Probabilistic Numerics for Differential Equations

Tim Sullivan[1] and **Jon Cockayne**[2]
with M. Girolami[2], H. C. Lie[1], C. J. Oates[3], and A. M. Stuart[4]

[1]Free University of Berlin / Zuse Institute Berlin, Germany
[2]The University of Warwick, UK
[3]University of Technology Sydney, Australia
[4]California Institute of Technology, USA

Scaling Cascades in Complex Systems Days
Ketzin, DE
10–12 October 2016

Overview

Overview

What is "Probabilistic Numerics"?

- Beginning with a seminal papers[1] of Kadane (1985), Diaconis (1988), O'Hagan (1992), and Skilling (1992) there has been interest in giving probabilistic answers to ostensibly deterministic problems, e.g. quadrature, optimisation, solution of differential equations.
- In some sense, this is a Bayesian statistician's natural approach to numerical analysis, phrasing computational tasks as inference problems using finite/incomplete/imperfect information.
- Disadvantage: costs more, and gives 'fuzzier' answers, so why bother?
- Advantage: fold uncertainty arising from numerical error into inferences, and propagate this uncertainty through later computations.
  - Replicability of results $\neq$ accuracy.
  - Good data + bad/overconfident model $\implies$ faulty inferences.
  - In many practical examples from physical, social, and data sciences, we *know* that our numerical solutions are coarse approximations of models that are themselves approximate.

[1] But actually going all the way back to Poincaré (1896).

Overview

Overview

Setting

- We consider the following autonomous ODE for $u\colon [0, T] \to \mathcal{S}$, $\mathcal{S} = \mathbb{R}^n$ or a separable Hilbert space, with vector field $f\colon \mathcal{S} \to \mathcal{S}$:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} u(t) &= f(u(t)), && \text{for } t \in \mathcal{T}, \\ u(0) &= u_0, \end{aligned} \tag{1}$$

  where the initial state $u_0 \in \mathcal{S}$ is given.
- Write $\Phi^t\colon \mathcal{S} \to \mathcal{S}$ for the flow: $u(t) = \Phi^t(u_0)$.
- Aim: Build a probabilistic numerical approximation to solutions of (1), with convergence guarantees.
- Motivation: The ODE solver is often a forward model (likelihood) in a Bayesian inverse problem, so a statistical description of the discretisation error is essential for valid inferences.
- Ours will be an ensemble-based approach, cf. the global Gaussian process approach of Schober et al. (2014).

We need assumptions on (1) the underlying exact flow, (2) the deterministic numerical method, and (3) the random perturbation.

Assumption 1 (re: exact flow)

Suppose that f is smooth enough that, for $|t|$ small enough, its flow map Φ^t is globally Lipschitz with Lipschitz constant $1 + L|t|$:

$$\|\Phi^t(u) - \Phi^t(v)\| \leq (1 + L|t|)\|u - v\|.$$

Assumption 1 holds if, e.g., f is one-sided Lipschitz:

$$\langle f(u) - f(v), u - v \rangle \leq \mu \|u - v\|^2 \tag{2}$$

for all $u, v \in \mathcal{S}$, for some constant $\mu \in \mathbb{R}$; in this case, Φ^t has Lipschitz constant $\exp(\mu|t|)$, which for small enough $|t|$ is dominated by $1 + 2|\mu||t|$.

PN for ODEs

- Let $\Psi^\tau : \mathcal{S} \to \mathcal{S}$ be a one-step numerical integrator for the ODE (1) with time step $\tau > 0$; set $t_k := k\tau$ and $u_k := u(t_k)$.
- This class includes all Runge–Kutta methods and Taylor methods.
- That is, Ψ^τ is a numerical flow map, an approximation to the exact flow Φ^τ. This numerical flow produces a sequence of deterministic approximations to the solution of the ODE (1),

$$U_{k+1} := \Psi^\tau(U_k) \approx u_{k+1} = \Phi^\tau(u_k).$$

Assumption 2 (re: numerical flow)

Suppose that the numerical flow-map Ψ^τ has uniform local truncation error of order $q+1$: for some constant $C \geq 0$,

$$\sup_{u \in \mathcal{S}} \|\Psi^\tau(u) - \Phi^\tau(u)\| \leq C\tau^{q+1}.$$

PN for ODEs (Euler)

- Integral formulation for $u(t)$, $t_k \leq t \leq t_{k+1}$:

$$u(t) = u_k + \int_{t_k}^{t} f(u(s))\,\mathrm{d}s = \int_{t_k}^{t} g(s)\,\mathrm{d}s.$$

- Numerical integrators amount to a choice of $g(\,\cdot\,)$, subject to the common-sense criterion that $g(t_k) = f(U_k)$.
- If we posit e.g. a Gaussian random field for g, then we get a sequence of random approximations to u:

$$\begin{aligned} U(t) &= U_k + (t - t_k) f(U_k) + \xi_k(t - t_k) \\ U_{k+1} &= U_k + \tau f(U_k) + \xi_k(\tau), \end{aligned}$$

where $\mathbb{E}g = f(U_k)$ and $\xi_k = g - \mathbb{E}g$.
- Extend this idea to more general means (deterministic integrators).

PN for ODEs (General)

- We now define a new randomised one-step integrator

$$U_{k+1} := \Psi^{\tau}(U_k) + \xi_k(\tau)$$

with $\xi_k(t) := \int_0^t \chi_k(s)\,\mathrm{d}s$, where $\chi_k \sim \mathcal{N}(0, C^{\tau})$ are Gaussian. The covariance structure should reflect the smoothness of f and the accuracy of Ψ^{τ} in terms of numbers of derivatives.

- This definition not only provides for forward propagation of the the numerical state U_k, but also a continuous output via

$$U(t) = \Psi^{t-t_k}(U_k) + \xi_k(t - t_k) \quad \text{for } t \in [t_k, t_{k+1}].$$

- Note: this approach is intended to model sub-grid effects rather than sub-floating point effects, though a comprehensive analysis would include the latter (Hairer et al., 2008; Mosbach and Turner, 2009).

Assumption 3 (re: random perturbation)

Suppose that $\xi_k(t) := \int_0^t \chi_k(s)\,\mathrm{d}s$, where $\chi_k \sim \mathcal{N}(0, C^\tau)$ are i.i.d., and that there are constants $C \geq 0$ and $p \geq 1$ such that, for all $t \in [0, \tau]$, $\mathbb{E}\|\xi_k(t) \otimes \xi_k(t)\| \leq Ct^{2p+1}$, and, in particular, $\mathbb{E}\|\xi_k(t)\|^2 \leq Ct^{2p+1}$.

- Prime Gaussian example is a scaled integrated Brownian motion:

$$\xi_k(t) = \tau^{p-1} \int_0^t B(s)\,\mathrm{d}s.$$

- Our $\tau \to 0$ convergence results only use the highlighted bound on the second moment, $\mathbb{E}\xi_k = 0$, and independence of the ξ_k, so the Gaussian structure is not essential to the construction.
- We can and should enlist the help of numerical analysis to inform other priors for the truncation error for finer analysis.

Strong Convergence Result

Theorem 4 (Lie, Stuart & S., in prep.)

Under Assumptions 1–3, even with non-Gaussian ξ, we get convergence in $L^2(\Theta, \mathbb{P}; L^\infty([0, T]; \mathcal{S}))$ with rate "min(classical, noise)":

$$\mathbb{E}\left[\max_{0 \leq k \leq T/\tau} \|u_k - U_k\|^2\right] \leq C\tau^{2p \wedge 2q}. \tag{3}$$

If Assumption 3 is strengthened to $\mathbb{E}\left[\sup_{0 \leq t \leq \tau} \|\xi_0(t)\|^2\right] \leq C\tau^{2p+1}$, then

$$\mathbb{E}\left[\sup_{0 \leq t \leq T} \|u(t) - U(t)\|^2\right] \leq C\tau^{2p \wedge 2q}. \tag{4}$$

- A natural choice of scaling for the noise is $p = q$, for maximal uncertainty consistent with the deterministic convergence rate.
- Plenty of scope for numerical analysis and domain expertise to inform the fine structure of ξ (covariance / non-Gaussian structure, ...).

Overview

1. Introduction

2. PN for ODEs
   - Sampling-Based PN for ODEs (Forward Problems)
   - PN for ODE Inverse Problems

3. Probabilistic Meshless Methods for PDEs
   - PMM for PDEs (Forward Problems)
   - PMM in PDE Inverse Problems

4. Closing Remarks

Example: FitzHugh–Nagumo

FitzHugh–Nagumo Oscillator

Nonlinear oscillator $u\colon [0, T] \to \mathbb{R}^2$:

$$\frac{\mathrm{d}u}{\mathrm{d}t} = f(u) := \begin{bmatrix} u_1 - \frac{u_1^3}{3} + u_2 \\ -\frac{1}{\theta_3}(u_1 - \theta_1 + \theta_2 u_2) \end{bmatrix}$$

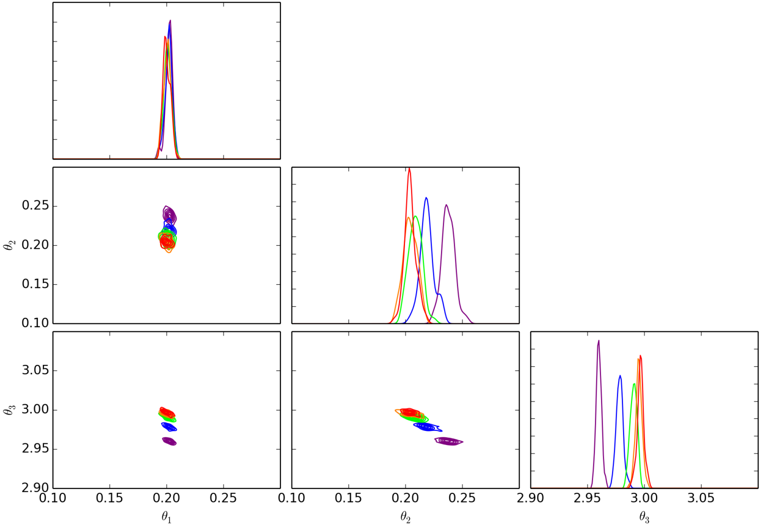
Note that f is not globally Lipschitz, but is one-sided Lipschitz!

- Aim: $\theta \in \mathbb{R}^3_{>0}$ from observations $y_i = u(t_i^{\mathrm{obs}}) + \eta_i$ at some discrete times $t_i^{\mathrm{obs}} = 0, 1, \dots, 40$, $\eta_i \sim \mathcal{N}(0, 10^{-3} I)$ i.i.d.
- Take ground truth $u(0) = (-1, 1)$ and $\theta = (0.2, 0.2, 3)$; generate data from a reference trajectory using RK4 with time step $\tau = 10^{-3}$.
- Infer θ using PN explicit Euler solvers with noise

$$\mathbb{E}\big[\xi_k(\tau) \otimes \xi_k(\tau)\big] = \sigma I \tau^{2p+1} \quad p = q = 1.$$

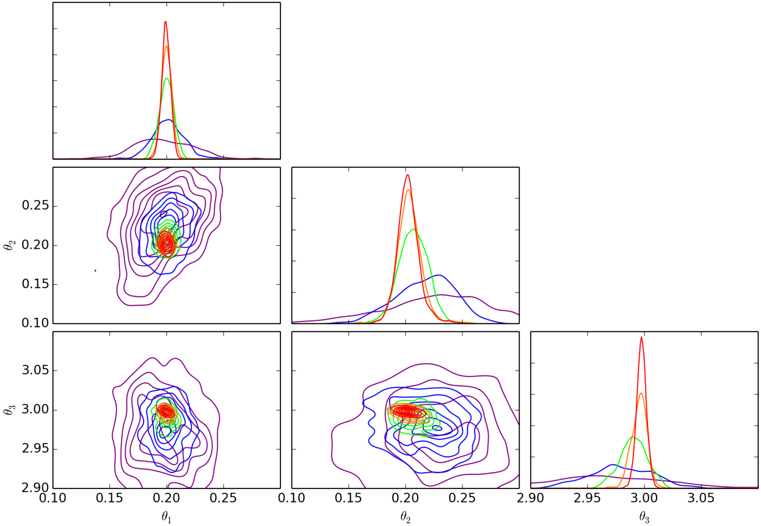
- Take log-normal prior for θ and compute the Bayesian posterior $\mathbb{E}_\xi \mathbb{P}[\theta | y, \tau, \xi]$ for various $\tau > 0$ and $\sigma \geq 0$.

Example: FitzHugh–Nagumo



The deterministic posteriors are over-confident at all values of the time step $\tau = 0.1, 0.05, 0.02, 0.01, 0.005$, do not overlap, and are biased.

Example: FitzHugh–Nagumo



The PN-Euler posteriors for $\tau = 0.1, 0.05, 0.02, 0.01, 0.005$ are less confident and overlap more, though are still biased.

Next Steps

- Relaxed regularity assumptions: convergence rates for locally Lipschitz flows and locally accurate deterministic solvers. Connections to numerical analysis for random dynamical systems.
- Construct structure-preserving PN integrators, e.g. for Hamiltonian dynamics the PN perturbation should not push the trajectory off the energy contours. Connections to stochastic analysis on manifolds, thermostats in molecular dynamics.
- 'On the fly' calibration of the noise covariance, and non-Gaussian structure. Connection to local error estimation and adaptivity, and to hierarchical Bayesian inversion.

Overview

Overview

Strong Formulation of Elliptic PDE

- General elliptic PDE on a bounded Lipschitz domain $D \subset \mathbb{R}^d$:

$$\begin{aligned} \mathcal{A}u(x) &= g(x) && \text{in } D, \\ \mathcal{B}u(x) &= b(x) && \text{on } \partial D. \end{aligned}$$

Example: Poisson's equation with Dirichlet BCs

$$\begin{aligned} \mathcal{A}u &:= -\nabla \cdot (\kappa \nabla u) \\ \mathcal{B}u &:= \operatorname{trace} u \end{aligned}$$

Prior Measure for u

Aim to infer a GP emulator for u given observations of the right-hand side g.

- Choose a Gaussian prior measure for u; $u \sim \mathcal{GP}(0, k)$

Prior Measure for u

Aim to infer a GP emulator for u given observations of the right-hand side g.

- Choose a Gaussian prior measure for u; $u \sim \mathcal{GP}(0, k)$
- Different choices for prior covariance k:
  - Involving the Green's function (Owhadi, 2016).

$$k(x, x') := \int_D \int_D G(x, z)\Lambda(z, z')G(z', x)\mathrm{d}z\mathrm{d}z'$$

  - More practical (Cialenco et al., 2012, Lemma 2.2):

$$\hat{k}(x, x') := \int_D \tilde{k}(x, z)\tilde{k}(z, x')\,\mathrm{d}z,$$

Posterior Measure for u

Assume linearity of $\mathcal{A}$, $\mathcal{B}$

- Now construct the posterior for u given observations $\boldsymbol{g}$, $\boldsymbol{b}$ of the RHS g, b at sets of locations $X_{\mathcal{A}}^{\circ} \subset D, X_{\mathcal{B}}^{\circ} \subset \partial D$, respectively.
- For subsets $A = \{a_i\}$, $B = \{b_j\}$ of D:

$$\begin{aligned} K(A, B) &= [k(a_i, b_j)]_{ij} \\ \mathcal{A}K(A, B) &= [\mathcal{A}k(a_i, b_j)]_{ij} \qquad \text{etc.} \end{aligned}$$

- $\bar{\mathcal{A}}K(A, B)$ denotes $\mathcal{A}$ applied to the second argument of k.

Posterior Measure for u

Assume linearity of $\mathcal{A}$, $\mathcal{B}$

- Now construct the posterior for u given observations $\boldsymbol{g}$, $\boldsymbol{b}$ of the RHS g, b at sets of locations $X^\circ_{\mathcal{A}} \subset D, X^\circ_{\mathcal{B}} \subset \partial D$, respectively.
- For subsets $A = \{a_i\}$, $B = \{b_j\}$ of D:

$$\begin{aligned} K(A,B) &= [k(a_i,b_j)]_{ij} \\ \mathcal{A}K(A,B) &= [\mathcal{A}k(a_i,b_j)]_{ij} \qquad \text{etc.} \end{aligned}$$

- $\bar{\mathcal{A}}K(A,B)$ denotes $\mathcal{A}$ applied to the second argument of k.

Theorem 5 (Posterior Measure for u)

$u|\boldsymbol{g},\boldsymbol{b} \sim \mathcal{N}(\mu, C)$, where, for $\mathcal{L} := \begin{bmatrix}\mathcal{A} & \mathcal{B}\end{bmatrix}^\top$ and $X \subset D$,

$$\begin{aligned} \mu(X) &:= \mathcal{L}\hat{K}(X^\circ, X)\big(\mathcal{L}\bar{\mathcal{L}}\hat{K}(X^\circ, X^\circ)\big)^{-1}[\boldsymbol{g}^\top, \boldsymbol{b}^\top]^\top \\ C(X) &:= \hat{K}(X,X) - \bar{\mathcal{L}}\hat{K}(X,X^\circ)\big(\mathcal{L}\bar{\mathcal{L}}\hat{K}(X^\circ, X^\circ)\big)^{-1}\mathcal{L}\hat{K}(X^\circ, X) \end{aligned}$$

Accuracy and Contraction of the Posterior for u

Theorem 6 (Cockayne et al., 2016, Prop. 6–Theorem 8)

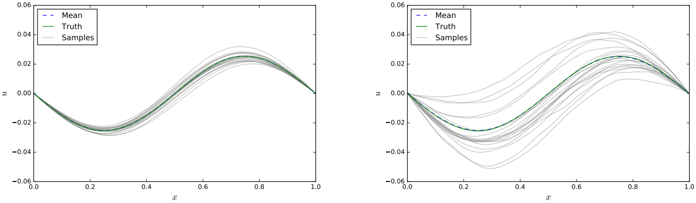
For a ball $B_\epsilon(u_0)$ of radius ϵ centered on the true solution $u^\dagger$ of the PDE, we have

$$\mathbb{P}\big[\|u - u^\dagger\|_{L^2(D)} > \varepsilon\big] = O\left(\frac{h^{\beta-\rho-d/2}}{\varepsilon}\right).$$

- *h the fill distance*
- *β the smoothness of the prior*
- *$\rho < \beta - d/2$ the order of the PDE*
- *d the input dimension*

Comparison of PN Solutions in 1d

$$-u''(x) = \sin(2\pi x) \quad \text{with Dirichlet BCs on } [0, 1]$$



(a) 'Natural' kernel, $\#X^\circ = 39$ (b) Integral kernel, $\#X^\circ = 39$

Overview

1. Introduction

2. PN for ODEs
   - Sampling-Based PN for ODEs (Forward Problems)
   - PN for ODE Inverse Problems

3. Probabilistic Meshless Methods for PDEs
   - PMM for PDEs (Forward Problems)
   - PMM in PDE Inverse Problems

4. Closing Remarks

Bayesian Parameter Inference with PN: Simple Example

Application to PDE-constrained inverse problems

$$-\theta u''(x) = \sin(2\pi x) \quad \text{with Dirichlet BCs on } [0,1]$$

Infer θ given observations of u with $\theta = \theta^\dagger = 1$ at $x = 0.25$ and 0.75, corrupted by additive noise $\mathcal{N}(0, 10^{-6} I)$.

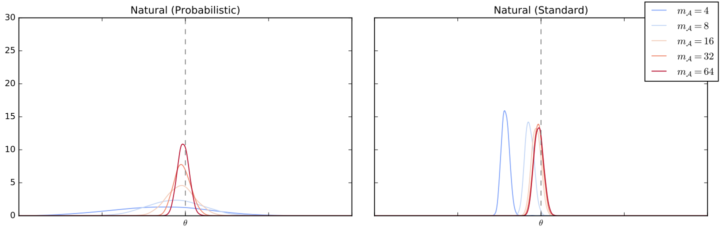


Figure: Posterior distributions for θ using optimal 'natural' kernel.

Bayesian Parameter Inference with PN: Simple Example

Application to PDE-constrained inverse problems

$$-\theta u''(x) = \sin(2\pi x) \quad \text{with Dirichlet BCs on } [0,1]$$

Infer θ given observations of u with $\theta = \theta^\dagger = 1$ at $x = 0.25$ and 0.75, corrupted by additive noise $\mathcal{N}(0, 10^{-6} I)$.

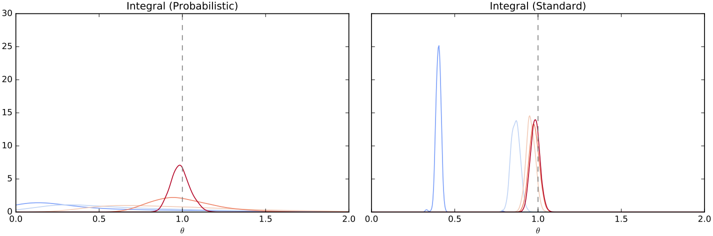


Figure: Posterior distributions for θ using integral Wendland kernel.

Bayesian Parameter Inference with PN: Simple Example

Application to PDE-constrained inverse problems

$$-\theta u''(x) = \sin(2\pi x) \quad \text{with Dirichlet BCs on } [0, 1]$$

Infer θ given observations of u with $\theta = \theta^\dagger = 1$ at $x = 0.25$ and 0.75, corrupted by additive noise $\mathcal{N}(0, 10^{-6} I)$.

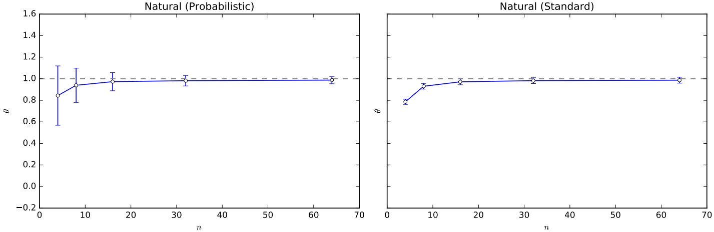


Figure: Posterior credible (1 s.d.) intervals for θ using optimal 'natural' kernel, as a function of $n = \#X^\circ$.

Bayesian Parameter Inference with PN: Simple Example

Application to PDE-constrained inverse problems

$$-\theta u''(x) = \sin(2\pi x) \quad \text{with Dirichlet BCs on } [0,1]$$

Infer θ given observations of u with $\theta = \theta^\dagger = 1$ at $x = 0.25$ and 0.75, corrupted by additive noise $\mathcal{N}(0, 10^{-6} I)$.

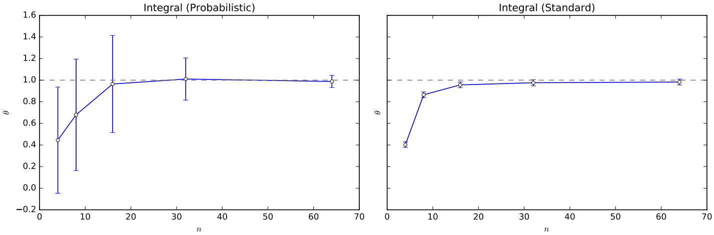


Figure: Posterior credible (1 s.d.) intervals for θ using integral Wendland kernel, as a function of $n = \#X^\circ$.

Posterior Contraction for the Inverse Problem

Theorem 7 (Cockayne et al., 2016, Theorem 11)

If the posterior for θ under the idealised exact solution $u^{\dagger}$ contracts in probability to $\delta_{\theta^{\dagger}}$, then so too does the posterior for θ under the PN solution $u^{\mathrm{PN}} := u | g(X^{\circ})$, provided the fill distance h of X° and the number of data points n scale as

$$h = o\left(\frac{1}{n^{1/(\beta - \rho - d/2)}} \right).$$

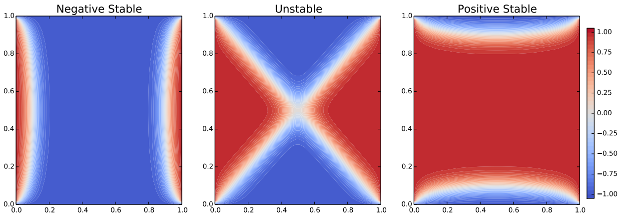
Semi-Linear Example: Steady-State Allen–Cahn

$$
\begin{aligned}
-\theta \Delta u + \theta^{-1}(u^3 - u) &= g && \text{in } D = (0,1)^2, \\
u(x_1, x_2) &= +1 && \text{for } x_1 = 0 \text{ or } 1; \\
u(x_1, x_2) &= -1 && \text{for } x_2 = 0 \text{ or } 1.
\end{aligned}
$$

- 'Mild' nonlinearity: linear + monotone.
- Extend PMM to handle this nonlinearity by introducing a latent variable z, which is later marginalised out:

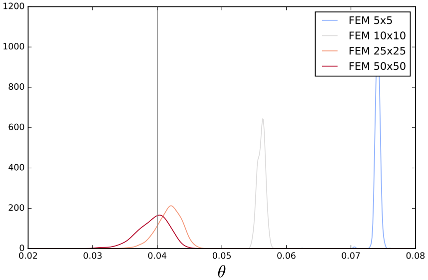
$$
-\theta \Delta u - \theta^{-1} u = z, \qquad \theta^{-1} u^3 = g - z.
$$

- Exhibits multiple solutions for $\delta \approx 0.04$, $g = 0$:

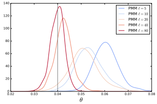


Allen–Cahn Parameter Inference

We try to recover θ from 16 observations of u on a 4×4 regular interior grid, corrupted by $\mathcal{N}(0, \frac{1}{10} I)$ noise.



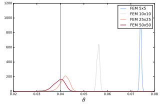
(a) FE



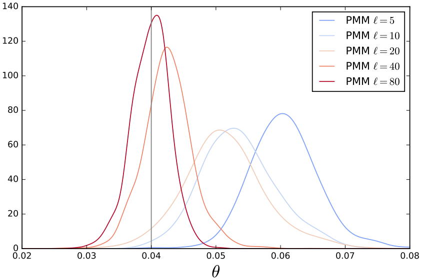
(b) PMM

Figure: Comparison of posteriors for θ with various forward models (likelihoods). Ground truth $\theta^\dagger = 0.04$, with prior $\theta \sim \mathrm{Unif}(0.02, 0.15)$. PN forward model uses squared exponential kernel, marginalising over a half-range Cauchy length scale parameter. Details of pseudomarginal MCMC etc. in Cockayne et al. (2016, Section 7).

We try to recover θ from 16 observations of u on a 4×4 regular interior grid, corrupted by $\mathcal{N}(0, \frac{1}{10} I)$ noise.



(a) FE



(b) PMM

Figure: Comparison of posteriors for θ with various forward models (likelihoods). Ground truth $\theta^\dagger = 0.04$, with prior $\theta \sim \mathrm{Unif}(0.02, 0.15)$. PN forward model uses squared exponential kernel, marginalising over a half-range Cauchy length scale parameter. Details of pseudomarginal MCMC etc. in Cockayne et al. (2016, Section 7).

Overview

General Comments

- PN offers ways to fold uncertainty arising from numerical error into inferences, and propagate this uncertainty to later computations.
  - Thus, we don't confuse the replicability of deterministic simulations with their accuracy.
  - Good data + appropriately skeptical model $\implies$ sound inferences.
- For both ODEs and PDEs, we have a good idea of how to proceed with Gaussian priors. In some cases, we can see how Gaussian and non-Gaussian priors have the same high-precision limits, but we can't expect this Gaussian universality to always hold true.
- Numerical analysis expertise is needed to build more realistic priors.
- Statistical expertise is needed to explore their posteriors.

`http://probabilistic-numerics.org`
`https://github.com/jcockayne/bayesian_pdes`

General Comments

- PN offers ways to fold uncertainty arising from numerical error into inferences, and propagate this uncertainty to later computations.
  - Thus, we don't confuse the replicability of deterministic simulations with their accuracy.
  - Good data + appropriately skeptical model $\implies$ sound inferences.
- For both ODEs and PDEs, we have a good idea of how to proceed with Gaussian priors. In some cases, we can see how Gaussian and non-Gaussian priors have the same high-precision limits, but we can't expect this Gaussian universality to always hold true.
- Numerical analysis expertise is needed to build more realistic priors.
- Statistical expertise is needed to explore their posteriors.

http://probabilistic-numerics.org

https://github.com/jcockayne/bayesian_pdes

Thank You

Bibliography I

I. Cialenco, G. E. Fasshauer, and Q. Ye. Approximation of stochastic partial differential equations by a kernel-based collocation method. *Int. J. Comput. Math.*, 89(18):2543–2561, 2012. URL `http://dx.doi.org/10.1080/00207160.2012.688111`.

J. Cockayne, C. J Oates, T. J. Sullivan, and M. Girolami. Probabilistic meshless methods for partial differential equations and Bayesian inverse problems, 2016. URL `http://arxiv.org/pdf/1605.07811`.

P. Diaconis. Bayesian numerical analysis. In *Statistical Decision Theory and Related Topics, IV, Vol. 1 (West Lafayette, Ind., 1986)*, pages 163–175. Springer, New York, 1988.

E. Hairer, R. I. McLachlan, and A. Razakarivony. Achieving Brouwer's law with implicit Runge–Kutta methods. *BIT*, 48(2):231–243, 2008. URL `http://dx.doi.org/10.1007/s10543-008-0170-3`.

J. B. Kadane. Parallel and sequential computation: a statistician's view. *J. Complexity*, 1(2): 256–263, 1985. URL `http://dx.doi.org/10.1016/0885-064X(85)90014-7`. Complexity of approximately solved problems (Morningside Heights, N.Y., 1985).

S. Mosbach and A. G. Turner. A quantitative probabilistic investigation into the accumulation of rounding errors in numerical ODE solution. *Comput. Math. Appl.*, 57(7):1157–1167, 2009. URL `http://dx.doi.org/10.1016/j.camwa.2009.01.020`.

A. O'Hagan. Some Bayesian numerical analysis. In *Bayesian statistics, 4 (Peñíscola, 1991)*, pages 345–363. Oxford Univ. Press, New York, 1992.

H. Owhadi. Multigrid with rough coefficients and multiresolution operator decomposition from hierarchical information games, 2016. URL `http://arxiv.org/pdf/1503.03467v4`.

Bibliography II

H. Poincaré. *Calcul des Probabilites*. Georges Carré, Paris, 1896.

M. Schober, D. Duvenaud, and P. Hennig. Probabilistic ODE solvers with Runge–Kutta means. In *Advances in Neural Information Processing Systems 25*, 2014.

J. Skilling. Bayesian solution of ordinary differential equations. In *Maximum Entropy and Bayesian Methods*, pages 23–37. Springer, 1992.