm{I}_d$. Similar results are obtained with other choices of F in terms of the distribution of samples u_n. The choice of F does, however, influence the conditioning of the correlation matrix $\mathbf{R}(\mathbf{u}_n)$, and the choice $F(x) = \exp(x)$, for example, can lead to numerical instabilities. As described in section 2.2, we will sample from the finite dimensional distributions obtained by sampling from the Gaussian process at a finite number of points in the domain D. To generate the samples, we use the command `mvnrnd` in MATLAB, and when plotting the samples, we use linear interpolation.

In Figure 1, we show four independent realizations of the first seven layers $u_0, \dots, u_6$, where u_0 is taken as a sample of the stationary Gaussian process with correlation kernel ρ_S. The domain D is here chosen as the interval $(0,1)$, and the sampling points are given by the uniform grid $x_i = \frac{i-1}{256}$, for $i = 1, \dots, 257$. Each column in Figure 1 corresponds to one realization, and each row corresponds to a given layer u_n, the first row showing u_0. We can clearly see the non-stationary behaviour in the samples when progressing through the levels. We note that the ergodicity of the chain is also reflected in the samples, with the distribution of the samples u_n looking similar for larger values of n.

Figure 2 shows the same information as Figure 1, in the case where the domain D is $(0,1)^2$ and the sampling points are the tensor product of the one-dimensional points $x_i^1 = \frac{i-1}{64}$, for $i = 1, \dots, 65$.

4.2 Covariance Operator

We now consider the covariance operator construction of the deep Gaussian process. In order to produce more interesting behaviour in the samples, we move away from the absolutely continuous setting considered in section 3.3 by introducing a rescaling of $C(u)$ that depends

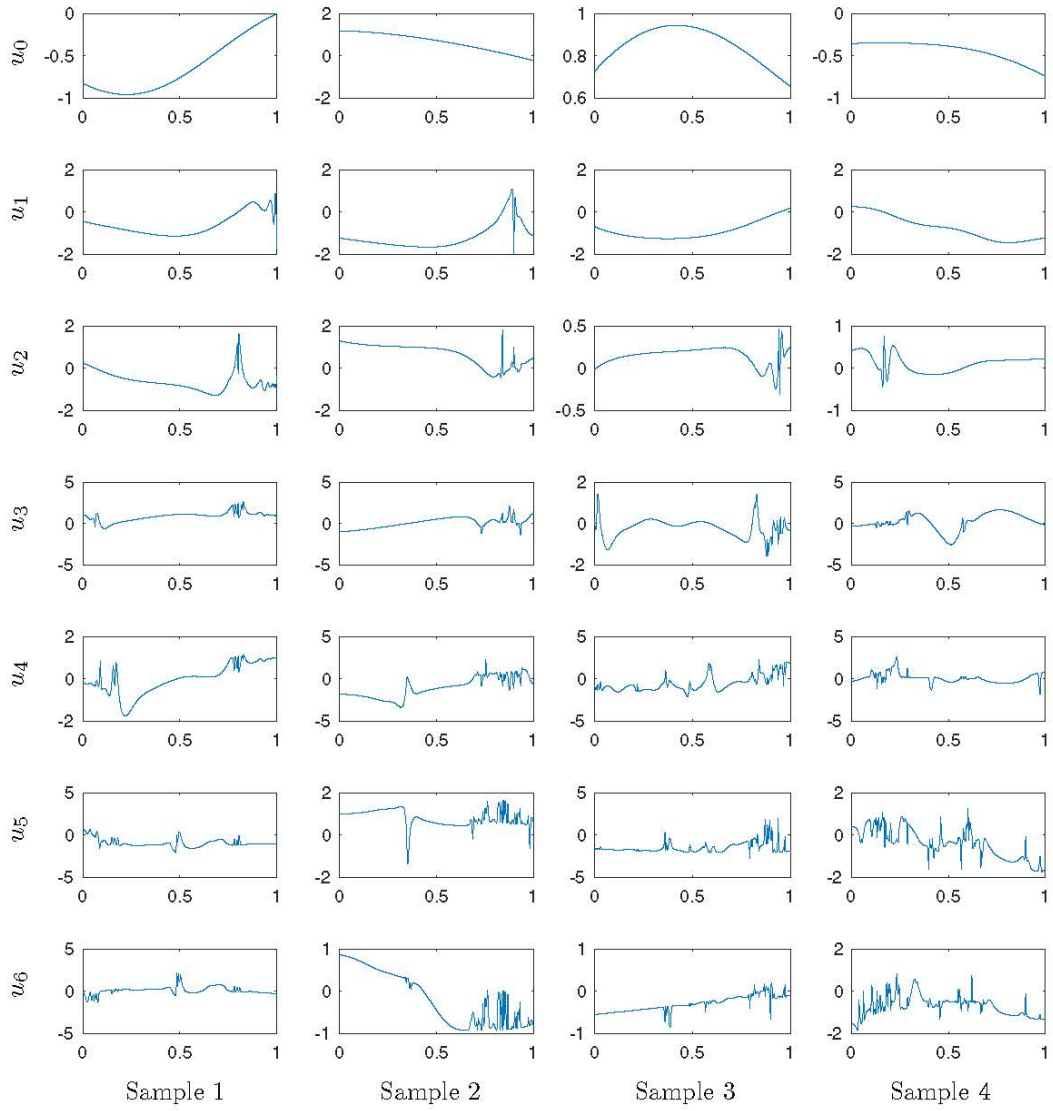

Figure 1: Four independent realizations of the first seven layers of a deep Gaussian process, in one spatial dimension, using the covariance kernel construction described in subsection 2.2. Each column corresponds to an independent chain, and layers $u_0, u_1, \ldots, u_6$ are shown from top-to-bottom.

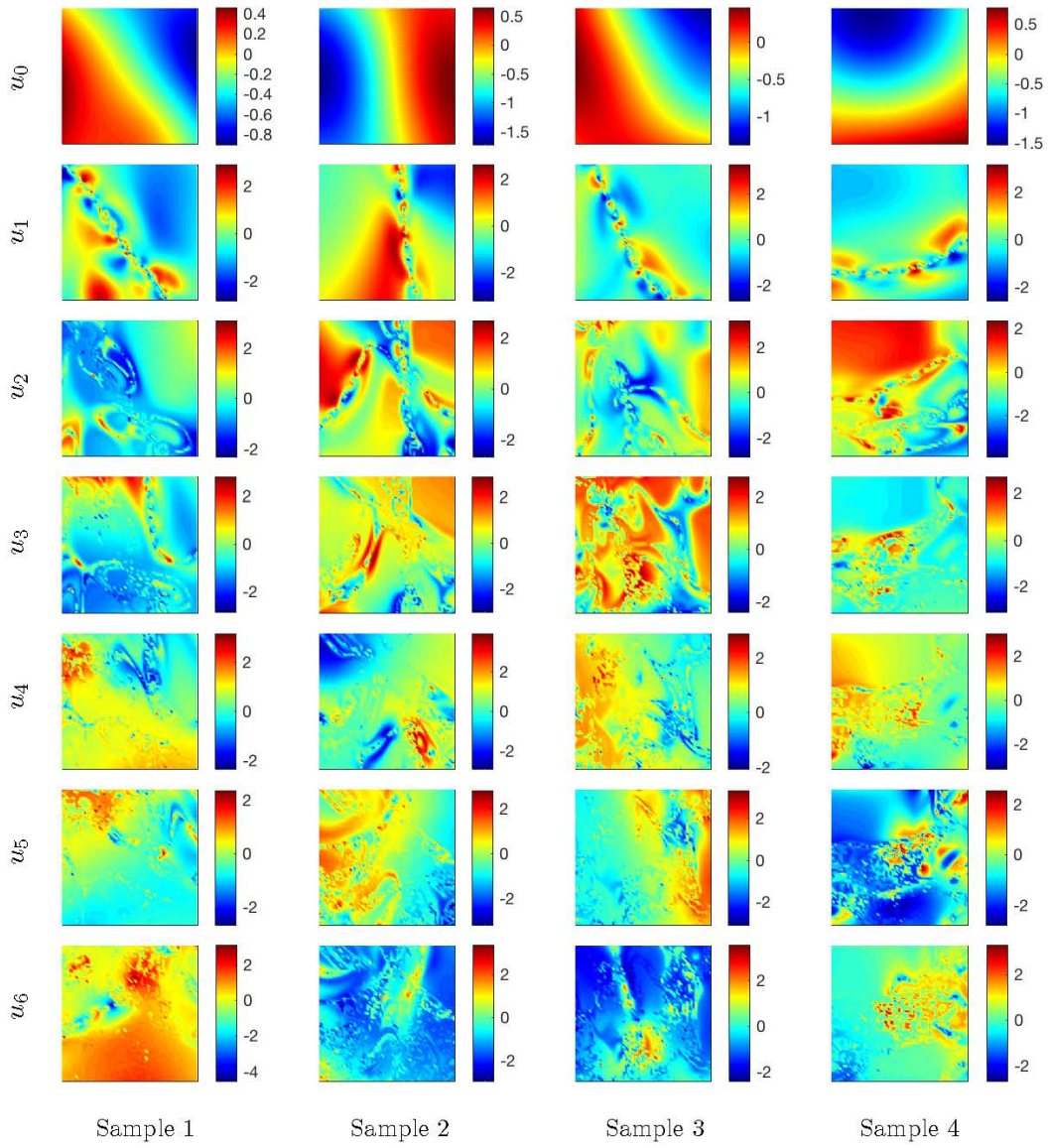

Figure 2: Four independent realizations of the first seven layers of a deep Gaussian process, in two spatial dimensions, using the covariance kernel construction described in subsection 2.2. Each column corresponds to an independent chain, and layers $u_0, u_1, \ldots, u_6$ are shown from top-to-bottom.

on u. This scaling is chosen so that the amplitude of samples is $\mathcal{O}(1)$ with respect to u. The rescaled family can be shown to satisfy Assumptions 3, and a minorization condition as in Lemma 15 can also be shown to hold when the state space is finite-dimensional. From this we can deduce that the resulting discretized process will still be ergodic.

Assume $D \subseteq \mathbb{R}^d$ and define the negative Laplacian $-\triangle$ on $D(-\triangle)$,

$$D(-\triangle) = \left\{ u \in H^2(D;\mathbb{R}) \;\middle|\; \frac{\mathrm{d}u}{\mathrm{d}\nu}(x) = 0 \text{ for } x \in \partial D \right\},$$

where ν is the outward normal to ∂D. Given $\alpha > d/2$, $\sigma > 0$, we define $P = -\triangle$ and

$$C(u)^{-1} = \sigma^{-2}(P + \Gamma(u))^{\alpha/2}\Gamma(u)^{d/2-\alpha}(P + \Gamma(u))^{\alpha/2} \tag{15}$$

where $\big(\Gamma(u)v\big)(x) = F\big(u(x)\big)v(x)$. The scaling introduced is inspired by the SPDE representation of Whittle-Matérn distributions (Lindgren et al., 2011); if $F(u) = \tau^2$ is chosen to be constant, then modulo boundary conditions, samples from a centred Gaussian distribution with covariance $C(u)$ are samples from a Whittle-Matérn distribution. In particular, τ corresponds to the inverse length-scale of samples, and samples almost-surely have s Sobolev and Hölder and derivatives for any $s < \alpha - d/2$.

For numerical experiments, we take

$$F(u) = \min\{F_- + ae^{bu^2}, F_+\}$$

for some $F_+, F_-, a, b > 0$. In particular, in one spatial dimension we take $F_+ = 150^2$, $F_- = 200$, $a = 100$ and $b = 2$. In two dimensions, we take $F_+ = 150^2$, $F_- = 50$, $a = 25$ and $b = 0.3$. We take $\alpha = 4$ in both cases, and choose σ such that $\mathbb{E}\big(u(x)^2\big) \approx 1$. These parameter choices were made empirically to ensure interesting structure of the samples. In order to generate samples at a given level, the negative Laplacian P is constructed using a finite-difference method. Given u, the operator $A(u)$ is then computed,

$$A(u) := \sigma^{-1}\Gamma(u)^{d/4-\alpha/2}(P + \Gamma(u))^{\alpha/2},$$

so that $v \sim N(0, C(u))$ solves the SPDE $A(u)v = \xi$, where ξ is white noise.

In Figure 3 we show samples of the deep Gaussian process on domain $D = (0, 1)$, sampled on the uniform grid $x_i = \frac{i-1}{1000}$, for $i = 1, \ldots, 1001$. We show 4 independent realizations of the first seven layers of the process—each row corresponds to a given layer u_n. The anisotropy of the length-scale is evident in levels beyond u_0, and the effect of ergodicity is evident, with deeper levels having similar properties. Compared to the covariance function construction, local effects are less prominent, though a greater level of anisotropy could potentially be obtained by making an alternative choice of $F(\cdot)$. Figure 4 shows the same experiments on domain $D = (0, 1)^2$, sampled on the tensor product of the one-dimensional points $x_i^1 = \frac{i-1}{150}$, for $i = 1, \ldots, 151$, and the same effects are observed. Figure 5 shows the trace of the norm of a DGP $\{u_n\}$ with $d = 1$, along with the running mean of these norms; the rapid convergence of the mean reflects the ergodicity of the chain.

We emphasize that our perspective on inference includes quite general inverse problems, and is not limited to the problems of regression and classification which dominate much of

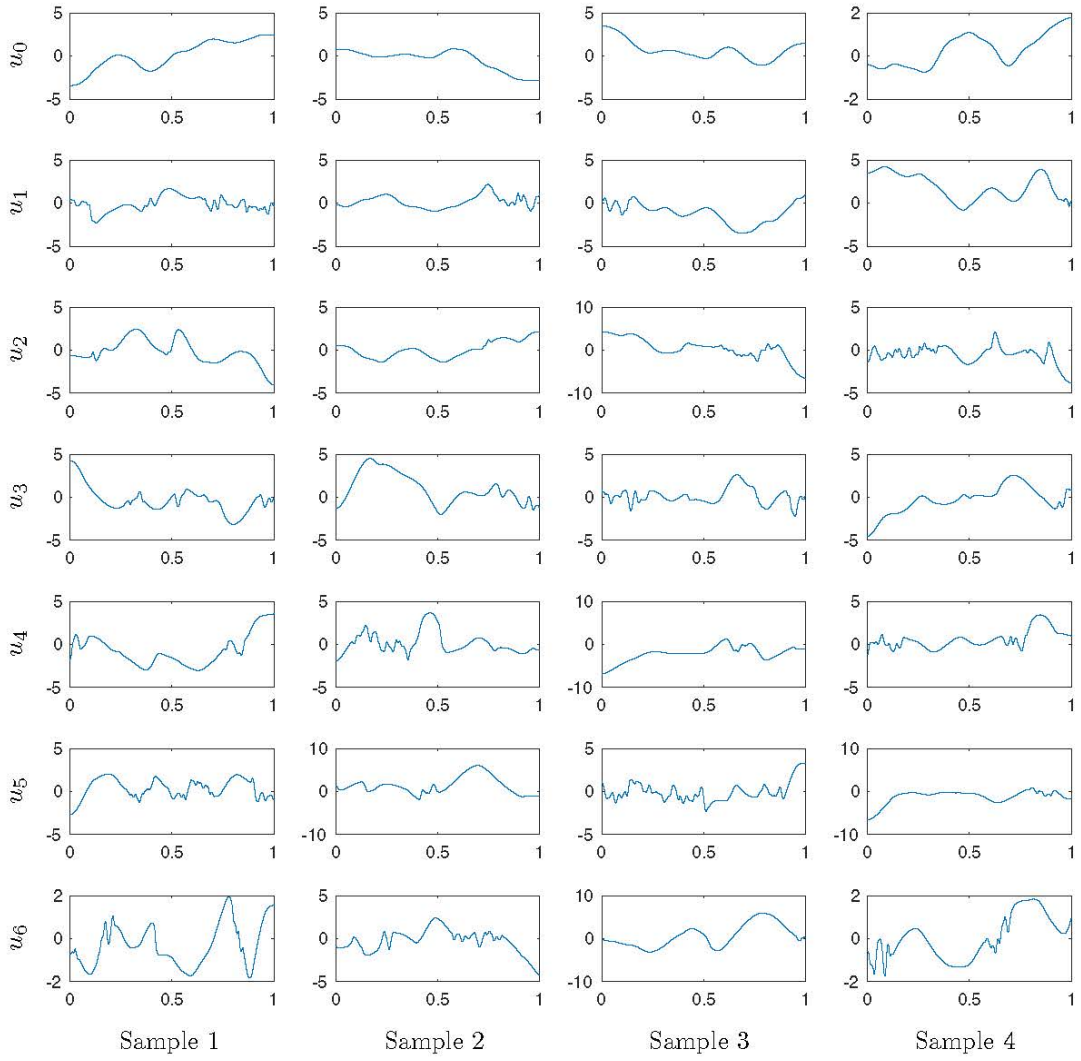

Figure 3: Four independent realizations of the first seven layers of a deep Gaussian process, in one spatial dimension, using the covariance operator construction described in subsection 4.2. Each column corresponds to an independent chain, and layers $u_0, u_1, \ldots, u_6$ are shown from top-to-bottom.

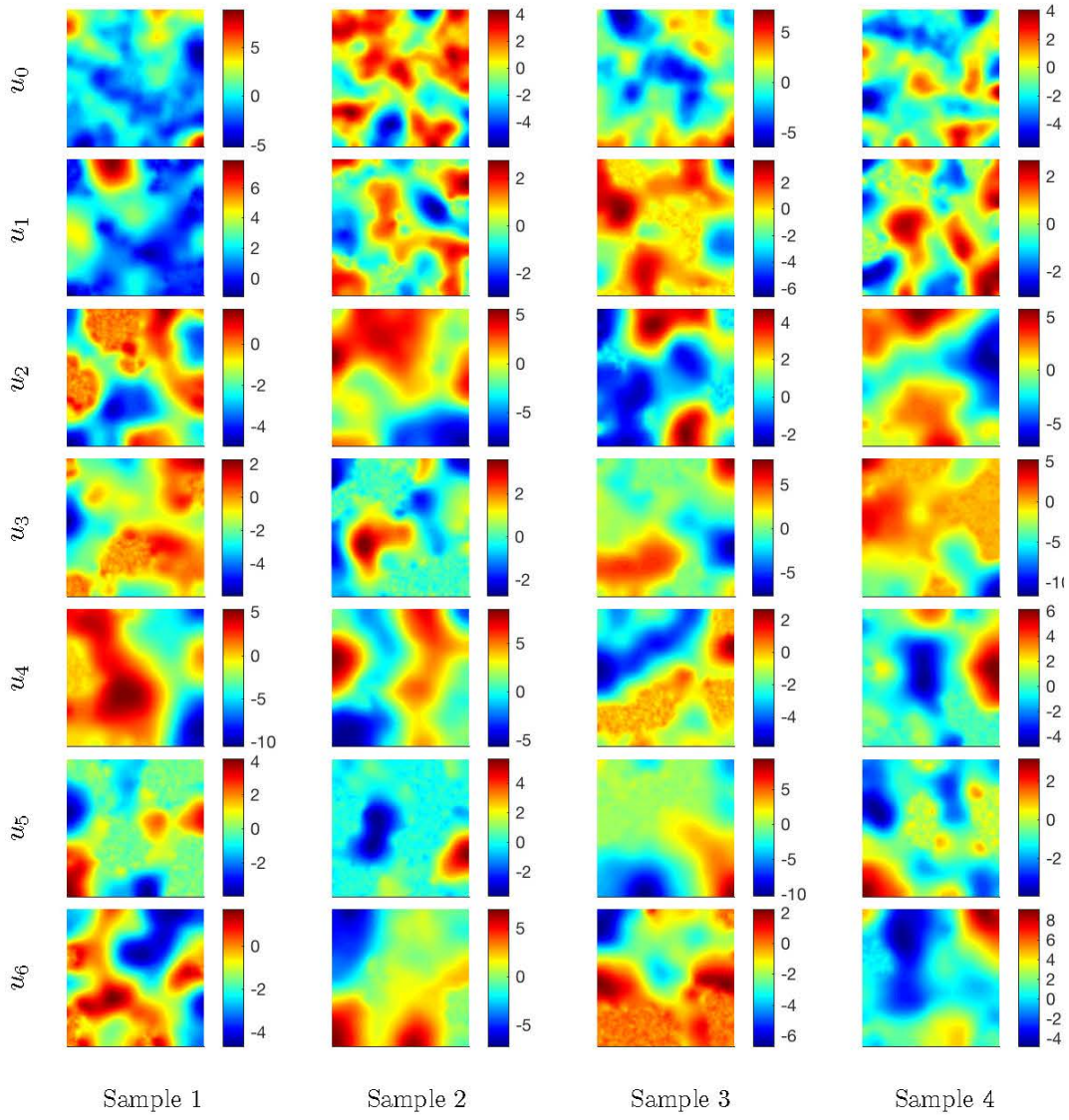

Figure 4: Four independent realizations of the first seven layers of a deep Gaussian process, in two spatial dimensions, using the covariance operator construction described in subsection 4.2. Each column corresponds to an independent chain, and layers $u_0, u_1, \ldots, u_6$ are shown from top-to-bottom.

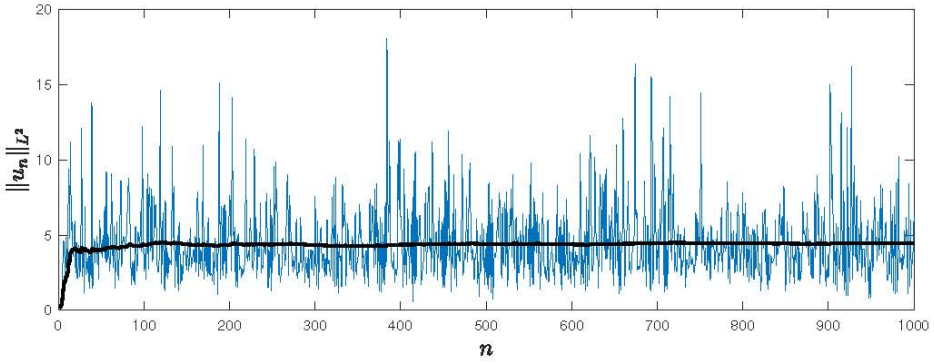

Figure 5: The trace of the norm of u_n versus n for a 1000 layer DGP $\{u_n\}$ as in Figure 3. The thick black curve shows the running mean of the norms.

classical machine learning; this broad perspective on the potential for the methodology affects the choice of algorithms that we study as we do not exploit any of the special structures that arise in regression and classification.

The deep Gaussian processes discussed in the previous sections were introduced with the idea of providing flexible prior distributions for inference, for example in inverse problems. The structure of such problems is as follows. We have data $y \in \mathbb{R}^J$ arising via the model

$$y = \mathcal{G}(u) + \eta \tag{16}$$

where η is a realization of some additive noise, and $\mathcal{G} : X \to \mathbb{R}^J$ is a (typically non-linear) forward map. The map $\mathcal{G}$ may involve, for example, solution of a partial differential equation which takes function u as input, or point evaluations of a function u, regression. In this paper we will fix $X = \mathcal{H}^N$, writing $u = (u_0, \ldots, u_{N-1}) \in X$; our prior beliefs on u will then be characterized by the first N states of a Markov chain of a form considered in the previous sections. Note that the map $\mathcal{G}$ could incorporate a projection map if the dependence is only upon a single state u_{N-1}; indeed this is the canonical example—the variables $(u_0, \ldots, u_{N-2})$ are viewed as hyperparameters in a prior on the parameter u_{N-1}.

4.2.1 ALGORITHMS

We now turn to the design of algorithms for the Bayesian inference problems of sampling $u|y$. As already mentioned above, we are typically only interested in sampling the deepest layer $u_{N-1}|y$. However, due to the hierarchical definition of u_{N-1} given all the components of u, our algorithms work with the full set of layers u. Since the components of u are functions, and hence infinite dimensional objects in general, a guiding principle is to design algorithms which are well-defined on function space, an approach to MCMC inference reviewed by Cotter et al. (2013); the value of this approach is that it leads to algorithms whose mixing time is not dependent on the number of mesh points used to represent the function to be inferred. For simplicity of exposition we assume that the observational noise η is distributed

as $N(0,\Gamma)$; this is not central to our developments but makes the exposition concrete. Recalling that the Markov chain defining the prior beliefs is given by (ZeroMean), we can consider the unknowns in the problem to be the variables $u = (u_0, \ldots, u_{N-1})$, which are correlated under the prior, or the variables $\xi = (\xi_0, \ldots, \xi_{N-1})$, where we define $\xi_0 = u_0$, which are independent under the prior. These variables are related via $u = T(\xi)$, where the components of the deterministic map $T : X \to X$ are defined iteratively by

$$\begin{aligned} T_1(\xi_0, \ldots, \xi_{N-1}) &= \xi_0, \\ T_{n+1}(\xi_0, \ldots, \xi_{N-1}) &= L(T_n(\xi_0, \ldots, \xi_{N-1}))\xi_n, \quad n = 1, \ldots, N-1. \end{aligned}$$

The data may then be expressed in terms of ξ rather than u:

$$y = \tilde{\mathcal{G}}(\xi) + \eta = \mathcal{G}(T(\xi)) + \eta \tag{17}$$

where our prior belief on ξ is that its components are i.i.d. Gaussians. To be consistent with the notation introduced by Papaspiliopoulos et al. (2007); Yu and Meng (2011), (16) will be referred to as the *centred* model and (17) will be referred to as the *non-centred* model. The space $\mathcal{H}$ may be chosen differently in the centred and non-centred cases.

Associated with the two data models are two likelihoods: $\mathbb{P}(y|u)$ and $\mathbb{P}(y|\xi)$. Assuming that the observational noise $\eta \sim N(0,\Gamma)$ is Gaussian, where $\Gamma \in \mathbb{R}^{J\times J}$ is a positive definite covariance matrix, the likelihoods are given by

$$\begin{aligned} \mathbb{P}(y|u) &= \frac{1}{Z(y)} \exp\big(-\Phi(u;y)\big), \quad \Phi(u;y) := \frac{1}{2}\big|\Gamma^{-\frac{1}{2}}(y - \mathcal{G}(u))\big|^2, \\ \mathbb{P}(y|\xi) &= \frac{1}{\tilde{Z}(y)} \exp\big(-\tilde{\Phi}(\xi;y)\big), \quad \tilde{\Phi}(\xi;y) := \frac{1}{2}\big|\Gamma^{-\frac{1}{2}}(y - \tilde{\mathcal{G}}(\xi))\big|^2. \end{aligned}$$

We may then apply Bayes' theorem to write down the posterior distributions $\mathbb{P}(u|y)$ and $\mathbb{P}(\xi|y)$:

$$\begin{aligned} \mathbb{P}(u|y) &\propto \mathbb{P}(y|u)\mathbb{P}(u) \propto \exp\big(-\Phi(u;y)\big)\mathbb{P}(u), \\ \mathbb{P}(\xi|y) &\propto \mathbb{P}(y|\xi)\mathbb{P}(\xi) \propto \exp\big(-\tilde{\Phi}(\xi;y)\big)\mathbb{P}(\xi). \end{aligned}$$

We know (Cotter et al., 2013) that it is straightforward to design algorithms to sample $\mathbb{P}(\xi|y)$ which are well-defined in infinite dimensions, exploiting the fact that $\mathbb{P}(\xi)$ is Gaussian. An example of such an algorithm given by Algorithm 1.

This algorithm produces a chain $\{\xi^{(k)}\}_{k\in\mathbb{N}}$ that samples $\mathbb{P}(\xi|y)$ in stationarity; and $\{T(\xi^{(k)})\}_{k\in\mathbb{N}}$ will be samples of $\mathbb{P}(u|y)$. By working in non-centred coordinates we have been able to design this algorithm which is well-defined on function space. If we were to work with the centred coordinates u directly, the algorithm would not be well-defined on function space: in infinite dimensions, each family of measures $\{\mathbb{P}(u_n|u_{n-1})\}_{u_{n-1}\in\mathcal{H}}$ will typically be mutually singular, and so a proposed update $u \mapsto \hat{u}$ will almost surely be rejected. To see why this rejection occurs in practice, in high finite dimensions K, notice that the acceptance probability for an update $u \mapsto \hat{u}$ will involve the ratios of the Gaussian densities $N(u_n; 0, C(u_{n-1}))$ and $N(u_n; 0, C(\hat{u}_{n-1}))$. These densities will decay to zero as the dimension K is increased, and their ratio will only be well-defined in the limit if the measures are equivalent; consequently, the Markov chain will mix very poorly. Working with the

Algorithm 1 Non-Centred Algorithm

1. Fix $\beta_0, \ldots, \beta_{N-1} \in (0, 1]$ and define $B = \mathrm{diag}(\beta_j)$. Choose initial state $\xi^{(0)} \in X$, and set $u^{(0)} = T(\xi^{(0)}) \in X$. Set $k = 0$.

2. Propose $\hat{\xi}^{(k)} = (I - B^2)^{\frac{1}{2}} \xi^{(k)} + B\zeta_j^{(k)}, \quad \zeta^{(k)} \sim N(0, I)$.

3. Set $\xi^{(k+1)} = \hat{\xi}^{(k)}$ with probability

$$\alpha_k = \min \left\{ 1, \exp \left(\Phi\big(T(\xi^{(k)}); y\big) - \Phi\big(T(\hat{\xi}^{(k)}); y\big) \right) \right\};$$

   otherwise set $\xi^{(k+1)} = \xi^{(k)}$.

4. Set $k \mapsto k + 1$ and go to 1.

non-centred coordinates ξ, the prior does not appear in the acceptance probability and so this issue is circumvented. Another advantage of using the non-centred coordinates is that there is no need to calculate the (divergent) log determinants which appear in the centred acceptance probability, avoiding potential numerical issues. These issues are discussed in greater depth and generality by Chen et al. (2018). For the reasons set-out in that paper, including those above, we have used only the non-centred algorithm in what follows. When the forward model $\mathcal{G}(u) = Au$ is linear, the non-centred algorithm can be combined with standard Gaussian process regression techniques via the identity

$$\mathbb{P}(\mathrm{d}u_N|y) = \int_X \mathbb{P}(\mathrm{d}u_N|u_{N-1}, y)\mathbb{P}(\mathrm{d}u_{N-1}|y).$$

The distribution $\mathbb{P}(\mathrm{d}u_N|u_{N-1}, y) = N(m_y(u_{N-1}), C_y(u_{N-1}))$ is Gaussian, where expressions for m_y, C_y are known, and so direct sampling methods are available. On the other hand, we have that $\mathbb{P}(y|u_{N-1}) = N(0, AC(u_{N-1})A^* + \Gamma)$, and so we may use the non-centred algorithm to robustly sample the measure

$$\begin{aligned}
\mathbb{P}(\mathrm{d}u_{N-1}|y) &= \exp(-\Psi(u_{N-1}; y))\mathbb{P}(\mathrm{d}u_{N-1}), \\
\Psi(u_{N-1}; y) &= \frac{1}{2}\|y\|^2_{AC(u_{N-1})A^*+\Gamma} + \frac{1}{2}\log\det(AC(u_{N-1})A^* + \Gamma),
\end{aligned}$$

after reparametrizing in terms of ξ. This approach can be viable even when the data is particularly informative so that Φ is very singular—this singularity does not in general pass to Ψ. It is this approach that we use for the simulations in the following subsections. An alternative approach not based on MCMC would be to use the non-centred parameterization of the Ensemble Kalman Filter (Chada et al., 2018) which we have successfully implemented in the context of the deep Gaussian processes of this paper, but do not show here for reasons of brevity.

4.3 Application to Regression

We consider the application of the non-centered algorithm described above to simple regression problems in one and two spatial dimensions.

4.3.1 One-Dimensional Simulations

We consider first the case $D = (0, 1)$, where the forward map is given by a number of point evaluations: $\mathcal{G}_j(u) = u(x_j)$ for some sequence $\{x_j\}_{j=1}^J \subseteq D$. We compare the quality of reconstruction versus both the number of point evaluations and the number of levels in the deep Gaussian prior. We use the same parameters for the family of covariance operators as in subsection 4.2. The base layer u_0 is taken to be Gaussian with covariance of the form (15), with $\Gamma(u) \equiv 20^2$.

The true unknown field $u^\dagger$ is given by the indicator function $u^\dagger = \mathbb{1}_{(0.3,0.7)}$, shown in Figure 6. It is generated on a mesh of 400 points, and three data sets are created wherein it is observed on uniform grids of $J = 25$, 50 and 100 points, and corrupted by white noise with standard deviation $\gamma = 0.02$. Sampling is performed on a mesh of 200 points to avoid an inverse crime (Kaipio and Somersalo, 2006). 10^6 samples are generated per chain, with the first 2×10^5 discarded as burn-in when calculating means. The jump parameters β_j are adaptively tuned to keep acceptance rates close to 30%.

In these experiments the deepest field is labelled as u_N, rather than as u_{N-1} as in the statement of the algorithm; this is purely for notational convenience, of course. In Figure 7 the means of the deepest field u_N and of the length-scales associated with each hidden layer are shown, that is, approximations to $\mathbb{E}(u_N)$ and $\mathbb{E}(F(u_j)^{\frac{1}{2}})$ for each $j = 0, \ldots, N-1$. We see that, in all cases, the reconstructions of $u^\dagger$ are visually similar when two or more layers are used, and similar length-scale fields $\mathbb{E}(F(u_{N-1})^{\frac{1}{2}})$ are obtained in these cases. The sharpness of these length-scale fields is related to the amount of data. Additionally, when $N = 4$ and $J = 100$ the location of the discontinuities is visible in the estimate for $\mathbb{E}(F(u_{N-2})^{\frac{1}{2}})$, suggesting the higher quality data can influence the process more deeply. When $J = 50$ or $J = 25$, this layer does not appear to be significantly informed. When a single layer prior is used, the reconstruction fails to accurately capture the discontinuities. Figure 7 also shows bands of quantiles of the values $u(x)$ under the posterior, illustrating their distribution; in particular the lack of symmetry and disagreement of the means and medians show that the posterior is clearly non-Gaussian. Uncertainty increases both as the number of observations J and the layer n in the chain is increased. Note in particular the over-confidence of the shallow Gaussian process posterior: the truth is not contained within 95% credible intervals in all cases.

In Table 1 we show the L^1-errors between the true field and the posterior means arising from the different setups. The errors decrease as the number of observation points is increased, as would be expected. Additionally, when $J = 100$ and $J = 50$, the accuracy of the reconstruction increases with the number of layers, though the most significant increase occurs when increasing from 1 to 2 layers. When $J = 25$, the error increases beyond 2 layers, suggesting that some balance is required between the quality of the data and the flexibility of the prior.

In Figure 8 we replace the uniformly spaced observations with 10^6 randomly placed observations, to illustrate the effect of very high quality data. With 3 or 4 layers, more anisotropic behavior is observed in the length-scale field. Additionally, the layer u_{N-2} is much more strongly informed than the cases with fewer observations, though the layer u_{N-3} in the case $N = 4$ does not appear to be informed at all, indicating a limitation on how deeply the process can be influenced by data. The corresponding errors are shown in Table 1—as

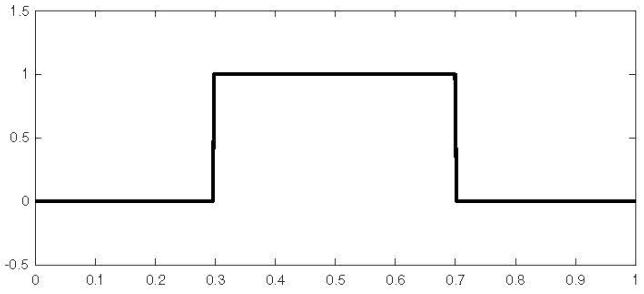

Figure 6: The true field used to generate the data for the one-dimensional inverse problem.

J	1 layer	2 layers	3 layers	4 layers
100	0.0485	0.0200	0.0198	**0.0196**
50	0.0568	0.0339	0.0339	**0.0337**
25	0.0746	**0.0658**	0.0667	0.0670
10^6	0.0131	0.000145	**0.000133**	**0.000133**

Table 1: The L^1-errors $\|u^\dagger - \mathbb{E}(u_N)\|_{L^1}$ between the true field and sample means for the one-dimensional simulations shown in Figure 7, for different numbers of data points J and layers N. Also shown are the corresponding errors for the simulations shown in Figure 8

.

in the cases $N = 50, 100$, more layers increases the accuracy of the mean, with diminishing returns for each additional layer. Note that higher accuracy could be attained in the single layer case by adjusting the constant length-scale parameter.

Finally, in Figure 9, we consider the same experiment as in Figure 7, except observations are limited to the subset $(0, 0.5)$ of the domain. Uncertainty is naturally higher in the unobserved portion of the domain. Uncertainty also increases in the observed layer u_N as N is increased; this could suggest that deep Gaussian processes may provide better generalization to unseen data than shallow Gaussian processes—note that the truth has much higher probability under the posterior with 4 layers versus just 1.

4.3.2 Two-Dimensional Simulations

We now consider the case $D = (0, 1)^2$, again where the forward map is given by a number of point evaluations. We fix the number of point observations $J = 2^{10}$, on a $2^5 \times 2^5$ uniform grid. We again compare quality of reconstruction versus the number of point evaluations and the number of levels in the deep Gaussian prior, and use the same parameters for the

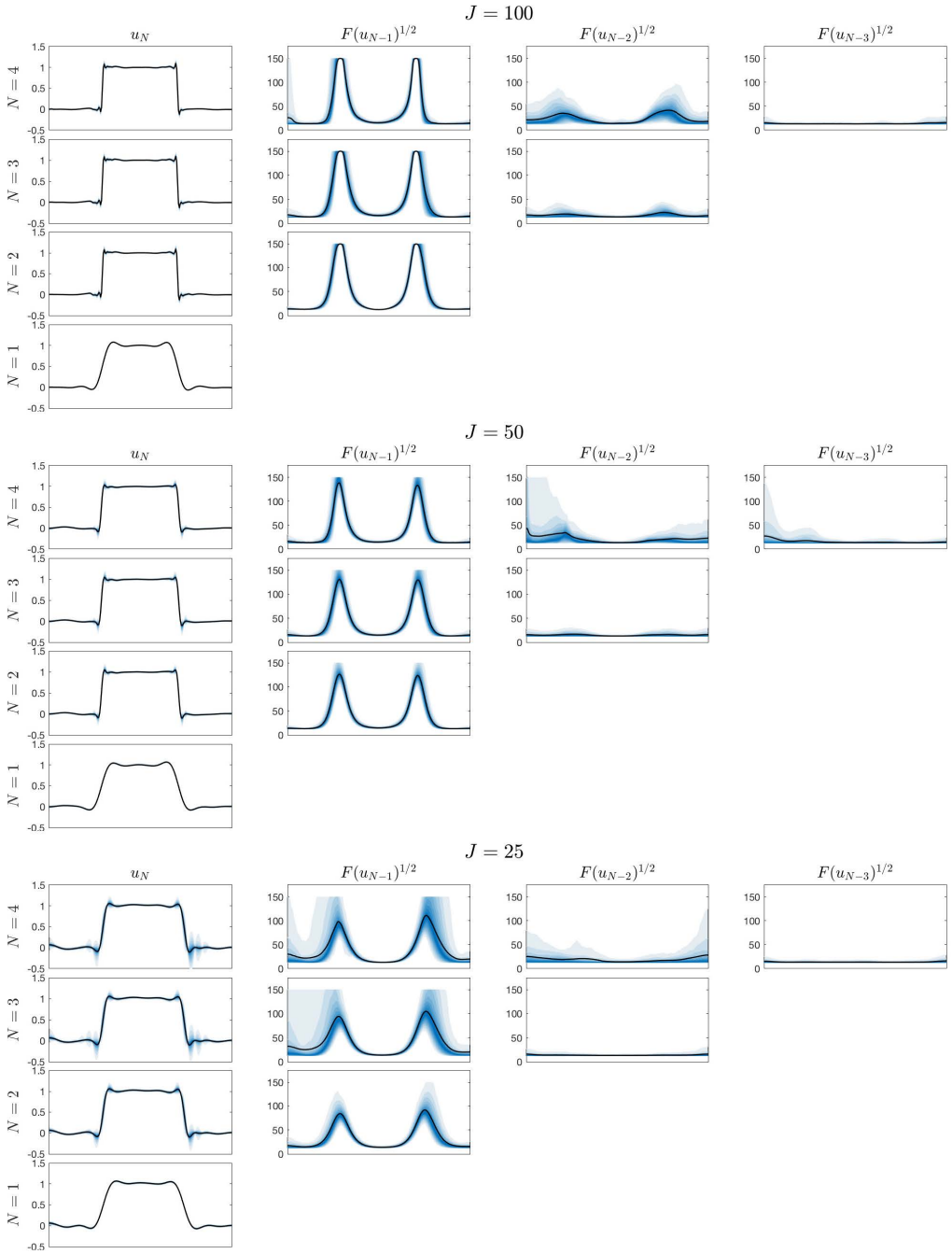

Figure 7: Estimates of posterior means (solid curves) and $5-95\%$ quantiles (shaded regions) arising from one-dimensional inverse problem. Number of data points taken are $J = 100$ (top block), $J = 50$ (middle block), $J = 25$ (bottom block). From left-to right, results for u_N, $F(u_{N-1})^{\frac{1}{2}}$, $F(u_{N-2})^{\frac{1}{2}}$, $F(u_{N-3})^{\frac{1}{2}}$ are shown. From top-to-bottom within each block, $N = 4, 3, 2, 1$.

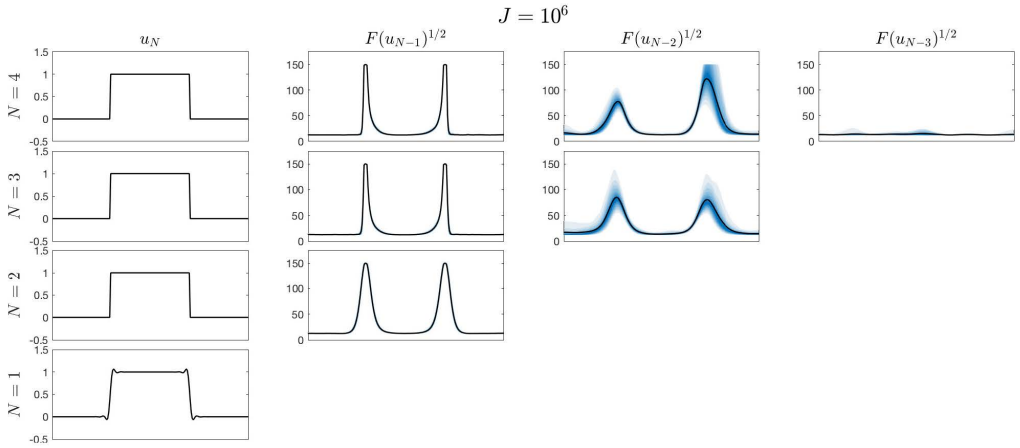

Figure 8: Estimates of posterior means (solid curves) and $5-95\%$ quantiles (shaded regions) arising from one-dimensional inverse problem, with $J = 10^6$ data points. From left-to right, results for u_N, $F(u_{N-1})^{\frac{1}{2}}$, $F(u_{N-2})^{\frac{1}{2}}$, $F(u_{N-3})^{\frac{1}{2}}$ are shown. From top-to-bottom, $N = 4, 3, 2, 1$.

family of covariance operators as in subsection 4.2. The base layer u_0 is taken to be Gaussian with covariance of the form (15), with $\Gamma(u) \equiv 20^2$.

The true unknown field $u^\dagger$ is constructed as a linear combination of truncated trigonometric functions with different length-scales, and shown in Figure 10 along with its contours. It is given by

$$\begin{aligned} u^\dagger(x,y) = &\cos(2\pi x)\cos(2\pi y) + \sin(4\pi x)\sin(4\pi y)\mathbb{1}_{(1/4,3/4)^2}(x,y) \\ &+ \sin(8\pi x)\sin(8\pi y)\mathbb{1}_{(1/2,3/4)^2}(x,y) \\ &+ \sin(16\pi x)\sin(16\pi y)\mathbb{1}_{(1/4,1/2)^2}(x,y). \end{aligned}$$

It is generated on a uniform square mesh of 2^{14} points, and two data sets are created wherein it is observed on uniform square grid of $J = 2^{10}, 2^8$ points, and corrupted by white noise with standard deviation $\gamma = 0.02$. Sampling is performed on a mesh of 2^{12} points to again avoid an inverse crime. 4×10^5 samples are generated per chain, with the first 2×10^5 discarded as burn-in when calculating means. Again the jump parameters β_j are adaptively tuned to keep acceptance rates close to 30%.

In Figure 11, analogously to Figure 7, the means of u_N and of the length-scales associated with each layer are shown, for $N = 1, 2, 3$. When $J = 2^{10}$, reconstructions are similar, though quality is generally proportional to the number of layers. In particular the, effect of too short a length-scale is evident in the case $N = 1$, in the regions where the length-scale should be larger, and conversely the effect of too long a length-scale is evident in the cases $N = 1, 2$ in the region where the length-scale should be the shortest. In the cases $N = 2, 3$, the length-scale fields $\mathbb{E}\big(F(u_{N-1})^{\frac{1}{2}}\big)$ are similar, though in the case $N = 3$ more accurately captures the true length-scales. When $J = 2^8$ the reconstructions are again similar, though

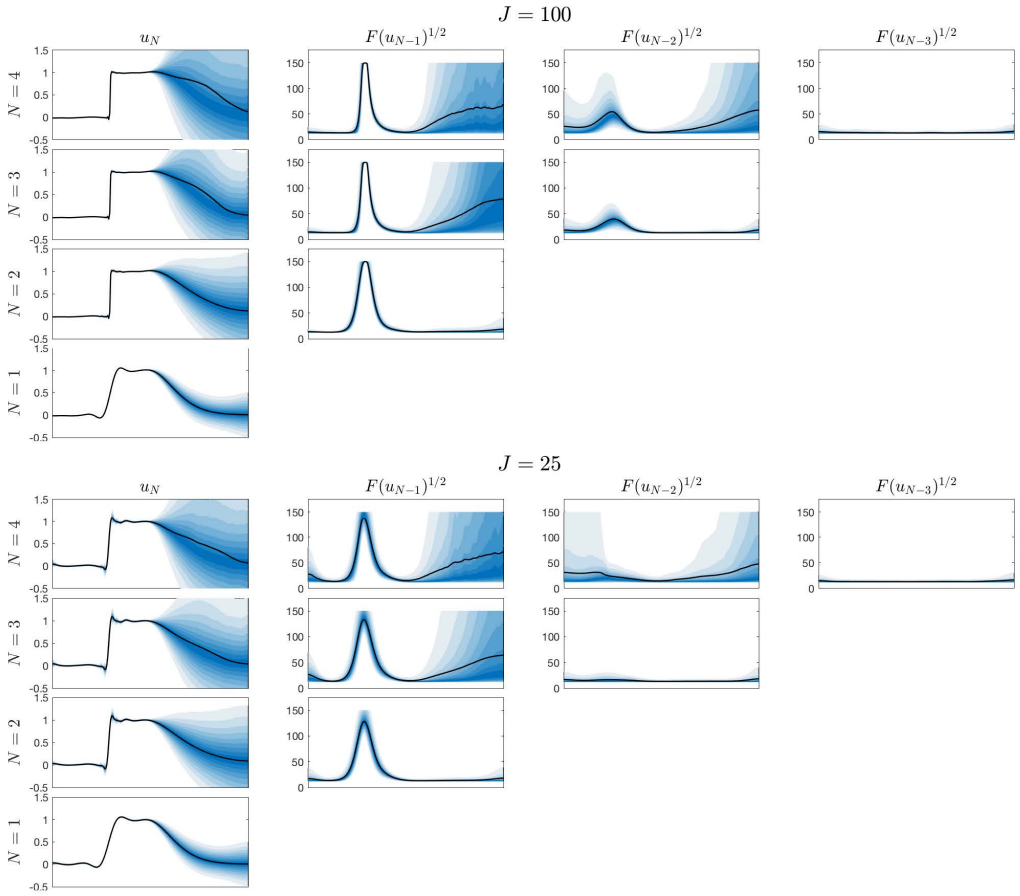

Figure 9: Estimates of posterior means (solid curves) and $5-95\%$ quantiles (shaded regions) arising from one-dimensional inverse problem. Number of data points taken are $J = 100$ (top block), $J = 25$ (bottom block). From left-to right, results for u_N, $F(u_{N-1})^{\frac{1}{2}}$, $F(u_{N-2})^{\frac{1}{2}}$, $F(u_{N-3})^{\frac{1}{2}}$ are shown. From top-to-bottom within each block, $N = 4, 3, 2, 1$.

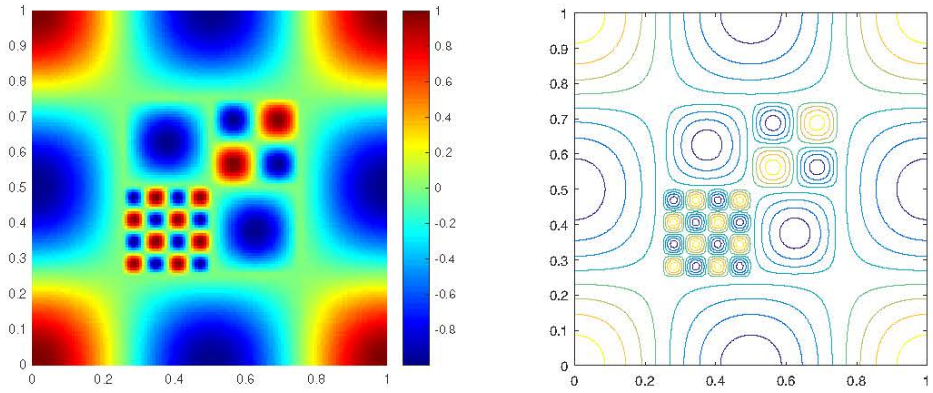

Figure 10: The true field used to generate the data for the two-dimensional inverse problem.

J	1 layer	2 layers	3 layers
2^{10}	0.0856	0.0813	**0.0681**
2^8	0.1310	**0.1260**	0.1279

Table 2: The L^2-errors $\|u^\dagger - \mathbb{E}(u_N)\|_{L^2}$ between the true field and sample means for the two-dimensional simulations shown in Figure 11, for different numbers of data points J and layers N.

there is now less accuracy in the shapes of the contours. In particular, the effect of too short a length-scale is especially evident in the case $N = 1$. The values of the reconstructed fields in the area of shortest length-scale are inaccurate in all cases—the positions of the observation points meant that the actual values of the peaks were not reflected in the data. The fields $\mathbb{E}\big(F(u_{N-1})^{\frac{1}{2}}\big)$ have similar structure to the case $J = 2^{10}$, though less accurately represent the true length scales. The L^2-errors between the means and the truth are shown in Table 2

5. Conclusions, Discussion, and Actionable Advice

In this section we provide an overview of the advantages and disadvantages of each of the four different DGP constructions considered in the paper, and summarize actionable advice that can be taken from the theoretical and numerical results that have been presented. We then outline a number of directions that would be interesting for future study.

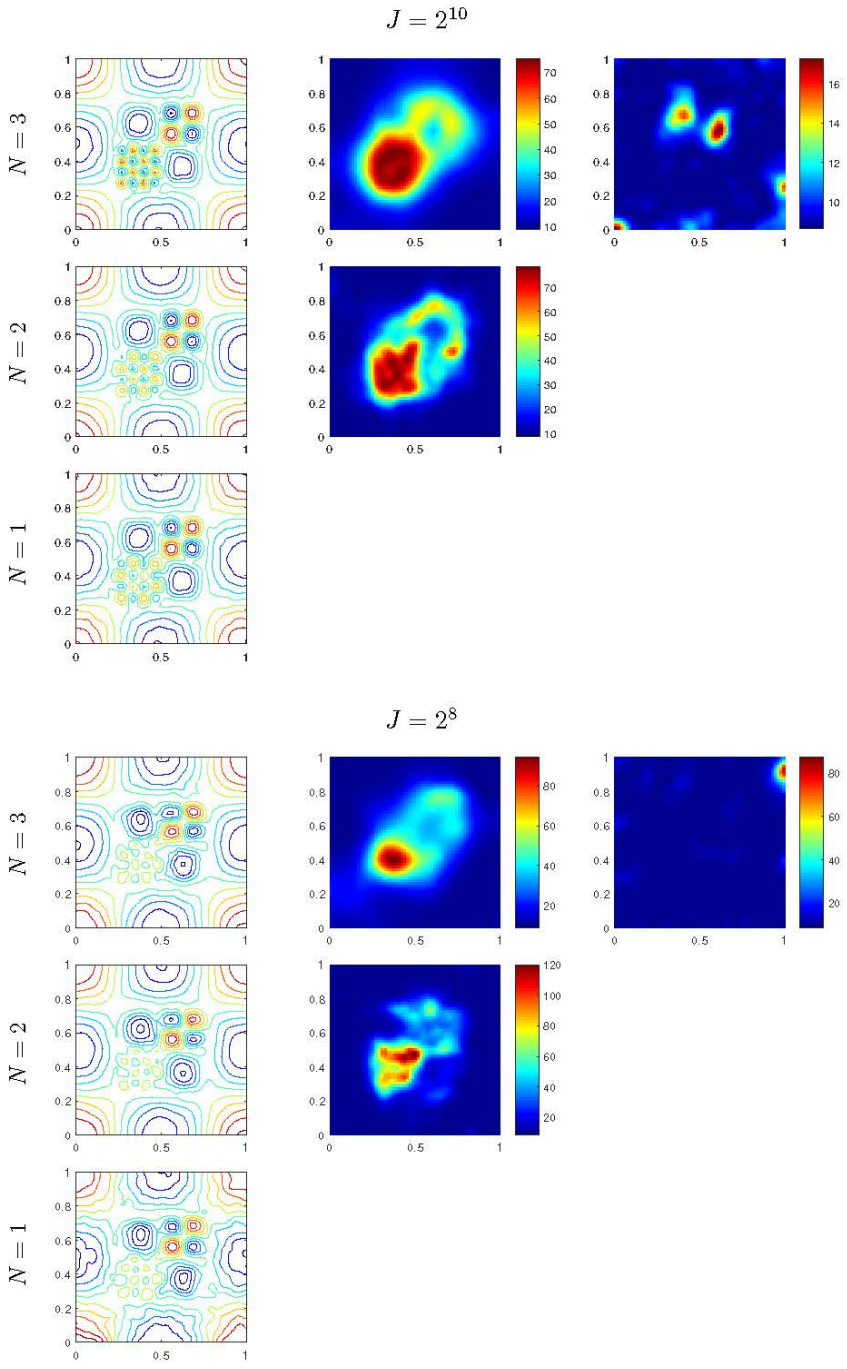

Figure 11: Estimates of posterior means arising from two-dimensional inverse problem. (Top block) $J = 2^{10}$, (Bottom block) $J = 2^8$. From left-to right, $\mathbb{E}(u_N)$, $\mathbb{E}(F(u_{N-1})^{\frac{1}{2}})$, $\mathbb{E}(F(u_{N-2})^{\frac{1}{2}})$. From top-to-bottom within each block, $N = 3, 2, 1$.

5.1 Comparison of Deep GP Constructions

We have considered four different constructions of deep GPs and we now discuss their relative merits. We also consider the context of variational inference which is popular in machine learning primarily because of its tractability. We emphasize however that it forms an uncontrolled approximation of the true posterior distribution and may fail to adequately represent the posterior distribution, and uncertainty in particular.

The **composition** construction is the classical construction introduced by Damianou and Lawrence (2013), building a hierarchy of layers using a stationary covariance function and composition. It has received the most study, and methods for variational inference have already been established. It has the advantage of scaling well with respect to data dimension d, however accurate sampling methods such as MCMC are intractable for large numbers of data points, due to the requirement to construct and factor dense covariance matrices at every step.

The **covariance function** construction builds the hierarchy using a stationary covariance function, and iteratively modifying its associated length scale. It has the advantage that each layer can be readily interpreted as the anisotropic length-scale field of the following layer. Its scaling properties are similar to those of the composition construction, however variational inference methods for this construction have not yet been studied.

The **covariance operator** construction builds the hierarchy using an SPDE representation of stationary Matèrn fields, and again iteratively modifies their associated length scale. It allows for fast sampling in low data dimension d via the use of PDE solvers, even when the number of data points is large. Accurate sampling via MCMC methods is tractable with this construction, due to the low cost of constructing and storing the inverse covariance (precision) matrix. Inference when d is large appears to be intractable at present, due to the requirement of dense meshes for PDE solvers.

Finally, the **convolution** construction builds the hierarchy via iterative convolution of Gaussian random fields. It has the advantage of being amenable to analysis, however the results of this analysis indicate that it would likely be a poor construction to use for inference due to trivial behaviour for large depth.

To summarize the numerical results on illustrative regression problems from the previous section, if the data is high quality, a small number of layers in the DGP will be sufficient as the problem becomes closer to interpolation. Conversely, if the data is low quality the likelihood is not strong enough to inform deeper layers in the DGP, and so a small number of layers is again sufficient. As a consequence, when the data lies between these two cases, and the truth has sufficiently rich structure, the use of deeper processes may be advantageous, but care is required to limit the number of layers employed.

5.2 Summary and Future Work

There are a number of interesting ways in which this work may be generalized. Within the context of covariance operators it is of interest to construct covariances $C(u)$ which are defined as $L^{-\alpha}$ with L being the divergence form elliptic operator

$$Lu = -\nabla \cdot \big(F(u)\nabla u\big).$$

Such a construction allows for the conditional distributions of the layers to be viewed as stationary on deformed spaces (Lindgren et al., 2011, §3.4), or to incorporate anisotropy in specific directions (Roininen et al., 2014, §3.1). Similar notions of anisotropy in different directions can be incorporated into the covariance function formulation by choosing the length scale $\Sigma(z)$ different to a multiple of the identity matrix. Additionally, we could consider a non-zero mean in the iteration (GP), as considered by Duvenaud et al. (2014); Salimbeni and Deisenroth (2017), allowing for forcing of the system. For example, with the choice $m(u_n) = u_n$ and a rescaling of the covariance, we obtain the ResNet-type iteration

$$u_{n+1} = u_n + \sqrt{\triangle t} L(u_n)\xi_{n+1}.$$

This may be viewed as a discretization of the continuous-time stochastic differential equation

$$\mathrm{d}u = L(u)L(u)^\top \mathrm{d}W,$$

analogously to what has been considered for neural networks (Haber and Ruthotto, 2017). Study of these systems could be insightful, for example deriving conditions to ensure a lack of ergodicity and hence arbitrary depth. As before $\top$ denotes the adjoint operation.

And finally it is possible to consider processes outside the four categories considered here; for example the one-step transition from u_n to u_{n+1} might be defined via stochastic integration against i.i.d. Brownian motions.

We have shown how a number of ideas in the literature may be recursed to produce deep Gaussian processes, different from those introduced by Damianou and Lawrence (2013). We have studied the effective depth of these processes, either through demonstrating ergodicity, or through showing convergence to a trivial solution (such as 0 or ∞). Together these results demonstrate that, as also shown by Duvenaud et al. (2014) for the original construction of deep Gaussian processes, care is needed in order to design processes with significant depth. Nonetheless, even a few layers can be useful for inference purposes, and we have demonstrated this also. It is an interesting question to ask precisely how the approximation power and effective depth are affected by the number of layers of the process, both in the non-ergodic case, and in the ergodic case before stationarity has been reached.

We also emphasize that the analysis in the paper is based solely on the deep Gaussian process u_n, and not the conditioned process $u_n|y$ in the inference problem with observed data y. The ergodicity properties of u_n do not directly carry over to $u_n|y$. As we have seen in the numerical experiments, the number of layers required in the inference problem in practice depends on the information content in the observed data y, and the analysis in this paper does not fully answer the question as to how many. The results in this paper do show, however, that in the case of ergodic constructions, the expressive power of the *prior* distribution in the inference problem does not increase past a certain number of layers. This provides some justification for using only a moderate number of layers in a deep Gaussian process prior in inference problems.

There are interesting approximation theory questions around deep processes, such as those identified in the context of neural networks by Pinkus (1999). There are also interesting questions around the use of these deep processes for inversion; in particular it seems hard to get significant value from using depth of more than two or three layers for noisy inverse problems. On the algorithmic side the issue of efficiently sampling these deep processes

(even over only two layers), when conditioned on possibly nonlinear observations remains open. We have used non-centred parameterizations because these may be sampled using function-space MCMC (Cotter et al., 2013; Chen et al., 2018); but centred methods, or mixtures, may be desirable for some applications.

Acknowledgments

MG is supported by EPSRC grants [EP/R034710/1, EP/R018413/1, EP/R004889/1, EP/P020720/1], an EPSRC Established Career Fellowship EP/J016934/3, a Royal Academy of Engineering Research Chair, and The Lloyds Register Foundation Programme on Data Centric Engineering. AMS is supported by AFOSR Grant FA9550-17-1-0185 and by US National Science Foundation (NSF) grant DMS 1818977. ALT is partially supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

Appendix A. Proofs for Section 2

Proof of Proposition 1 The stationary kernel ρ_{S} is positive definite by Assumption 1, and so by (Wendland, 2004, Theorem 7.14), we have

$$\rho_{\mathrm{S}}(r) = \int_0^\infty \exp(-r^2 t)\,\mathrm{d}\nu(t) \quad \text{for all } r \in [0,\infty),$$

for a finite, non-negative Borel measure ν on $[0,\infty)$ that is not concentrated at 0 (i.e. it is not a multiple of the Dirac measure centred at 0).

For any $x \in \mathbb{R}^d$ and $t \in [0,\infty)$, let us now define the matrix $\tilde{\Sigma}_t(x) := (4t)^{-1}\Sigma(x)$ and the functions

$$K_{x,t}(z) = \frac{1}{(2\pi)^{d/2}|\tilde{\Sigma}_t(x)|^{1/2}} \exp\Big(-\frac{1}{2}(x-z)^T \tilde{\Sigma}_t(x)^{-1}(x-z)\Big).$$

Here $|\cdot|$ denotes determinant and so the preceding is simply an expression for a normal density with mean x and covariance matrix $\tilde{\Sigma}_t(x)$ when $t > 0$; at $t = 0$, we simply have $K_{x,t}(z) = 0$, for all $x, z \in \mathbb{R}^d$. Then $\rho(x,x')$ is given by

$$\begin{aligned}
&\frac{2^{\frac{d}{2}}|\Sigma(x)|^{\frac{1}{4}}|\Sigma(x')|^{\frac{1}{4}}}{|\Sigma(x)+\Sigma(x')|^{\frac{1}{2}}}\rho_{\mathrm{S}}\Big(\sqrt{Q(x,x')}\Big) = \frac{2^{\frac{d}{2}}|\Sigma(x)|^{\frac{1}{4}}|\Sigma(x')|^{\frac{1}{4}}}{|\Sigma(x)+\Sigma(x')|^{\frac{1}{2}}}\int_0^\infty \exp\big(-tQ(x,x')\big)\,\mathrm{d}\nu(t)\\
&= \frac{2^{\frac{d}{2}}|\Sigma(x)|^{\frac{1}{4}}|\Sigma(x')|^{\frac{1}{4}}}{|\Sigma(x)+\Sigma(x')|^{\frac{1}{2}}}\int_0^\infty \exp\left(-t(x-x')^T\left(\frac{\Sigma(x)+\Sigma(x')}{2}\right)^{-1}(x-x')\right)\mathrm{d}\nu(t)\\
&= 2^{\frac{d}{2}}\int_0^\infty \frac{|\tilde{\Sigma}_t(x)|^{\frac{1}{4}}|\tilde{\Sigma}_t(x')|^{\frac{1}{4}}}{|\tilde{\Sigma}_t(x)+\tilde{\Sigma}_t(x')|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-x')^T\left(\tilde{\Sigma}_t(x)+\tilde{\Sigma}_t(x')\right)^{-1}(x-x')\right)\mathrm{d}\nu(t)\\
&= (2\pi)^{\frac{d}{2}}2^{\frac{d}{2}}\int_0^\infty |\tilde{\Sigma}_t(x)|^{\frac{1}{4}}|\tilde{\Sigma}_t(x')|^{\frac{1}{4}}\int_{\mathbb{R}^d} K_{x,t}(z)K_{x',t}(z)\,\mathrm{d}z\,\mathrm{d}\nu(t),
\end{aligned}$$

where in the last step, we have used the fact that the convolution $\int_{\mathbb{R}^d} K_{x,t}(z)K_{x',t}(z)\,\mathrm{d}z$ can be calculated explicitly using properties of normal random variables. More precisely, we

have

$$\int_{\mathbb{R}^d} K_{x,t}(z)K_{x',t}(z)\,\mathrm{d}z = \int_{\mathbb{R}^d} p_X(z-x)p_{X'}(z)\,\mathrm{d}z = \int_{\mathbb{R}^d} p_{X,X'}(z-x,z)\,\mathrm{d}z,$$

where p_X is the density of $X \sim N(0, \tilde{\Sigma}_t(x))$, $p_{X'}$ is the density of $X' \sim N(x', \tilde{\Sigma}_t(x'))$ and X and X' are independent. The change of variable from X, X' to W, X', where $W = X' - X$, has Jacobian 1, and so

$$\int_{\mathbb{R}^d} p_{X,X'}(z-x,z)\,\mathrm{d}z = \int_{\mathbb{R}^d} p_{W,X'}(z-(z-x),z)\,\mathrm{d}z = \int_{\mathbb{R}^d} p_{W,X'}(x,z)\,\mathrm{d}z = p_W(x).$$

Since $W = X' - X \sim N(x', \tilde{\Sigma}_t(x) + \tilde{\Sigma}_t(x'))$, we hence have

$$\begin{aligned}
&\int_{\mathbb{R}^d} K_{x,t}(z)K_{x',t}(z)\,\mathrm{d}z \\
&\quad = \frac{1}{(2\pi)^{\frac{d}{2}}|\tilde{\Sigma}_t(x) + \tilde{\Sigma}_t(x')|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}(x-x')^T \left(\tilde{\Sigma}_t(x) + \tilde{\Sigma}_t(x')\right)^{-1}(x-x')\Big),
\end{aligned}$$

as required.

Now, for any $b \in \mathbb{R}^N$ and pairwise distinct $\{x_i\}_{i=1}^N$, we then have

$$\begin{aligned}
&\sum_{i=1}^N \sum_{j=1}^N b_i b_j \rho(x_i, x_j) \\
&= (2\pi)^{\frac{d}{2}} 2^{\frac{d}{2}} \sum_{i=1}^N \sum_{j=1}^N b_i b_j \int_0^\infty \int_{\mathbb{R}^d} |\tilde{\Sigma}_t(x_i)|^{\frac{1}{4}} K_{x_i,t}(z) |\tilde{\Sigma}_t(x_j)|^{\frac{1}{4}} K_{x_j,t}(z)\,\mathrm{d}z\,\mathrm{d}\nu(t) \\
&= (2\pi)^{\frac{d}{2}} 2^{\frac{d}{2}} \int_0^\infty \int_{\mathbb{R}^d} \left(\sum_{i=1}^N b_i |\tilde{\Sigma}_t(x_i)|^{\frac{1}{4}} K_{x_i,t}(z) \right)^2 \mathrm{d}z\,\mathrm{d}\nu(t) \\
&\geq 0,
\end{aligned}$$

since the Borel measure ν is finite and non-negative. It remains to show that strict inequality also holds.

Firstly, we note that $|\tilde{\Sigma}_0(x_i)|^{\frac{1}{4}} K_{x_i,0}(z) = 0$, for all $x_i, z \in \mathbb{R}^d$, which means that the integrand with respect to t is identically equal to zero at $t = 0$. Secondly, we note that the points $\{x_i\}_{i=1}^N$ are pairwise distinct and the functions $\{|\tilde{\Sigma}_t(x_i)|^{\frac{1}{4}} K_{x_i,t}(\cdot)\}_{i=1}^N$ are hence linearly independent for any $t \in (0, \infty)$. It is thus impossible to make the integrand with respect to z identically equal to 0 for a.e. $z \in \mathbb{R}^d$. As a consequence the integrand with respect to t is positive for all $t \in (0, \infty)$. Since we know that the measure ν is not concentrated at 0 this completes the proof that ρ is positive definite on $\mathbb{R}^d \times \mathbb{R}^d$, for any $d \in \mathbb{N}$.

Finally, we note that the kernel ρ is clearly non-stationary, and is a correlation function since $\rho(x, x) = 1$, for any $x \in \mathbb{R}^d$. ■

Proof of Proposition 2 We note that the definition of positive definite in Assumptions 1(i) refers only to behaviour of the kernel on a finite set of pairwise distinct points $\{x_i\}_{i=1}^N$. By Assumption 2(i), the function G is non-negative and bounded. If $G(z) > 0$ for all $z \in \mathbb{R}^d$,

then the matrix $\Sigma(z)$ is positive definite for all $z \in \mathbb{R}^d$, and the fact that $\rho(\cdot,\cdot)$ is positive definite follows directly from Proposition 1.

It remains to investigate the case where $G(z) = 0$ for some $z \in \mathbb{R}^d$. We will prove that $\rho(\cdot,\cdot)$ is positive definite by showing that the correlation matrix $\mathbf{R}$, with entries $\mathbf{R}_{ij} = \rho(x_i, x_j)$, is positive definite for any pairwise disjoint points $\{x_i\}_{i=1}^N$. Without loss of generality, we will study the case $G(x_1) = 0$; the proof easily adapts to the case where $G(x_i) = 0$, for $i \neq 1$. To define $\rho(x_1, x_j)$ in this case, we start by assuming $G(x_1) > 0, G(x_j) > 0$, and then take limits.

With $\Sigma(z) = G(z)\mathrm{I}_d$, we have

$$\begin{aligned} Q(x_1, x_j) &= (x_1 - x_j)^T \left(\frac{\Sigma(x_1) + \Sigma(x_j)}{2}\right)^{-1} (x_1 - x_j) \\ &= 2\|x_1 - x_j\|_2^2 \Big(G(x_1) + G(x_j)\Big)^{-1}, \end{aligned}$$

where $\|\cdot\|_2$ is the Euclidean norm, and

$$\frac{2^{\frac{d}{2}} \det(\Sigma(x_1))^{\frac{1}{4}} \det(\Sigma(x_j))^{\frac{1}{4}}}{\det(\Sigma(x_1) + \Sigma(x_j))^{\frac{1}{2}}} = \left(\frac{4G(x_1)G(x_j)}{\big(G(x_1) + G((x_j)\big)^2}\right)^{\frac{d}{4}}.$$

We now study separately three cases:

i) $x_j = x_1$: we have

$$\lim_{G(x_1)\to 0} \left(\frac{4G(x_1)G(x_1)}{\big(G(x_1) + G(x_1)\big)^2}\right)^{\frac{d}{4}} = \lim_{G(x_1)\to 0} 1 = 1, \tag{18}$$

and so using the algebra of limits, the continuity of ρ_{S}, (18) and the fact that $\rho_{\mathrm{S}}(0) = 1$, we have

$$\lim_{G(x_1)\to 0} \rho(x_1, x_1) = \lim_{G(x_1)\to 0} \rho_{\mathrm{S}}\Big(\sqrt{Q(x_1, x_1)}\Big) = \rho_{\mathrm{S}}(0) = 1.$$

ii) $x_j \neq x_1$ and $G(x_j) > 0$: we have

$$\lim_{G(x_1)\to 0} Q(x_1, x_j) = 2\|x_1 - x_j\|_2^2 \Big(G(x_j)\Big)^{-1},$$

and

$$\lim_{G(x_1)\to 0} \left(\frac{4G(x_1)G(x_j)}{\big(G(x_1) + G(x_j)\big)^2}\right)^{\frac{d}{4}} = 0. \tag{19}$$

Thus, using the continuity of ρ_{S}, together with (19) and the algebra of limits, we have

$$\lim_{G(x_1)\to 0} \rho(x_1, x_j) = 0.$$

iii) $x_j \neq x_1, G(x_j) = 0$: we obtain

$$\lim_{G(x_1),G(x_j)\to 0} Q(x_1, x_j) = \infty,$$

which by Assumptions 2(ii) implies that

$$\lim_{G(x_1),G(x_j)\to 0} \rho_S\left(\sqrt{Q(x_1, x_j)}\right) = 0.$$

Since $(a+b)^2 \geq 4ab$ for any positive numbers a and b, we have

$$0 \leq \left(\frac{4G(x_1)G(x_j)}{\left(G(x_1) + G((x_j)\right)^2}\right)^{\frac{d}{4}} \leq 1,$$

for any $G(x_1) > 0, G(x_j) > 0$, and hence

$$\lim_{G(x_1),G(x_j)\to 0} \rho(x_1, x_j) = 0.$$

Hence, when $G(x_i) > 0$, for $i = 2, \ldots, N$, we have $\lim_{G(x_1)\to 0} \mathbf{R} = \mathbf{R}^*$, where the matrix $\mathbf{R}^*$ has the first row and column equal to the first basis vector $e_1 = (1, 0, 0, \ldots, 0) \in \mathbb{R}^N$, and the remaining submatrix $\mathbf{R}^*_{N-1} \in \mathbb{R}^{N-1\times N-1}$ with entries $\rho(x_i, x_j)$, for $i, j = 2, \ldots, N$. The matrix $\mathbf{R}^*_{N-1}$ is positive definite by Proposition 1, from which we can conclude that $\mathbf{R}^*$ is positive definite also. A similar argument holds when $G(x_i) = 0$ for one or more indices $i \in \{2, \ldots, N\}$. ■

References

Neil K Chada, Marco A Iglesias, Lassi Roininen, and Andrew M Stuart. Parameterizations for ensemble Kalman inversion. *Inverse Problems*, 34(5):055009, 2018.

Victor Chen, Matthew M Dunlop, Omiros Papaspiliopoulos, and Andrew M Stuart. Robust MCMC sampling with non-Gaussian and hierarchical priors in high dimensions. arXiv preprint arXiv:1803.03344, 2018.

Simon L Cotter, Gareth O Roberts, Andrew M Stuart, and David White. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28(3): 424–446, 2013.

Kurt Cutajar, Edwin V Bonilla, Pietro Michiardi, and Maurizio Filippone. Random feature expansions for deep Gaussian processes. In *International Conference on Machine Learning*, pages 884–893, 2017.

Zhenwen Dai, Andreas Damianou, Javier González, and Neil Lawrence. Variational auto-encoded deep Gaussian processes. arXiv preprint arXiv:1511.06455, 2015.

Andreas C Damianou and Neil D Lawrence. Deep Gaussian Processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013.

Yair Daon and Georg Stadler. Mitigating the influence of the boundary on PDE-based covariance operators. *Inverse Problems and Imaging*, 12(5), 2018.

Masoumeh Dashti and Andrew M Stuart. The Bayesian approach to inverse problems. *Handbook of Uncertainty Quantification*, 2017.

Persi Diaconis and David Freedman. Iterated random functions. *SIAM Review*, 41(1):45–76, 1999.

David K Duvenaud, Oren Rippel, Ryan P Adams, and Zoubin Ghahramani. Avoiding pathologies in very deep networks. In *Artificial Intelligence and Statistics*, pages 202–210, 2014.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004, 2017.

Martin Hairer. An introduction to stochastic PDEs. arXiv preprint arXiv:0907.4178, 2009.

Martin Hairer, Andrew M Stuart, Jochen Voss, and Petter Wiberg. Analysis of SPDEs arising in path sampling. Part I: The Gaussian case. *Communications in Mathematical Sciences*, 3(4):587–603, 2005.

Markus Heinonen, Henrik Mannerström, Juho Rousu, Samuel Kaski, and Harri Lähdesmäki. Non-stationary Gaussian process regression with Hamiltonian Monte Carlo. In *Artificial Intelligence and Statistics*, pages 732–740, 2016.

Dave Higdon, Marc Kennedy, James C Cavendish, John A Cafeo, and Robert D Ryne. Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26(2):448–466, 2004.

Marco A Iglesias, Yulong Lu, and Andrew M Stuart. A Bayesian level set method for geometric inverse problems. *Interfaces and Free Boundaries*, 18(2):181–217, 2016.

Jari Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*, volume 160. Springer Science & Business Media, 2006.

Gopinath Kallianpur. *Stochastic Filtering Theory*, volume 13. Springer Science & Business Media, 2013.

Marc C Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

Jonathan C Mattingly, Andrew M Stuart, and Desmond J Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101(2):185–232, 2002.

Sean P Meyn and Richard L Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, 2012.

Radford M Neal. *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, 1995.

Radford M Neal. Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. arXiv preprint physics/9701026, 1997.

Christopher J Paciorek. *Nonstationary Gaussian processes for regression and spatial modelling*. PhD thesis, Carnegie Mellon University, 2003.

Christopher J Paciorek and Mark J Schervish. Nonstationary covariance functions for Gaussian process regression. *Advances in Neural Information Processing Systems*, 16:273–280, 2004.

Omiros Papaspiliopoulos, Gareth O Roberts, and Martin Sköld. A general framework for the parametrization of hierarchical models. *Statistical Science*, pages 59–73, 2007.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999.

Frank J Pinski, Gideon Simpson, Andrew M Stuart, and Hendrik Weber. Kullback–Leibler approximation for probability measures on infinite dimensional spaces. *SIAM Journal on Mathematical Analysis*, 47(6):4091–4122, 2015.

Carl E Rasmussen and Christopher K I Williams. Gaussian processes for machine learning. *the MIT Press*, 2(3):4, 2006.

James C Robinson. *Infinite-Dimensional Dynamical Systems: An Introduction to Dissipative Parabolic PDEs and the Theory of Global Attractors*, volume 28. Cambridge University Press, 2001.

Lassi Roininen, Janne M J Huttunen, and Sari Lasanen. Whittle-Matérn priors for Bayesian statistical inversion with applications in electrical impedance tomography. *Inverse Problems and Imaging*, 8(2):561–586, 2014.

Lassi Roininen, Mark Girolami, Sari Lasanen, and Markku Markkanen. Hyperpriors for Matérn fields with applications in Bayesian inversion. arXiv preprint arXiv:1612.02989, 2016.

Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 4588–4599, 2017.

Alexandra M Schmidt and Anthony O'Hagan. Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):743–758, 2003.

Edward Snelson, Zoubin Ghahramani, and Carl E Rasmussen. Warped Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 337–344, 2004.

Michael L Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

Holger Wendland. *Scattered Data Approximation*, volume 17. Cambridge University Press, 2004.

Yaming Yu and Xiao-Li Meng. To center or not to center: That is not the question—an Ancillarity–Sufficiency Interweaving Strategy (ASIS) for boosting MCMC efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570, 2011.