Bayesian Probabilistic Numerical Methods

J. Cockayne[1] M. Girolami[2,3] C. Oates[3,4] **T. J. Sullivan**[5,6]

Bayes Forum
Garching bei München, DE, 12 October 2018

[1]University of Warwick, UK
[2]Imperial College London, UK
[3]Alan Turing Institute, London, UK
[4]Newcastle University, UK
[5]**Free University of Berlin, DE**
[6]**Zuse Institute Berlin, DE**

A Probabilistic Treatment of Numerics?

- The last 5 years have seen a renewed interest in probabilistic perspectives on numerical tasks — e.g. quadrature, ODE and PDE solution, optimisation — continuing a theme with a long heritage: Poincaré (1896); Larkin (1970); Diaconis (1988); Skilling (1992).
- There are many ways to motivate this modelling choice:
    - To a statistician's eye, numerical tasks look like inverse problems.
    - Worst-case errors are often too pessimistic — perhaps we should adopt an average-case viewpoint (Traub et al., 1988; Ritter, 2000; Trefethen, 2008)?
    - "Big data" problems often require (random) subsampling.
    - If discretisation error is not properly accounted for, then biased and over-confident inferences result (Conrad et al., 2016). However, the necessary numerical analysis in nonlinear and evolutionary contexts can be **hard**!
    - Accounting for the impact of discretisation error in a statistical way allows forward and Bayesian inverse problems to speak a common statistical language.
- To make these ideas precise and to relate them to one another, some concrete definitions are needed!

Outline

1. Numerics: An Inference Perspective

2. Bayes' Theorem via Disintegration

3. Optimal Information

4. Numerical Disintegration

5. Coherent Pipelines of BPNMs

6. Applications

7. Closing Remarks

An Inference Perspective on Numerical Tasks

An Abstract View of Numerical Methods I

An abstract setting for numerical tasks consists of three spaces and two functions:

- $\mathcal{X}$, where an unknown/variable object x or u lives; $\dim \mathcal{X} = \infty$
- $\mathcal{A}$, where we observe information $A(x)$, via a function $A \colon \mathcal{X} \to \mathcal{A}$; $\dim \mathcal{A} < \infty$
- $\mathcal{Q}$, with a quantity of interest $Q \colon \mathcal{X} \to \mathcal{Q}$.

An Abstract View of Numerical Methods I

An abstract setting for numerical tasks consists of three spaces and two functions:

- $\mathcal{X}$, where an unknown/variable object x or u lives; $\dim \mathcal{X} = \infty$
- $\mathcal{A}$, where we observe information $A(x)$, via a function $A\colon \mathcal{X} \to \mathcal{A}$; $\dim \mathcal{A} < \infty$
- $\mathcal{Q}$, with a quantity of interest $Q\colon \mathcal{X} \to \mathcal{Q}$.

Example (Quadrature)

$$\mathcal{X} = C^0([0,1];\mathbb{R}) \qquad \mathcal{A} = ([0,1] \times \mathbb{R})^m \qquad \mathcal{Q} = \mathbb{R}$$

$$A(u) = (t_i, u(t_i))_{i=1}^m \qquad Q(u) = \int_0^1 u(t) \,\mathrm{d}t$$

An Abstract View of Numerical Methods I

An abstract setting for numerical tasks consists of three spaces and two functions:

- $\mathcal{X}$, where an unknown/variable object x or u lives; $\dim \mathcal{X} = \infty$
- $\mathcal{A}$, where we observe information $A(x)$, via a function $A \colon \mathcal{X} \to \mathcal{A}$; $\dim \mathcal{A} < \infty$
- $\mathcal{Q}$, with a quantity of interest $Q \colon \mathcal{X} \to \mathcal{Q}$.

Example (Quadrature)

$$\mathcal{X} = C^0([0,1];\mathbb{R}) \qquad \mathcal{A} = ([0,1] \times \mathbb{R})^m \qquad \mathcal{Q} = \mathbb{R}$$

$$A(u) = (t_i, u(t_i))_{i=1}^m \qquad Q(u) = \int_0^1 u(t)\,\mathrm{d}t$$

- Conventional numerical methods are cleverly-designed functions $b \colon \mathcal{A} \to \mathcal{Q}$: they estimate $Q(x)$ by $b(A(x))$.
- N.B. *Some* methods try to invert A, form an estimate of x, then apply Q, but e.g. vanilla Monte Carlo — $b((t_i, y_i)_{i=1}^n) := \frac{1}{n} \sum_{i=1}^n y_i$ — does not! (cf. O'Hagan, 1987)

An Abstract View of Numerical Methods II

- Question: What makes for a "good" numerical method? (Larkin, 1970)
- Answer 1, Gauss: $b \circ A = Q$ on a "large" finite-dimensional subspace of $\mathcal{X}$.
- Answer 2, Sard (1949): $b \circ A - Q$ is "small" on $\mathcal{X}$. In what sense?
  - The **worst-case error**:

$$e_{\mathrm{WC}} := \sup_{x \in \mathcal{X}} \| b(A(x)) - Q(x) \|_{\mathcal{Q}}.$$

  - The **average-case error** with respect to a probability measure μ on $\mathcal{X}$:

$$e_{\mathrm{AC}} := \int_{\mathcal{X}} \| b(A(x)) - Q(x) \|_{\mathcal{Q}} \, \mu(\mathrm{d}x).$$

  - To a **Bayesian**, seeing the additional structure of μ, there is only one way forward: if $x \sim \mu$, then $b(A(x))$ should be obtained by conditioning μ and then applying Q. But is this Bayesian solution always well-defined, and what are its error properties?

Rev. Bayes Does Some Numerics I

$$\begin{tikzcd}
\mathcal{X} \arrow[rr, bend left, "A"] \arrow[d, bend right, "Q"'] & & \mathcal{A} \\
\mathcal{Q} & &
\end{tikzcd}$$

Rev. Bayes Does Some Numerics I

$$\begin{tikzcd}
\mathcal{X} \arrow[r, bend left, "A"] \arrow[d, bend right, "Q"'] & \mathcal{A} \arrow[dl, squiggly, "b"] \\
\mathcal{Q} &
\end{tikzcd}$$

$$b \colon \mathcal{A} \to \mathcal{Q}$$

Example (Quadrature)

$$\mathcal{X} = C^0([0,1];\mathbb{R}) \qquad \mathcal{A} = ([0,1] \times \mathbb{R})^m \qquad \mathcal{Q} = \mathbb{R}$$

$$A(u) = (t_i, u(t_i))_{i=1}^m \qquad Q(u) = \int_0^1 u(t)\,\mathrm{d}t$$

A deterministic numerical method uses only the spaces and data to produce a point estimate of the integral.

Rev. Bayes Does Some Numerics I

$$\begin{tikzcd}
\mathcal{X} \arrow[r, bend left, "A"] \arrow[d, bend right, "Q"'] & \mathcal{A} \arrow[dl, squiggly, "b"] \\
\mathcal{Q} &
\end{tikzcd}$$

Go
Probabilistic!

$$\begin{tikzcd}
\mathcal{P}_{\mathcal{X}} \arrow[r, bend left, "A_{\#}"] \arrow[d, bend right, "Q_{\#}"'] & \mathcal{P}_{\mathcal{A}} \\
\mathcal{P}_{\mathcal{Q}} & \mathcal{A} \arrow[u, bend right, "\delta"']
\end{tikzcd}$$

$$b \colon \mathcal{A} \to \mathcal{Q}$$

Example (Quadrature)

$$\mathcal{X} = C^0([0,1];\mathbb{R}) \qquad \mathcal{A} = ([0,1] \times \mathbb{R})^m \qquad \mathcal{Q} = \mathbb{R}$$

$$A(u) = (t_i, u(t_i))_{i=1}^m \qquad Q(u) = \int_0^1 u(t)\,\mathrm{d}t$$

A deterministic numerical method uses only the spaces and data to produce a point estimate of the integral.

Rev. Bayes Does Some Numerics I

$$\begin{tikzcd}
\mathcal{X} \arrow[r, bend left, "A"] \arrow[d, bend right, "Q"'] & \mathcal{A} \arrow[dl, squiggly, "b"] \\
\mathcal{Q} &
\end{tikzcd}$$

$$b \colon \mathcal{A} \to \mathcal{Q}$$

Go Probabilistic!

$$\begin{tikzcd}
\mathcal{P}_{\mathcal{X}} \arrow[r, bend left, "A_{\#}"] \arrow[d, bend right, "Q_{\#}"'] & \mathcal{P}_{\mathcal{A}} \\
\mathcal{P}_{\mathcal{Q}} & \mathcal{A} \arrow[l, squiggly, "a \mapsto B(\mu, a)"] \arrow[u, bend right, "\delta"']
\end{tikzcd}$$

$$B \colon \mathcal{P}_{\mathcal{X}} \times \mathcal{A} \to \mathcal{P}_{\mathcal{Q}}$$

Example (Quadrature)

$$\mathcal{X} = C^0([0,1]; \mathbb{R}) \qquad \mathcal{A} = ([0,1] \times \mathbb{R})^m \qquad \mathcal{Q} = \mathbb{R}$$

$$A(u) = (t_i, u(t_i))_{i=1}^m \qquad Q(u) = \int_0^1 u(t) \, \mathrm{d}t$$

A deterministic numerical method uses only the spaces and data to produce a point estimate of the integral.

A probabilistic numerical method converts an additional belief about the integrand into a belief about the integral.

Rev. Bayes Does Some Numerics I

$$\begin{tikzcd}
\mathcal{X} \arrow[rr, bend left, "A"] \arrow[d, bend right, "Q"'] & & \mathcal{A} \arrow[dll, squiggly, "b"] \\
\mathcal{Q} & &
\end{tikzcd}$$

$b \colon \mathcal{A} \to \mathcal{Q}$

Go Probabilistic!

$$\begin{tikzcd}
\mathcal{P}_{\mathcal{X}} \arrow[loop left, dashed, "\text{`Bayes'}"] \arrow[d, bend right, "Q_{\#}"'] \arrow[rr, dotted, no head, "A_{\#}"] & & \mathcal{P}_{\mathcal{A}} \arrow[d, dotted, no head, "\delta"] \\
\mathcal{P}_{\mathcal{Q}} & & \mathcal{A} \arrow[llu, dashed, "a \mapsto \mu^a"'] \arrow[ll, squiggly, "a \mapsto B(\mu,a)"]
\end{tikzcd}$$

$B \colon \mathcal{P}_{\mathcal{X}} \times \mathcal{A} \to \mathcal{P}_{\mathcal{Q}}$

Definition (Bayesian PNM)

A PNM $B(\mu, \cdot) \colon \mathcal{A} \to \mathcal{P}_{\mathcal{Q}}$ with prior $\mu \in \mathcal{P}_{\mathcal{X}}$ is **Bayesian** for a QoI $Q \colon \mathcal{X} \to \mathcal{Q}$ and information operator $A \colon \mathcal{X} \to \mathcal{A}$ if the bottom-left $\mathcal{A}$-$\mathcal{P}_{\mathcal{X}}$-$\mathcal{P}_{\mathcal{Q}}$ triangle commutes, i.e. the output of B is the push-forward of the conditional distribution μ^a through Q:

$$B(\mu, a) = Q_{\#}\mu^a, \qquad \text{for } A_{\#}\mu\text{-almost all } a \in \mathcal{A},$$

$$(A_{\sharp}\mu)(E) := \mu \circ A^{-1}(E) = \mu\{x \in \mathcal{X} \mid A(x) \in E\} \qquad \text{for } E \subseteq \mathcal{A}.$$

Rev. Bayes Does Some Numerics II

Definition (Bayesian PNM)

A PNM B with prior $\mu \in \mathcal{P}_{\mathcal{X}}$ is **Bayesian** for a quantity of interest Q and information A if its output is exactly the image of the conditional distribution $\mu^a = \mu|[A = a]$ under Q:

$$B(\mu, a) = Q_{\#}\mu^a, \qquad \text{for } A_{\#}\mu\text{-almost all } a \in \mathcal{A}.$$

Rev. Bayes Does Some Numerics II

Definition (Bayesian PNM)

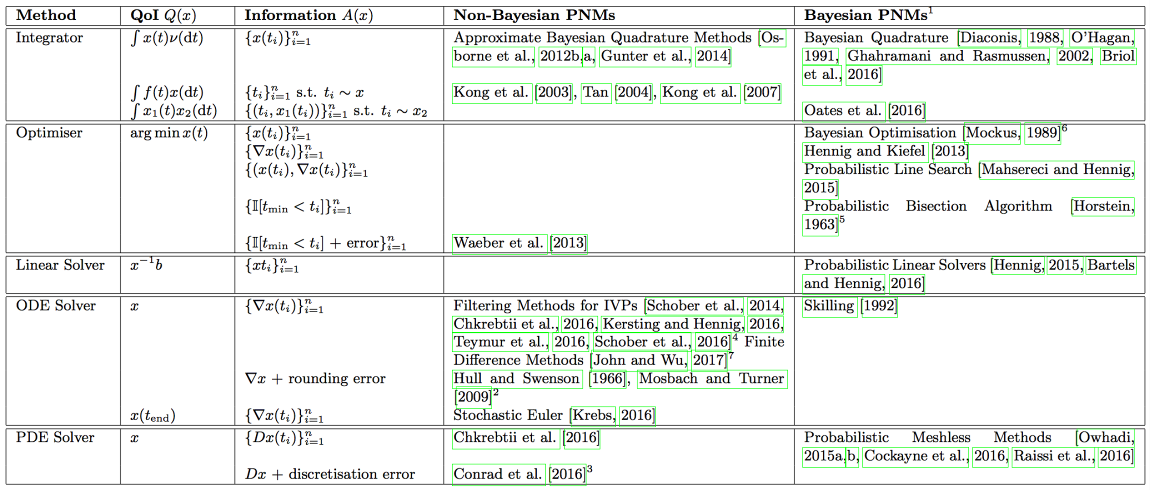
A PNM B with prior $\mu \in \mathcal{P}_{\mathcal{X}}$ is **Bayesian** for a quantity of interest Q and information A if its output is exactly the image of the conditional distribution $\mu^a = \mu|[A = a]$ under Q:

$$B(\mu, a) = Q_{\#}\mu^a, \qquad \text{for } A_{\#}\mu\text{-almost all } a \in \mathcal{A}.$$

Example

- Under the Gaussian Brownian motion prior on $\mathcal{X} = C^0([0,1];\mathbb{R})$, the posterior mean / MAP estimator for the definite integral is the trapezoidal rule, i.e. integration using linear interpolation (Sul′din, 1959, 1960).
- The integrated Brownian motion prior corresponds to integration using cubic spline interpolation.

A Rogue's Gallery of Bayesian and non-Bayesian PNMs



Method	QoI $Q(x)$	Information $A(x)$	Non-Bayesian PNMs	Bayesian PNMs[1]
Integrator	$\int x(t)\nu(\mathrm{d}t)$	$\{x(t_i)\}_{i=1}^n$	Approximate Bayesian Quadrature Methods [Osborne et al., 2012b,a, Gunter et al., 2014]	Bayesian Quadrature [Diaconis, 1988, O'Hagan, 1991, Ghahramani and Rasmussen, 2002, Briol et al., 2016]
	$\int f(t)x(\mathrm{d}t)$	$\{t_i\}_{i=1}^n$ s.t. $t_i \sim x$	Kong et al. [2003], Tan [2004], Kong et al. [2007]	
	$\int x_1(t)x_2(\mathrm{d}t)$	$\{(t_i, x_1(t_i))\}_{i=1}^n$ s.t. $t_i \sim x_2$		Oates et al. [2016]
Optimiser	$\arg\min x(t)$	$\{x(t_i)\}_{i=1}^n$ $\{\nabla x(t_i)\}_{i=1}^n$ $\{(x(t_i), \nabla x(t_i)\}_{i=1}^n$		Bayesian Optimisation [Mockus, 1989][6] Hennig and Kiefel [2013] Probabilistic Line Search [Mahsereci and Hennig, 2015]
		$\{\mathbb{I}[t_{\min} < t_i]\}_{i=1}^n$		Probabilistic Bisection Algorithm [Horstein, 1963][5]
		$\{\mathbb{I}[t_{\min} < t_i] + \text{error}\}_{i=1}^n$	Waeber et al. [2013]	
Linear Solver	$x^{-1}b$	$\{xt_i\}_{i=1}^n$		Probabilistic Linear Solvers [Hennig, 2015, Bartels and Hennig, 2016]
ODE Solver	x	$\{\nabla x(t_i)\}_{i=1}^n$	Filtering Methods for IVPs [Schober et al., 2014, Chkrebtii et al., 2016, Kersting and Hennig, 2016, Teymur et al., 2016, Schober et al., 2016][4] Finite Difference Methods [John and Wu, 2017][7]	Skilling [1992]
		∇x + rounding error	Hull and Swenson [1966], Mosbach and Turner [2009][2]	
	$x(t_{\text{end}})$	$\{\nabla x(t_i)\}_{i=1}^n$	Stochastic Euler [Krebs, 2016]	
PDE Solver	x	$\{Dx(t_i)\}_{i=1}^n$	Chkrebtii et al. [2016]	Probabilistic Meshless Methods [Owhadi, 2015a,b, Cockayne et al., 2016, Raissi et al., 2016]
		Dx + discretisation error	Conrad et al. [2016][3]	

Generalising Bayes' Theorem via Disintegration

Bayes' Theorem I

- Thus, we are expressing PNMs in terms of Bayesian inverse problems for functional unknowns (Stuart, 2010 + "the Finnish school").
- When describing prior and posterior distributions over infinite-dimensional quantities x (e.g. integrands, O/PDE solutions) one cannot work in terms of Lebesgue densities.
- At least for finite-dimensional data a with $a|x \sim \rho(a|x)\,\mathrm{d}a$, the fix is to describe the posterior μ^a in terms of its *density with respect to the prior*:

$$\frac{\mathrm{d}\mu^a}{\mathrm{d}\mu} \colon \mathcal{X} \to \mathbb{R} \qquad\qquad \frac{\mathrm{d}\mu^a}{\mathrm{d}\mu}(x) = \frac{\rho(a|x)}{\mathbb{E}_\mu[\rho(a|\,\cdot\,)]}.$$

Bayes' Theorem II

- Even the Stuart-style formulation of Bayes' rule does not work for PN, because

  $$\operatorname{supp}(\mu^a) \subseteq \mathcal{X}^a := \{x \in \mathcal{X} \mid A(x) = a\},$$

  typically $\mu(\mathcal{X}^a) = 0$, and — in contrast to typical statistical inverse problems — we think of the observation process as noiseless.
- We cannot take the Stuart-style approach of defining μ^a via its prior density as

  $$\frac{\mathrm{d}\mu^a}{\mathrm{d}\mu}(x) \propto \rho(a|x)$$

  because this density "wants" to be the indicator function $\mathbb{1}[u \in \mathcal{X}^a]$, which typically vanishes μ-a.e.
- E.g. quadrature example from earlier, with $A(u) = (t_i, u(t_i))_{i=1}^m$.
- While linear-algebraic tricks work for linear conditioning of Gaussians, in general we condition on events of measure zero using disintegration (or *regular conditional probabilities*).

Disintegration I

Write

$$\mu(f) \equiv \mathbb{E}_\mu[f] \equiv \int_{\mathcal{X}} f(x)\,\mu(\mathrm{d}x)$$

Definition (Disintegration)

A **disintegration** of $\mu \in \mathcal{P}_{\mathcal{X}}$ with respect to a measurable map $A\colon \mathcal{X} \to \mathcal{A}$ is a map $\mathcal{A} \to \mathcal{P}_{\mathcal{X}}, a \mapsto \mu^a$, such that

- $\mu^a(\mathcal{X} \setminus \mathcal{X}^a) = 0$ for $A_\# \mu$-almost all $a \in \mathcal{A}$; (support)

and, for each measurable $f\colon \mathcal{X} \to [0, \infty)$,

- $a \mapsto \mu^a(f)$ is measurable; (measurability)
- $\mu(f) = A_\# \mu\big(\mu^a(f)\big)$, (conditioning/reconstruction)

$$\text{i.e.} \quad \int_{\mathcal{X}} f(x)\,\mu(\mathrm{d}x) = \int_{\mathcal{A}} \left[\int_{\mathcal{X}^a} f(x)\,\mu^a(\mathrm{d}x) \right] (A_\# \mu)(\mathrm{d}a).$$

Disintegration II

Theorem (Disintegration theorem (Chang and Pollard, 1997, Thm. 1))

Let $\mathcal{X}$ be a metric space and let $\mu \in \mathcal{P}_{\mathcal{X}}$ be inner regular. If the Borel σ-algebra on $\mathcal{X}$ is countably generated and contains all singletons $\{a\}$ for $a \in \mathcal{A}$, then there is an essentially unique disintegration $\{\mu^a\}_{a \in \mathcal{A}}$ of μ with respect to A. (If $\{\nu^a\}_{a \in \mathcal{A}}$ is another such disintegration, then $\{a \in \mathcal{A} : \mu^a \neq \nu^a\}$ is an $A_{\#}\mu$-null set.)

Example

For $\mu \in \mathcal{P}_{\mathbb{R}^2}$ with continuous Lebesgue density $\rho \colon \mathbb{R}^2 \to [0, \infty)$, i.e. $\mathrm{d}\mu(x_1, x_2) = \rho(x_1, x_2)\, \mathrm{d}(x_1, x_2)$, the disintegration of μ with respect to vertical projection $A(x_1, x_2) := x_1$ is just the family of measures μ^a, where μ^a has Lebesgue density $\rho(a, \cdot)/Z^a$ on the vertical line $\{(a, x_2) \mid x_2 \in \mathbb{R}\}$, and $Z^a := \int_{\mathbb{R}} \rho(a, x_2)\, \mathrm{d}x_2$.

Except for nice situations like this, Gaussian measures, etc. (Owhadi and Scovel, 2015), disintegrations cannot be computed exactly — we have to work approximately.

Optimal Information: the Worst, the Average, and the Bayesian

Measures of Error

Suppose we have a **loss function** $L \colon \mathcal{Q} \times \mathcal{Q} \to \mathbb{R}$, e.g. $L(q, q') := \| q - q' \|_{\mathcal{Q}}^2$.

- The **worst-case loss** for a classical numerical method $b \colon \mathcal{A} \to \mathcal{Q}$ is

$$e_{\mathrm{WC}}(A, b) := \sup_{x \in \mathcal{X}} L\big(b(A(x)), Q(x)\big).$$

- The **average-case loss** under a probability measure $\mu \in \mathcal{P}_{\mathcal{X}}$ is

$$e_{\mathrm{AC}}(A, b) := \int_{\mathcal{X}} L\big(b(A(x)), Q(x)\big)\, \mu(\mathrm{d}x),$$

$$e_{\mathrm{AC}}(A, B) := \int_{\mathcal{X}} \left[\int_{\mathcal{Q}} L(q, Q(x))\, B(\mu, A(x))(\mathrm{d}q) \right] \mu(\mathrm{d}x).$$

- Kadane and Wasilkowski (1985) show that the minimiser B is a non-random Bayes decision rule b, and the minimiser A is “optimal information” for this task.
- A BPNM B has “no choice” but to be $Q_{\sharp}\mu^a$ once $A(x) = a$ is given; optimality of A means minimising the **Bayesian loss**

$$e_{\mathrm{BPN}}(A) := \int_{\mathcal{X}} \left[\int_{\mathcal{Q}} L(q, Q(x))\, (Q_{\sharp}\mu^{A(x)})(\mathrm{d}q) \right] \mu(\mathrm{d}x).$$

Optimal Information: AC = BPN?

Theorem (AC = BPN for quadratic loss; Cockayne et al., 2017b)

For a quadratic loss $L(q, q') := \|q - q'\|_{\mathcal{Q}}^2$ on a Hilbert space $\mathcal{Q}$, optimal information for BPNM and ACE coincide (though the minimal values may differ).

Optimal Information: AC = BPN?

Theorem (AC = BPN for quadratic loss; Cockayne et al., 2017b)

For a quadratic loss $L(q, q') := \|q - q'\|^2_{\mathcal{Q}}$ on a Hilbert space $\mathcal{Q}$, optimal information for BPNM and ACE coincide (though the minimal values may differ).

Example (AC = BPN in general?)

Decide whether or not a card drawn fairly at random is ♦, incurring unit loss if you guess wrongly; can choose to be told whether the card is red (A_1) or is non-♣ (A_2).

$$\begin{array}{lll} \mathcal{X} = \{♣, ♦, ♥, ♠\} & \mu = \mathrm{Unif}_{\mathcal{X}} & \mathcal{Q} = \{0, 1\} \subset \mathbb{R} \\ \mathcal{A}_1 = \{0, 1\} & A_1(x) = \mathbb{1}[x \in \{♦, ♥\}] & Q(x) = \mathbb{1}[x = ♦] \\ \mathcal{A}_2 = \{0, 1\} & A_2(x) = \mathbb{1}[x \in \{♦, ♥, ♠\}] & L(q, q') = \mathbb{1}[q \neq q'] \end{array}$$

Which information operator, A_1 or A_2, is better? (Note that $e_{\mathrm{WC}}(A_i, b) = 1$ for all deterministic b!)

Optimal Information: AC $\neq$ BPN!

$\mathcal{X} = \{\clubsuit, \textcolor{red}{\blacklozenge}, \textcolor{red}{\heartsuit}, \spadesuit\}$ $\qquad$ $\mu = \mathrm{Unif}_{\mathcal{X}}$ $\qquad$ $\mathcal{Q} = \{0, 1\} \subset \mathbb{R}$

$A_1(x) = \blacksquare$ vs. $\textcolor{red}{\blacksquare}$ $\qquad$ $Q(x) = \mathbb{1}[x = \textcolor{red}{\blacklozenge}]$

$A_2(x) = \neg\clubsuit$ vs. $\clubsuit$ $\qquad$ $L(q, q') = \mathbb{1}[q \neq q']$

$x =$	$\clubsuit$		$\textcolor{red}{\blacklozenge}$		$\textcolor{red}{\heartsuit}$		$\spadesuit$	
$e_{\mathrm{AC}}(A_1, b) = \frac{1}{4}\Big($	$L(b(\blacksquare), 0)$	$+$	$L(b(\textcolor{red}{\blacksquare}), 1)$	$+$	$L(b(\textcolor{red}{\blacksquare}), 0)$	$+$	$L(b(\blacksquare), 0)$	$\Big)$

Optimal Information: AC $\neq$ BPN!

$\mathcal{X} = \{$♣, ♦, ♥, ♠$\}$ $\mu = \mathrm{Unif}_{\mathcal{X}}$ $\mathcal{Q} = \{0, 1\} \subset \mathbb{R}$

$A_1(x) =$ ■ vs. 🟥 $Q(x) = \mathbb{1}[x =$ ♦$]$

$A_2(x) = \neg$♣ vs. ♣ $L(q, q') = \mathbb{1}[q \neq q']$

$x =$	♣		♦		♥		♠		
$e_{\mathrm{AC}}(A_1, b) = \frac{1}{4}($	$L(b($■$), 0)$	$+$	$L(b($🟥$), 1)$	$+$	$L(b($🟥$), 0)$	$+$	$L(b($■$), 0)$	$)$	
$e_{\mathrm{AC}}(A_1, 0) = \frac{1}{4}($	0	$+$	1	$+$	0	$+$	0	$)$	$= \frac{1}{4}$
$e_{\mathrm{AC}}(A_1, \mathrm{id}) = \frac{1}{4}($	0	$+$	0	$+$	1	$+$	0	$)$	$= \frac{1}{4}$

Optimal Information: AC $\neq$ BPN!

$$\mathcal{X} = \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\} \qquad \mu = \mathrm{Unif}_{\mathcal{X}} \qquad \mathcal{Q} = \{0, 1\} \subset \mathbb{R}$$

$$A_1(x) = \text{■ vs. 🟥} \qquad Q(x) = \mathbb{1}[x = \diamondsuit]$$

$$A_2(x) = \neg\clubsuit \text{ vs. } \clubsuit \qquad L(q, q') = \mathbb{1}[q \neq q']$$

$$\begin{array}{rccccccccl}
x = & \clubsuit & & \diamondsuit & & \heartsuit & & \spadesuit & & \\
\hline\hline
e_{\mathrm{AC}}(A_1, b) = \frac{1}{4}(& L(b(\text{■}), 0) & + & L(b(\text{🟥}), 1) & + & L(b(\text{🟥}), 0) & + & L(b(\text{■}), 0) &) & \\
e_{\mathrm{AC}}(A_1, 0) = \frac{1}{4}(& 0 & + & 1 & + & 0 & + & 0 &) & = \frac{1}{4} \\
e_{\mathrm{AC}}(A_1, \mathrm{id}) = \frac{1}{4}(& 0 & + & 0 & + & 1 & + & 0 &) & = \frac{1}{4} \\
\hline
e_{\mathrm{AC}}(A_2, b) = \frac{1}{4}(& L(b(\clubsuit), 0) & + & L(b(\neg\clubsuit), 1) & + & L(b(\neg\clubsuit), 0) & + & L(b(\neg\clubsuit), 0) &) & \\
e_{\mathrm{AC}}(A_2, 0) = \frac{1}{4}(& 0 & + & 1 & + & 0 & + & 0 &) & = \frac{1}{4} \\
\hline\hline
\end{array}$$

Optimal Information: AC $\neq$ BPN!

$$\mathcal{X} = \{\clubsuit, \blacklozenge, \heartsuit, \spadesuit\} \qquad \mu = \mathrm{Unif}_{\mathcal{X}} \qquad \mathcal{Q} = \{0,1\} \subset \mathbb{R}$$

$$A_1(x) = \blacksquare \text{ vs. } {\color{red}\blacksquare} \qquad Q(x) = \mathbb{1}[x = \blacklozenge]$$

$$A_2(x) = \neg\clubsuit \text{ vs. } \clubsuit \qquad L(q,q') = \mathbb{1}[q \neq q']$$

$$\begin{array}{rlccccccccl}
x = & & \clubsuit & & \blacklozenge & & \heartsuit & & \spadesuit & & \\
\hline\hline
e_{\mathrm{AC}}(A_1, b) = \frac{1}{4} & (& L(b(\blacksquare),0) & + & L(b({\color{red}\blacksquare}),1) & + & L(b({\color{red}\blacksquare}),0) & + & L(b(\blacksquare),0) &) & \\
e_{\mathrm{AC}}(A_1, 0) = \frac{1}{4} & (& 0 & + & 1 & + & 0 & + & 0 &) & = \frac{1}{4} \\
e_{\mathrm{AC}}(A_1, \mathrm{id}) = \frac{1}{4} & (& 0 & + & 0 & + & 1 & + & 0 &) & = \frac{1}{4} \\
\hline
e_{\mathrm{AC}}(A_2, b) = \frac{1}{4} & (& L(b(\clubsuit),0) & + & L(b(\neg\clubsuit),1) & + & L(b(\neg\clubsuit),0) & + & L(b(\neg\clubsuit),0) &) & \\
e_{\mathrm{AC}}(A_2, 0) = \frac{1}{4} & (& 0 & + & 1 & + & 0 & + & 0 &) & = \frac{1}{4} \\
\hline\hline
e_{\mathrm{BPN}}(A_1) = \frac{1}{4} & (& \mathbb{E}_{Q_\sharp \mu_{\blacksquare}} L(\cdot,0) & + & \mathbb{E}_{Q_\sharp \mu_{\color{red}\blacksquare}} L(\cdot,1) & + & \mathbb{E}_{Q_\sharp \mu_{\color{red}\blacksquare}} L(\cdot,0) & + & \mathbb{E}_{Q_\sharp \mu_{\blacksquare}} L(\cdot,0) &) & \\
= \frac{1}{4} & (& (\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0) & + & (\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 1) & + & (\frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 0) & + & (\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0) &) & = \frac{1}{4} \\
\hline
\end{array}$$

Optimal Information: AC $\neq$ BPN!

$$\mathcal{X} = \{\clubsuit, {\color{red}\diamondsuit}, {\color{red}\heartsuit}, \spadesuit\} \qquad \mu = \mathrm{Unif}_{\mathcal{X}} \qquad \mathcal{Q} = \{0,1\} \subset \mathbb{R}$$

$$A_1(x) = \blacksquare \text{ vs. } {\color{red}\blacksquare} \qquad Q(x) = \mathbb{1}[x = {\color{red}\diamondsuit}]$$

$$A_2(x) = \neg\clubsuit \text{ vs. } \clubsuit \qquad L(q,q') = \mathbb{1}[q \neq q']$$

$$
\begin{array}{rcccccccccl}
x = & & \clubsuit & & {\color{red}\diamondsuit} & & {\color{red}\heartsuit} & & \spadesuit & & \\
\hline\hline
e_{\mathrm{AC}}(A_1, b) = \frac{1}{4}\Big(& & L(b(\blacksquare),0) & + & L(b({\color{red}\blacksquare}),1) & + & L(b({\color{red}\blacksquare}),0) & + & L(b(\blacksquare),0) & \Big) & \\
e_{\mathrm{AC}}(A_1, 0) = \frac{1}{4}\Big(& & 0 & + & 1 & + & 0 & + & 0 & \Big) & = \frac{1}{4} \\
e_{\mathrm{AC}}(A_1, \mathrm{id}) = \frac{1}{4}\Big(& & 0 & + & 0 & + & 1 & + & 0 & \Big) & = \frac{1}{4} \\
\hline
e_{\mathrm{AC}}(A_2, b) = \frac{1}{4}\Big(& & L(b(\clubsuit),0) & + & L(b(\neg\clubsuit),1) & + & L(b(\neg\clubsuit),0) & + & L(b(\neg\clubsuit),0) & \Big) & \\
e_{\mathrm{AC}}(A_2, 0) = \frac{1}{4}\Big(& & 0 & + & 1 & + & 0 & + & 0 & \Big) & = \frac{1}{4} \\
\hline\hline
e_{\mathrm{BPN}}(A_1) = \frac{1}{4}\Big(& & \mathbb{E}_{Q_\sharp \mu^{\blacksquare}} L(\cdot,0) & + & \mathbb{E}_{Q_\sharp \mu^{\color{red}\blacksquare}} L(\cdot,1) & + & \mathbb{E}_{Q_\sharp \mu^{\color{red}\blacksquare}} L(\cdot,0) & + & \mathbb{E}_{Q_\sharp \mu^{\blacksquare}} L(\cdot,0) & \Big) & \\
= \frac{1}{4}\Big(& & (\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0) & + & (\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 1) & + & (\frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 0) & + & (\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0) & \Big) & = \frac{1}{4} \\
\hline
e_{\mathrm{BPN}}(A_2) = \frac{1}{4}\Big(& & \mathbb{E}_{Q_\sharp \mu^{\clubsuit}} L(\cdot,0) & + & \mathbb{E}_{Q_\sharp \mu^{\neg\clubsuit}} L(\cdot,1) & + & \mathbb{E}_{Q_\sharp \mu^{\neg\clubsuit}} L(\cdot,0) & + & \mathbb{E}_{Q_\sharp \mu^{\neg\clubsuit}} L(\cdot,0) & \Big) & \\
= \frac{1}{4}\Big(& & (1\cdot 0) & + & (\frac{1}{3}\cdot 0 + \frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 1) & + & (\frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 0 + \frac{1}{3}\cdot 0) & + & (\frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 0 + \frac{1}{3}\cdot 0) & \Big) & = {\color{red}\frac{1}{3}}
\end{array}
$$

Numerical Disintegration

Numerical Disintegration I

- The exact disintegration "$\mu^a(\mathrm{d}x) \propto \mathbb{1}[A(x) = a]\, \mu(\mathrm{d}x)$" can be accessed numerically via relaxation, with approximation guarantees provided $a \mapsto \mu^a$ is "nice", e.g. $A_{\#}\mu \in \mathcal{P}_{\mathcal{A}}$ has a smooth Lebesgue density.
- Consider relaxed posterior $\mu^a_\delta(\mathrm{d}x) \propto \phi(\|A(x) - a\|_{\mathcal{A}} / \delta)\, \mu(\mathrm{d}x)$ with $0 < \delta \ll 1$.
  - Essentially any $\phi \colon [0, \infty) \to [0, 1]$ tending continuously to 1 at 0 and decaying quickly enough to 0 at ∞ will do.
  - E.g. $\phi(r) := \mathbb{1}[r < 1]$ or $\phi(r) := \exp(-r^2)$.

Definition

The **integral probability metric** on $\mathcal{P}_{\mathcal{X}}$ associated to a normed space $\mathcal{F}$ of test functions $f \colon \mathcal{X} \to \mathbb{R}$ is

$$d_{\mathcal{F}}(\mu, \nu) := \sup\big\{ |\mu(f) - \nu(f)| \,\big|\, \|f\|_{\mathcal{F}} \leq 1 \big\}.$$

- $\mathcal{F} =$ bounded continuous functions with uniform norm ↔ total variation.
- $\mathcal{F} =$ bounded Lipschitz continuous functions with Lipschitz norm ↔ Wasserstein.

Numerical Disintegration II

$$\text{“}\mu^a(\mathrm{d}x) \propto \mathbb{1}[A(x) = a]\,\mu(\mathrm{d}x)\text{”}$$

$$\mu^a_\delta(\mathrm{d}x) \propto \phi(\|A(x) - a\|_{\mathcal{A}}/\delta)\,\mu(\mathrm{d}x)$$

$$d_{\mathcal{F}}(\mu, \nu) := \sup\{|\mu(f) - \nu(f)| \,|\, \|f\|_{\mathcal{F}} \leq 1\}$$

Theorem (Cockayne, Oates, Sullivan, and Girolami, 2017b, Theorem 4.3)

If $a \mapsto \mu^a$ is γ-Hölder from $(\mathcal{A}, \|\cdot\|_{\mathcal{A}})$ into $(\mathcal{P}_{\mathcal{X}}, d_{\mathcal{F}})$, then so too is the approximation $\mu^a_\delta \approx \mu^a$ as a function of δ. That is,

$$\begin{aligned} & d_{\mathcal{F}}(\mu^a, \mu^{a'}) \leq C \cdot \|a - a'\|^\gamma && \text{for } a, a' \in \mathcal{A} \\ \Longrightarrow \quad & d_{\mathcal{F}}(\mu^a, \mu^a_\delta) \leq C \cdot C_\phi \cdot \delta^\gamma && \text{for } A_{\#}\mu\text{-almost all } a \in \mathcal{A}. \end{aligned}$$

Open question: when does the hypothesis, a quantitative version of the Tjur property (Tjur, 1980), actually hold?

Numerical Disintegration III

To evaluate expectations against μ^a we can extrapolate expectations against μ^a_δ (Schillings and Schwab, 2016).

To sample μ^a_δ we take inspiration from rare event simulation and use tempering schemes to sample the posterior: we set $\delta_0 > \delta_1 > \ldots > \delta_N$ and consider

$$\mu^a_{\delta_0},\ \mu^a_{\delta_1},\ \ldots,\ \mu^a_{\delta_N}$$

- $\mu^a_{\delta_0}$ is easy to sample — often $\mu^a_{\delta_0} = \mu$.
- $\mu^a_{\delta_N}$ has δ_N close to zero and is hard to sample.
- Intermediate distributions define a "ladder" which takes us from prior to posterior.
- Even within this framework, there is considerable choice of sampling scheme, e.g. brute-force MCMC, SMC, QMC, pCN, ...

Example: Painlevé's First Transcendental I

A multivalent boundary value problem:

$$\begin{aligned} u''(t) - u(t)^2 &= -t && \text{for } t \geq 0 \\ u(0) &= 0 \\ u(t)/\sqrt{t} &\to 1 && \text{as } t \to +\infty \end{aligned}$$

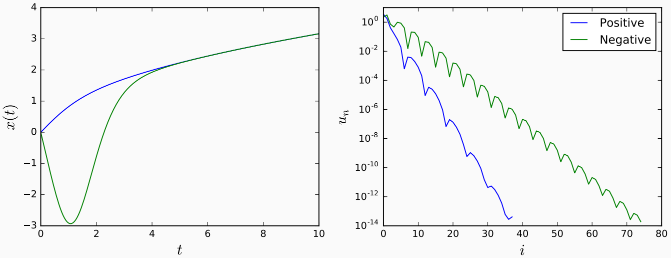


Figure 1: The two solutions of Painlevé's first transcendental and their spectra in the orthonormal Chebyshev polynomial basis over $[0, 10]$.

Example: Painlevé's First Transcendental I

A multivalent boundary value problem:

$$
\begin{aligned}
u''(t) - u(t)^2 &= -t && \text{for } t \geq 0 \\
u(0) &= 0 \\
u(10) &= \sqrt{10}
\end{aligned}
$$

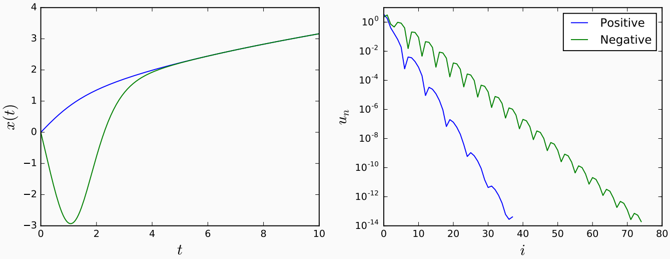
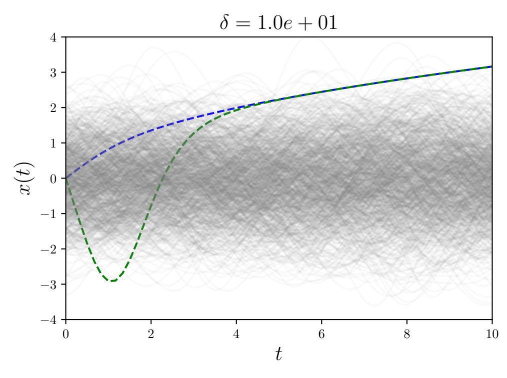


Figure 1: The two solutions of Painlevé's first transcendental and their spectra in the orthonormal Chebyshev polynomial basis over $[0, 10]$.

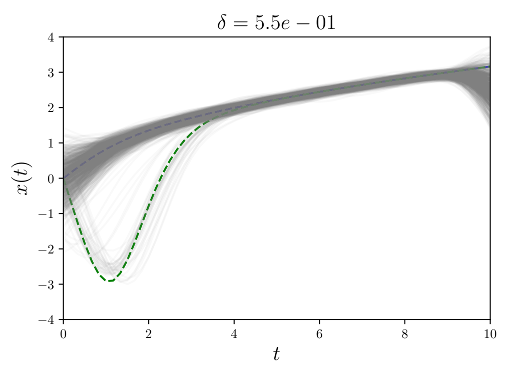
Example: Painlevé's First Transcendental III

- Parallel tempered pCN with 100 δ-values log-spaced from $\delta = 10$ to $\delta = 10^{-4}$ and 10^8 iterations recovers both solutions in approximately the same proportions as the posterior densities at the two exact solutions. ✓
- SMC reliably recovers one solution, but not both simultaneously. ⁉
- Of course, this comes at the price of MCMC... ✗



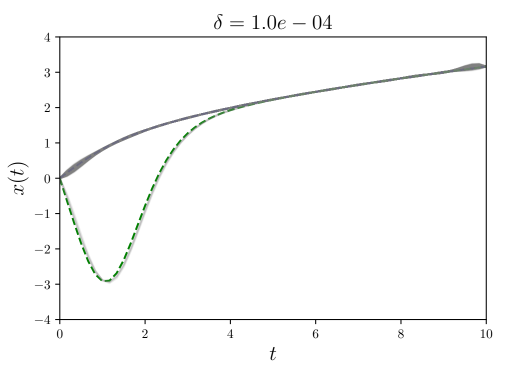
Example: Painlevé's First Transcendental III

- Parallel tempered pCN with 100 δ-values log-spaced from $\delta = 10$ to $\delta = 10^{-4}$ and 10^8 iterations recovers both solutions in approximately the same proportions as the posterior densities at the two exact solutions. ✓
- SMC reliably recovers one solution, but not both simultaneously. ⁉
- Of course, this comes at the price of MCMC… ✗



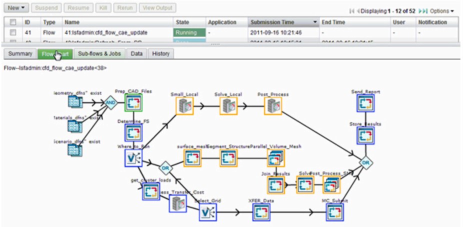
Example: Painlevé's First Transcendental III

- Parallel tempered pCN with 100 δ-values log-spaced from $\delta = 10$ to $\delta = 10^{-4}$ and 10^8 iterations recovers both solutions in approximately the same proportions as the posterior densities at the two exact solutions. ✓
- SMC reliably recovers one solution, but not both simultaneously. ⁉
- Of course, this comes at the price of MCMC… ✗



Coherent Pipelines of BPNMs

Computational Pipelines



□	ID	Type	Name	State	Application	Submission Time	End Time	User	Notification
□	41	Flow	41:lsfadmin:cfd_flow_cae_update	Running	-	2011-09-16 10:21:46	-	-	-

Flow--lsfadmin:cfd_flow_cae_update<38>

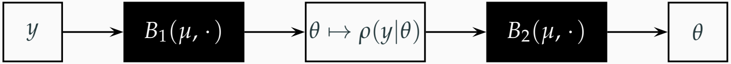
- Numerical methods usually form part of pipelines.
- Prime example: a PDE solve is a *forward model* in an *inverse problem*.
- Motivation for PNMs in the context of Bayesian inverse problems:

  Make the forward and inverse problem
  speak the same statistical language!

- We can compose PNMs in series, e.g. $B_2(B_1(\mu, a_1), a_2)$ is formally $B(\mu, (a_1, a_2))$... although figuring out what the spaces $\mathcal{X}_i$, $\mathcal{A}_i$ and operators A_i etc. are is a headache!

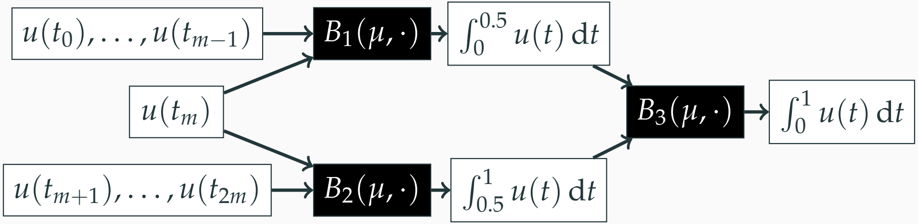
Pipeline Example i: Bayesian Inverse Problems

- A “simple” example of a computational pipeline is a Bayesian inverse problem for recovering parameters $\theta \in \Theta$ from data $y \in \mathcal{Y}$. This can be represented as the two-stage computational pipeline



- B_1 is the method that converts data y into the likelihood function for parameters θ, and hence incorporates any forward model such as an O/PDE solver.
- B_2 is the method that converts the prior on θ and the likelihood into a joint distribution for (θ, y), then conditions upon the actual observation — it returns something in $\mathcal{P}_{\Theta}$.
- B_1 conventionally has deterministic output in $\mathbb{R}^{\Theta}$, but in the world of PN, it could return a non-trivial probability distribution in $\mathcal{P}_{\mathbb{R}^{\Theta}}$, i.e. a randomised likelihood.
- Lie et al. (2018) analyse how the stochastic variability in the forward model / likelihood propagates to the (randomised or marginal) Bayesian posterior on θ.

Pipeline Example ii: Split Integration



$u(t_0), \dots, u(t_{m-1})$

$u(t_m)$

$u(t_{m+1}), \dots, u(t_{2m})$

$B_1(\mu, \cdot)$

$B_2(\mu, \cdot)$

$\int_0^{0.5} u(t)\,\mathrm{d}t$

$\int_{0.5}^{1} u(t)\,\mathrm{d}t$

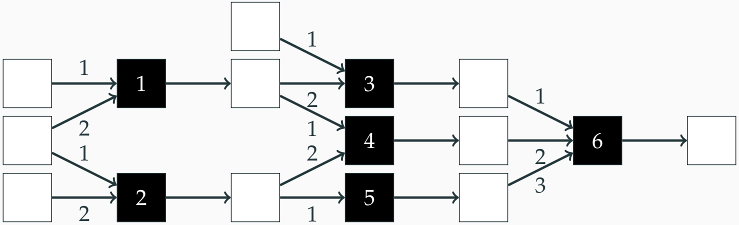
$B_3(\mu, \cdot)$

$\int_0^{1} u(t)\,\mathrm{d}t$

- Integrate a function over $[0, 1]$ in two steps using nodes $0 \leq t_0 < \cdots < t_{m-1} < 0.5$, $t_m = 0.5$, and $t_{m+1} < \cdots < t_{2m} \leq 1$.
- For example, the two nodal sets are very large, and so two are handled by two different processors with non-shared memory.
- A third processor handles the (easy!) task of aggregating the two estimates of the two integrals $\int_0^{0.5} u(t)\,\mathrm{d}t$ and $\int_{0.5}^{1} u(t)\,\mathrm{d}t$ into an estimate of $\int_0^{1} u(t)\,\mathrm{d}t$.

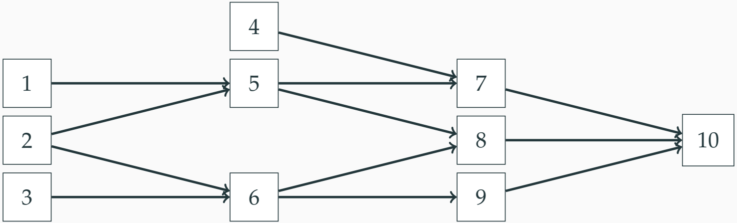
Coherence I

- We compose PNMs in a graphical way by allowing input information nodes (□) to feed into method nodes (■), which in turn output new information.
- N.B. one should at first think of having *deterministic* data at the left-most □ nodes, then *random variables* as outputs, *realisations* of which get fed into the next ■.



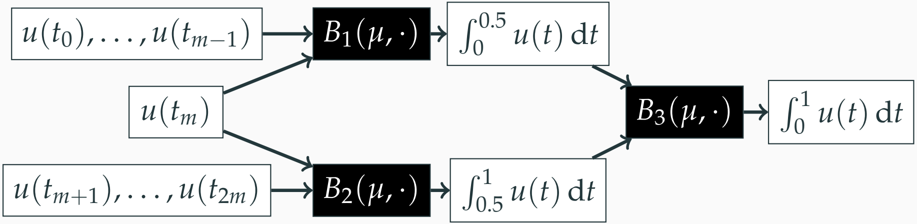
Coherence I

- We compose PNMs in a graphical way by allowing input information nodes (□) to feed into method nodes (■), which in turn output new information.
- N.B. one should at first think of having *deterministic* data at the left-most □ nodes, then *random variables* as outputs, *realisations* of which get fed into the next ■.



- We define the corresponding **dependency graph** by replacing each □→■→□ by □→□, and number the vertices in an increasing fashion, so that $\boxed{i} \to \boxed{i'}$ implies $i < i'$.
- The independence properties of the random variables at each node are crucial.

Coherence II


$u(t_0), \dots, u(t_{m-1})$
$u(t_m)$
$u(t_{m+1}), \dots, u(t_{2m})$
$B_1(\mu, \cdot)$
$B_2(\mu, \cdot)$
$\int_0^{0.5} u(t) \, \mathrm{d}t$
$\int_{0.5}^{1} u(t) \, \mathrm{d}t$
$B_3(\mu, \cdot)$
$\int_0^1 u(t) \, \mathrm{d}t$

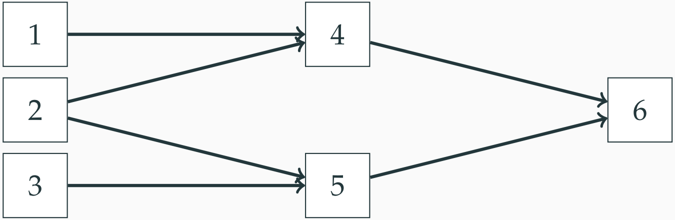
Definition

A prior μ is **coherent** for the dependency graph if — when the "leaf" input nodes are $A_\sharp \mu$-distributed and the remaining nodes are $B(\mu, \text{parents})$-distributed — every node Y_k is conditionally independent of all older non-parent nodes Y_i given its direct parents Y_j:

$$Y_k \perp\!\!\!\perp Y_{\{1,\dots,k-1\} \setminus \text{parents}(k)} \mid Y_{\text{parents}(k)}$$

This is weaker than the Markov condition for directed acyclic graphs (Lauritzen, 1991): we do not insist that the variables at the source nodes are independent.

Coherence II



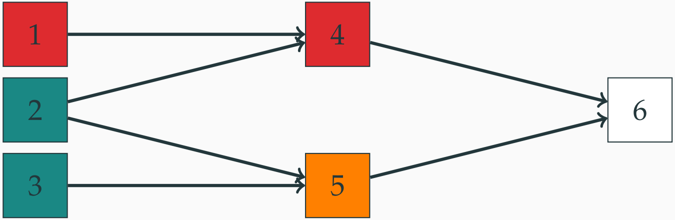
Definition

A prior μ is **coherent** for the dependency graph if — when the “leaf” input nodes are $A_\sharp\mu$-distributed and the remaining nodes are $B(\mu, \text{parents})$-distributed — every node Y_k is conditionally independent of all older non-parent nodes Y_i given its direct parents Y_j:

$$Y_k \perp\!\!\!\perp Y_{\{1,\dots,k-1\}\setminus \text{parents}(k)} \;\big|\; Y_{\text{parents}(k)}$$

This is weaker than the Markov condition for directed acyclic graphs (Lauritzen, 1991): we do not insist that the variables at the source nodes are independent.

Coherence II



Definition

A prior μ is **coherent** for the dependency graph if — when the "leaf" input nodes are $A_\sharp\mu$-distributed and the remaining nodes are $B(\mu, \text{parents})$-distributed — every node Y_k is conditionally independent of all older non-parent nodes Y_i given its direct parents Y_j:

$$Y_k \perp\!\!\!\perp Y_{\{1,\ldots,k-1\}\setminus\text{parents}(k)} \mid Y_{\text{parents}(k)}$$

This is weaker than the Markov condition for directed acyclic graphs (Lauritzen, 1991): we do not insist that the variables at the source nodes are independent.

Coherency Theorem

Theorem (Cockayne, Oates, Sullivan, and Girolami, 2017b, Theorem 5.9)

If a pipeline of PNMs is such that

- *the prior is coherent for the dependency graph, and*
- *the component PNMs are all Bayesian*

then the pipeline is the Bayesian pipeline $\boxed{\text{data at leaves}} \rightarrow \blacksquare \rightarrow \boxed{\text{final output}}$.

Coherency Theorem

Theorem (Cockayne, Oates, Sullivan, and Girolami, 2017b, Theorem 5.9)

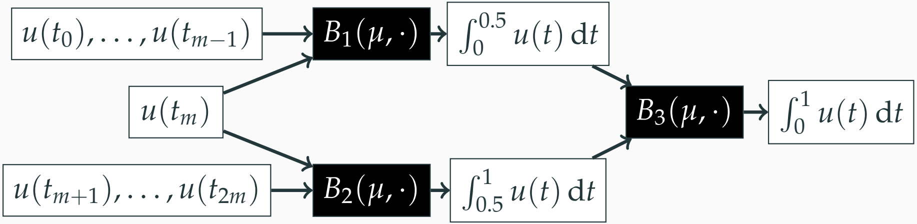
If a pipeline of PNMs is such that

- *the prior is coherent for the dependency graph, and*
- *the component PNMs are all Bayesian*

then the pipeline is the Bayesian pipeline $\boxed{\text{data at leaves}} \to \blacksquare \to \boxed{\text{final output}}$.

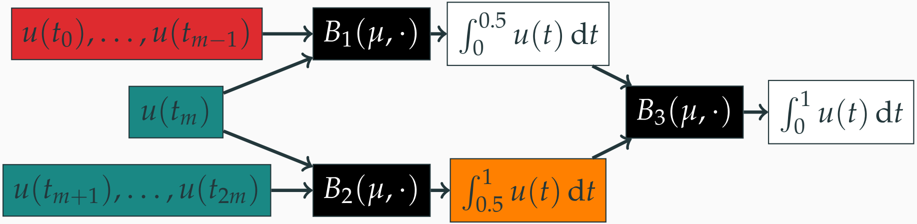
- Redundant structure in the pipeline (recycled information) will break coherence, and hence Bayesianity of the pipeline.
- In principle, coherence and hence being Bayesian depend upon the prior.
- This should not be surprising — as a loose analogy, one doesn't expect the trapezoidal rule to be a good way to integrate very smooth functions.

Split Integration: Coherence


$u(t_0), \ldots, u(t_{m-1})$
$B_1(\mu, \cdot)$
$\int_0^{0.5} u(t)\,\mathrm{d}t$
$u(t_m)$
$B_3(\mu, \cdot)$
$\int_0^1 u(t)\,\mathrm{d}t$
$u(t_{m+1}), \ldots, u(t_{2m})$
$B_2(\mu, \cdot)$
$\int_{0.5}^1 u(t)\,\mathrm{d}t$

- Integrate a function over $[0, 1]$ in two steps using nodes $0 \leq t_0 < \cdots < t_{m-1} < 0.5$, $t_m = 0.5$, and $t_{m+1} < \cdots < t_{2m} \leq 1$.

Split Integration: Coherence



$u(t_0), \ldots, u(t_{m-1})$

$u(t_m)$

$u(t_{m+1}), \ldots, u(t_{2m})$

$B_1(\mu, \cdot)$

$B_2(\mu, \cdot)$

$\int_0^{0.5} u(t) \, \mathrm{d}t$

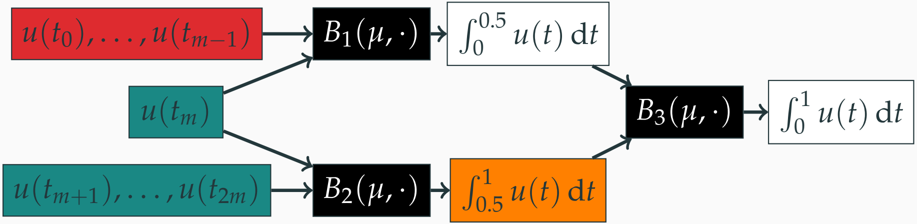
$\int_{0.5}^{1} u(t) \, \mathrm{d}t$

$B_3(\mu, \cdot)$

$\int_0^{1} u(t) \, \mathrm{d}t$

- Integrate a function over $[0, 1]$ in two steps using nodes $0 \leq t_0 < \cdots < t_{m-1} < 0.5$, $t_m = 0.5$, and $t_{m+1} < \cdots < t_{2m} \leq 1$.
- Is $\int_{0.5}^{1} u(t) \, \mathrm{d}t$ independent of $u(t_0), \ldots, u(t_{m-1})$ given $u(t_m), \ldots, u(t_{2m})$?

Split Integration: Coherence



$u(t_0), \ldots, u(t_{m-1})$

$u(t_m)$

$u(t_{m+1}), \ldots, u(t_{2m})$

$B_1(\mu, \cdot)$

$B_2(\mu, \cdot)$

$\int_0^{0.5} u(t)\, \mathrm{d}t$

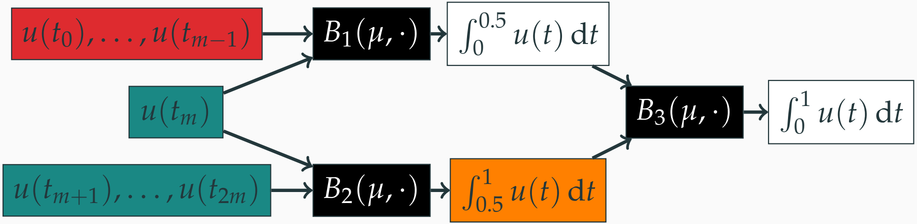
$\int_{0.5}^{1} u(t)\, \mathrm{d}t$

$B_3(\mu, \cdot)$

$\int_0^1 u(t)\, \mathrm{d}t$

- Integrate a function over $[0, 1]$ in two steps using nodes $0 \leq t_0 < \cdots < t_{m-1} < 0.5$, $t_m = 0.5$, and $t_{m+1} < \cdots < t_{2m} \leq 1$.
- Is $\int_{0.5}^{1} u(t)\, \mathrm{d}t$ independent of $u(t_0), \ldots, u(t_{m-1})$ given $u(t_m), \ldots, u(t_{2m})$?
- For a Brownian motion prior on the integrand u, yes.
- For an integrated BM prior on u, i.e. a BM prior on u', no.

Split Integration: Coherence



$u(t_0), \dots, u(t_{m-1})$

$u(t_m)$

$u(t_{m+1}), \dots, u(t_{2m})$

$B_1(\mu, \cdot)$

$B_2(\mu, \cdot)$

$\int_0^{0.5} u(t) \, \mathrm{d}t$

$\int_{0.5}^{1} u(t) \, \mathrm{d}t$

$B_3(\mu, \cdot)$

$\int_0^1 u(t) \, \mathrm{d}t$

- Integrate a function over $[0, 1]$ in two steps using nodes $0 \leq t_0 < \dots < t_{m-1} < 0.5$, $t_m = 0.5$, and $t_{m+1} < \dots < t_{2m} \leq 1$.
- Is $\int_{0.5}^{1} u(t) \, \mathrm{d}t$ independent of $u(t_0), \dots, u(t_{m-1})$ given $u(t_m), \dots, u(t_{2m})$?
- For a Brownian motion prior on the integrand u, yes.
- For an integrated BM prior on u, i.e. a BM prior on u', no.
- So how do we elicit an appropriate prior that respects the problem's structure? ⁉
- And is being *fully* Bayesian worth it in terms of cost and robustness? Cf. Owhadi et al. (2015), Jacob et al. (2017), and Lie et al. (2018). ⁉

Applications

Example 1: FitzHugh–Nagumo ODE inference

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: recover $\theta \in \mathbb{R}^3_{>0}$ 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–Euler solvers with local noise ξ of variance $\propto \sigma\tau^3$ and hence strong error $\mathbb{E}\left[\sup_{0 \leq t \leq T} \|u(t) - u^{\mathrm{PN}}(t)\|^2\right] \leq C\tau^2$ (Lie et al., 2017).
- Take log-normal prior for θ and compute the marginal Bayesian posterior $\mathbb{E}_\xi\left[\mathbb{P}[\theta | y, \tau, \xi]\right]$ for various $\tau > 0$ and $\sigma \geq 0$.

Example i: FitzHugh–Nagumo ODE inference

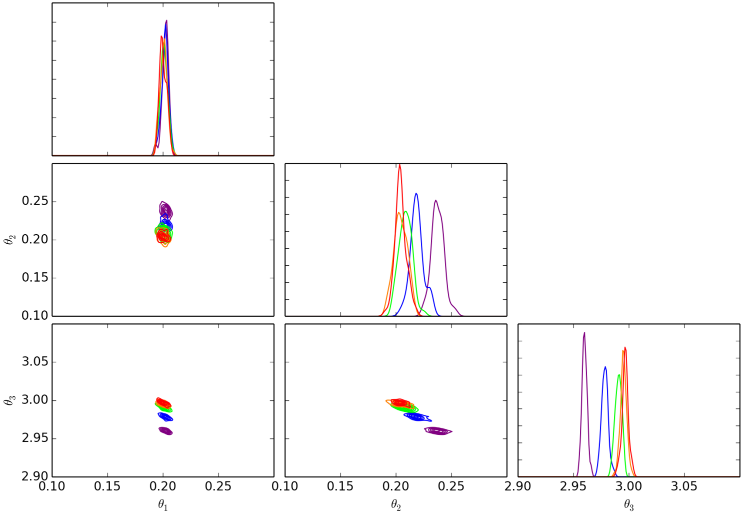


Figure 2: The deterministic posteriors (i.e. $\sigma = 0$) are over-confident at all values of the time step $\tau = 0.1, 0.05, 0.02, 0.01, 0.005$, do not overlap, and are biased.

Example i: FitzHugh–Nagumo ODE inference

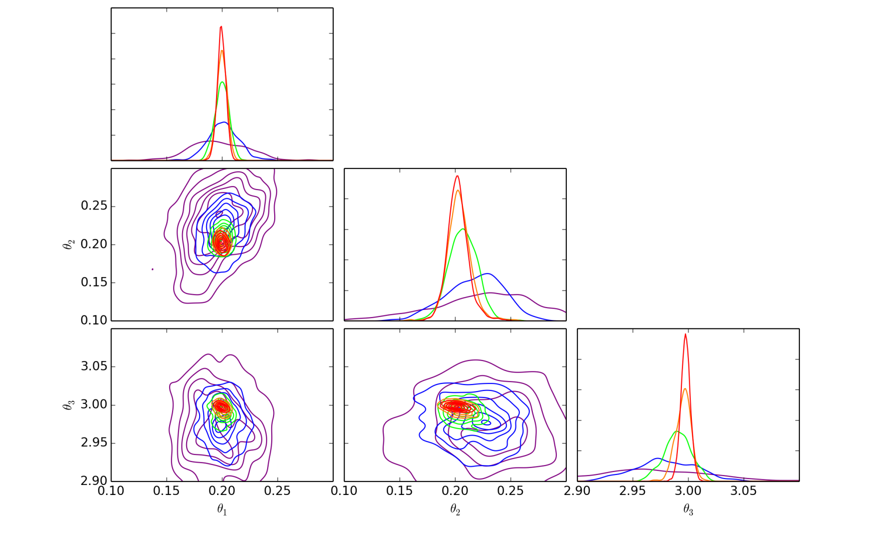
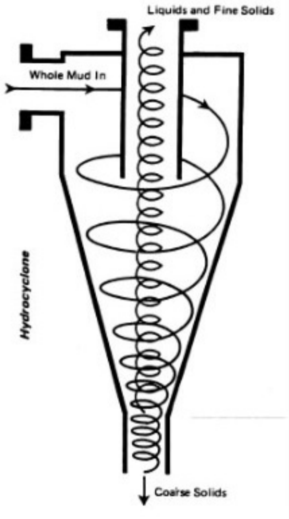


Figure 2: In contrast, the PN-Euler posteriors (here with $\sigma = 1/5$) for $\tau = 0.1, 0.05, 0.02, 0.01, 0.005$ are less confident and overlap more, though are still biased.

Example ii: Hydrocyclones (Oates, Cockayne, and Ackroyd, 2017)

- Hydrocyclones are used in industry as an alternative to centrifuges or filtration systems to separate fluids of different densities or particulate matter from a fluid.
- Monitoring is an essential control component, but usually cannot be achieved visually: Gutierrez et al. (2000) propose electrical impedance tomography as an alternative.
- EIT is an indirect imaging technique in which the conductivity field in the interior — which correlates with many material properties of interest — is inferred from current and voltage boundary conditions.
- In its Bayesian formulation, this is a well-posed inverse problem (Dunlop and Stuart, 2016a,b) closely related to Calderón's problem (Uhlmann, 2009).



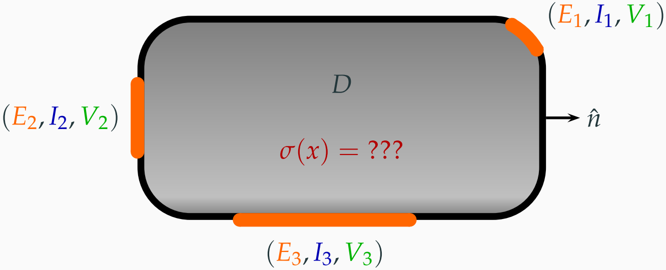
Complete Electrode Model (Cheng et al., 1989; Somersalo et al., 1992)

The interior conductivity field σ and electrical potential field v and the applied boundary currents I_i, measured voltages V_i, and known contact impedances ζ_i are related by

$$-\nabla \cdot \sigma(x) \nabla v(x) = 0 \qquad x \in D; \qquad \int_{E_i} \sigma(x) \frac{\partial v(x)}{\partial \hat{n}} \, \mathrm{d}x = I_i \qquad x \in E_i, i = 1, \dots, m;$$

$$v(x) + \zeta_i \sigma(x) \frac{\partial v(x)}{\partial \hat{n}} = V_i \qquad x \in E_i; \qquad \sigma(x) \frac{\partial v(x)}{\partial \hat{n}} = 0 \qquad x \in \partial D \setminus \bigcup_{i=1}^{m} E_i.$$

Furthermore, we consider a vector of such models, with multiple current stimulation patterns, at multiple points in time, for a time-dependent field $\sigma(t, x)$.



EIT Forward Problem

- Sampling from the posterior(s) requires repeatedly solving the forward PDE.
- We use the probabilistic meshless method of Cockayne et al. (2016, 2017a):
  - a Gaussian process extension of symmetric collocation;
  - a BPNM for a Gaussian prior and linear elliptic PDEs of this type.
- PMM allows us to:
  - account for uncertainty arising from the PDE having no explicit solution;
  - use coarser discretisations of the PDE to solve the problem faster while still providing meaningful UQ.

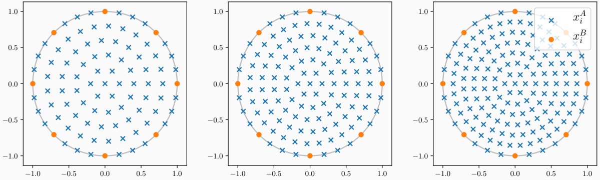


Figure 3: Like collocation, PMM imposes the PDE relation at $n_{\mathcal{A}}$ interior nodes and boundary conditions at $n_{\mathcal{B}}$ boundary nodes.

EIT Inverse Problem

- For the inverse problem we use a Karhunen–Loève series prior:

  $$\log \sigma(t, x; \omega) = \sum_{k=1}^{\infty} k^{-\alpha} \psi_k(t; \omega) \phi_k(x),$$

  with the ψ_k being a-priori independent Brownian motions in t.
- Like Dunlop and Stuart (2016a), we assume additive Gaussian observational noise with variance $\gamma^2 > 0$, independently on each E_i.
- We adopt a filtering formulation, inferring $\sigma(t_i, \cdot; \cdot)$ sequentially.
- Within each data assimilation step, the Bayesian update is performed by SMC with $P \in \mathbb{N}$ weighted particles and a pCN transition kernel (which uses point evaluations of σ directly and avoids truncation of the KL expansion).
- Real-world data obtained at 49 regular time intervals: rapid injection between frames 10 and 11, followed by diffusion and rotation of the liquids.

EIT Static Recovery I

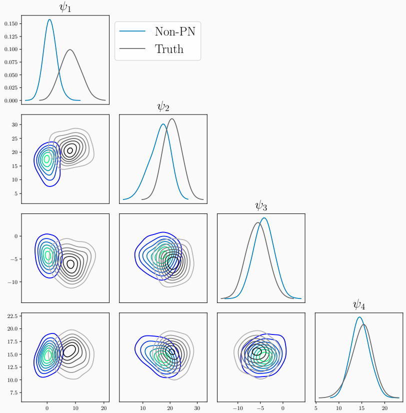


Figure 4: A small number $n_{\mathcal{A}} + n_{\mathcal{B}} = 71$ of collocation points was used to discretise the PDE, but the uncertainty due to discretisation was not modelled. The reference posterior distribution over the coefficients ψ_k is plotted (grey) and compared to the approximation to the posterior obtained when the PDE is discretised and the discretisation error is not modelled (blue, 'Non-PN'). The approximate posterior is highly biased.

EIT Static Recovery ii

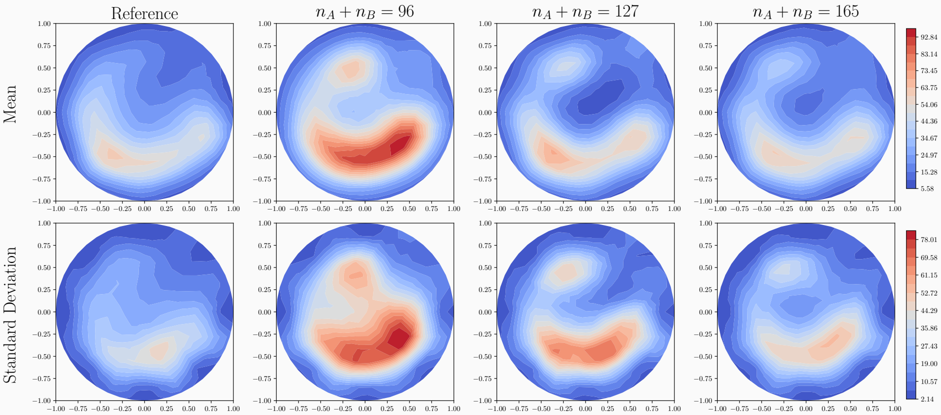


Figure 5: Posterior means and standard-deviations for the recovered conductivity field at $t = 14$. The first column shows the reference solution, obtained using symmetric collocation with a large number of collocation points. The remaining columns show the recovered field when PMM is used with $n_{\mathcal{A}} + n_{\mathcal{B}}$ collocation points.

EIT Dynamic Recovery

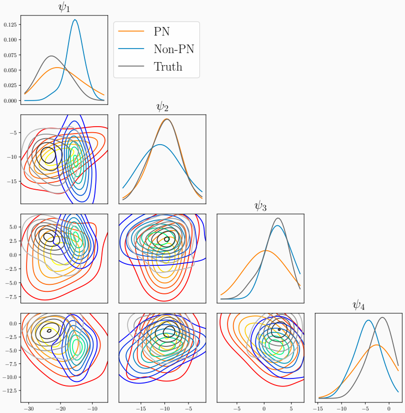


Figure 6: Posterior distribution over the coefficients ψ_k at the final time. A small number $n_{\mathcal{A}} + n_{\mathcal{B}} = 71$ of collocation points was used to discretise the PDE. The reference posterior distribution over the coefficients ψ_k is plotted (grey) and compared to the approximation to the posterior obtained when discretisation of the PDE is not modelled (blue, 'Non-PN') and modelled (orange, 'PN').

EIT Comments

- Typically PDE discretisation error in BIPs is ignored, or its contribution is bounded through detailed numerical analysis (Schwab and Stuart, 2012). Theoretical bounds are difficult in the temporal setting due to propagation and accumulation of errors
- As a modelling choice, the PN approach eases these difficulties. As with the Painlevé example, this is a statistically correct implementation of the assumptions, but it is (at present) costly. ✓/✗
- Furthermore, Markov temporal evolution of the conductivity field was assumed; this is likely incorrect, since time derivatives of this field will vary continuously. Even a-priori knowledge about the spin direction is neglected at present. ✗
- Again, we see a need for priors that are ‘physically reasonable’ and statistically/computationally appropriate. ⁉

Closing Remarks

Closing Remarks

- Numerical methods can be characterised in a Bayesian fashion. ✓
- This does not coincide with average-case analysis and IBC. ✓
- BPNMs can be composed into pipelines, e.g. for inverse problems. ✓
- Bayes' rule as disintegration → (expensive!) numerical implementation. ✓/✗
    - Lots of room to improve computational cost and bias. ⁉
    - Departures from the "Bayesian gold standard" can be assessed in terms of cost-accuracy tradeoff. ⁉
- How to choose/design an appropriate (numerically-analytically right) prior? ⁉

Cockayne, Oates, Sullivan, and Girolami (2017b) arXiv:1702.03673.

Bayesian probabilistic numerical methods

Lie, Sullivan, and Teckentrup (2018) arXiv:1712.05717.

Random forward models and log-likelihoods in Bayesian inverse problems *SIAM/ASA J. Uncertain. Quantif.* To appear.

Closing Remarks

- Numerical methods can be characterised in a Bayesian fashion. ✓
- This does not coincide with average-case analysis and IBC. ✓
- BPNMs can be composed into pipelines, e.g. for inverse problems. ✓
- Bayes' rule as disintegration → (expensive!) numerical implementation. ✓/✗
  - Lots of room to improve computational cost and bias. ⁉
  - Departures from the "Bayesian gold standard" can be assessed in terms of cost-accuracy tradeoff. ⁉
- How to choose/design an appropriate (numerically-analytically right) prior? ⁉

Cockayne, Oates, Sullivan, and Girolami (2017b) arXiv:1702.03673.

Bayesian probabilistic numerical methods

Lie, Sullivan, and Teckentrup (2018) arXiv:1712.05717.

Random forward models and log-likelihoods in Bayesian inverse problems *SIAM/ASA J. Uncertain. Quantif.* To appear.

Thank You

References i

N. L. Ackerman, C. E. Freer, and D. M. Roy. On computability and disintegration. *Math. Structures Comput. Sci.*, 27(8): 1287–1314, 2017. doi:10.1017/S0960129516000098.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 78(5):1103–1130, 2016. doi:10.1111/rssb.12158.

J. T. Chang and D. Pollard. Conditioning as disintegration. *Statist. Neerlandica*, 51(3):287–317, 1997. doi:10.1111/1467-9574.00056.

K.-S. Cheng, D. Isaacson, J. C. Newell, and D. G. Gisser. Electrode models for electric current computed tomography. *IEEE Trans. Biomed. Eng.*, 36(9), 1989. doi:10.1109/10.35300.

J. Cockayne, C. J. Oates, T. J. Sullivan, and M. Girolami. Probabilistic meshless methods for partial differential equations and Bayesian inverse problems, 2016. arXiv:1605.07811.

J. Cockayne, C. J. Oates, T. J. Sullivan, and M. Girolami. Probabilistic numerical methods for PDE-constrained Bayesian inverse problems. In G. Verdoolaege, editor, *Proceedings of the 36th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 1853 of *AIP Conference Proceedings*, pages 060001–1–060001–8, 2017a. doi:10.1063/1.4985359.

J. Cockayne, C. J. Oates, T. J. Sullivan, and M. Girolami. Bayesian probabilistic numerical methods, 2017b. arXiv:1702.03673.

P. R. Conrad, M. Girolami, S. Särkkä, A. M. Stuart, and K. C. Zygalakis. Statistical analysis of differential equations: introducing probability measures on numerical solutions. *Stat. Comput.*, 27(4), 2016. doi:10.1007/s11222-016-9671-0.

References ii

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.

M. M. Dunlop and A. M. Stuart. The Bayesian formulation of EIT: analysis and algorithms. *Inv. Probl. Imaging*, 10(4): 1007–1036, 2016a. doi:10.3934/ipi.2016030.

M. M. Dunlop and A. M. Stuart. MAP estimators for piecewise continuous inversion. *Inv. Probl.*, 32(10):105003, 50, 2016b. doi:10.1088/0266-5611/32/10/105003.

M. Giry. A categorical approach to probability theory. In *Categorical aspects of topology and analysis (Ottawa, Ont., 1980)*, volume 915 of *Lecture Notes in Math.*, pages 68–85. Springer, Berlin-New York, 1982.

J. Gutierrez, T. Dyakowski, M. Beck, and R. Williams. Using electrical impedance tomography for controlling hydrocyclone underflow discharge. 108(2):180–184, 2000.

P. E. Jacob, L. M. Murray, C. C. Holmes, and C. P. Robert. Better together? Statistical learning in models made of modules, 2017. arXiv:1708.08719.

J. B. Kadane and G. W. Wasilkowski. Average case ϵ-complexity in computer science. A Bayesian view. In *Bayesian Statistics, 2 (Valencia, 1983)*, pages 361–374. North-Holland, Amsterdam, 1985.

F. M. Larkin. Optimal approximation in Hilbert spaces with reproducing kernel functions. *Math. Comp.*, 24:911–921, 1970. doi:10.2307/2004625.

S. Lauritzen. *Graphical Models*. Oxford University Press, 1991.

H. C. Lie, A. M. Stuart, and T. J. Sullivan. Strong convergence rates of probabilistic integrators for ordinary differential equations, 2017. arXiv:1703.03680.

References iii

H. C. Lie, T. J. Sullivan, and A. L. Teckentrup. Random forward models and log-likelihoods in Bayesian inverse problems. *SIAM/ASA J. Uncertain. Quantif.*, 2018. To appear. arXiv:1712.05717.

C. J. Oates, J. Cockayne, and R. G. Ackroyd. Bayesian probabilistic numerical methods for industrial process monitoring, 2017. arXiv:1707.06107.

A. O'Hagan. Monte Carlo is fundamentally unsound. *Statistician*, 36(2/3):247–249, 1987. doi:10.2307/2348519.

H. Owhadi and C. Scovel. Conditioning Gaussian measure on Hilbert space, 2015. arXiv:1506.04208.

H. Owhadi, C. Scovel, and T. J. Sullivan. Brittleness of Bayesian inference under finite information in a continuous world. *Electron. J. Stat.*, 9(1):1–79, 2015. doi:10.1214/15-EJS989.

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

K. Ritter. *Average-Case Analysis of Numerical Problems*, volume 1733 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000. doi:10.1007/BFb0103934.

A. Sard. Best approximate integration formulas; best approximation formulas. *Amer. J. Math.*, 71:80–91, 1949. doi:10.2307/2372095.

C. Schillings and C. Schwab. Scaling limits in computational Bayesian inversion. *ESAIM Math. Model. Numer. Anal.*, 50(6): 1825–1856, 2016. doi:10.1051/m2an/2016005.

C. Schwab and A. M. Stuart. Sparse deterministic approximation of Bayesian inverse problems. *Inv. Probl.*, 28(4):045003, 32, 2012. doi:10.1088/0266-5611/28/4/045003.

References IV

J. Skilling. Bayesian solution of ordinary differential equations. In C. R. Smith, G. J. Erickson, and P. O. Neudorfer, editors, *Maximum Entropy and Bayesian Methods*, volume 50 of *Fundamental Theories of Physics*, pages 23–37. Springer, 1992. doi:10.1007/978-94-017-2219-3.

E. Somersalo, M. Cheney, and D. Isaacson. Existence and uniqueness for electrode models for electric current computed tomography. *SIAM J. Appl. Math.*, 52(4):1023–1040, 1992. doi:10.1137/0152060.

A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451–559, 2010. doi:10.1017/S0962492910000061.

A. V. Sul′din. Wiener measure and its applications to approximation methods. I. *Izv. Vysš. Učebn. Zaved. Matematika*, 6(13): 145–158, 1959.

A. V. Sul′din. Wiener measure and its applications to approximation methods. II. *Izv. Vysš. Učebn. Zaved. Matematika*, 5(18): 165–179, 1960.

T. Tjur. *Probability Based on Radon Measures*. John Wiley & Sons, Ltd., Chichester, 1980. Wiley Series in Probability and Mathematical Statistics.

J. F. Traub, G. W. Wasilkowski, and H. Woźniakowski. *Information-Based Complexity*. Computer Science and Scientific Computing. Academic Press, Inc., Boston, MA, 1988. With contributions by A. G. Werschulz and T. Boult.

L. N. Trefethen. Is Gauss quadrature better than Clenshaw–Curtis? *SIAM Rev.*, 50(1):67–87, 2008. doi:10.1137/060659831.

G. Uhlmann. Electrical impedance tomography and Calderón's problem. *Inv. Probl.*, 25(12):123011, 39, 2009. doi:10.1088/0266-5611/25/12/123011.

A. Zellner. Optimal information processing and Bayes's theorem. *Amer. Statist.*, 42(4):278–284, 1988. doi:10.2307/2685143.