

Gaussian process emulators in Bayesian inverse problems

Aretha Teckentrup

School of Mathematics, University of Edinburgh
Alan Turing Institute, London

LMS-EPSRC Durham symposium on *Model Order Reduction*



THE UNIVERSITY of EDINBURGH
School of Mathematics

**The
Alan Turing
Institute**

Outline

- 1 Bayesian Inverse Problems
- 2 Gaussian Process Regression
- 3 Approximations of the Posterior
- 4 Conclusions

Bayesian Inverse Problems

Mathematical Formulation [Stuart '10] [Kaipio, Somersalo '04]

We are given

- a model \mathcal{F} of a physical process depending on parameters $u \in U \subset \mathbb{R}^{d_u}$ for some $d_u \in \mathbb{N}$ and compact U ,
 - evaluation of \mathcal{F} typically involves the solution of a PDE
- observations/data $y = \mathcal{O}(\mathcal{F}(u)) + \eta$, with $y \in \mathbb{R}^{d_y}$ and η a realisation of a $\mathcal{N}(0, \sigma_\eta^2 \mathbf{I})$ random variable

We are interested in the inverse problem of finding u given y .

Bayesian Inverse Problems

Mathematical Formulation [Stuart '10] [Kaipio, Somersalo '04]

We are given

- a model \mathcal{F} of a physical process depending on parameters $u \in U \subset \mathbb{R}^{d_u}$ for some $d_u \in \mathbb{N}$ and compact U ,
 - evaluation of \mathcal{F} typically involves the solution of a PDE
- observations/data $y = \mathcal{O}(\mathcal{F}(u)) + \eta$, with $y \in \mathbb{R}^{d_y}$ and η a realisation of a $\mathcal{N}(0, \sigma_\eta^2 \mathbf{I})$ random variable

We are interested in the inverse problem of finding u given y .

Following the Bayesian approach, we

- assign a prior distribution μ_0 to u ;
- determine the data likelihood $\mathcal{P}(y|u) \approx \exp\left(-\frac{1}{2\sigma_\eta^2} \|y - \mathcal{O}(\mathcal{F}(u))\|_2^2\right)$;
- want to determine the posterior distribution μ^y on $u|y$.

Bayesian Inverse Problems

Computational Challenges

- Using Bayes' Theorem, the posterior distribution μ^y is given by

$$\frac{d\mu^y}{d\mu_0}(u) = \frac{1}{Z} \exp(-\Phi(u)), \quad \left(\pi^y(u) = \frac{1}{Z} \exp(-\Phi(u)) \pi_0(u) \right)$$

where $\Phi(u) = \frac{1}{2\sigma_\eta^2} \|y - \mathcal{O}(\mathcal{F}(u))\|_2^2$ and $Z = \mathbb{E}_{\mu_0} \left(\exp(-\Phi(u)) \right)$.

Bayesian Inverse Problems

Computational Challenges

- Using Bayes' Theorem, the posterior distribution μ^y is given by

$$\frac{d\mu^y}{d\mu_0}(u) = \frac{1}{Z} \exp(-\Phi(u)), \quad \left(\pi^y(u) = \frac{1}{Z} \exp(-\Phi(u)) \pi_0(u) \right)$$

where $\Phi(u) = \frac{1}{2\sigma_\eta^2} \|y - \mathcal{O}(\mathcal{F}(u))\|_2^2$ and $Z = \mathbb{E}_{\mu_0}(\exp(-\Phi(u)))$.

- Sampling methods such as Markov chain Monte Carlo require repeated evaluation of the data likelihood $\exp(-\Phi(u))$, easily in the order of millions of evaluations.
- Since the computation of Φ involves evaluating \mathcal{F} , this is typically very costly. We approximate Φ by a surrogate model (emulator, reduced order model, ...).
- We will use Gaussian process emulators, but other choices are possible.

Gaussian Process Regression

Simple Derivation [Rasmussen, Williams '06]

- We treat Φ as unknown, and assign a probability distribution to it: we model Φ as a Gaussian process, with mean 0 and (positive definite) covariance kernel $k : U \times U \rightarrow \mathbb{R}$:

$$\Phi_0 \sim \text{GP}(0, k(u, u'))$$

For every set $\{u_i\}_{i=1}^m \subseteq U$, the random variables $\{\Phi_0(u_i)\}_{i=1}^m$ are multivariate Gaussian with $\mathbb{E}(\Phi(u_i)) = 0$ and $\mathbb{E}(\Phi_0(u_i)\Phi_0(u_j)) = k(u_i, u_j)$. The kernel k incorporates information such as smoothness and typical length scales.

Gaussian Process Regression

Simple Derivation [Rasmussen, Williams '06]

- We treat Φ as unknown, and assign a probability distribution to it: we model Φ as a Gaussian process, with mean 0 and (positive definite) covariance kernel $k : U \times U \rightarrow \mathbb{R}$:

$$\Phi_0 \sim \text{GP}(0, k(u, u'))$$

For every set $\{u_i\}_{i=1}^m \subseteq U$, the random variables $\{\Phi_0(u_i)\}_{i=1}^m$ are multivariate Gaussian with $\mathbb{E}(\Phi(u_i)) = 0$ and $\mathbb{E}(\Phi_0(u_i)\Phi_0(u_j)) = k(u_i, u_j)$. The kernel k incorporates information such as smoothness and typical length scales.

- We evaluate Φ at design points $D = \{u^n\}_{n=1}^N \subseteq U$, obtaining function values $\{\Phi(u^n)\}_{n=1}^N$.

Gaussian Process Regression

Simple Derivation [Rasmussen, Williams '06]

- We treat Φ as unknown, and assign a probability distribution to it: we model Φ as a Gaussian process, with mean 0 and (positive definite) covariance kernel $k : U \times U \rightarrow \mathbb{R}$:

$$\Phi_0 \sim \text{GP}(0, k(u, u'))$$

For every set $\{u_i\}_{i=1}^m \subseteq U$, the random variables $\{\Phi_0(u_i)\}_{i=1}^m$ are multivariate Gaussian with $\mathbb{E}(\Phi(u_i)) = 0$ and $\mathbb{E}(\Phi_0(u_i)\Phi_0(u_j)) = k(u_i, u_j)$. The kernel k incorporates information such as smoothness and typical length scales.

- We evaluate Φ at design points $D = \{u^n\}_{n=1}^N \subseteq U$, obtaining function values $\{\Phi(u^n)\}_{n=1}^N$.
- Conditioning Φ_0 on given function values $\{\Phi(u^n)\}_{n=1}^N$ leads to $\Phi_N \sim \text{GP}(m_N^\Phi(u), k_N(u, u'))$, with

$$m_N^\Phi(u) = k_*(u)^T K_*^{-1} \Phi_*, \quad k_N(u, u') = k(u, u') - k_*(u)^T K_*^{-1} k_*(u'),$$

$$\text{and } (k_*(u))_n = k(u, u^n), (K_*)_{nm} = k(u^n, u^m) \text{ and } (\Phi_*)_n = \Phi(u^n).$$

Gaussian Process Regression

Relation to Kernel Interpolation

- The predictive mean is a linear combination of kernel evaluations

$$m_N^{\Phi}(u) = \sum_{n=1}^N \alpha_n k(u, u^n), \quad \alpha = K_*^{-1} \Phi_*.$$

- We have $m_N^{\Phi}(u^n) = \Phi(u^n)$ and $k_N(u^n, u^n) = 0$, for $n = 1, \dots, N$.
 $\Rightarrow \Phi_N(u^n) \equiv m_N^{\Phi}(u^n) = \Phi(u^n)$, for $n = 1, \dots, N$.

Gaussian Process Regression

Relation to Kernel Interpolation

- The predictive mean is a linear combination of kernel evaluations

$$m_N^\Phi(u) = \sum_{n=1}^N \alpha_n k(u, u^n), \quad \alpha = K_*^{-1} \Phi_*.$$

- We have $m_N^\Phi(u^n) = \Phi(u^n)$ and $k_N(u^n, u^n) = 0$, for $n = 1, \dots, N$.
 $\Rightarrow \Phi_N(u^n) \equiv m_N^\Phi(u^n) = \Phi(u^n)$, for $n = 1, \dots, N$.
- The predictive mean m_N^Φ is a **kernel interpolant** of Φ , and in the special case of isotropic kernels $k(u, u') = k(\|u - u'\|)$, a **radial basis function interpolant**.
- The emulator Φ_N is a **random interpolant** of Φ , reflecting the uncertainty in Φ away from the design points D .

Gaussian Process Regression

Matèrn Kernels

- Examples of kernels frequently used are the family of Matèrn covariances

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

with smoothness parameter $\nu > 0$, marginal variance σ^2 , correlation length λ , Γ the gamma function and B_ν the modified Bessel function of the second kind.

$$\begin{aligned} \nu = 1/2 & : k_{\nu, \lambda, \sigma^2}(u, u') = \sigma^2 \exp \left(- \frac{\|u - u'\|}{\lambda} \right), \\ \nu = \infty & : k_{\nu, \lambda, \sigma^2}(u, u') = \exp \left(- \frac{\|u - u'\|^2}{\lambda^2} \right). \end{aligned}$$

Gaussian Process Emulators

Scattered Data Approximation [Wendland '04]

With design points $D = \{u^n\}_{n=1}^N$, define the fill distance

$$h_D = \sup_{u \in U} \inf_{u^n \in D} \|u - u^n\|.$$

Theorem (see e.g. [Scheuerer, Schaback, Schlather '13], [Stuart, ALT '17])

Suppose U satisfies an interior cone condition. With covariance kernel $k_{\nu, \lambda, \sigma^2}$, we have for h_D sufficiently small

$$\|\Phi - m_N^\Phi\|_{L^2(U)} \leq C h_D^{\nu+d_u/2} \|\Phi\|_{H^{\nu+d_u/2}(U)},$$

with C independent of D and Φ . Furthermore,

$$\|k_N^{\frac{1}{2}}\|_{L^2(U)} \leq C h_D^\nu.$$

Approximations of the Posterior

- Recall:

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

- We now use the emulator Φ_N to build different approximations μ_N^y to the posterior distribution μ^y .
- We focus on bounding the Hellinger distance

$$d_{\text{hell}}(\mu^y, \mu_N^y) = \left(\frac{1}{2} \int_U \left(\sqrt{\frac{d\mu^y}{d\mu_0}} - \sqrt{\frac{d\mu_N^y}{d\mu_0}} \right)^2 d\mu_0 \right)^{1/2}.$$

Approximations of the Posterior

Using the mean m_N^Φ , we define the mean based approximation

$$\frac{d\mu_{\text{mean}}^{y,N}}{d\mu_0}(u) = \frac{1}{Z_N^{\text{mean}}} \exp(-m_N^\Phi(u)).$$

Approximations of the Posterior

Using the mean m_N^Φ , we define the mean based approximation

$$\frac{d\mu_{\text{mean}}^{y,N}}{d\mu_0}(u) = \frac{1}{Z_N^{\text{mean}}} \exp(-m_N^\Phi(u)).$$

Lemma [Stuart, ALT '17]

There exist a positive constants C_1, C_2 , independent of N , such that

$$C_1 \leq Z_N^{\text{mean}} \leq C_2.$$

Proof: Uses convergence of m_N^Φ to Φ .

Theorem [Stuart, ALT '17]

There exists a constant C , independent of N , such that

$$d_{\text{hell}}(\mu^y, \mu_{\text{mean}}^{y,N}) \leq C \|\Phi - m_N^\Phi\|_{L^2(U)}.$$

Proof: Uses Lemma and Lipschitz continuity of likelihood.

Approximations of the Posterior

Using the process Φ_N , we define the the **random approximation**

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

and the **marginal approximation**

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

- The emulator Φ_N includes the **uncertainty** in $\Phi(u)$ for $u \notin D$.
- $\mathbb{E}\left(\exp(-\Phi_N(u, \cdot))\right)$ is the **optimal approximation** of $\exp(-\Phi(u))$ in an L^2 -sense.

Approximations of the Posterior

Lemma [Stuart, ALT '17]

There exist positive constants C_1, C_2, C_3 , independent of N , s.t.

$$C_1 \leq \mathbb{E}((Z_N^{\text{sample}})^p) \leq C_2, \quad \text{and} \quad C_2 \leq \mathbb{E}((Z_N^{\text{sample}})^{-p}) \leq C_3,$$

for all $1 \leq p < \infty$ and N sufficiently large.

Proof: Uses convergence of m_N^Φ and k_N , Fernique's Theorem, Borell-TIS inequality and Sudakov-Fernique inequality.

Approximations of the Posterior

Theorem [Stuart, ALT '17]

There exists a constant C , independent of N , such that

$$d_{\text{hell}}(\mu^y, \mu_{\text{marginal}}^{y,N}) \leq C \left\| \mathbb{E} \left(|\Phi - \Phi_N|^{1+\delta} \right)^{\frac{1}{1+\delta}} \right\|_{L^2(U)},$$

for any $\delta > 0$.

Proof: Uses Lemma, Lipschitz cont'y of likelihood, Fernique's Theorem.

Approximations of the Posterior

Theorem [Stuart, ALT '17]

There exists a constant C , independent of N , such that

$$d_{\text{hell}}(\mu^y, \mu_{\text{marginal}}^{y,N}) \leq C \left\| \mathbb{E} \left(|\Phi - \Phi_N|^{1+\delta} \right)^{\frac{1}{1+\delta}} \right\|_{L^2(U)},$$

for any $\delta > 0$.

Proof: Uses Lemma, Lipschitz cont'y of likelihood, Fernique's Theorem.

Theorem [Stuart, ALT '17]

There exists a constant C , independent of N , such that

$$\mathbb{E} \left(d_{\text{hell}}(\mu^y, \mu_{\text{sample}}^{y,N})^2 \right)^{1/2} \leq C \left\| \left(\mathbb{E} \left(|\Phi - \Phi_N|^{2+\delta} \right) \right)^{\frac{1}{2+\delta}} \right\|_{L^2(U)}.$$

for any $\delta > 0$.

Proof: Uses Lemma, Lipschitz cont'y of likelihood, Fernique's Theorem.

Approximations of the Posterior

Extensions

- Combining the two types of error estimates, we get convergence rates in terms of h_D for $d_{\text{hell}}(\mu^y, \mu_N^y)$.
- Instead of emulating Φ , we can also emulate $\mathcal{O}(\mathcal{F})$, with similar error bounds.
- Numerical computations confirm the rates proved.

Approximations of the Posterior

Extensions

- Combining the two types of error estimates, we get convergence rates in terms of h_D for $d_{\text{hell}}(\mu^y, \mu_N^y)$.
- Instead of emulating Φ , we can also emulate $\mathcal{O}(\mathcal{F})$, with similar error bounds.
- Numerical computations confirm the rates proved.
- We are currently investigating the influence of the choice of hyper-parameters ν, λ, σ^2 , and how to choose these optimally (joint with A Stuart).
- We are devising a general framework for random approximations of Bayesian posterior distributions (joint with H Lie and T Sullivan).

Approximations of the Posterior

Dimension reduction in U

- The error estimates in terms of h_D yield strong dependence on the dimension of $U \subseteq \mathbb{R}^{d_u}$.
- For a uniform tensor grid with N points in d_u dimensions, we have







$$h_D = \sqrt{d_u} (N^{1/d_u} - 1)^{-1}.$$

- Incorporating dimensionality reduction of U in the definition of the Gaussian process emulator should alleviate this?
- Covariance kernels are frequently defined in terms of $\|u - u'\|_2$ - use distance preserving methods?

Conclusions

- Gaussian process emulators can be used in inverse problems to approximate the mathematical model.
- The error between the true and approximate posterior can be bounded by moments of the GP emulation error.
- Our theory does not make any assumptions on the GP emulator other than convergence as $N \rightarrow \infty$.
- The only assumptions on the mathematical model are in terms of its (Sobolev) smoothness.

References

-  J. KAIPIO AND E. SOMERSALO, *Statistical and computational inverse problems*, Springer, 2004.
-  C. E. RASMUSSEN AND C. K. WILLIAMS, *Gaussian processes for machine learning*, (2006).
-  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.
-  A. M. STUART, *Inverse Problems: A Bayesian Perspective*, Acta Numerica, 19 (2010), pp. 451–559.
-  A. M. STUART AND A. L. TECKENTRUP, *Posterior Consistency for Gaussian Process Approximations of Bayesian Posterior Distributions*. ArXiv preprint 1603.02004, to appear in *Mathematics of Computation*, 2017.
-  H. WENDLAND, *Scattered Data Approximation*, vol. 17, Cambridge University Press, 2004.