Bayesian inference for complex models

Lecture 3

Aretha Teckentrup

School of Mathematics, University of Edinburgh

SFB1294 Data assimilation Spring school - March 21, 2022

THE UNIVERSITY *of* EDINBURGH
School of Mathematics

Motivation

- So far, we have mainly studied algorithms to infer a finite set of parameters $u \in \mathbb{R}^{d_u}$.
- Today, we will focus on inferring an unknown function $f : X \to \mathbb{R}$.

Outline

Problem formulation

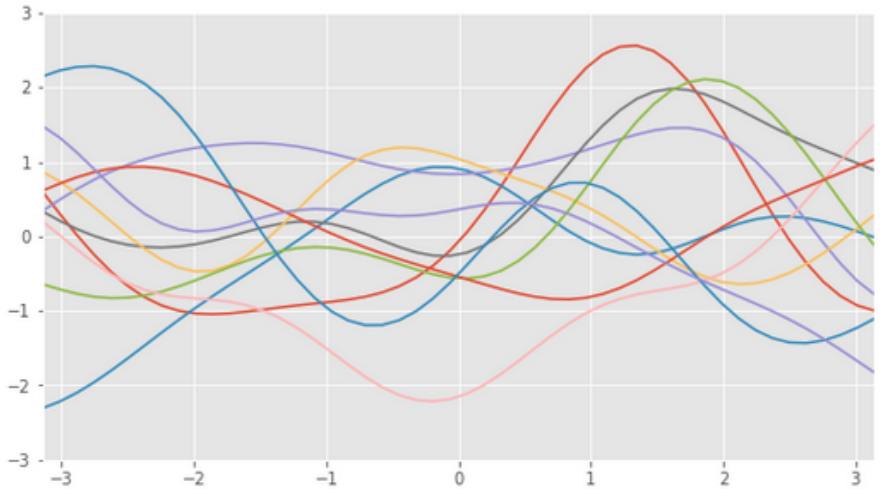
- Given N function evaluations $y = \{x^n, f(x^n)\}_{n=1}^N$, we want to reconstruct the underlying function $f : X \to \mathbb{R}$, $X \subseteq \mathbb{R}^{d_x}$.
  - This problem appears in many contexts including machine learning and reduced order modelling.
  - For general f, this is not an easy problem to solve.
- Solving this problem in a Bayesian statistical framework results in a posterior measure on $f|y$, allowing for uncertainty quantification in the reconstruction.

Gaussian process regression

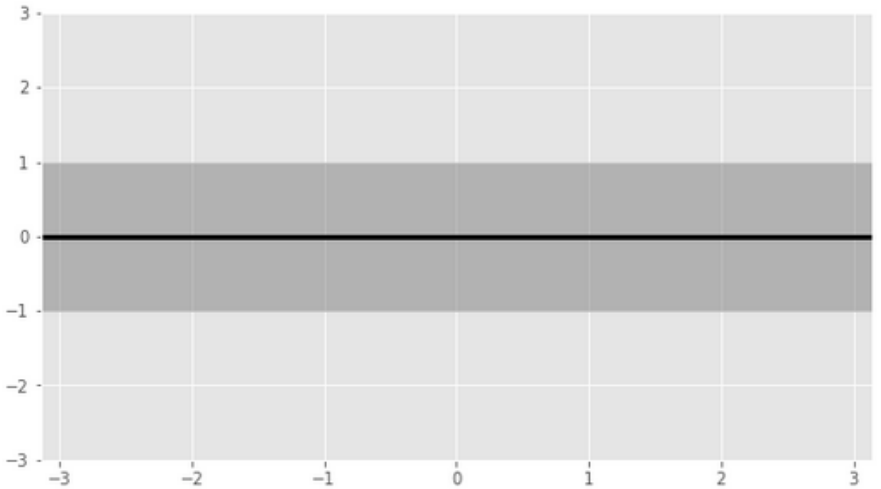
Set-up [Rasmussen, Williams '06]

- Gaussian process regression is an instance of the Bayesian framework.
- We put a Gaussian process prior $\mathrm{GP}(0, k)$ on f, where k is chosen to reflect properties of f.

  For $\{x_i\}_{i=1}^m \subseteq X$, the random variables $\{f(x_i)\}_{i=1}^m$ follow a joint Gaussian distribution with $\mathbb{E}[f(x_i)] = 0$ and $\mathbb{C}[f(x_i), f(x_j)] = k(x_i, x_j)$.



Sample paths



Mean and standard deviation

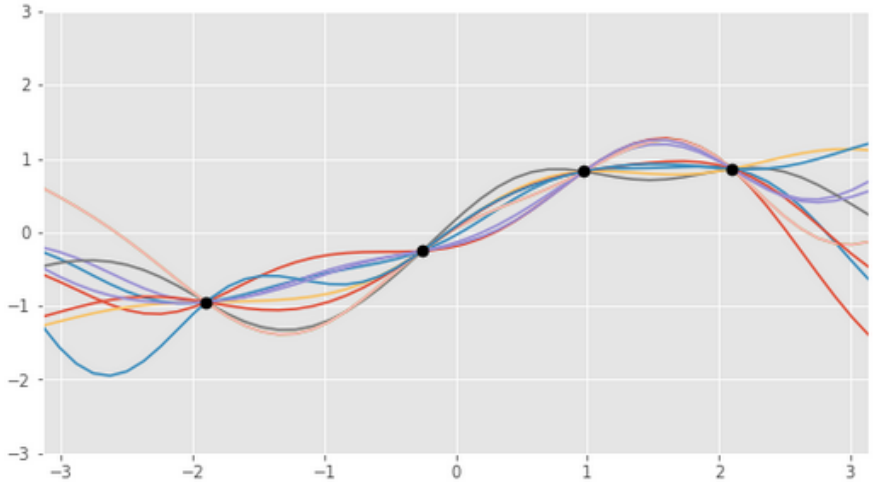
Gaussian process regression

Set-up [Rasmussen, Williams '06]

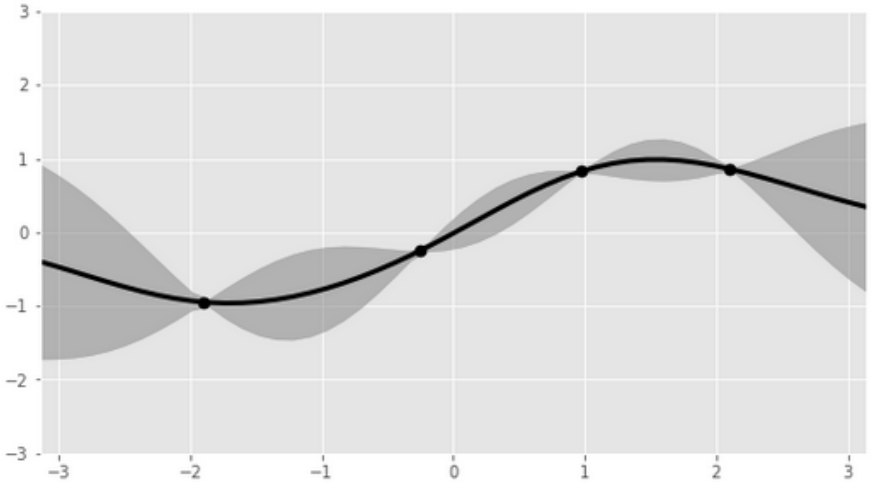
- The Gaussian process posterior $\mathrm{GP}(m_N^f, k_N)$ on $f|y$ is obtained by conditioning the prior on the observed data $y = \{x^n, f(x^n)\}_{n=1}^N$. Here

$$m_N^f(x) = k(x, D_N)^T K(D_N, D_N)^{-1} f(D_N),$$
$$k_N(x, x') = k(x, x') - k(x, D_N)^T K(D_N, D_N)^{-1} k(x', D_N),$$

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



Sample paths



Mean and standard deviation

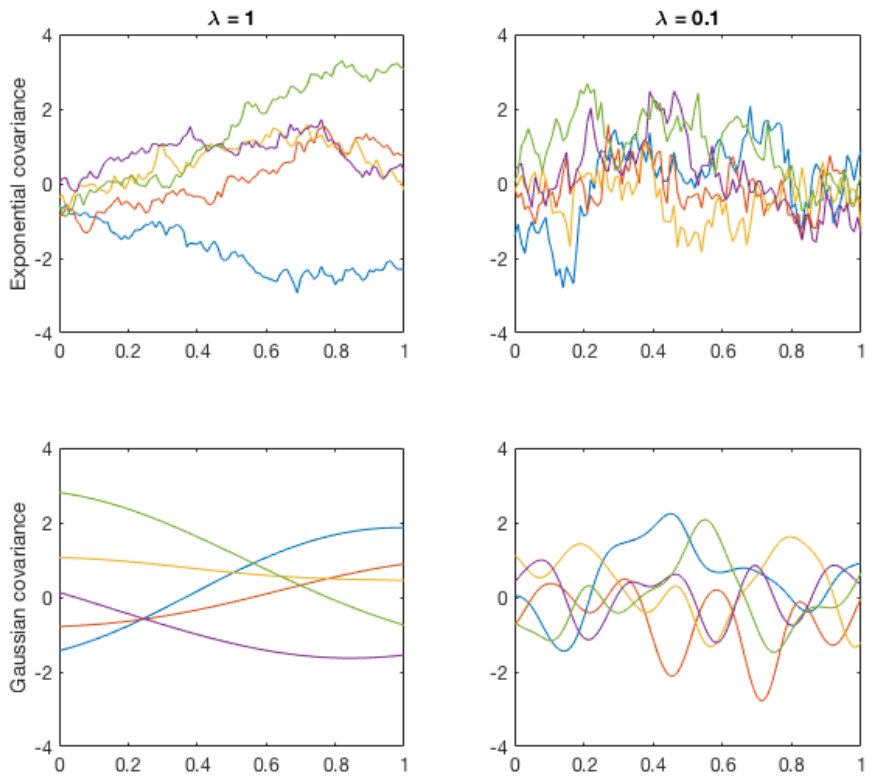
Gaussian process regression

Choice of prior distribution

- The prior $\mathrm{GP}(0, k)$ needs to be chosen to reflect properties of f.

The covariance kernel k determines properties of the Gaussian process and its sample paths:

- smoothness (differentiability),
- contrast,
- length scales of fluctuations,
- periodicity,
- stationarity.



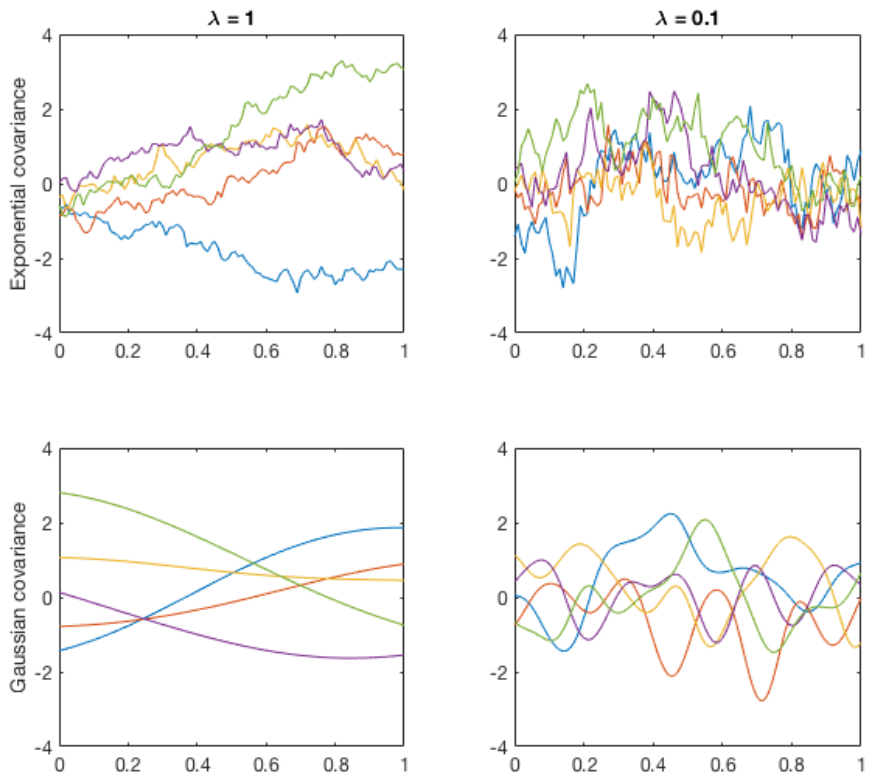
Gaussian process regression

Choice of prior distribution

- The prior $\mathrm{GP}(0, k)$ needs to be chosen to reflect properties of f.

The covariance kernel k determines properties of the Gaussian process and its sample paths:

- smoothness (differentiability),
- contrast,
- length scales of fluctuations,
- periodicity,
- stationarity.



- Challenge: hyper-parameters θ are usually unknown a-priori!

Gaussian process regression

Empirical Bayes'

- In a hierarchical Bayesian approach, we obtain the posterior $f|y$ as a marginal distribution of the joint posterior $f, \theta|y$. This is often intractable.

- We use an empirical Bayes' (or plug-in) approach, where we estimate values of any hyper-parameters θ from $y = \{x^n, f(x^n)\}_{n=1}^N$ and plug the estimate $\widehat{\theta}_N$ into the prior distribution.

Gaussian process regression

Empirical Bayes'

- In a hierarchical Bayesian approach, we obtain the posterior $f|y$ as a marginal distribution of the joint posterior $f, \theta|y$. This is often intractable.
- We use an empirical Bayes' (or plug-in) approach, where we estimate values of any hyper-parameters θ from $y = \{x^n, f(x^n)\}_{n=1}^N$ and plug the estimate $\widehat{\theta}_N$ into the prior distribution.
- The sequence of estimates $\widehat{\theta}_N$ can be found via maximum likelihood estimation, maximum a-posteriori estimation, cross validation, ...
- Under what conditions do we get posterior consistency for the Gaussian process posterior $\mathrm{GP}(m_N^f(\widehat{\theta}_N), k_N(\widehat{\theta}_N))$ as $N \to \infty$?

Convergence of Gaussian process regression

Literature review

- Earlier references (not exhaustive!) on this question include:
  - [Stein '88] and [Stein '93]
  - [Choi and Schervish '07]
  - [van der Vaart and van Zanten '11]
  - [Scheuerer, Schaback, Schlather '13]
  - ...
- These results do not apply in our current setting where:
  - the function f is a given deterministic function,
  - the design points are deterministic,
  - there is no noise in the training data,
  - we want to measure the error in $L^2(X)$-norms, ,
  - and the estimated hyper-parameters $\widehat{\theta}_N$ change with N.

Convergence of Gaussian process regression

Relation to Kernel Interpolation

- To prove convergence as $N \to \infty$, we can make use of results from numerical analysis. We want $m_N^f \to f$ and $k_N \to 0$.

Convergence of Gaussian process regression

Relation to Kernel Interpolation

- To prove convergence as $N \to \infty$, we can make use of results from numerical analysis. We want $m_N^f \to f$ and $k_N \to 0$.
- We have $m_N^f(x^n) = f(x^n)$ and $k_N(x^n, x^n) = 0$, for $n = 1, \ldots, N$.
- The predictive mean m_N^f is a linear combination of kernel functions:

$$m_N^f(x) = \sum_{n=1}^{N} \alpha_n k(x, x^n), \qquad \text{for known } \alpha \in \mathbb{R}^N.$$

Convergence of Gaussian process regression

Relation to Kernel Interpolation

- To prove convergence as $N \to \infty$, we can make use of results from numerical analysis. We want $m_N^f \to f$ and $k_N \to 0$.
- We have $m_N^f(x^n) = f(x^n)$ and $k_N(x^n, x^n) = 0$, for $n = 1, \ldots, N$.
- The predictive mean m_N^f is a linear combination of kernel functions:

$$m_N^f(x) = \sum_{n=1}^{N} \alpha_n k(x, x^n), \qquad \text{for known } \alpha \in \mathbb{R}^N.$$

- The predictive mean m_N^f is a kernel interpolant of f, and in the special case of isotropic kernels $k(x, x') = k(\|x - x'\|_2)$, a radial basis function interpolant.
- Convergence properties will depend on the specific choice of k.

Convergence of Gaussian process regression

Scattered Data Approximation [ALT, '20], [Wendland '04]

- Suppose we use the family of Matèrn covariances

$$k_{\text{Mat}}(x, x') = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}} \left(\frac{\|x - x'\|_2}{\lambda} \right)^{\nu} B_\nu \left(\frac{\|x - x'\|_2}{\lambda} \right).$$

  Special cases: $\nu = \frac{1}{2} \Rightarrow \sigma^2 \exp(-\frac{\|x-x'\|_2}{\lambda})$ and $\nu = \infty \Rightarrow \sigma^2 \exp(-\frac{\|x-x'\|_2^2}{\lambda^2})$.

- $\sigma^2 \to$ contrast, $\lambda \to$ fluctuation length scale and $\nu \to$ smoothness.

Convergence of Gaussian process regression

Scattered Data Approximation

Theorem [ALT '20]

With covariance kernel k_{Mat}, we have with $\nu^* = \lim_{N\to\infty} \widehat{\nu}_N$

$$\|f - m_N^f(\widehat{\theta}_N)\|_{L^2(X)} \leq C \underbrace{h_{D_N}^{\min\{\nu,\nu^*\}+\frac{d_u}{2}}}_{\text{decreasing in } N \atop \to \textit{convergence}} \underbrace{\rho_{D_N}^{\max\{\nu^*-\nu,0\}}}_{\text{non-decreasing in } N \atop \to \textit{stability}} \|f\|_{H^{\nu+d_u/2}(X)}.$$

Furthermore, $\|k_N(\widehat{\theta}_N)^{\frac{1}{2}}\|_{L^2(X)} \leq C' h_{D_N}^{\min\{\nu,\nu^*\}} \rho_{D_N}^{\max\{\nu^*-\nu,0\}}$.

- With design points $D_N = \{x^n\}_{n=1}^N$, define the fill distance

$$h_{D_N} = \sup_{x\in X} \inf_{x^n \in D_N} \|x - x^n\|_2. \qquad h_{D_N} \sim N^{-1/d_x}$$

  and mesh ratio

$$\rho_{D_N} = \frac{\sup_{x\in X} \inf_{x^n\in D_N} \|x - x^n\|_2}{\min_{n\neq m} \|x^n - x^m\|_2} \qquad \rho_{D_N} \geq 1$$

- $h_{D_N} \to$ space filling, $\rho_{D_N} \to$ quasi-uniformity.

Convergence of Gaussian process regression

Scattered Data Approximation

Theorem [ALT '20]

With covariance kernel k_{Mat}, we have with $\nu^* = \lim_{N\to\infty} \widehat{\nu}_N$

$$\|f - m_N^f(\widehat{\theta}_N)\|_{L^2(X)} \leq C \underbrace{h_{D_N}^{\min\{\nu,\nu^*\}+\frac{d_u}{2}}}_{\text{decreasing in } N \atop \to \textit{convergence}} \underbrace{\rho_{D_N}^{\max\{\nu^*-\nu,0\}}}_{\text{non-decreasing in } N \atop \to \textit{stability}} \|f\|_{H^{\nu+d_u/2}(X)}.$$

Furthermore, $\|k_N(\widehat{\theta}_N)^{\frac{1}{2}}\|_{L^2(X)} \leq C' h_{D_N}^{\min\{\nu,\nu^*\}} \rho_{D_N}^{\max\{\nu^*-\nu,0\}}$.

- This result holds for any $f \in H^{\nu+d_u/2}(X)$.
- We require $0 < \widehat{\sigma}_N^2, \widehat{\lambda}_N, \widehat{\nu}_N < \infty$, but estimates need not converge.
  The dependency of C, C' on $\sigma_N^2, \widehat{\lambda}_N, \widehat{\nu}_N$ can be tracked explicitly.
- Optimal convergence rates $N^{-\frac{\nu+d_x/2}{d_x}}$ are obtained with $\nu^* = \nu$.
- Penalty for over-estimating the smoothness only if ρ_{D_N} grows with N, i.e. if D_N is not quasi-uniform.

Conclusions

- Using Gaussian process regression, we can determine a posterior measure on an unknown function f given N function values.
- We get posterior consistency as $N \to \infty$: $\|m_N^f - f\|_{L^2(X)} \to 0$ and $\|k_N^{1/2}\|_{L^2(X)} \to 0$.
- The choice of L^2-norm comes from applications - eg approximating the solution operator of a PDE in an inverse problem.
  We also get point-wise error estimates.
- Posterior consistency is guaranteed under mild assumptions, and is robust to prior mis-specification and learning of hyper-parameters.

Deep Gaussian process regression

General framework [Dunlop et al, '18]

- Deep Gaussian processes can be used as prior distributions in regression.
- A general framework allows for a large variety of constructions, which can easily model complex behaviour and tune hyper-parameters.

Deep Gaussian process regression

General framework [Dunlop et al, '18]

- Deep Gaussian processes can be used as prior distributions in regression.
- A general framework allows for a large variety of constructions, which can easily model complex behaviour and tune hyper-parameters.
- We consider sequences $\{f_\ell\}_{\ell \in \mathbb{N}_0}$ of functions that are conditionally Gaussian:

$$f_0 \sim \mathsf{GP}\big(0, k(x, x')\big),$$
$$f_{\ell+1} | f_\ell \sim \mathsf{GP}\big(0, k(x, x'; f_\ell)\big).$$

- We refer to f_{L-1} as a deep Gaussian process with L layers.
- In practice the number of layers L is often quite small. Some mathematical justification for this is given in [Dunlop et al, '18], which shows that many constructions converge as $L \to \infty$.

Deep Gaussian process regression

Example: Composition

- A particular example is the construction in [Damaniou, Lawrence '13], which involves composition of Gaussian processes:

$$f_{\ell+1}|f_\ell \sim \mathsf{GP}\Big(0, k\big(f_\ell(x), f_\ell(x')\big)\Big),$$

for $f_\ell : \mathbb{R} \to \mathbb{R}$. (Similar structure to deep neural networks).

Deep Gaussian process regression

Example: Composition

- A particular example is the construction in [Damaniou, Lawrence '13], which involves composition of Gaussian processes:

$$f_{\ell+1}|f_\ell \sim \mathsf{GP}\Big(0, k\big(f_\ell(x), f_\ell(x')\big)\Big),$$

  for $f_\ell : \mathbb{R} \to \mathbb{R}$. (Similar structure to deep neural networks).

- The general case $f_\ell : X \subseteq \mathbb{R}^d \to \mathbb{R}$ can be formulated as

$$f_{\ell+1}|f_\ell \sim \mathsf{GP}\left(0, k\Big(G\big(f_\ell(x)\big), G\big(f_\ell(x')\big)\Big)\right),$$

  for suitable $G : \mathbb{R} \to X$.

- This construction is known to display pathologies for large L, see [Duvenaud et al '14].

Deep Gaussian process regression

Example: Covariance kernel

- Another example is based on the non-stationary covariance kernel construction introduced in [Paciorek, Schervish '05].
- Given
  - a stationary covariance kernel $k_{\mathrm{S}}(\|x - x'\|_2)$ and
  - a function $\lambda(x)$ representing a (non-stationary) correlation length,

  we define the non-stationary covariance kernel

$$k_{\mathrm{NS}}(x, x'; \lambda) = \frac{2^{d/2}\lambda(x)^{d/4}\lambda(x')^{d/4}}{(\lambda(x) + \lambda(x'))^{d/2}} \, k_{\mathrm{S}}\left(\frac{\|x - x'\|_2}{\sqrt{(\lambda(x) + \lambda(x'))/2}}\right).$$

Deep Gaussian process regression

Example: Covariance kernel

- Another example is based on the non-stationary covariance kernel construction introduced in [Paciorek, Schervish '05].
- Given
  - a stationary covariance kernel $k_{\mathrm{S}}(\|x - x'\|_2)$ and
  - a function $\lambda(x)$ representing a (non-stationary) correlation length,

  we define the non-stationary covariance kernel

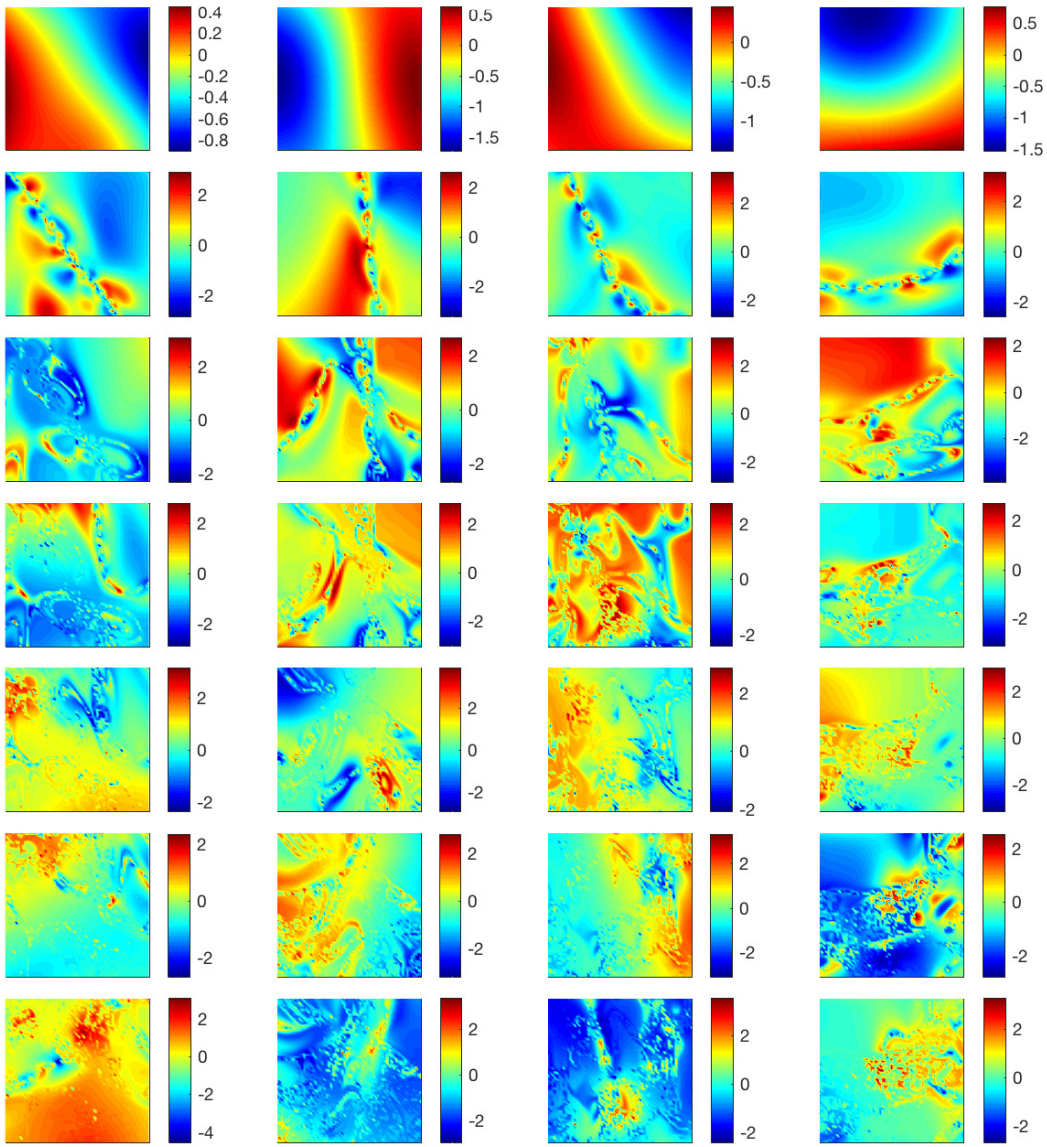
$$k_{\mathrm{NS}}(x, x'; \lambda) = \frac{2^{d/2}\lambda(x)^{d/4}\lambda(x')^{d/4}}{(\lambda(x) + \lambda(x'))^{d/2}}\, k_{\mathrm{S}}\left(\frac{\|x - x'\|_2}{\sqrt{(\lambda(x) + \lambda(x'))/2}}\right).$$

- For a deep Gaussian process, choose $\lambda(x) = F(f_\ell(x))$ for some non-negative function F, e.g. $\lambda(x) = f_\ell(x)^2$ or $\lambda(x) = \exp(f_\ell(x))$:

$$f_{\ell+1}|f_\ell \sim \mathsf{GP}\Big(0, k_{\mathrm{NS}}(x, x'; F(f_\ell))\Big).$$

Deep Gaussian process regression

Numerical examples: Covariance kernel



Deep Gaussian process regression

Algorithmic considerations

- We now want to use deep Gaussian processes as prior distribution.

Deep Gaussian process regression

Algorithmic considerations

- We now want to use deep Gaussian processes as prior distribution.
- The iteration

$$f_{\ell+1}|f_\ell \sim \mathsf{GP}\big(0, k(x, x'; f_\ell)\big)$$

can be written in the form

$$f_{\ell+1} = R(f_\ell)\xi_{\ell+1},$$

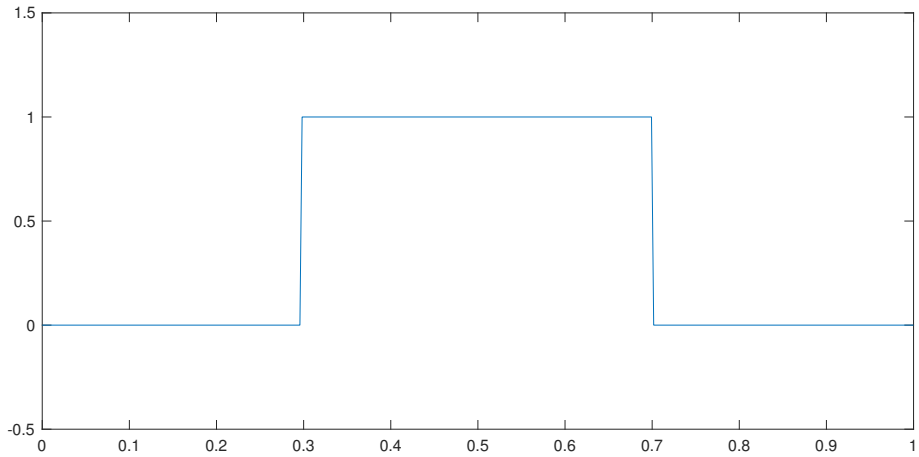
where $\{\xi_\ell\}$ form an i.i.d. Gaussian sequence and, for each f, $R(x)$ is a linear operator. (Think $X = RZ$ for $X \sim N(0, C), C = RR^T$.)

- To sample from the posterior distribution $f_L|y$, we use a Metropolis-Hastings algorithm on $\xi = [\xi_0, \ldots, \xi_L]$ that is well-defined in function space.

Deep Gaussian process regression

Numerical Example: Regression in 1d

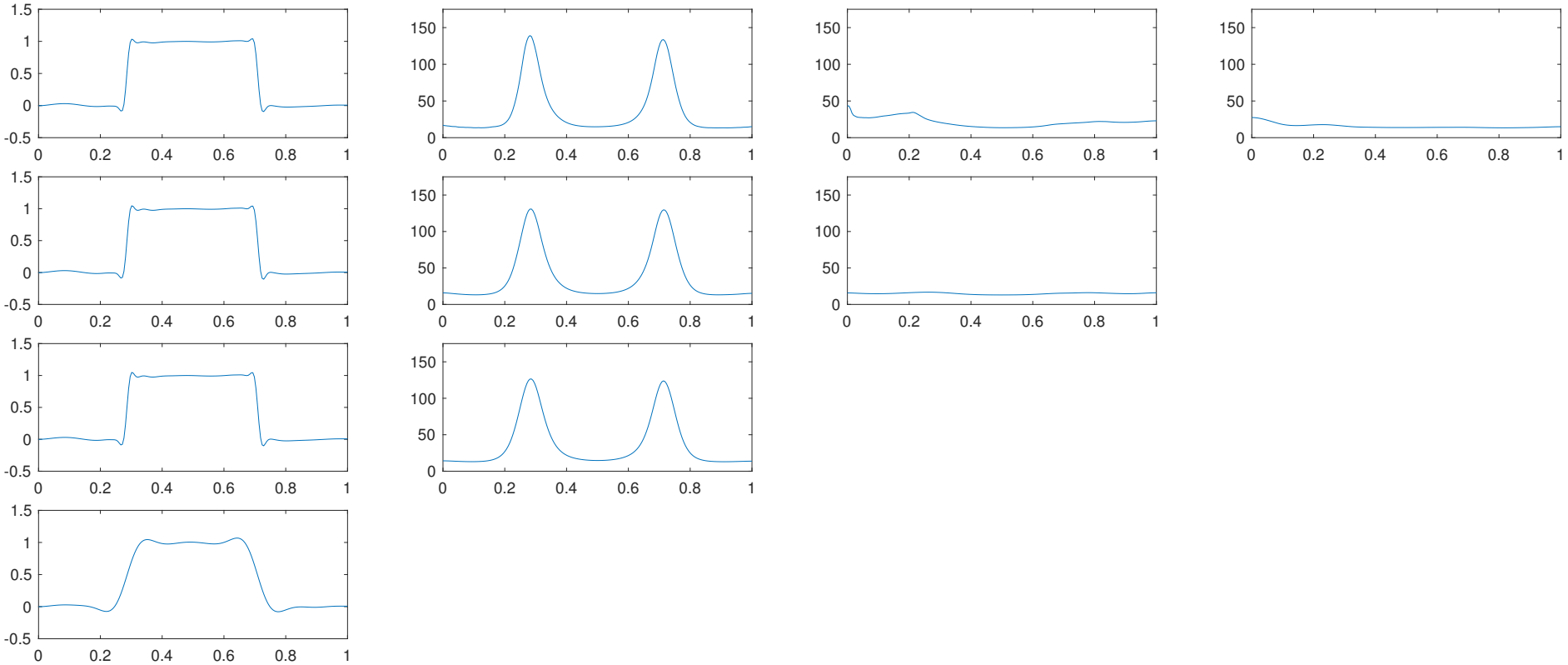
We now want to use deep Gaussian processes as prior distribution.



Given data: noisy point evaluations at $J = 50$ equispaced points.

Deep Gaussian process regression

Numerical Example: Regression in 1d

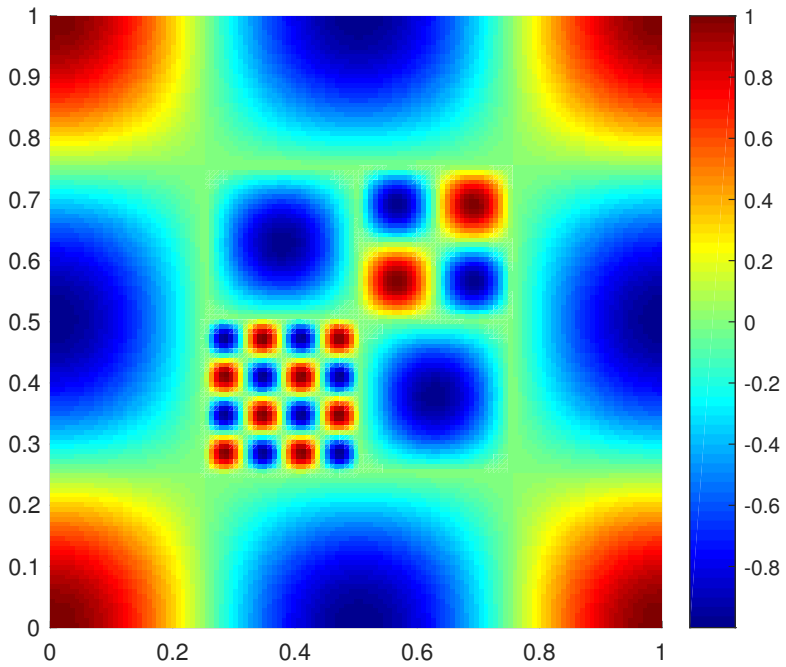


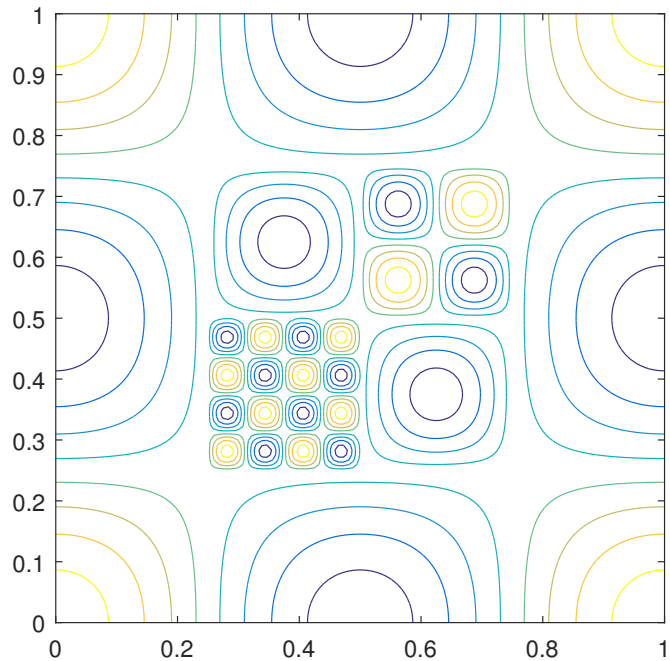
Posterior means computed using MCMC.

Rows correspond to f_3, f_2, f_1 and f_0 respectively, from top to bottom.

Deep Gaussian process regression

Numerical Example: Regression in 2d

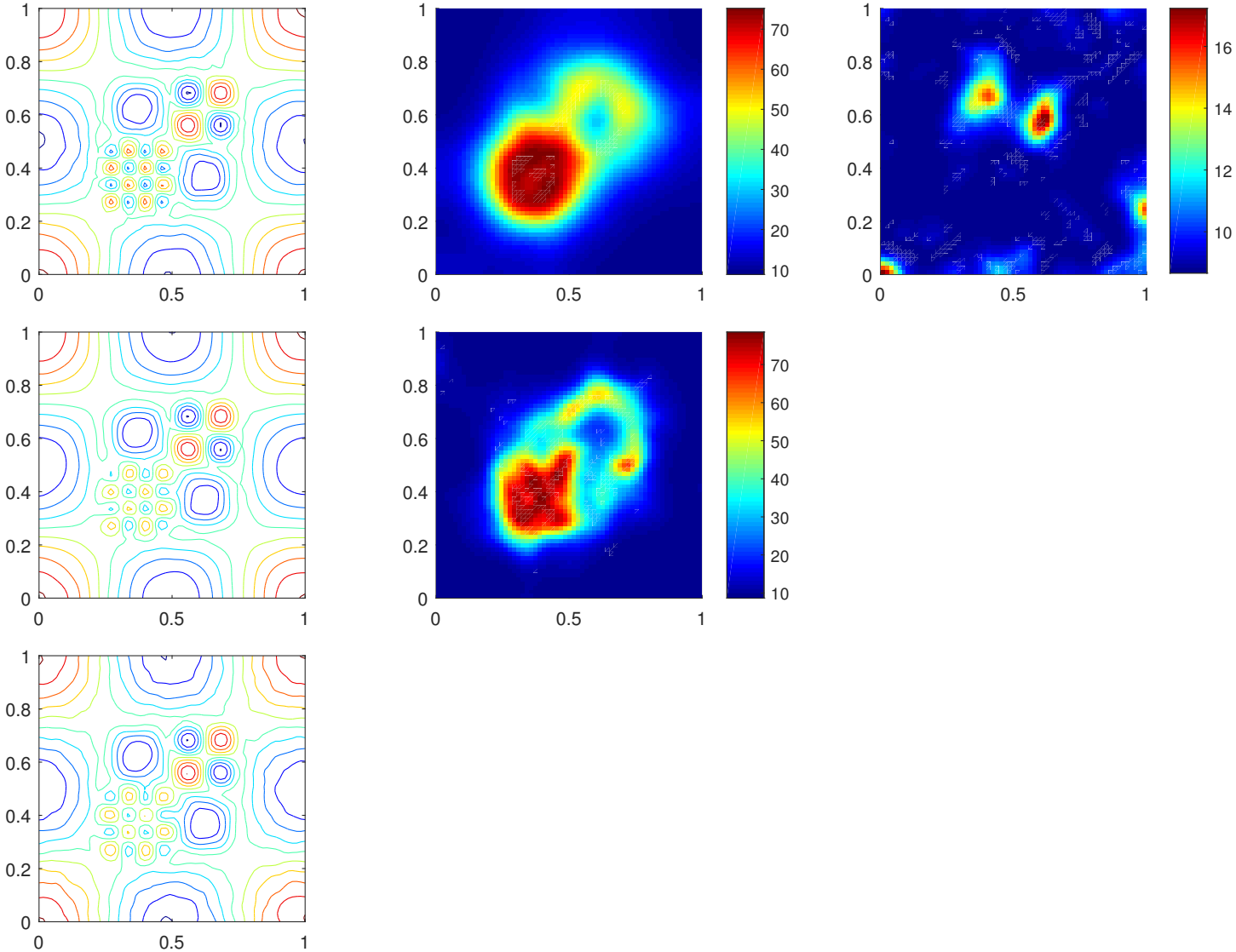




Given data: noisy point evaluations at $J = 1024$ equispaced points.

Deep Gaussian process regression

Numerical Example: Regression in 2d



Deep Gaussian process regression

Numerical Example: Regression in 2d

Table: L^2-errors $\|f^\dagger - \mathbb{E}(f_L|y)\|_{L^2}$ between the true field and sample means.

J	1 layer	2 layers	3 layers
1024	0.0856	0.0813	0.0681
256	0.1310	0.1260	0.1279

Conclusions

- We presented a general framework for constructing deep Gaussian processes, which includes known constructions as particular examples.
- Deep Gaussian processes can be used as priors in Bayesian inference tasks including regression and classification. They offer very flexible priors, which can easily model complex functions.
- Compared to other non-stationary Gaussian processes, deep GPs require less prior information and learn everything as part of the inference.

References I


https://towardsdatascience.com/an-intuitive-guide-to-gaussian-processes-ec2f0b45c71d.


T. Choi and M. J. Schervish, *On posterior consistency in nonparametric regression problems*, Journal of Multivariate Analysis, 98 (2007), pp. 1969–1987.


P. Diamond and M. Armstrong, *Robustness of variograms and conditioning of kriging matrices*, Journal of the International Association for Mathematical Geology, 16 (1984), pp. 809–822.


F. Nobile, R. Tempone, and S. Wolfers, *Sparse approximation of multilinear problems with applications to kernel-based methods in UQ*, Numerische Mathematik, (2017), pp. 1–34.


H. Putter and G. A. Young, *On the effect of covariance function estimation on the accuracy of kriging predictors*, Bernoulli, (2001), pp. 421–438.


C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*, (2006).

References II


M. Scheuerer, R. Schaback, and M. Schlather, *Interpolation of spatial data–A stochastic or a deterministic problem?*, European Journal of Applied Mathematics, 24 (2013), pp. 601–629.


M. L. Stein, *Asymptotically efficient prediction of a random field with a misspecified covariance function*, The Annals of Statistics, (1988), pp. 55–63.


———, *A simple condition for asymptotic optimality of linear predictions of random fields*, Statistics & Probability Letters, 17 (1993), pp. 399–404.


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


A. L. Teckentrup, *Convergence of Gaussian process regression with estimated hyper-parameters and applications in Bayesian inverse problems*, SIAM/ASA Journal on Uncertainty Quantification, 8 (2020), pp. 1310–1337.


A. Van Der Vaart and H. Van Zanten, *Information rates of nonparametric gaussian process methods.*, Journal of Machine Learning Research, 12 (2011).

References III

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

G. Wynne, F.-X. Briol, and M. Girolami, *Convergence guarantees for Gaussian process means with misspecified likelihoods and smoothness*, Journal of Machine Learning Research, 22 (2021), pp. 1–40.