# Deep Gaussian processes for PDE inverse problems

Kweku Abraham
Joint work in progress with Neil Deo

University of Cambridge

# Outline

1. Why use Bayesian methods for inverse problems?
2. Why use Gaussian process priors?
3. Why use deep Gaussian process priors?
4. Work still to be done

# Darcy's problem

**Background physics**

$$\begin{aligned} \nabla \cdot (f \nabla u) &= g \quad \text{in } D \\ u &= 0 \quad \text{on } \partial D \end{aligned}$$

**Target** The diffusivity/conductivity $f$

**Observations** $(X_i, Y_i)$, $i \leq n$, with

$$\begin{aligned} &Y_i = u(X_i) + \sigma \xi_i, \\ &X_i \overset{iid}{\sim} \mathrm{Unif}(D), \quad \xi_i \overset{iid}{\sim} N(0,1). \end{aligned}$$

The "source" $g \in C^\infty(D)$ and the "noise level" $\sigma$ are assumed known.

# The Bayesian approach to statistical inverse problems

We can recast the problem of estimating $f$ in a more general way:

**Forward map** $\mathcal{G} : f \mapsto \mathcal{G}(f) = u$ the solution to the PDE.

**Aim** Invert $\mathcal{G}$ in a way robust to noise: find an estimator $\hat{f}$ based on $n$ noisy observations of $\mathcal{G}(f)$ which gets close to $f$ in some norm as $n \to \infty$.

# The Bayesian approach to statistical inverse problems

We can recast the problem of estimating $f$ in a more general way:

**Forward map** $\mathcal{G} : f \mapsto \mathcal{G}(f) = u$ the solution to the PDE.

**Aim** Invert $\mathcal{G}$ in a way robust to noise: find an estimator $\hat{f}$ based on $n$ noisy observations of $\mathcal{G}(f)$ which gets close to $f$ in some norm as $n \to \infty$.

If we place a prior on $f$, we derive a posterior via Bayes' rule

$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

By sampling from the posterior, we posit a solution to the inverse problem.

# The Bayesian approach to statistical inverse problems

We can recast the problem of estimating $f$ in a more general way:

**Forward map** $\mathcal{G} : f \mapsto \mathcal{G}(f) = u$ the solution to the PDE.

**Aim** Invert $\mathcal{G}$ in a way robust to noise: find an estimator $\hat{f}$ based on $n$ noisy observations of $\mathcal{G}(f)$ which gets close to $f$ in some norm as $n \to \infty$.

If we place a prior on $f$, we derive a posterior via Bayes' rule

$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

By sampling from the posterior, we posit a solution to the inverse problem. *This proposed solution only requires calls to the forward operator, not its inverse, and so is computationally feasible.*

# GP priors are computationally feasible

The posterior associated with a Gaussian process prior can generically be computed via Markov Chain Monte Carlo methods, for example by Metropolis–Hastings using a preconditioned Crank–Nicholson (pCN) proposal.

**The pCN algorithm**

Pick $\theta^{(0)}$ and $\beta \in (0,1)$, then for $i \leq k$:

- Propose $\phi^{(i)} = \sqrt{1-\beta^2}\theta^{(i)} + \beta\xi^{(i)}$, with $\xi^{(i)}$ drawn from the prior
- Set $\theta^{(i)} = \phi^{(i)}$ with probability $\min\{1, \exp(\ell_N(\phi^{(i)}) - \ell_n(\theta^{(i)}))\}$, set $\theta^{(i)} = \theta^{(i-1)}$ otherwise.

Output $\theta^{(0)}, \dots, \theta^{(k)}$.

# Theoretical guarantees for the posterior come from continuity results

Key to obtaining guarantees for the posterior are continuity results for $\mathcal{G}$ and $\mathcal{G}^{-1}$. Continuity properties are well understood in Darcy's problem:

**Forward continuity**

$$\|u_{f_1} - u_{f_2}\|_{L^2(D)} \leq C\|f_1 - f_2\|_{(H^1(D))^*} \leq C\|f_1 - f_2\|_{L^\infty}.$$

**Inverse continuity (stability)** $\|f_1 - f_2\|_{L^2(D)} \leq C\|u_{f_1} - u_{f_2}\|_{L^2(D)}^{(\beta-1)/(\beta+1)}$ if $f_1, f_2 \in H^\beta(D)$, $\beta > d/2 + 1$ (and mild extra conditions ensuring uniqueness).

# Theoretical guarantees for the posterior come from continuity results

Key to obtaining guarantees for the posterior are continuity results for $\mathcal{G}$ and $\mathcal{G}^{-1}$. Continuity properties are well understood in Darcy's problem:

**Forward continuity**

$$\|u_{f_1} - u_{f_2}\|_{L^2(D)} \leq C\|f_1 - f_2\|_{(H^1(D))^*} \leq C\|f_1 - f_2\|_{L^\infty}.$$

**Inverse continuity (stability)** $\|f_1 - f_2\|_{L^2(D)} \leq C\|u_{f_1} - u_{f_2}\|_{L^2(D)}^{(\beta-1)/(\beta+1)}$ if $f_1, f_2 \in H^\beta(D)$, $\beta > d/2 + 1$ (and mild extra conditions ensuring uniqueness).

Consequently, Bayesian methods can be shown to work well.

**Theorem (Giordano + Nickl 2020)**

*Let $f$ be in $H^\alpha(D)$ and choose $1 \leq \beta < \alpha - d/2$. For a suitable scaled Gaussian process prior on $f$, the posterior mean $\hat{f}$ satisfies*

$$\|\hat{f} - f\|_{L^2(D)} \leq Cn^{-\lambda} \quad \textit{with probabality tending to 1,}$$

$$\lambda = \frac{(\alpha+1)(\beta-1)}{(2\alpha+2+d)(\beta+1)}.$$

# Whittle–Matérn processes model additive functions poorly

Giordano + Nickl consider priors of the form

$$f = \Phi \circ \theta, \quad \theta = n^{-d/(4\alpha+4+2d)}\theta',$$

where $\theta'$ is a Whittle–Matérn process with reproducing kernel Hilbert space $H^\alpha(D)$, with $\alpha$ chosen to match the smoothness of the true diffusivity, and where $\Phi$ is a 'link function' $\Phi : \mathbb{R} \to (m, \infty)$ for some $m > 0$, say $\Phi(x) = m + e^x$.

Suppose the true diffusivity $f_0$ is of the form

$$f_0(x_1, \dots, x_d) = h(x_1 + \cdots + x_d), \quad h \in C^\alpha(\mathbb{R}).$$

Then Giordano + Nickl achieve the rate $n^{-\lambda}$, $\lambda = \frac{\alpha+1}{2\alpha+2+d}\frac{\beta-1}{\beta+1}$.

# Whittle–Matérn processes model additive functions poorly

Giordano + Nickl consider priors of the form

$$f = \Phi \circ \theta, \quad \theta = n^{-d/(4\alpha+4+2d)}\theta',$$

where $\theta'$ is a Whittle–Matérn process with reproducing kernel Hilbert space $H^\alpha(D)$, with $\alpha$ chosen to match the smoothness of the true diffusivity, and where $\Phi$ is a 'link function' $\Phi : \mathbb{R} \to (m, \infty)$ for some $m > 0$, say $\Phi(x) = m + e^x$.

Suppose the true diffusivity $f_0$ is of the form

$$f_0(x_1, \dots, x_d) = h(x_1 + \cdots + x_d), \quad h \in C^\alpha(\mathbb{R}).$$

Then Giordano + Nickl achieve the rate $n^{-\lambda}$, $\lambda = \frac{\alpha+1}{2\alpha+2+d}\frac{\beta-1}{\beta+1}$.

Because $h$ is univariate it should be possible to replace $d$ by 1 (e.g. Schmidt-Hieber 2020).

**Proposition**

*No Gaussian process prior with RKHS equal to $H^\gamma(D)$ for some $\gamma$ is able to achieve a rate $n^{-\lambda}$ with $\lambda = \frac{\alpha+1}{2\alpha+3}$.*

# Modelling the compositional structure can improve the rate

Note that $f(x) = h(x_1 + \cdots + x_d)$ is of the form $\zeta_2 \circ \zeta_1$ with $\zeta_1(x) = x_1 + \cdots + x_d \in C^\infty(D)$ and $\zeta_2 = h \in C^\alpha(\mathbb{R})$.

## Proposition

*Suppose the true $f_0 \in H^\alpha(D)$ can be written as $f_0 = \zeta_2 \circ \zeta_1$ with $\zeta_1 \in H^{\alpha_1}(D)$, $\zeta_2 \in H^{\alpha_2}(\mathbb{R})$. Consider a deep Gaussian process prior*

$$f = \Phi \circ Z_2 \circ Z_1,$$
$$Z_i = N^{-\gamma_1} Z_i',$$

*where $Z_2', Z_1'$ are Whittle-Matérn processes, with $Z_2$ having RKHS $H^{\alpha_2}(\mathbb{R})$ and $Z_1$ having RKHS $H^{\alpha_1}(D)$ and where $\gamma_i = d/(4\alpha_i + 2d)$, $\gamma_2 = 1/(4\alpha_2 + 2)$. Then $\|\hat{f} - f\|_{L^\infty} \leq Cn^{-\lambda}$ with probability tending to 1 for a constant $C$, where $\lambda = \frac{\beta-1}{\beta+1} \max\left(\frac{\alpha_1}{2\alpha_1+d}, \frac{\alpha_2}{2\alpha_2+1}\right)$.*

Compare to $\lambda = \frac{\beta-1}{\beta+1} \frac{\alpha+1}{2\alpha+2+d}$ obtainable with a single GP prior.

# Deep GPs arise as limit of Bayesian neural networks

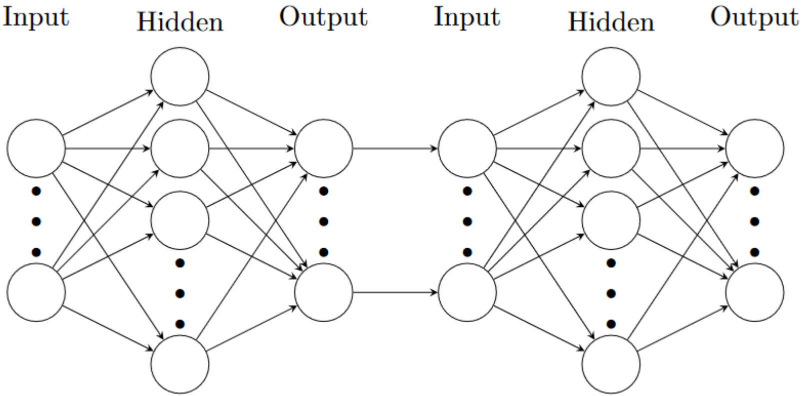


Figure: Figure 3 from Finocchio + Schmidt-Hieber 2021: schematic stacking of two shallow neural networks.

# To do...

- Adaptivity!
- Improve the rate?