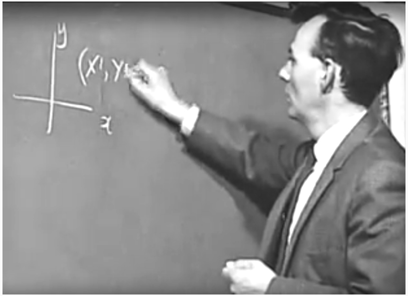
Bayesian probabilistic numerical methods

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

Uncertainty Quantification, Machine Learning and Bayesian Statistics in Scientific Computing
Ruprecht-Karls-Universität Heidelberg, DE, 1–5 July 2019

[1]University of Warwick, UK
[2]University of Cambridge, UK
[3]Alan Turing Institute, London, UK
[4]Newcastle University, UK
[5]**Freie Universität Berlin, DE**
[6]**Zuse Institute Berlin, DE**

> *“Numerical analysts and statisticians are both in the business of estimating parameter values from incomplete information. The two disciplines have separately developed their own approaches to formalizing strangely similar problems and their own solution techniques; the author believes they have much to offer each other.”*
> — F. M. Larkin (1979)



OVERVIEW: PNMS AND BIPS

- There are many reasons to consider a **probabilistic/statistical perspective on the analysis and design of numerical methods**, and even to return probabilistic solutions to deterministic forward problems like quadrature / DE solution.
- In various forms, these ideas have a long history.

  → Oates and Sullivan (2019) *Stat. Comp.* arXiv:1901.04457
- What are probabilistic numerical methods (PNMs) and in what sense can they be **Bayesian**?

  → Cockayne et al. (2019) *SIAM Rev.* arXiv:1702.03673
- A Bayesian interpretation of forward problems is especially appealing for Bayesian inverse problems (BIPs), since then both the forward and inverse problem "speak the same language", without spurious posterior over-concentration.
- How does their use connect to established theory for BIPs?

  → Lie et al. (2018) *SIAM/ASA JUQ* arXiv:1712.05717

MOTIVATING EXAMPLE: 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., 2019).
- 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 $\sigma \geq 0$.

MOTIVATING EXAMPLE: FITZHUGH–NAGUMO ODE INFERENCE

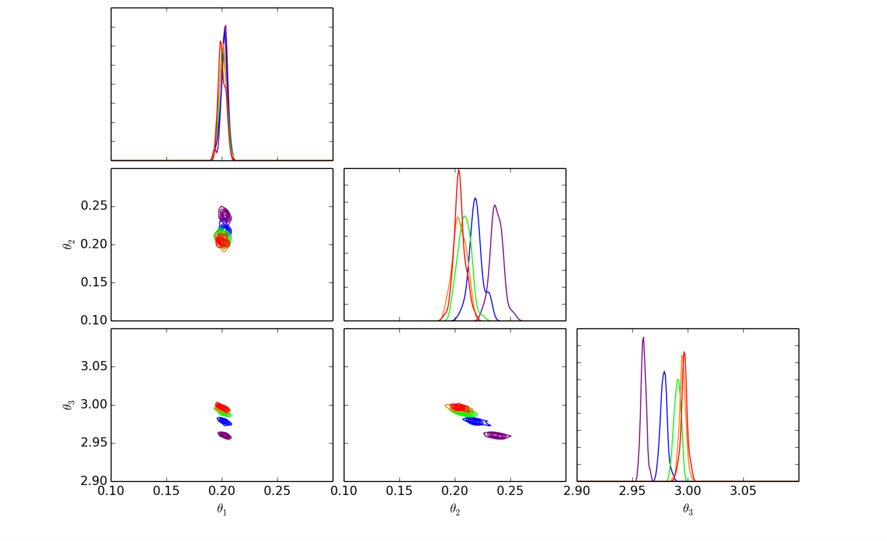


Figure 1: The deterministic posteriors (i.e. $\sigma = 0$) are over-confident at all values of the time step $\tau = 0.1, 0.05, 0.02, 0.01, 0.005$, often do not overlap, and are biased.

MOTIVATING EXAMPLE: FITZHUGH–NAGUMO ODE INFERENCE

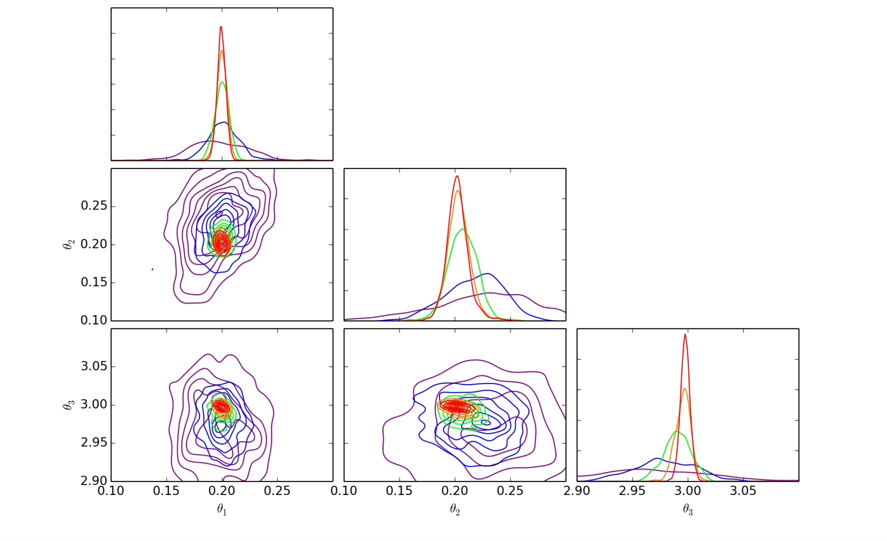


Figure 1: In contrast, the PN-Euler posteriors (here with $\sigma = 1/5$) for $\tau = 0.1, 0.05, 0.02, 0.01, 0.005$ are less confident and overlap more, though are still biased.

OUTLINE

1. Numerics: An inference perspective
2. Optimal information
3. Disintegration
4. Coherent pipelines of PNMs, and Bayesian inverse problems
5. Applications
6. Closing remarks

AN INFERENCE PERSPECTIVE ON NUMERICAL TASKS

AN ABSTRACT VIEW OF NUMERICAL METHODS I

An abstraction of a numerical task consists of three spaces and three functions:

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

AN ABSTRACT VIEW OF NUMERICAL METHODS I

An abstraction of a numerical task consists of three spaces and three functions:

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

Example (Quadrature)

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

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

AN ABSTRACT VIEW OF NUMERICAL METHODS I

An abstraction of a numerical task consists of three spaces and three functions:

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

Example (Quadrature)

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

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

- Conventional numerical methods are cleverly-designed functions $\mathsf{B} \colon \mathcal{Y} \to \mathcal{Q}$: such a method **“believes”** that $\mathsf{Q}(u) \approx \mathsf{B}(\mathsf{Y}(u))$.

AN ABSTRACT VIEW OF NUMERICAL METHODS II

Example (Quadrature)

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

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

- Conventional numerical methods are cleverly-designed functions $\mathsf{B} \colon \mathcal{Y} \to \mathcal{Q}$: such a method **“believes”** that $\mathsf{Q}(u) \approx \mathsf{B}(\mathsf{Y}(u))$.

AN ABSTRACT VIEW OF NUMERICAL METHODS II

Example (Quadrature)

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

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

- Conventional numerical methods are cleverly-designed functions $\mathsf{B} \colon \mathcal{Y} \to \mathcal{Q}$: such a method **“believes”** that $\mathsf{Q}(u) \approx \mathsf{B}(\mathsf{Y}(u))$.
- N.B. *Some* methods try to invert Y, form an estimate of u, then apply Q, but not all do!
  - E.g. the trapezoidal rule does estimate u:

$$\mathsf{B}_{\mathrm{trap}}\big((t_j, z_j)_{j=1}^J\big) := \sum_{j=1}^{J-1} \frac{z_{j+1} + z_j}{2}(t_{j+1} - t_j) = z_1 \frac{t_2 - t_1}{2} + \sum_{j=2}^{J-1} z_j \frac{t_{j+1} - t_{j-1}}{2} + z_J \frac{t_J - t_{J-1}}{2}.$$

  - E.g. vanilla Monte Carlo does not estimate u! (cf. O’Hagan, 1987)

$$\mathsf{B}_{\mathrm{MC}}\big((t_i, z_i)_{i=1}^n\big) := \frac{1}{n} \sum_{i=1}^n z_i$$

AN ABSTRACT VIEW OF NUMERICAL METHODS III

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

$$e_{\mathrm{WC}} := \sup_{u \in \mathcal{U}} \| \mathsf{B}(\mathsf{Y}(u)) - \mathsf{Q}(u) \|_{\mathcal{Q}}.$$

    - The **average-case error** (Ritter, 2000) with respect to a probability measure $\mu \in \mathcal{P}_{\mathcal{U}}$:

$$e_{\mathrm{AC}} := \int_{\mathcal{U}} \| \mathsf{B}(\mathsf{Y}(u)) - \mathsf{Q}(u) \|_{\mathcal{Q}} \, \mu(\mathrm{d}u).$$

    - To a **Bayesian**, seeing the additional structure of μ, there is only one way forward: if $u \sim \mu$, then $\mathsf{B}(\mathsf{Y}(u))$ 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

$$\begin{tikzcd}
\mathcal{U} \arrow[r, bend left, "\mathsf{Y}"] \arrow[d, bend right, "\mathsf{Q}"'] & \mathcal{Y} \\
\mathcal{Q} &
\end{tikzcd}$$

REV. BAYES DOES SOME NUMERICS I

$$\begin{tikzcd}
\mathcal{U} \arrow[r, bend left, "\mathsf{Y}"] \arrow[d, bend right, "\mathsf{Q}"'] & \mathcal{Y} \arrow[dl, dashed, "\mathsf{B}"] \\
\mathcal{Q} &
\end{tikzcd}$$

$$\mathsf{B} \colon \mathcal{Y} \to \mathcal{Q}$$

Example (Quadrature)

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

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

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

Rev. Bayes does some numerics I

$$\begin{tikzcd}
\mathcal{U} \arrow[r, bend left, "\mathsf{Y}"] \arrow[d, bend right, "\mathsf{Q}"'] & \mathcal{Y} \arrow[dl, dashed, "\mathsf{B}"] \\
\mathcal{Q} &
\end{tikzcd}$$

$$\mathsf{B} \colon \mathcal{Y} \to \mathcal{Q}$$

Go Probabilistic!

$$\mu \in \mathcal{P}_{\mathcal{U}}$$

$$(\mathsf{Y}_\sharp \mu)(E) := \mu(\mathsf{Y}^{-1}(E))$$

$$\begin{tikzcd}
\mathcal{P}_{\mathcal{U}} \arrow[r, bend left, "\mathsf{Y}_\sharp"] \arrow[d, bend right, "\mathsf{Q}_\sharp"'] & \mathcal{P}_{\mathcal{Y}} \arrow[dl, dashed, "\mathsf{B}_\sharp"] \\
\mathcal{P}_{\mathcal{Q}} & \mathcal{Y} \arrow[u, bend right, "\delta"']
\end{tikzcd}$$

average-case performance of B?

Example (Quadrature)

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

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

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

REV. BAYES DOES SOME NUMERICS I

$$\begin{tikzcd}
\mathcal{U} \arrow[r, bend left, "\mathsf{Y}"] \arrow[d, bend right, "\mathsf{Q}"'] & \mathcal{Y} \arrow[dl, dashed, "\mathsf{B}"] \\
\mathcal{Q} &
\end{tikzcd}$$

$$\mathsf{B} \colon \mathcal{Y} \to \mathcal{Q}$$

Go Probabilistic!

$$\mu \in \mathcal{P}_{\mathcal{U}}$$

$$(\mathsf{Y}_\sharp \mu)(E) := \mu(\mathsf{Y}^{-1}(E))$$

$$\begin{tikzcd}
\mathcal{P}_{\mathcal{U}} \arrow[r, dotted, bend left, "\mathsf{Y}_\sharp"] \arrow[d, bend right, "\mathsf{Q}_\sharp"'] & \mathcal{P}_{\mathcal{Y}} \\
\mathcal{P}_{\mathcal{Q}} & \mathcal{Y} \arrow[l, dashed, "y \mapsto \beta(\mu, y)"] \arrow[u, dotted, "\delta"']
\end{tikzcd}$$

$$\beta(\mu, \cdot) \colon \mathcal{Y} \to \mathcal{P}_{\mathcal{Q}}$$

Example (Quadrature)

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

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

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

$$\begin{tikzcd}
\mathcal{U} \arrow[r, bend left, "\mathsf{Y}"] \arrow[d, bend right, "\mathsf{Q}"'] & \mathcal{Y} \arrow[dl, dashed, "\mathsf{B}"] \\
\mathcal{Q} &
\end{tikzcd}$$

$$\mathsf{B} \colon \mathcal{Y} \to \mathcal{Q}$$

Go Probabilistic!

$$\mu \in \mathcal{P}_{\mathcal{U}}$$

$$(\mathsf{Y}_{\sharp}\mu)(E) := \mu(\mathsf{Y}^{-1}(E))$$

$$\begin{tikzcd}
\mathcal{P}_{\mathcal{U}} \arrow[r, dotted, bend left, "\mathsf{Y}_{\sharp}"] \arrow[d, bend right, "\mathsf{Q}_{\sharp}"'] & \mathcal{P}_{\mathcal{Y}} \\
\mathcal{P}_{\mathcal{Q}} & \mathcal{Y} \arrow[ul, "y \mapsto \mu^{y}"'] \arrow[l, dashed, "y \mapsto \beta(\mu, y)"] \arrow[u, dotted, "\delta"']
\end{tikzcd}$$

$$\beta(\mu, \cdot) \colon \mathcal{Y} \to \mathcal{P}_{\mathcal{Q}}$$

Definition (Bayesian PNM)

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

$$\beta(\mu, y) = \mathsf{Q}_{\sharp}\mu^{y}, \qquad \text{for } \mathsf{Y}_{\sharp}\mu\text{-almost all } y \in \mathcal{Y}.$$

REV. BAYES DOES SOME NUMERICS II

Definition (Bayesian PNM)

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

$$\beta(\mu, y) = \mathsf{Q}_{\sharp}\mu^{y}, \qquad \text{for } \mathsf{Y}_{\sharp}\mu\text{-almost all } y \in \mathcal{Y}.$$

Rev. Bayes does some numerics II

Definition (Bayesian PNM)

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

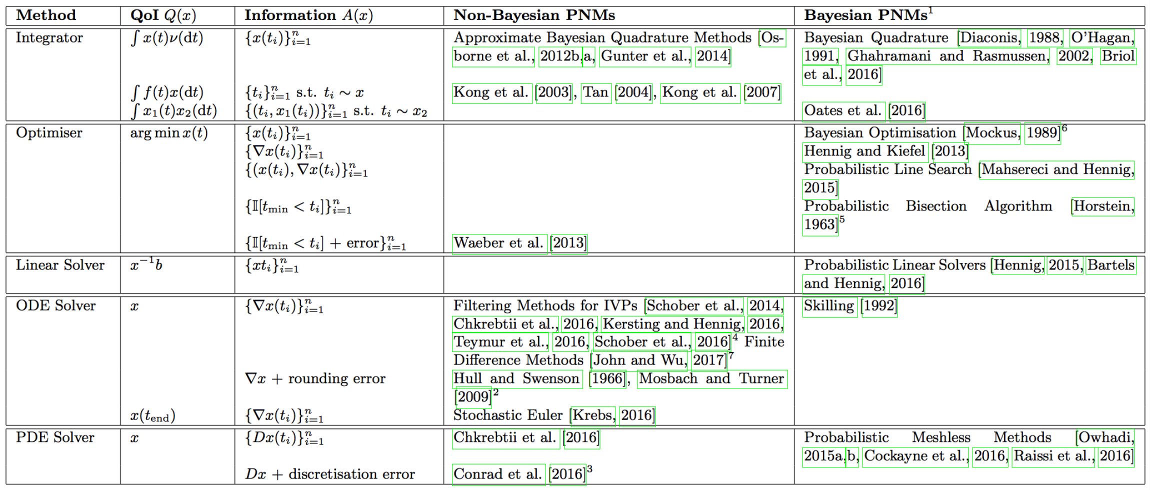
$$\beta(\mu, y) = \mathsf{Q}_\sharp \mu^y, \qquad \text{for } \mathsf{Y}_\sharp \mu\text{-almost all } y \in \mathcal{Y}.$$

Example

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

For technical reasons, “conditioning” here is meant in the sense of disintegration, as advocated by e.g. Chang and Pollard (1997).

A ROGUE'S GALLERY OF BAYESIAN AND NON-BAYESIAN PNMS (2017)



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]	

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

MEASURES OF ERROR / LOSS

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 $\mathsf{B} \colon \mathcal{Y} \to \mathcal{Q}$ is

$$e_{\mathrm{WC}}(\mathsf{Y}, \mathsf{B}) := \sup_{u \in \mathcal{U}} L\bigl(\mathsf{B}(\mathsf{Y}(u)), \mathsf{Q}(u)\bigr).$$

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

$$e_{\mathrm{AC}}(\mathsf{Y}, \mathsf{B}) := \int_{\mathcal{U}} L\bigl(\mathsf{B}(\mathsf{Y}(u)), \mathsf{Q}(u)\bigr)\, \mu(\mathrm{d}u),$$

$$e_{\mathrm{AC}}(\mathsf{Y}, \beta) := \int_{\mathcal{U}} \left[\int_{\mathcal{Q}} L(q, \mathsf{Q}(u))\, \beta(\mu, \mathsf{Y}(u))(\mathrm{d}q) \right] \mu(\mathrm{d}u).$$

- Kadane and Wasilkowski (1985) show that the minimisers are *deterministic* decision rules B, and the minimiser Y is “optimal information” for this task.
- A BPNM β has “no choice” but to be $\mathsf{Q}_{\sharp}\mu^{y}$ once $\mathsf{Y}(u) = y$ is given; optimality of Y means minimising the **Bayesian loss**

$$e_{\mathrm{BPN}}(\mathsf{Y}) := \int_{\mathcal{U}} \left[\int_{\mathcal{Q}} L(q, \mathsf{Q}(u))\, (\mathsf{Q}_{\sharp}\mu^{\mathsf{Y}(u)})(\mathrm{d}q) \right] \mu(\mathrm{d}u).$$

OPTIMAL INFORMATION: AC = BPN?

Theorem (AC = BPN for quadratic loss; Cockayne et al., 2019)

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

OPTIMAL INFORMATION: AC = BPN?

Theorem (AC = BPN for quadratic loss; Cockayne et al., 2019)

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

Theorem (AC ≠ BPN in general; Oates et al. (2019b))

If $\mathcal{U}$ *can be partitioned into three sets of positive probability, then there exists a choice of QoI and loss so that optimal information for BPNM and AC differ.*

OPTIMAL INFORMATION: AC = BPN?

Theorem (AC = BPN for quadratic loss; Cockayne et al., 2019)

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

Example (AC ≠ BPN in general; Oates et al. (2019b))

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 (Y_1) or is non-♣ (Y_2).

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

$$\mathcal{Y}_1 = \{0, 1\} \qquad \mathsf{Y}_1(u) = \mathbb{1}[u \in \{♦, ♥\}] \qquad \mathsf{Q}(u) = \mathbb{1}[u = ♦]$$

$$\mathcal{Y}_2 = \{0, 1\} \qquad \mathsf{Y}_2(u) = \mathbb{1}[u \in \{♦, ♥, ♠\}] \qquad L(q, q') = \mathbb{1}[q \neq q']$$

Which information operator, Y_1 or Y_2, is better? (Note that $e_{\mathrm{WC}}(\mathsf{Y}_i, \mathsf{B}) = 1$ for all deterministic b!)

OPTIMAL INFORMATION: AC $\neq$ BPN!

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

$\mathsf{Y}_1(u) = \blacksquare$ vs. 🟥 $\quad$ $\mathsf{Y}(u) = \mathbb{1}[u = \blacklozenge]$

$\mathsf{Y}_2(u) = \neg\clubsuit$ vs. $\clubsuit$ $\quad$ $L(q, q') = \mathbb{1}[q \neq q']$

$u =$	$\clubsuit$	$\blacklozenge$	$\heartsuit$	$\spadesuit$	
$e_{\mathrm{AC}}(\mathsf{Y}_1, \mathsf{B}) = \frac{1}{4}\Big($	$L(\mathsf{B}(\blacksquare), 0)$ +	$L(\mathsf{B}($🟥$), 1)$ +	$L(\mathsf{B}($🟥$), 0)$ +	$L(\mathsf{B}(\blacksquare), 0)$	$\Big)$

OPTIMAL INFORMATION: AC $\neq$ BPN!

$\mathcal{U} = \{\clubsuit, \text{♦}, \text{♥}, \spadesuit\}$ $\quad \mu = \mathrm{Unif}_{\mathcal{U}}$ $\quad \mathcal{Q} = \{0, 1\} \subset \mathbb{R}$

$\mathsf{Y}_1(u) = \text{⬛ vs. 🟥}$ $\quad \mathsf{Y}(u) = \mathbb{1}[u = \text{♦}]$

$\mathsf{Y}_2(u) = \neg\clubsuit \text{ vs. } \clubsuit$ $\quad L(q, q') = \mathbb{1}[q \neq q']$

$u =$	$\clubsuit$		♦		♥		$\spadesuit$	
$e_{\mathrm{AC}}(\mathsf{Y}_1, \mathsf{B}) = \frac{1}{4}\big($	$L(\mathsf{B}(\text{⬛}), 0)$	$+$	$L(\mathsf{B}(\text{🟥}), 1)$	$+$	$L(\mathsf{B}(\text{🟥}), 0)$	$+$	$L(\mathsf{B}(\text{⬛}), 0)$	$\big)$
$e_{\mathrm{AC}}(\mathsf{Y}_1, 0) = \frac{1}{4}\big($	0	$+$	1	$+$	0	$+$	0	$\big) = \frac{1}{4}$
$e_{\mathrm{AC}}(\mathsf{Y}_1, \mathrm{id}) = \frac{1}{4}\big($	0	$+$	0	$+$	1	$+$	0	$\big) = \frac{1}{4}$

Optimal information: AC $\neq$ BPN!

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

$$\mathsf{Y}_1(u) = \blacksquare \text{ vs. } {\color{red}\blacksquare} \qquad \mathsf{Y}(u) = \mathbb{1}[u = \diamondsuit]$$

$$\mathsf{Y}_2(u) = \neg\clubsuit \text{ vs. } \clubsuit \qquad L(q,q') = \mathbb{1}[q \neq q']$$

$$\begin{array}{rcccccccccc}
u = & & \clubsuit & & \diamondsuit & & \heartsuit & & \spadesuit & & \\
\hline\hline
e_{\mathrm{AC}}(\mathsf{Y}_1, \mathsf{B}) = \frac{1}{4}(& & L(\mathsf{B}(\blacksquare), 0) & + & L(\mathsf{B}({\color{red}\blacksquare}), 1) & + & L(\mathsf{B}({\color{red}\blacksquare}), 0) & + & L(\mathsf{B}(\blacksquare), 0) &) & \\
e_{\mathrm{AC}}(\mathsf{Y}_1, 0) = \frac{1}{4}(& & 0 & + & 1 & + & 0 & + & 0 &) & = \frac{1}{4} \\
e_{\mathrm{AC}}(\mathsf{Y}_1, \mathrm{id}) = \frac{1}{4}(& & 0 & + & 0 & + & 1 & + & 0 &) & = \frac{1}{4} \\
\hline
e_{\mathrm{AC}}(\mathsf{Y}_2, \mathsf{B}) = \frac{1}{4}(& & L(\mathsf{B}(\clubsuit), 0) & + & L(\mathsf{B}(\neg\clubsuit), 1) & + & L(\mathsf{B}(\neg\clubsuit), 0) & + & L(\mathsf{B}(\neg\clubsuit), 0) &) & \\
e_{\mathrm{AC}}(\mathsf{Y}_2, 0) = \frac{1}{4}(& & 0 & + & 1 & + & 0 & + & 0 &) & = \frac{1}{4} \\
\hline\hline
\end{array}$$

Optimal information: AC $\neq$ BPN!

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

$$\mathsf{Y}_1(u) = \blacksquare \text{ vs. } \textcolor{red}{\blacksquare} \qquad \mathsf{Y}(u) = \mathbb{1}[u = \diamondsuit]$$

$$\mathsf{Y}_2(u) = \neg\clubsuit \text{ vs. } \clubsuit \qquad L(q,q') = \mathbb{1}[q \neq q']$$

$$\begin{array}{rccccccccccc}
u = & & \clubsuit & & \diamondsuit & & \heartsuit & & \spadesuit & & \\
\hline\hline
e_{\mathrm{AC}}(\mathsf{Y}_1, \mathsf{B}) = & \frac{1}{4}(& L(\mathsf{B}(\blacksquare),0) & + & L(\mathsf{B}(\textcolor{red}{\blacksquare}),1) & + & L(\mathsf{B}(\textcolor{red}{\blacksquare}),0) & + & L(\mathsf{B}(\blacksquare),0) &) & \\
e_{\mathrm{AC}}(\mathsf{Y}_1, 0) = & \frac{1}{4}(& 0 & + & 1 & + & 0 & + & 0 &) & = \frac{1}{4} \\
e_{\mathrm{AC}}(\mathsf{Y}_1, \mathrm{id}) = & \frac{1}{4}(& 0 & + & 0 & + & 1 & + & 0 &) & = \frac{1}{4} \\
\hline
e_{\mathrm{AC}}(\mathsf{Y}_2, \mathsf{B}) = & \frac{1}{4}(& L(\mathsf{B}(\clubsuit),0) & + & L(\mathsf{B}(\neg\clubsuit),1) & + & L(\mathsf{B}(\neg\clubsuit),0) & + & L(\mathsf{B}(\neg\clubsuit),0) &) & \\
e_{\mathrm{AC}}(\mathsf{Y}_2, 0) = & \frac{1}{4}(& 0 & + & 1 & + & 0 & + & 0 &) & = \frac{1}{4} \\
\hline\hline
e_{\mathrm{BPN}}(\mathsf{Y}_1) = & \frac{1}{4}(& \mathbb{E}_{\mathsf{Q}_\sharp \mu \blacksquare} L(\cdot, 0) & + & \mathbb{E}_{\mathsf{Q}_\sharp \mu \textcolor{red}{\blacksquare}} L(\cdot, 1) & + & \mathbb{E}_{\mathsf{Q}_\sharp \mu \textcolor{red}{\blacksquare}} L(\cdot, 0) & + & \mathbb{E}_{\mathsf{Q}_\sharp \mu \blacksquare} L(\cdot, 0) &) & \\
= & \frac{1}{4}(& (\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0) & + & (\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1) & + & (\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 0) & + & (\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0) &) & = \frac{1}{4} \\
\hline
\end{array}$$

Optimal information: AC $\neq$ BPN!

$$\mathcal{U} = \{\clubsuit, \vardiamondsuit, \varheartsuit, \spadesuit\} \qquad \mu = \mathrm{Unif}_{\mathcal{U}} \qquad \mathcal{Q} = \{0,1\} \subset \mathbb{R}$$

$$\mathsf{Y}_1(u) = \blacksquare \text{ vs. } \textcolor{red}{\blacksquare} \qquad \mathsf{Y}(u) = \mathbb{1}[u = \vardiamondsuit]$$

$$\mathsf{Y}_2(u) = \neg\clubsuit \text{ vs. } \clubsuit \qquad L(q,q') = \mathbb{1}[q \neq q']$$

$$
\begin{array}{rcccccccccc}
u = & & \clubsuit & & \vardiamondsuit & & \varheartsuit & & \spadesuit & & \\
\hline\hline
e_{\mathrm{AC}}(\mathsf{Y}_1, \mathsf{B}) = \frac{1}{4} & (& L(\mathsf{B}(\blacksquare), 0) & + & L(\mathsf{B}(\textcolor{red}{\blacksquare}), 1) & + & L(\mathsf{B}(\textcolor{red}{\blacksquare}), 0) & + & L(\mathsf{B}(\blacksquare), 0) &) & \\
e_{\mathrm{AC}}(\mathsf{Y}_1, 0) = \frac{1}{4} & (& 0 & + & 1 & + & 0 & + & 0 &) & = \frac{1}{4} \\
e_{\mathrm{AC}}(\mathsf{Y}_1, \mathrm{id}) = \frac{1}{4} & (& 0 & + & 0 & + & 1 & + & 0 &) & = \frac{1}{4} \\
\hline
e_{\mathrm{AC}}(\mathsf{Y}_2, \mathsf{B}) = \frac{1}{4} & (& L(\mathsf{B}(\clubsuit), 0) & + & L(\mathsf{B}(\neg\clubsuit), 1) & + & L(\mathsf{B}(\neg\clubsuit), 0) & + & L(\mathsf{B}(\neg\clubsuit), 0) &) & \\
e_{\mathrm{AC}}(\mathsf{Y}_2, 0) = \frac{1}{4} & (& 0 & + & 1 & + & 0 & + & 0 &) & = \frac{1}{4} \\
\hline\hline
e_{\mathrm{BPN}}(\mathsf{Y}_1) = \frac{1}{4} & (& \mathbb{E}_{\mathsf{Q}_\sharp \mu^{\blacksquare}} L(\cdot, 0) & + & \mathbb{E}_{\mathsf{Q}_\sharp \mu^{\textcolor{red}{\blacksquare}}} L(\cdot, 1) & + & \mathbb{E}_{\mathsf{Q}_\sharp \mu^{\textcolor{red}{\blacksquare}}} L(\cdot, 0) & + & \mathbb{E}_{\mathsf{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
e_{\mathrm{BPN}}(\mathsf{Y}_2) = \frac{1}{4} & (& \mathbb{E}_{\mathsf{Q}_\sharp \mu^{\clubsuit}} L(\cdot, 0) & + & \mathbb{E}_{\mathsf{Q}_\sharp \mu^{\neg\clubsuit}} L(\cdot, 1) & + & \mathbb{E}_{\mathsf{Q}_\sharp \mu^{\neg\clubsuit}} L(\cdot, 0) & + & \mathbb{E}_{\mathsf{Q}_\sharp \mu^{\neg\clubsuit}} L(\cdot, 0) &) & \\
= \frac{1}{4} & (& (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) &) & = \frac{1}{3}
\end{array}
$$

Disintegration:

Exact and numerical

DEFINING THE POSTERIOR

- The posterior μ^y is subtle to define precisely, since *heuristically* it is given by

$$\mu^y(\mathrm{d}u) \propto \mathbb{1}[\mathsf{Y}(u) = y]\, \mu(\mathrm{d}u)$$

- We have a 0-1 likelihood, and moreover the likelihood is zero μ-a.e.!
    - Numerical analysts usually think of function evaluations as noiseless, in contrast to the noisy observations that are typical in statistics.
    - E.g. what is the prior probability that a Brownian path interpolates given data?
- We cannot even express Bayes' formula in the form favoured by Stuart (2010),

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu}(u) = \frac{\mathbb{1}[\mathsf{Y}(u) = y]}{Z(y)},$$

  because μ^y is singular with respect to μ, the density on the LHS does not exist, and $Z(y) = 0$.
- One way to consistently condition on events of measure zero is to define the conditioning operation in terms of **disintegration**.

DISINTEGRATION I

Definition (Disintegration)

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

- (support) $\mu^{y}(\{u \in \mathcal{U} \mid \mathsf{Y}(u) = y\}) = 1$ for $\mathsf{Y}_{\#}\mu$-almost all $y \in \mathcal{Y}$;

and, for each measurable $f \colon \mathcal{U} \to [0, \infty)$, ($f = \mathbb{1}_{E}$, $E \subseteq \mathcal{U}$ will do)

- (measurability) $y \mapsto \int_{\mathcal{U}} f(u) \, \mu^{y}(\mathrm{d}u)$ is
- (conditioning/reconstruction/law of total probability)

$$\int_{\mathcal{U}} f(u) \, \mu(\mathrm{d}u) = \int_{\mathcal{Y}} \left[\int_{\mathcal{U}} f(u) \, \mu^{y}(\mathrm{d}u) \right] (\mathsf{Y}_{\#}\mu)(\mathrm{d}y).$$

(Closely related concept: a regular conditional probability is basically the same thing, but in a different coordinate system.)

DISINTEGRATION II

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

Let $\mathcal{U}$ be a metric space and let $\mu \in \mathcal{P}_{\mathcal{U}}$ be inner regular. If the Borel σ-algebra on $\mathcal{U}$ is countably generated and contains all singletons $\{y\}$ for $y \in \mathcal{Y}$, then there is an essentially unique disintegration $\{\mu^y\}_{y \in \mathcal{Y}}$ of μ with respect to Y. (If $\{\nu^y\}_{y \in \mathcal{Y}}$ is another such disintegration, then $\{y \in \mathcal{Y} \mid \mu^y \neq \nu^y\}$ is an $\mathsf{Y}_{\#}\mu$-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 $\mathsf{Y}(x_1, x_2) := x_1$ is just the family of measures μ^y, where μ^y has Lebesgue density $\rho(a, \cdot\,)/Z^y$ on the vertical line $\{(y, x_2) \mid x_2 \in \mathbb{R}\}$, and $Z^y := \int_{\mathbb{R}} \rho(y, x_2)\, \mathrm{d}x_2$.

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

NUMERICAL DISINTEGRATION I

- The exact disintegration “$\mu^{y}(\mathrm{d}u) \propto \mathbb{1}[\mathsf{Y}(u) = y]\, \mu(\mathrm{d}u)$” can be accessed numerically via relaxation, with approximation guarantees provided $y \mapsto \mu^{y}$ is “nice”, e.g. $\mathsf{Y}_{\sharp}\mu \in \mathcal{P}_{\mathcal{Y}}$ has a smooth Lebesgue density.
- Consider relaxed posterior $\mu_{\delta}^{y}(\mathrm{d}u) \propto \phi(\|\mathsf{Y}(u) - y\|_{\mathcal{Y}}/\delta)\, \mu(\mathrm{d}u)$ 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 $\mathcal{P}_{\mathcal{U}}$ associated to a normed space $\mathcal{F}$ of test functions $f \colon \mathcal{U} \to \mathbb{R}$ is

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

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

NUMERICAL DISINTEGRATION II

$$\text{``}\mu^{y}(\mathrm{d}u) \propto \mathbb{1}[\mathsf{Y}(u) = y]\, \mu(\mathrm{d}u)\text{''}$$
$$\mu^{y}_{\delta}(\mathrm{d}u) \propto \phi(\|\mathsf{Y}(u) - y\|_{\mathcal{Y}} / \delta)\, \mu(\mathrm{d}u)$$
$$d_{\mathcal{F}}(\mu, \nu) := \sup\{ |\mu(f) - \nu(f)| \,|\, \|f\|_{\mathcal{F}} \leq 1 \}$$

Theorem (Cockayne et al., 2019, Theorem 4.3)

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

$$\begin{aligned} & d_{\mathcal{F}}(\mu^{y}, \mu^{y'}) \leq C \cdot \|y - y'\|^{\gamma} && \text{for } y, y' \in \mathcal{Y} \\ \Longrightarrow \quad & d_{\mathcal{F}}(\mu^{y}, \mu^{y}_{\delta}) \leq C \cdot C_{\phi} \cdot \delta^{\gamma} && \text{for } \mathsf{Y}_{\sharp}\mu\text{-almost all } y \in \mathcal{Y}. \end{aligned}$$

Open question: when does the hypothesis, a quantitative version of the Tjur property (Tjur, 1980), actually hold? (Fixing y and varying y' is ok; having both y and y' free is hard.)

EXAMPLE: PAINLEVÉ'S FIRST TRANSCENDENTAL I

A simple but multivalent boundary value problem:

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

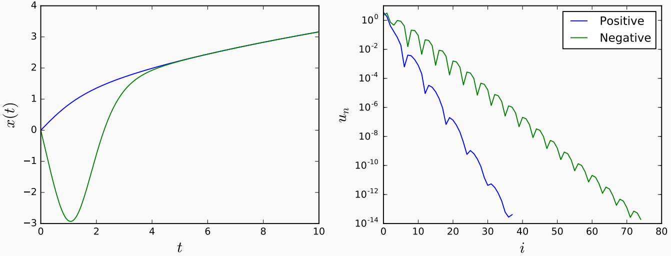


Figure 2: The two solutions of Painlevé's first transcendental and their spectra in the orthonormal Chebyshev polynomial basis over $[0, 10]$.

EXAMPLE: PAINLEVÉ'S FIRST TRANSCENDENTAL I

A simple but multivalent boundary value problem:

$$u''(t) - u(t)^2 = -t \qquad \text{for } t \geq 0$$

$$u(0) = 0$$

$$u(10) = \sqrt{10}$$

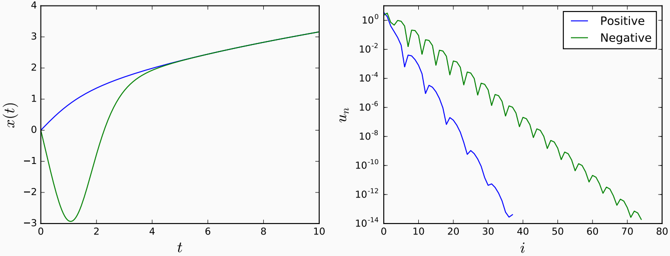
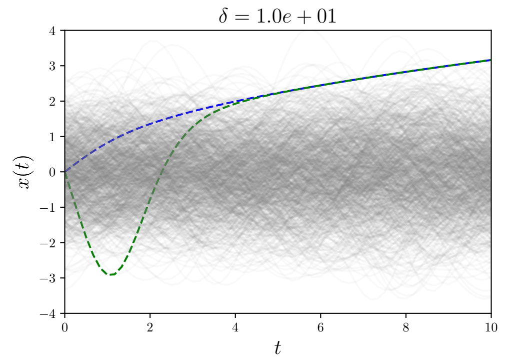


Figure 2: The two solutions of Painlevé's first transcendental and their spectra in the orthonormal Chebyshev polynomial basis over $[0, 10]$.

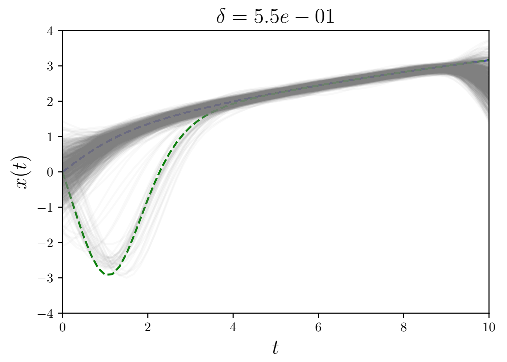
EXAMPLE: PAINLEVÉ'S FIRST TRANSCENDENTAL II

- 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… ✗



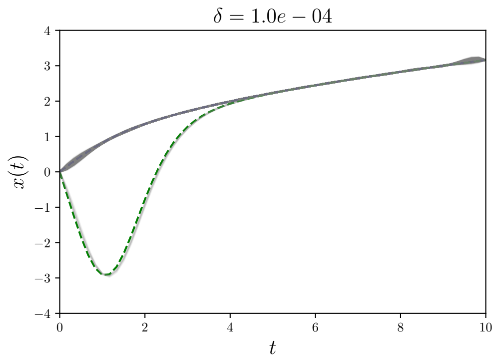
EXAMPLE: PAINLEVÉ'S FIRST TRANSCENDENTAL II

- 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... ✗



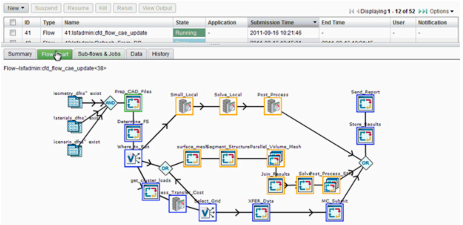
EXAMPLE: PAINLEVÉ'S FIRST TRANSCENDENTAL II

- 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 PNMs, and Bayesian inverse problems

COMPUTATIONAL PIPELINES



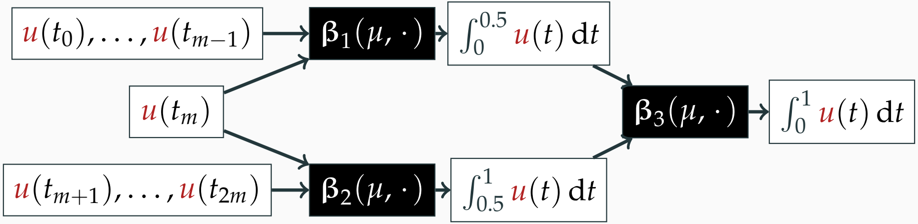
☐	ID	Type	Name	State	Application	Submission Time ▾	End Time	User	Notification
☐	41	Flow	41:lsfadmin:cfd_flow_cae_update	Running	-	2011-09-16 10:21:46	-	-	-

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. $\beta_2(\beta_1(\mu, y_1), y_2)$ is formally $\beta(\mu, (y_1, y_2))$… although figuring out what the spaces $\mathcal{U}_i$, $\mathcal{Y}_i$ and operators Y_i etc. are is a headache!

PIPELINE EXAMPLE: SPLIT INTEGRATION



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

$u(t_m)$

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

$\beta_1(\mu, \cdot)$

$\beta_2(\mu, \cdot)$

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

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

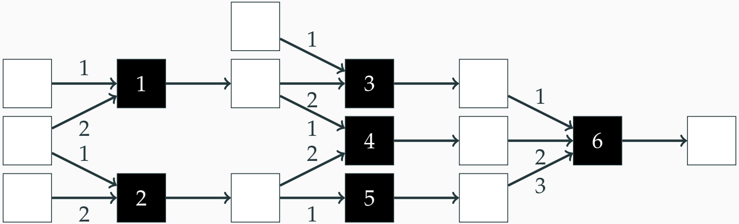
$\beta_3(\mu, \cdot)$

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

- Integrate a function over $[0, 1]$ in two steps using nodes $0 \le t_0 < \cdots < t_{m-1} < 0.5$, $t_m = 0.5$, and $t_{m+1} < \cdots < t_{2m} \le 1$.
- 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 $\int_0^{1} u(t)\,\mathrm{d}t$.

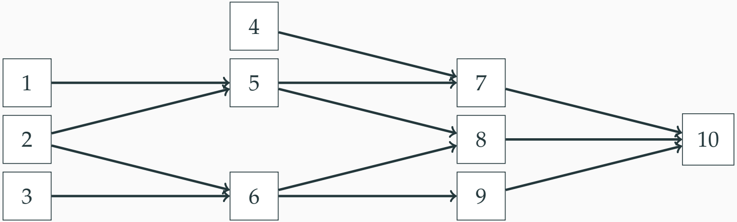
Coherence I

- 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 ■.



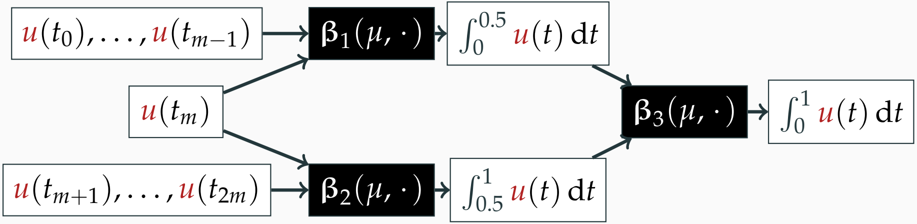
COHERENCE I

- 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



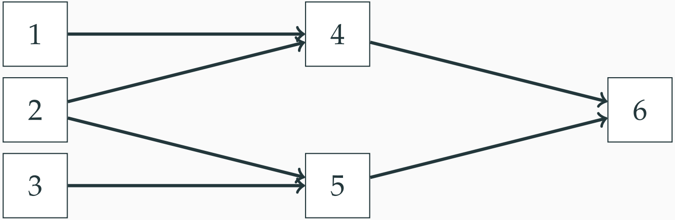
Definition

A prior μ is **coherent** for the dependency graph if — when the "leaf" input nodes are $\mathsf{Y}_\sharp \mu$-distributed and the remaining nodes are $\beta(\mu, \text{parents})$-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



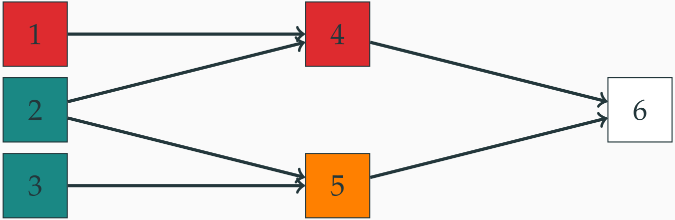
Definition

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

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

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 $\mathsf{Y}_\sharp\mu$-distributed and the remaining nodes are $\beta(\mu, \text{parents})$-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.

COHERENCY THEOREM

Theorem (Cockayne et al., 2019, 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 et al., 2019, Theorem 5.9)

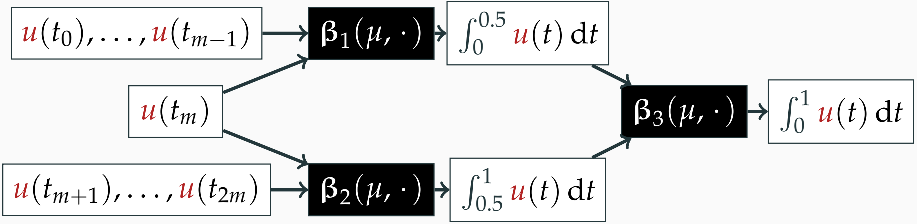
If a pipeline of PNMs is such that

- *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



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

$u(t_m)$

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

$\beta_1(\mu, \cdot)$

$\beta_2(\mu, \cdot)$

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

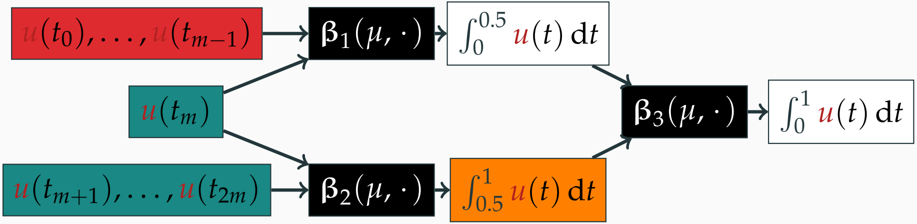
$\int_{0.5}^{1} u(t)\,\mathrm{d}t$

$\beta_3(\mu, \cdot)$

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

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

SPLIT INTEGRATION: COHERENCE



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

$u(t_m)$

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

$\beta_1(\mu, \cdot)$

$\beta_2(\mu, \cdot)$

$\beta_3(\mu, \cdot)$

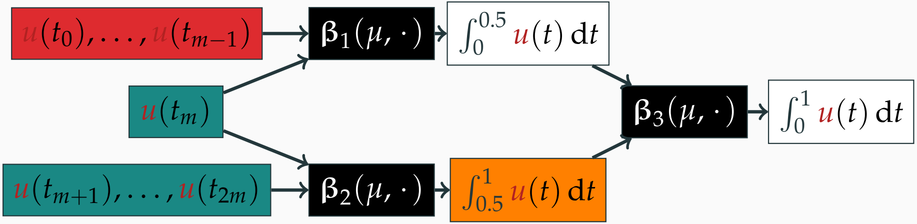
$\int_0^{0.5} u(t)\, \mathrm{d}t$

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

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

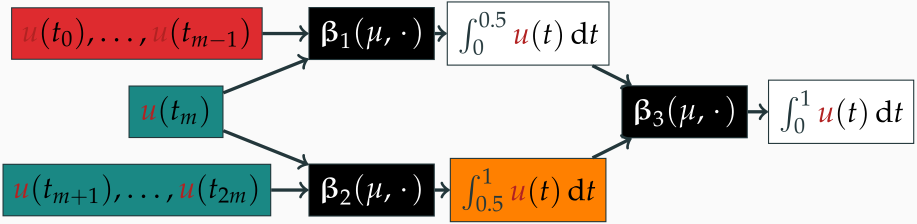
- Integrate a function over $[0, 1]$ in two steps using nodes $0 \leq t_0 < \cdots < t_{m-1} < 0.5$, $t_m = 0.5$, and $t_{m+1} < \cdots < t_{2m} \leq 1$.
- Is 🟧 $(\int_{0.5}^{1} u(t)\, \mathrm{d}t)$ independent of 🟥 $(u(t_0), \ldots, u(t_{m-1}))$ given 🟩 $(u(t_m), \ldots, u(t_{2m}))$?

Split integration: Coherence



- 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 🟥 $(u(t_0), \dots, u(t_{m-1}))$ 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.

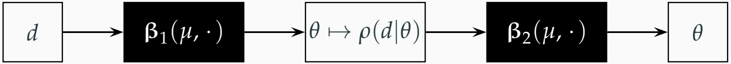
SPLIT INTEGRATION: COHERENCE



- 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. Jacob et al. (2017), and Lie et al. (2018). ⁉

SHORT PIPELINES: BAYESIAN INVERSE PROBLEMS

- A Bayesian inverse problem for recovering parameters $\theta \in \Theta$ from data $d \in \mathcal{D}$ can be represented as the *automatically coherent* two-stage computational pipeline



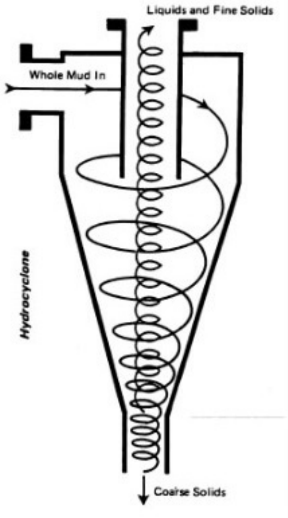
$$d \longrightarrow \beta_1(\mu, \cdot) \longrightarrow \theta \mapsto \rho(d|\theta) \longrightarrow \beta_2(\mu, \cdot) \longrightarrow \theta$$

- β_1 converts data d into the likelihood function for parameters θ, and hence incorporates any forward model such as an O/PDE solver.
- β_2 converts the prior on θ and the likelihood into a joint distribution for (θ, d), then conditions upon the actual observation — it returns something in $\mathcal{P}_{\Theta}$.
- β_1 conventionally has deterministic output in $\mathbb{R}^{\Theta}$; a bona fide PNM would 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.
- Alternative approach: assess sufficiency of forward solver accuracy for BIP purposes using Bayes factors (Capistrán et al., 2016; Christen et al., 2017).

APPLICATIONS

EXAMPLE: HYDROCYCLONES (OATES ET AL., 2019A)

- 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).



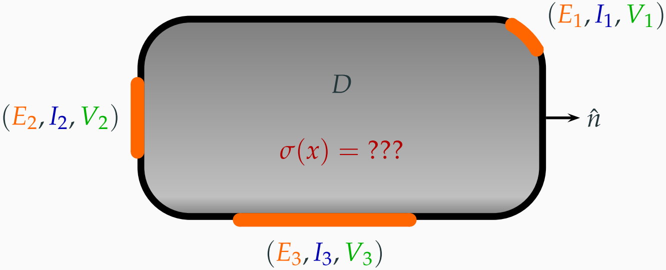
COMPLETE ELECTRODE MODEL (CHENG ET AL., 1989; SOMERSALO ET AL., 1992)

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

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

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

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 (PMM) of Cockayne et al. (2016, 2017):
  - a Gaussian process extension of symmetric collocation;
  - a Bayesian PNM 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 for the inverse problem, cf. Capistrán et al. (2016); Christen et al. (2017).

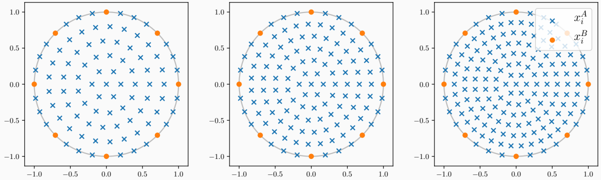


Figure 3: Like collocation, PMM imposes the PDE relation at $n_{\mathcal{A}}$ interior nodes and boundary conditions at $n_{\mathcal{B}}$ boundary nodes.

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

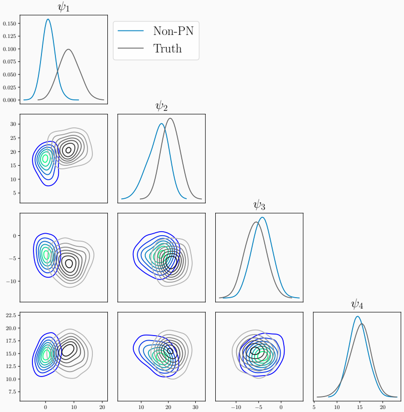


Figure 4: A small number $n_{\mathcal{A}} + n_{\mathcal{B}} = 71$ of collocation points was used to discretise the PDE, but the uncertainty due to discretisation was not modelled. The reference posterior distribution over the coefficients ψ_k is plotted (grey) and compared to the approximation to the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

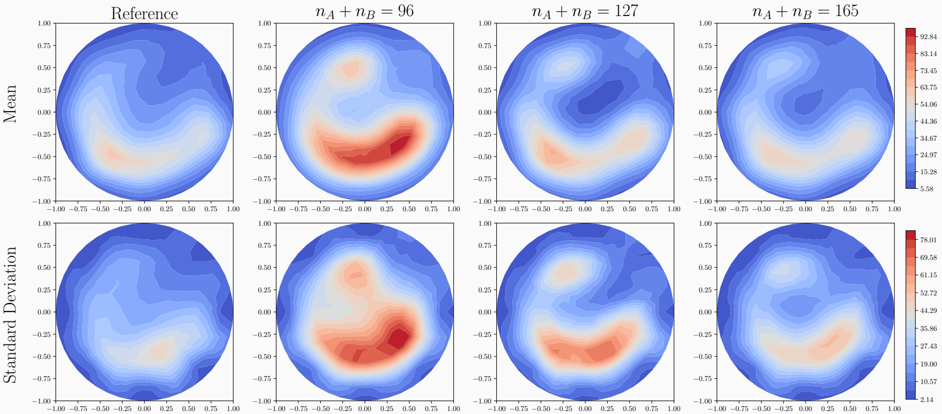


Figure 5: Posterior means and standard-deviations for the recovered conductivity field at $t = 14$. The first column shows the reference solution, obtained using symmetric collocation with a large number of collocation points. The remaining columns show the recovered field when PMM is used with $n_{\mathcal{A}}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

EIT DYNAMIC RECOVERY

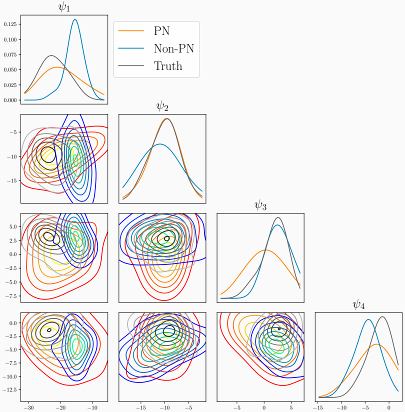


Figure 6: Posterior distribution over the coefficients ψ_k at the final time. A small number $n_{\mathcal{A}} + n_{\mathcal{B}} = 71$ of collocation points was used to discretise the PDE. The reference posterior distribution over the coefficients ψ_k is plotted (grey) and compared to the approximation 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, distinct from ACA. ✓
- 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? ⁉

- Foundations: Cockayne et al. (2019) arXiv:1702.03673
- Optimality: Oates et al. (2019b) arXiv:1901.04326
- BIPs: Lie et al. (2018) arXiv:1712.05717
- Industrial applications: Oates et al. (2019a) arXiv:1707.06107
- History: Oates and Sullivan (2019) arXiv:1901.04457

CLOSING REMARKS

- Numerical methods can be characterised in a Bayesian fashion, distinct from ACA. ✓
- 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? ⁉

- Foundations: Cockayne et al. (2019) arXiv:1702.03673
- Optimality: Oates et al. (2019b) arXiv:1901.04326
- BIPs: Lie et al. (2018) arXiv:1712.05717
- Industrial applications: Oates et al. (2019a) arXiv:1707.06107
- History: Oates and Sullivan (2019) arXiv:1901.04457

Thank You

REFERENCES I

M. A. Capistrán, J. A. Christen, and S. Donnet. Bayesian analysis of ODEs: solver optimal accuracy and Bayes factors. *SIAM/ASA J. Uncertain. Quantif.*, 4(1):829–849, 2016. doi:10.1137/140976777.

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. A. Christen, M. A. Capistrán, M. L. Daza-Torres, H. Flores-Argüedas, and J. C. Montesinos-López. Posterior distribution existence and error control in Banach spaces in the Bayesian approach to UQ in inverse problems, 2017. arXiv:1712.03299.

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, 2017. doi:10.1063/1.4985359.

J. Cockayne, C. J. Oates, T. J. Sullivan, and M. Girolami. Bayesian probabilistic numerical methods. *SIAM Rev.*, 2019. To appear. arXiv:1702.03673.

REFERENCES II

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.

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.

F. M. Larkin. Probabilistic estimation of poles or zeros of functions. *J. Approx. Theory*, 27(4):355–371, 1979. doi:10.1016/0021-9045(79)90124-2.

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

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.*, 6(4):1600–1629, 2018. doi:10.1137/18M1166523.

H. C. Lie, A. M. Stuart, and T. J. Sullivan. Strong convergence rates of probabilistic integrators for ordinary differential equations. *Stat. Comput.*, 2019. To appear. arXiv:1703.03680.

REFERENCES III

C. J. Oates and T. J. Sullivan. A modern retrospective on probabilistic numerics. *Stat. Comp.*, 2019. To appear. arXiv:1901.04457.

C. J. Oates, J. Cockayne, R. G. Aykroyd, and M. Girolami. Bayesian probabilistic numerical methods in time-dependent state estimation for industrial hydrocyclone equipment. *J. Amer. Stat. Assoc.*, pages 1–27, 2019a. doi:10.1080/01621459.2019.1574583.

C. J. Oates, J. Cockayne, D. Prangle, T. J. Sullivan, and M. Girolami. Optimality criteria for probabilistic numerical methods. In F. J. Hickernell and P. Kritzer, editors, *Multivariate Algorithms and Information-Based Complexity*. Berlin/Boston: De Gruyter, 2019b. To appear. arXiv:1901.04326.

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.

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. 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.

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.

REFERENCES IV

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.

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