Bayesian Probabilistic Numerical Methods

J. Cockayne[1] M. Girolami[2,3] H. C. Lie[4,5] C. Oates[3,6] **T. J. Sullivan**[4,5] A. Teckentrup[3,7]

SAMSI–Lloyds–Turing Workshop on Probabilistic Numerical Methods
Alan Turing Institute, London, UK, 11 April 2018

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

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

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 1 (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 1 (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.
- 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 1

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

Rev. Bayes Does Some Numerics 1

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

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

Rev. Bayes Does Some Numerics 1

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

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

Go
Probabilistic!

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

Rev. Bayes Does Some Numerics 1

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

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

Go
Probabilistic!

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

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

Rev. Bayes Does Some Numerics 1

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

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

Go Probabilistic!

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

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

Example 2 (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.

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

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

Go Probabilistic!

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

'Bayes' $\mid$

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

Definition 2 (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 quantity of interest $Q\colon \mathcal{X} \to \mathcal{Q}$ 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:

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

Zellner (1988) calls B an "information processing rule".

Definition 3 (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 the push-forward of the conditional distribution μ^a through Q:

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

Definition 3 (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 the push-forward of the conditional distribution μ^a through Q:

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

Example 4

- 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	$\arg\min x(t)$	$\{x(t_i)\}_{i=1}^n$		Bayesian Optimisation [Mockus, 1989][6]
		$\{\nabla x(t_i)\}_{i=1}^n$		Hennig and Kiefel [2013]
		$\{(x(t_i), \nabla x(t_i)\}_{i=1}^n$		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]	Skilling [1992]
			Finite Difference Methods [John and Wu, 2017][7]	
	∇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

- Thus, we are expressing PNMs in terms of Bayesian inverse problems (Stuart, 2010).
- But a naïve interpretation of Bayes' rule makes no sense here, 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.
- E.g. quadrature example from earlier, with $A(u) = (t_i, u(t_i))_{i=1}^m$.
- Thus, we cannot take the usual approach of defining μ^a via its prior density as

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

  because this density "wants" to be the indicator function $\mathbb{1}[x \in \mathcal{X}^a]$.
- While linear-algebraic tricks work for linear conditioning of Gaussians, in general we condition on events of measure zero using disintegration.

Disintegration I

Write

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

Definition 5 (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 6 (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 7

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

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

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

- 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 8 (AC = BPN for quadratic loss; Cockayne, Oates, Sullivan, and Girolami, 2017)

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 8 (AC = BPN for quadratic loss; Cockayne, Oates, Sullivan, and Girolami, 2017)

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).

Example 9 (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).

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

$$\mathcal{A}_1 = \{0, 1\} \qquad A_1(x) = \mathbb{1}[x \in \{\diamondsuit, \heartsuit\}] \qquad Q(x) = \mathbb{1}[x = \diamondsuit]$$

$$\mathcal{A}_2 = \{0, 1\} \qquad A_2(x) = \mathbb{1}[x \in \{\diamondsuit, \heartsuit, \spadesuit\}] \qquad L(q, q') = \mathbb{1}[q \neq q']$$

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!

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

$A_1(x) = \blacksquare$ vs. $\blacksquare$ $\quad$ $Q(x) = \mathbb{1}[x = \diamondsuit]$

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

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

Optimal Information: AC $\neq$ BPN!

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

$A_1(x) = \blacksquare$ vs. $\blacksquare$ $\quad Q(x) = \mathbb{1}[x = \diamondsuit]$

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

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

Optimal Information: AC $\neq$ BPN!

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

$A_1(x) = \blacksquare$ vs. $\blacksquare$ $\quad Q(x) = \mathbb{1}[x = \diamondsuit]$

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

$x =$	$\clubsuit$		$\diamondsuit$		$\heartsuit$		$\spadesuit$		
$e_{\mathrm{AC}}(A_1, b) = \frac{1}{4}($	$L(b(\blacksquare), 0)$	$+$	$L(b(\blacksquare), 1)$	$+$	$L(b(\blacksquare), 0)$	$+$	$L(b(\blacksquare), 0)$	$)$	
$e_{\mathrm{AC}}(A_1, 0) = \frac{1}{4}($	0	$+$	1	$+$	0	$+$	0	$)$	$= \frac{1}{4}$
$e_{\mathrm{AC}}(A_1, \mathrm{id}) = \frac{1}{4}($	0	$+$	0	$+$	1	$+$	0	$)$	$= \frac{1}{4}$
$e_{\mathrm{AC}}(A_2, b) = \frac{1}{4}($	$L(b(\clubsuit), 0)$	$+$	$L(b(\neg\clubsuit), 1)$	$+$	$L(b(\neg\clubsuit), 0)$	$+$	$L(b(\neg\clubsuit), 0)$	$)$	
$e_{\mathrm{AC}}(A_2, 0) = \frac{1}{4}($	0	$+$	1	$+$	0	$+$	0	$)$	$= \frac{1}{4}$

Optimal Information: AC $\neq$ BPN!

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

$A_1(x) = \blacksquare$ vs. $\blacksquare$ $\quad Q(x) = \mathbb{1}[x = \diamondsuit]$

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

$x =$	$\clubsuit$		$\diamondsuit$		$\heartsuit$		$\spadesuit$		
$e_{\mathrm{AC}}(A_1, b) = \frac{1}{4}($	$L(b(\blacksquare), 0)$	$+$	$L(b(\blacksquare), 1)$	$+$	$L(b(\blacksquare), 0)$	$+$	$L(b(\blacksquare), 0)$	$)$	
$e_{\mathrm{AC}}(A_1, 0) = \frac{1}{4}($	0	$+$	1	$+$	0	$+$	0	$)$	$= \frac{1}{4}$
$e_{\mathrm{AC}}(A_1, \mathrm{id}) = \frac{1}{4}($	0	$+$	0	$+$	1	$+$	0	$)$	$= \frac{1}{4}$
$e_{\mathrm{AC}}(A_2, b) = \frac{1}{4}($	$L(b(\clubsuit), 0)$	$+$	$L(b(\neg\clubsuit), 1)$	$+$	$L(b(\neg\clubsuit), 0)$	$+$	$L(b(\neg\clubsuit), 0)$	$)$	
$e_{\mathrm{AC}}(A_2, 0) = \frac{1}{4}($	0	$+$	1	$+$	0	$+$	0	$)$	$= \frac{1}{4}$
$e_{\mathrm{BPN}}(A_1) = \frac{1}{4}($	$\mathbb{E}_{Q_\sharp\mu\blacksquare} L(\cdot, 0)$	$+$	$\mathbb{E}_{Q_\sharp\mu\blacksquare} L(\cdot, 1)$	$+$	$\mathbb{E}_{Q_\sharp\mu\blacksquare} L(\cdot, 0)$	$+$	$\mathbb{E}_{Q_\sharp\mu\blacksquare} L(\cdot, 0)$	$)$	
$= \frac{1}{4}($	$(\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0)$	$+$	$(\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1)$	$+$	$(\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 0)$	$+$	$(\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0)$	$)$	$= \frac{1}{4}$

Optimal Information: AC $\neq$ BPN!

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

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

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

$x =$	$\clubsuit$	$\diamondsuit$	$\heartsuit$	$\spadesuit$	
$e_{\mathrm{AC}}(A_1, b) = \frac{1}{4}($	$L(b(\blacksquare),0)$ +	$L(b(\blacksquare),1)$ +	$L(b(\blacksquare),0)$ +	$L(b(\blacksquare),0)$ $)$	
$e_{\mathrm{AC}}(A_1, 0) = \frac{1}{4}($	0 +	1 +	0 +	0 $)$	$= \frac{1}{4}$
$e_{\mathrm{AC}}(A_1, \mathrm{id}) = \frac{1}{4}($	0 +	0 +	1 +	0 $)$	$= \frac{1}{4}$
$e_{\mathrm{AC}}(A_2, b) = \frac{1}{4}($	$L(b(\clubsuit),0)$ +	$L(b(\neg\clubsuit),1)$ +	$L(b(\neg\clubsuit),0)$ +	$L(b(\neg\clubsuit),0)$ $)$	
$e_{\mathrm{AC}}(A_2, 0) = \frac{1}{4}($	0 +	1 +	0 +	0 $)$	$= \frac{1}{4}$
$e_{\mathrm{BPN}}(A_1) = \frac{1}{4}($	$\mathbb{E}_{Q_\sharp \mu_{\blacksquare}} L(\cdot,0)$ +	$\mathbb{E}_{Q_\sharp \mu_{\blacksquare}} L(\cdot,1)$ +	$\mathbb{E}_{Q_\sharp \mu_{\blacksquare}} L(\cdot,0)$ +	$\mathbb{E}_{Q_\sharp \mu_{\blacksquare}} L(\cdot,0)$ $)$	
$= \frac{1}{4}($	$(\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0)$ +	$(\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 1)$ +	$(\frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 0)$ +	$(\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0)$ $)$	$= \frac{1}{4}$
$e_{\mathrm{BPN}}(A_2) = \frac{1}{4}($	$\mathbb{E}_{Q_\sharp \mu_{\neg\clubsuit}} L(\cdot,0)$ +	$\mathbb{E}_{Q_\sharp \mu_{\neg\clubsuit}} L(\cdot,1)$ +	$\mathbb{E}_{Q_\sharp \mu_{\neg\clubsuit}} L(\cdot,0)$ +	$\mathbb{E}_{Q_\sharp \mu_{\neg\clubsuit}} L(\cdot,0)$ $)$	
$= \frac{1}{4}($	$(1\cdot 0)$ +	$(\frac{1}{3}\cdot 0 + \frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 1)$ +	$(\frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 0 + \frac{1}{3}\cdot 0)$ +	$(\frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 0 + \frac{1}{3}\cdot 0)$ $)$	$= \frac{1}{3}$

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 10

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

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

- $\mathcal{F}$ = bounded continuous functions with uniform norm $\leftrightarrow$ total variation.
- $\mathcal{F}$ = bounded Lipschitz continuous functions with Lipschitz norm $\leftrightarrow$ 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 11 (Cockayne, Oates, Sullivan, and Girolami, 2017, 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,

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

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.

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:

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

$$u(0) = 0$$

$$u(t)/\sqrt{t} \to 1 \qquad \text{as } t \to +\infty$$

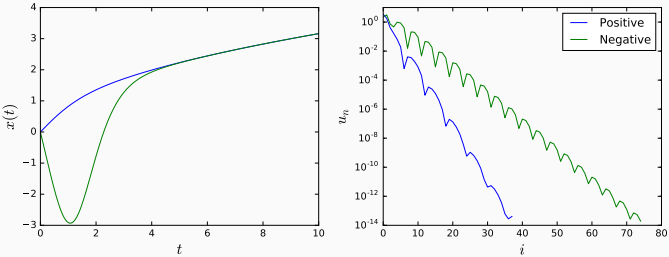


Figure 1: The two solutions of Painlevé's first transcendental and their spectra in the orthonormal Chebyshev polynomial basis over $[0, 10]$.

Example: Painlevé's First Transcendental I

A multivalent boundary value problem:

$$u''(t) - u(t)^2 = -t \qquad \text{for } t \geq 0$$
$$u(0) = 0$$
$$u(10) = \sqrt{10}$$

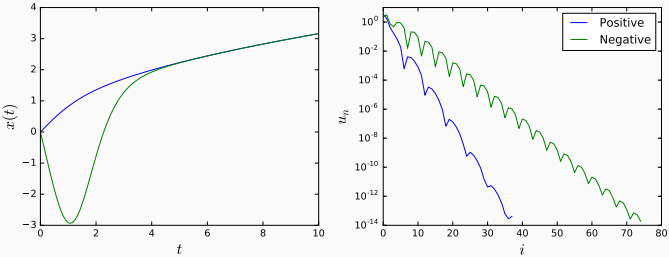
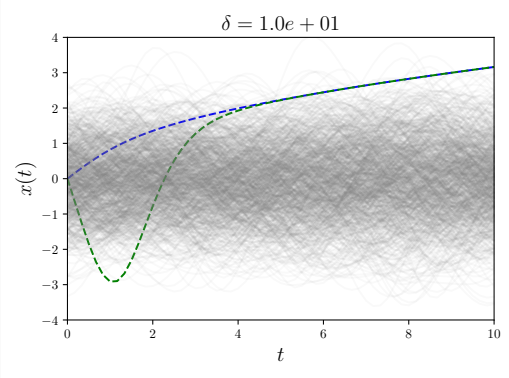


Figure 1: The two solutions of Painlevé's first transcendental and their spectra in the orthonormal Chebyshev polynomial basis over $[0, 10]$.

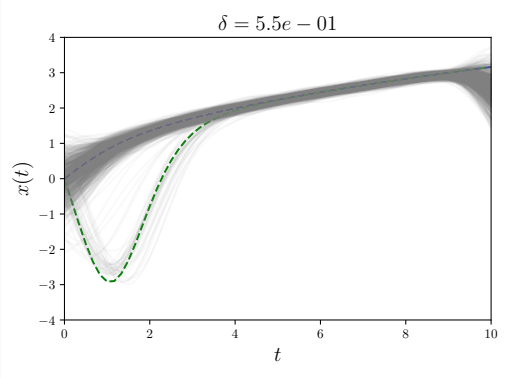
Example: Painlevé's First Transcendental iii

- 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… ✗



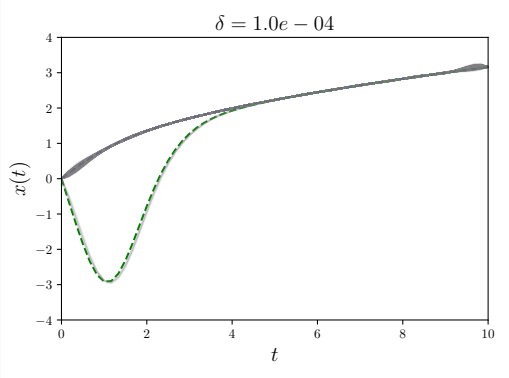
Example: Painlevé's First Transcendental iii

- 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… ✗



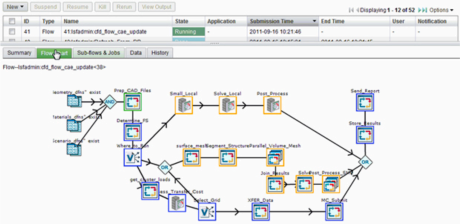
Example: Painlevé's First Transcendental iii

- 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



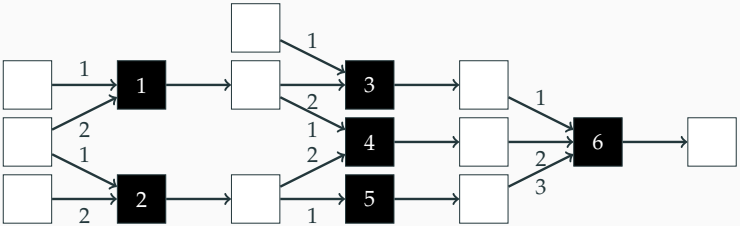
- Numerical methods usually form part of pipelines.
- Prime example: a PDE solve is a *forward model* in an *inverse problem*.
- Motivation for PNMs in the context of Bayesian inverse problems:

  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 $B(\mu, (a_1, a_2))\ldots$ although figuring out what the spaces $\mathcal{X}_i$, $\mathcal{A}_i$ and operators A_i etc. are is a headache!

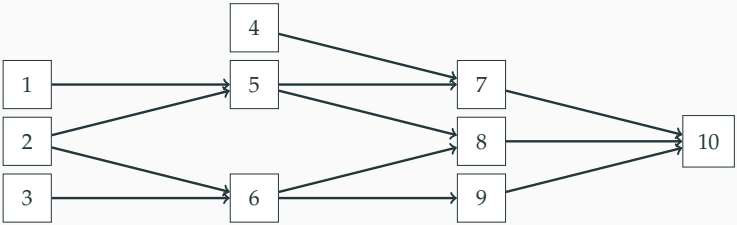
Coherence I

- More generally, 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. Deterministic data at the left-most □ nodes, then random variables as outputs, realisations of which get fed into the next ■.



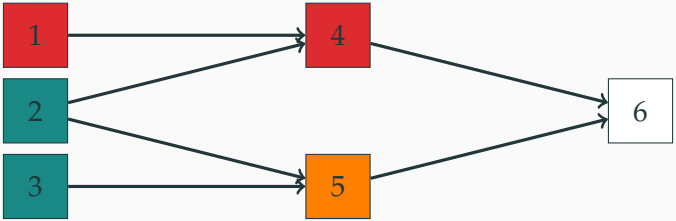
Coherence I

- More generally, 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. 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 12

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

If a pipeline of PNMs is such that

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

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

Coherency Theorem

Theorem 13 (Cockayne, Oates, Sullivan, and Girolami, 2017, Theorem 5.9)

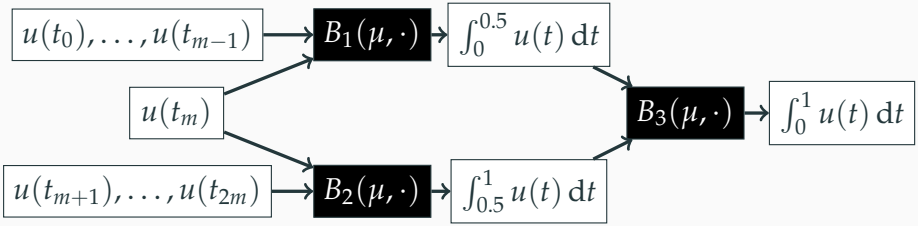
If a pipeline of PNMs is such that

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

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

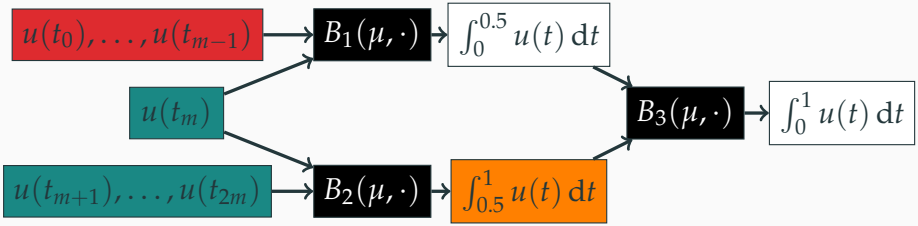
- Redundant structure in the pipeline (recycled information) will break coherence, and hence Bayesianity of the pipeline.
- In principle, coherence and hence being Bayesian depend upon the prior.
- This should not be surprising — as a loose analogy, one doesn't expect the trapezoidal rule to be a good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

Split Integration: Coherence



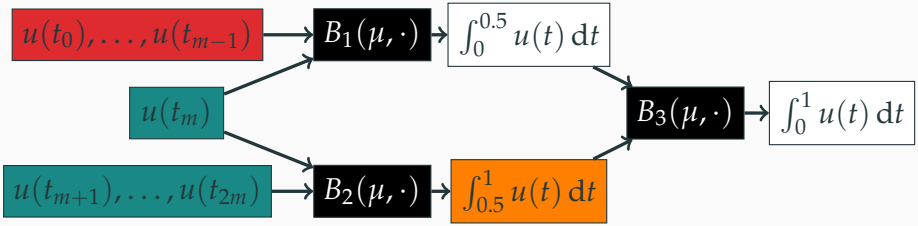
- Integrate a function over $[0, 1]$ in two steps using nodes $0 \leq t_0 < \cdots < t_{m-1} < 0.5$, $t_m = 0.5$, and $t_{m+1} < \cdots < t_{2m} \leq 1$.

Split Integration: Coherence



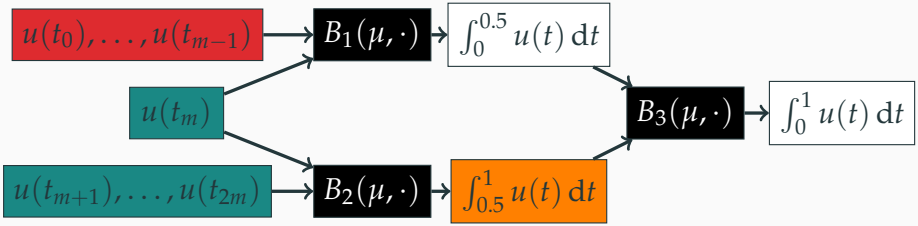
- Integrate a function over $[0, 1]$ in two steps using nodes $0 \leq t_0 < \cdots < t_{m-1} < 0.5$, $t_m = 0.5$, and $t_{m+1} < \cdots < t_{2m} \leq 1$.
- Is $\int_{0.5}^1 u(t)\,\mathrm{d}t$ independent of $u(t_0), \ldots, u(t_{m-1})$ given $u(t_m), \ldots, u(t_{2m})$?

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



- 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.
- 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. (2015a,b) and Jacob et al. (2017). ⁉

Randomised Bayesian Inverse Problems

Randomised Bayesian inverse problems

- A Bayesian inverse problem can be seen as a simple pipeline of the form

$$\square \to \blacksquare \to \square \to \blacksquare \to \square$$

  with the first method being the forward solve, the second the inversion.
- How much does replacing the traditional deterministic forward solve by a PNM (or just a random surrogate) affect the solution of the BIP, à la Stuart (2010)?
- Usual posterior μ^y over $\mathcal{U}$ with prior μ_0 and negative log-likelihood $\Phi = \Phi(\,\cdot\,; y) \colon \mathcal{U} \to \mathbb{R}$:

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

- Let now $\Phi_N \colon \Omega \times \mathcal{U} \to \mathbb{R}$ be a measurable function that provides a random approximation to Φ, and denote by ν_N the distribution of Φ_N.

Random and deterministic approximate posteriors

- Replacing Φ by Φ_N in (1), we obtain a random approximation μ_N^{samp} of μ:

$$\frac{\mathrm{d}\mu_N^{\text{samp}}}{\mathrm{d}\mu_0}(u) := \frac{\exp(-\Phi_N(u))}{Z_N^{\text{samp}}}, \tag{2}$$
$$Z_N^{\text{samp}} := \mathbb{E}_{\mu_0}\left[\exp(-\Phi_N(\cdot))\right].$$

- Taking the expectation of the random likelihood gives a deterministic approximation:

$$\frac{\mathrm{d}\mu_N^{\text{marg}}}{\mathrm{d}\mu_0}(u) := \frac{\mathbb{E}_{\nu_N}\left[\exp(-\Phi_N(u))\right]}{\mathbb{E}_{\nu_N}\left[Z_N^{\text{samp}}\right]}. \tag{3}$$

- An alternative deterministic approximation can be obtained by taking the expected value of the density $(Z_N^{\text{samp}})^{-1}e^{-\Phi_N(u)}$ in (2). However, μ_N^{marg} provides a clear interpretation as the posterior obtained by the approximation of the true data likelihood $e^{-\Phi(u)}$ by $\mathbb{E}_{\nu_N}\left[e^{-\Phi_N(u)}\right]$, and is more amenable to sampling methods such as pseudo-marginal MCMC (Beaumont, 2003; Andrieu and Roberts, 2009).

Summary of convergence rates

Theorem 14 (Lie, Sullivan, and Teckentrup, 2017)

For suitable Hölder exponents $p_1, p_1', p_2, \ldots$ quantifying the integrability of Φ and Φ_N, we obtain deterministic convergence $\mu_N^{\mathrm{marg}} \to \mu$ and mean-square convergence $\mu_N^{\mathrm{samp}} \to \mu$ in the Hellinger metric:

$$d_{\mathrm{H}}\left(\mu, \mu_N^{\mathrm{marg}}\right) \leq C \left\| \mathbb{E}_{\nu_N}\left[|\Phi - \Phi_N|^{p_2'} \right]^{1/p_2'} \right\|_{L^{2p_1'p_3'}_{\mu_0}(\mathcal{U})},$$

$$\mathbb{E}_{\nu_N}\left[d_{\mathrm{H}}\left(\mu, \mu_N^{\mathrm{samp}}\right)^2 \right]^{1/2} \leq D \left\| \mathbb{E}_{\nu_N}\left[|\Phi - \Phi_N|^{2q_1'} \right]^{1/2q_1'} \right\|_{L^{2q_2'}_{\mu_0}(\mathcal{U})}.$$

Theorem 14 (Lie, Sullivan, and Teckentrup, 2017)

For suitable Hölder exponents $p_1, p_1', p_2, \ldots$ quantifying the integrability of Φ and Φ_N, we obtain deterministic convergence $\mu_N^{\text{marg}} \to \mu$ and mean-square convergence $\mu_N^{\text{samp}} \to \mu$ in the Hellinger metric:

$$d_{\mathrm{H}}\left(\mu, \mu_N^{\text{marg}}\right) \leq C \left\| \mathbb{E}_{\nu_N}\left[|\Phi - \Phi_N|^{p_2'} \right]^{1/p_2'} \right\|_{L_{\mu_0}^{2p_1'p_3'}(\mathcal{U})},$$

$$\mathbb{E}_{\nu_N}\left[d_{\mathrm{H}}\left(\mu, \mu_N^{\text{samp}}\right)^2 \right]^{1/2} \leq D \left\| \mathbb{E}_{\nu_N}\left[|\Phi - \Phi_N|^{2q_1'} \right]^{1/2q_1'} \right\|_{L_{\mu_0}^{2q_2'}(\mathcal{U})}.$$

Skip to Example or **Show Gory Details** ?

Convergence of μ_N^{marg}

Theorem 15 (Deterministic convergence of the marginal posterior)

Suppose there exist positive scalars C_1, C_2, C_3, that do not depend on N, such that for the Hölder conjugate exponent pairs (p_1, p_1'), (p_2, p_2'), and (p_3, p_3'), we have

- $\min\left\{ \left\| \mathbb{E}_{\nu_N}[e^{-\Phi_N}]^{-1} \right\|_{L^{p_1}_{\mu_0}(\mathcal{U})}, \left\| e^{\Phi} \right\|_{L^{p_1}_{\mu_0}(\mathcal{U})} \right\} \leq C_1(p_1);$
- $\left\| \mathbb{E}_{\nu_N}\left[\left(e^{-\Phi} + e^{-\Phi_N}\right)^{p_2} \right]^{1/p_2} \right\|_{L^{2p_1'p_3}_{\mu_0}(\mathcal{U})} \leq C_2(p_1, p_2, p_3);$
- $C_3^{-1} \leq \mathbb{E}_{\nu_N}[Z_N^{\text{samp}}] \leq C_3.$

Then there exists a scalar $C = C(C_1, C_2, C_3, Z) > 0$ that does not depend on N, such that

$$d_{\mathrm{H}}(\mu, \mu_N^{\text{marg}}) \leq C \left\| \mathbb{E}_{\nu_N}\left[|\Phi - \Phi_N|^{p_2'} \right]^{1/p_2'} \right\|_{L^{2p_1'p_3'}_{\mu_0}(\mathcal{U})},$$

$$C(C_1, C_2, C_3, Z) = \left(\frac{C_1(p_1)}{Z} + C_3 \max\left\{ Z^{-3}, C_3^3 \right\} \right) C_2^2(p_1, p_2, p_3).$$

Convergence of μ_N^{samp}

Theorem 16 (Mean-square convergence of the sample posterior)

Suppose there exist positive scalars D_1, D_2, that do not depend on N, such that for Hölder conjugate exponent pairs (q_1, q_1') and (q_2, q_2'), we have

- $\left\| \mathbb{E}_{\nu_N}\left[\left(e^{-\Phi/2} + e^{-\Phi_N/2}\right)^{2q_1}\right]^{1/q_1} \right\|_{L^{q_2}_{\mu_0}(\mathcal{U})} \leq D_1(q_1, q_2);$
- $\left\| \mathbb{E}_{\nu_N}\left[\left(Z_N^{\text{samp}} \max\{Z^{-3}, (Z_N^{\text{samp}})^{-3}\}\left(e^{-\Phi} + e^{-\Phi_N}\right)^2\right)^{q_1}\right]^{1/q_1} \right\|_{L^{q_2}_{\mu_0}(\mathcal{U})} \leq D_2(q_1, q_2).$

Then

$$\mathbb{E}_{\nu_N}\left[d_{\mathrm{H}}\left(\mu, \mu_N^{\text{samp}}\right)^2\right]^{1/2} \leq (D_1 + D_2)\left\| \mathbb{E}_{\nu_N}\left[|\Phi - \Phi_N|^{2q_1'}\right]^{1/2q_1'} \right\|_{L^{2q_2'}_{\mu_0}(\mathcal{U})},$$

Applicability of convergence theorems I

The assumptions of Theorems 15 and 16 are satisfied when the exact potential Φ and the approximation quality $\Phi_N \approx \Phi$ are suitably well behaved. Recall from (1) that Z is the normalisation constant of μ. Therefore, for μ to be well-defined, we must have that $0 < Z < \infty$. In particular, there exists $0 < C_3 < \infty$ such that $C_3^{-1} < Z < C_3$.

Assumption 17

There exists $C_0 \in \mathbb{R}$ that does not depend on N such that, for all $N \in \mathbb{N}$,

$$\Phi \geq -C_0 \quad \text{and} \quad \nu_N(\{\Phi_N \mid \Phi_N \geq -C_0\}) = 1, \tag{4}$$

and for any $0 < C_3 < +\infty$ with the property that $C_3^{-1} < Z < C_3$, there exists $N^*(C_3) \in \mathbb{N}$ such that, for all $N \geq N^*$,

$$\mathbb{E}_{\mu_0}[\mathbb{E}_{\nu_N}[|\Phi_N - \Phi|]] \leq \frac{1}{2e^{C_0}} \min\left\{Z - \frac{1}{C_3}, C_3 - Z\right\}. \tag{5}$$

Applicability of convergence theorems II

Lemma 18

Suppose that Assumption 17 holds with C_0 as in (4) and C_3 and $N^*(C_3)$ as in (5), that $\exp(\Phi) \in L^{p^*}_{\mu_0}(\mathcal{U})$ for some $1 \leq p^* \leq +\infty$ with conjugate exponent $(p^*)'$, and there exists some $C_4 \in \mathbb{R}$ that does not depend on N, such that, for all $N \in \mathbb{N}$,

$$\nu_N\big(\{\Phi_N \mid \mathbb{E}_{\mu_0}[\Phi_N] \leq C_4\}\big) = 1.$$

Then the hypotheses of Theorem 15 hold, with

$$p_1 = p^*,\ p_2 = p_3 = +\infty,\ C_1 = \|e^{\Phi}\|_{L^{p^*}_{\mu_0}},\ C_2 = 2e^{C_0},$$

and C_3 as above. Moreover, the hypotheses of Theorem 16 hold, with

$$q_1 = q_2 = \infty,\ D_1 = 4e^{C_0},\ D_2 = 4e^{3C_0}\max\{C_3^{-3}, e^{3C_4}\}.$$

Lemma 19

Suppose that Assumption 17 holds with C_0 as in (4) and C_3 and $N^*(C_3)$ as in (5), and that there exists some $2 < \rho^* < +\infty$ such that $\mathbb{E}_{\nu_N}[\exp(\rho^*\Phi_N)] \in L^1_{\mu_0}$. Then the hypotheses of Theorem 15 hold, with

$$p_1 = \rho^*,\ p_2 = p_3 = +\infty,\ C_1 = \|\mathbb{E}_{\nu_N}[\exp(\rho^*\Phi_N)]\|_{L^1_{\mu_0}}^{1/\rho^*},\ C_2 = 2e^{C_0},$$

and C_3 as above. Moreover, the hypotheses of Theorem 16 hold, with

$$q_1 = \frac{\rho^*}{2},\ q_2 = +\infty,\ D_1 = 4e^{C_0},\ D_2 = 4e^{2C_0}\left(C_3^{-3}e^{C_0} + \|\mathbb{E}_{\nu_N}[e^{\rho^*\Phi_N}]\|_{L^1_{\mu_0}}^{2/\rho^*}\right).$$

Example: Monte Carlo approximation of high-dimensional misfits

We consider a Monte Carlo approximation Φ_N of a quadratic potential Φ (Nemirovski et al., 2008; Shapiro et al., 2009), further applied and analysed in the MAP estimator context by Le et al. (2017). This approximation is particularly useful for data $y \in \mathbb{R}^J$, $J \gg 1$.

$$
\begin{aligned}
\Phi(u) &:= \frac{1}{2}\left\|\Gamma^{-1/2}(y - G(u))\right\|^2 \\
&= \frac{1}{2}\left(\Gamma^{-1/2}(y - G(u))\right)^{\mathrm{T}}\mathbb{E}[\sigma\sigma^{\mathrm{T}}]\left(\Gamma^{-1/2}(y - G(u))\right) && \text{where } \mathbb{E}[\sigma] = 0 \in \mathbb{R}^J, \mathbb{E}[\sigma\sigma^{\mathrm{T}}] = I_{J\times J} \\
&= \frac{1}{2}\mathbb{E}\left[\left|\sigma^{\mathrm{T}}\left(\Gamma^{-1/2}(y - G(u))\right)\right|^2\right] \\
&\approx \frac{1}{2N}\sum_{i=1}^{N}\left|\sigma^{(i)\mathrm{T}}\left(\Gamma^{-1/2}(y - G(u))\right)\right|^2 && \text{for i.i.d. } \sigma^{(1)}, \dots, \sigma^{(N)} \overset{\mathrm{d}}{=} \sigma \\
&= \frac{1}{2}\left\|\Sigma_N^{\mathrm{T}}\left(\Gamma^{-1/2}(y - G(u))\right)\right\|^2 && \text{for } \Sigma_N := \frac{1}{\sqrt{N}}[\sigma^{(1)} \cdots \sigma^{(N)}] \in \mathbb{R}^{J\times N} \\
&=: \Phi_N(u).
\end{aligned}
$$

Example: Monte Carlo approximation of high-dimensional misfits

The analysis and numerical studies in Le et al. (2017, Sections 3–4) suggest that a good choice for the $\mathbb{R}^J$-valued random vector σ would be one with independent and identically distributed (i.i.d.) entries from a sub-Gaussian probability distribution. Examples of sub-Gaussian distributions considered include

- the Gaussian distribution: $\sigma_j \sim \mathcal{N}(0,1)$, for $j = 1, \ldots, J$; and
- the ℓ-sparse distribution: for $\ell \in [0,1)$, let $s := \frac{1}{1-\ell} \geq 1$ and set, for $j = 1, \ldots, J$,

$$\sigma_j := \sqrt{s} \begin{cases} 1, & \text{with probability } \frac{1}{2s}, \\ 0, & \text{with probability } \ell = 1 - \frac{1}{s}, \\ -1, & \text{with probability } \frac{1}{2s}. \end{cases}$$

Example: Monte Carlo approximation of high-dimensional misfits

- Le et al. (2017) observe that, for large J and moderate $N \approx 10$, the random potential Φ_N and the original potential Φ are very similar, in particular having approximately the same minimisers and minimum values.
- Statistically, these correspond to the maximum likelihood estimators under Φ and Φ_N being very similar; after weighting by a prior, this corresponds to similarity of maximum a posteriori (MAP) estimators.
- Here, we study the BIP instead of the MAP problem, and thus the corresponding conjecture is that the deterministic posterior $\mathrm{d}\mu(u) \propto \exp(-\Phi(u))\,\mathrm{d}\mu_0(u)$ is well approximated by the random posterior $\mathrm{d}\mu_N^{\mathrm{samp}}(u) \propto \exp(-\Phi_N(u))\,\mathrm{d}\mu_0(u)$.

Example: Well-posedness of BIPs with Monte Carlo misfits

Applying the general Theorem 16 from earlier gives the following transfer of the Monte Carlo convergence rate from the approximation of Φ to the approximation of μ: ✓

Proposition 20

Suppose that the entries of σ are i.i.d. ℓ-sparse, for some $\ell \in [0,1)$, and that $\Phi \in L^2_{\mu_0}(\mathcal{U})$. Then there exists a constant C, independent of N, such that

$$\left(\mathbb{E}_\sigma\left[d_{\mathrm{H}}\left(\mu, \mu_N^{\mathrm{samp}}\right)^2\right]\right)^{1/2} \leq \frac{C}{\sqrt{N}}.$$

For technical reasons to do with the non-compactness of the support and finiteness of MGFs of maxima, the current proof technique does not work for the Gaussian case. ⁉

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 $\rightarrow$ (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? ⁉
- Full details and further applications in

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

Lie, Sullivan, and Teckentrup (2017) arXiv:1712.05717.

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 $\rightarrow$ (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? ⁉
- Full details and further applications in

Cockayne, Oates, Sullivan, and Girolami (2017) arXiv:1702.03673.
Lie, Sullivan, and Teckentrup (2017) arXiv:1712.05717.

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.

C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2): 697–725, 2009. doi:10.1214/07-AOS574.

M. A. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3): 1139–1160, 2003.

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.

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

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. 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.

References II

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.

E. B. Le, A. Myers, T. Bui-Thanh, and Q. P. Nguyen. A data-scalable randomized misfit approach for solving large-scale PDE-constrained inverse problems. *Inverse Probl.*, 33(6):065003, 2017. doi:10.1088/1361-6420/aa6cbd.

H. C. Lie, T. J. Sullivan, and A. L. Teckentrup. Random forward models and log-likelihoods in Bayesian inverse problems, 2017. arXiv:1712.05717.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2008. doi:10.1137/070704277.

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, 2015a. doi:10.1214/15-EJS989.

H. Owhadi, C. Scovel, and T. J. Sullivan. On the brittleness of Bayesian inference. *SIAM Rev.*, 57(4):566–582, 2015b. doi:10.1137/130938633.

References III

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.

A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*, volume 9 of *MPS/SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Programming Society (MPS), Philadelphia, PA, 2009. doi:10.1137/1.9780898718751.

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.

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.

References iv

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.

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