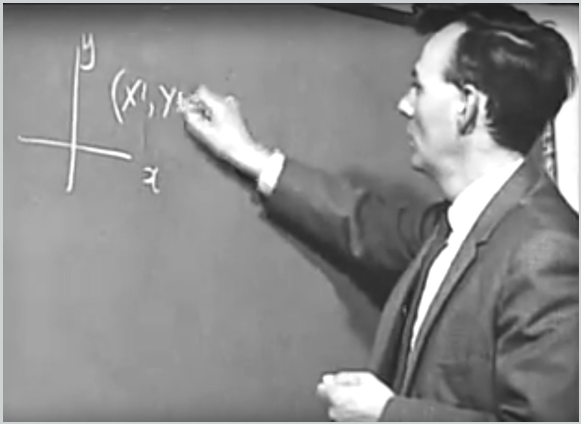
Bayesian probabilistic numerical methods

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

Applied and Computational Mathematics Seminar
University of Edinburgh, UK, 30 October 2019

[1]Alan Turing Institute, London, UK
[2]University of Cambridge, UK
[3]Newcastle University, UK
[4]**Freie Universität Berlin, DE**
[5]**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 (1979b)

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

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

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

- Aim: recover $\theta \in \mathbb{R}^3_{>0}$ from observations $y_i = u(t_i^{\mathrm{obs}}) + \eta_i$ at some discrete times $t_i^{\mathrm{obs}} = 0, 1, \ldots, 40$, $\eta_i \sim \mathcal{N}(0, 10^{-3} I)$ 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 $\mathbb{E}\left[\sup_{0\le t\le T} \|u(t) - u^{\mathrm{PN}}(t)\|^2\right] \le C\tau^2$ (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 $\sigma \ge 0$.

Motivating example: FitzHugh–Nagumo ODE inference

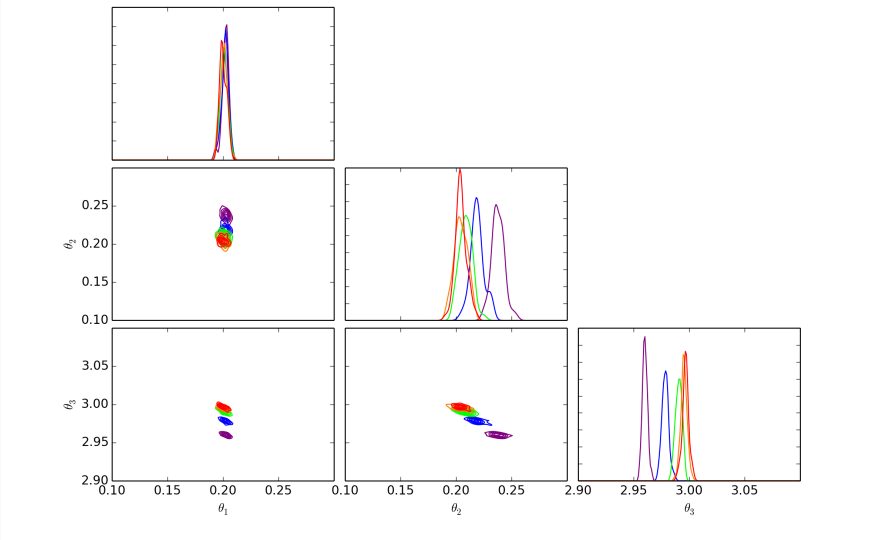


Figure 1: The deterministic posteriors (i.e. $\sigma = 0$) are over-confident at all values of the time step $\tau = 0.1, 0.05, 0.02, 0.01, 0.005$, often do not overlap, and are biased.

Motivating example: FitzHugh–Nagumo ODE inference

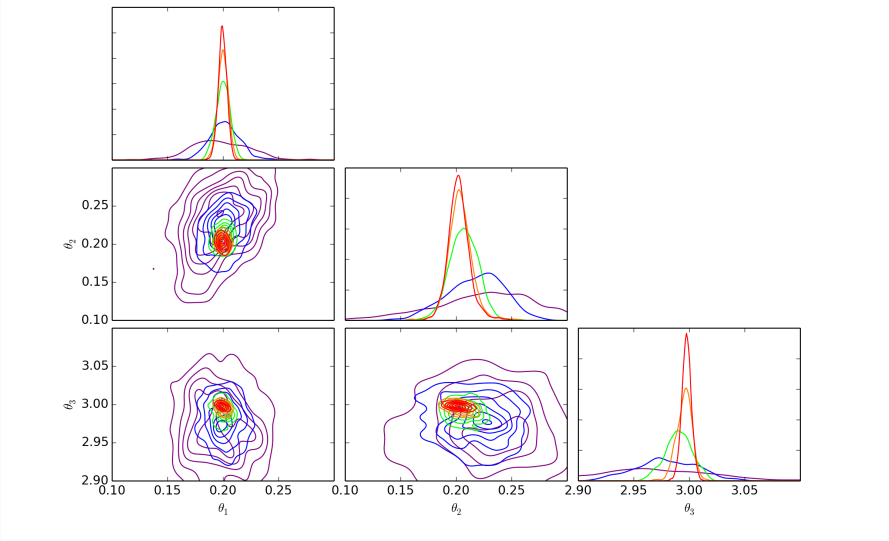


Figure 1: In contrast, the PN-Euler posteriors (here with $\sigma = 1/5$) for $\tau = 0.1, 0.05, 0.02, 0.01, 0.005$ are less confident and overlap more, though are still biased.

Outline

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

A little history

Poincaré (1912) blends interpolation and probability

"Je suppose que l'on sache a priori que la fonction $f(x)$ est développable, dans une certain domaine, suivant les puissances croissantes des x,

$$f(x) = A_0 + A_1 x + \dots.$$

Nous ne savons rien sur les A, sauf que la probabilité pour que l'un d'eux, A_i, soit compris entre certaines limites, y et $y + \mathrm{d}y$, est

$$\sqrt{\frac{h_i}{\pi}} e^{-h_i y^2} \, \mathrm{d}y.$$

Nous connaissons par n observations

$$\begin{aligned} f(a_1) &= B_1, \\ f(a_2) &= B_2, \\ &\dots\dots\dots\dots \\ f(a_n) &= B_n. \end{aligned}$$

Nous cherchons la valeur probable de $f(x)$ pour une autre valeur de x."

- What about probabilistic numerical methods for use on a computer?
- The limited nature of the earliest computers led authors to focus initially on the phenomenon of round-off error (Henrici, 1962; Hull and Swenson, 1966; von Neumann and Goldstine, 1947), whether of fixed-point or floating-point type, without any particular statistical *inferential* motivation; indeed, this aspect is still alive (Barlow and Bareiss, 1985; Chatelin and Brunet, 1990; Tienari, 1970).
- One early, utilitarian view is that probabilistic models in computation are just useful shortcuts:

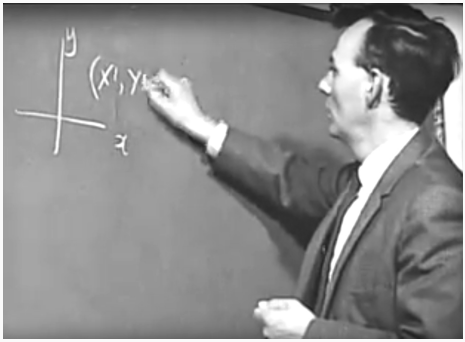
  "[Round-off errors] are strictly very complicated but uniquely defined number theoretical functions [of the inputs], yet our ignorance of their true nature is such that we best treat them as random variables."

  — von Neumann and Goldstine (1947, p. 1027)

- One of the earliest attempts to statistically motivate a numerical algorithm was due to A. V. Sul′din (1924–1996), working at Kazan State University in the USSR Norden et al. (1978); Zabotin et al. (1996).
- After first making contributions to the study of Lie algebras, towards the end of the 1950s Sul′din turned his attention to computational and applied mathematics, and in particular to probabilistic and statistical methodology.
- His work in this direction led to the establishment of the Faculty of Computational Mathematics and Cybernetics in Kazan, of which he was the founding Dean.

Al′bert Valentinovich Sul′din (1924–1996)
© Kazan Federal University, reproduced with permission.

- On the other side of the Iron Curtain, between 1957 and 1969, Frederick Michael ("Mike") Larkin (1936–1982) worked for the UK Atomic Energy Authority in its laboratories at Harwell and Culham, as well as working for two years at Rolls Royce; from 1969, he was at Queen's University in Kingston, Ontario, Canada.
- Following a parallel path to that of Sul′din, over the next decade Larkin would further blend numerical analysis and statistical thinking Kuelbs et al. (1972); Larkin (1969, 1972, 1974, 1979c,a,b), arguably laying the foundations on which modern PN would be developed.



Frederick Michael Larkin (1936–1982) © (Larkin et al., 1967, reproduced with permission).

Larkin

- Larkin worked on building some of the first graphical calculators, called GHOST (short for graphical output system), and the GHOUL (graphical output language) — perhaps a motivation for seeking a richer description of numerical error.
- The perspective developed by Larkin was fundamentally statistical and, in modern terminology, the probabilistic numerical methods he developed would be described as *Bayesian* — though Larkin used the term *relative likelihood* for the prior.
- Larkin's perspective on quadrature: consider the Wiener measure as a prior, the information $(t_j, u(t_j))_{j=1}^{J}$ as (noiseless) data, and output the posterior marginal for $\int_a^b u(t)\,\mathrm{d}t$ — what we would now recognise as a probabilistic numerical method:

  "Among other things, this permits, at least in principle, the derivation of joint probability density functions for [*both observed and unobserved*] *functionals on the space and also allows us to evaluate confidence limits on the estimate of a required functional* (*in terms of given values of other functionals*)*."* — Larkin (1972)

- We wish to approximate the definite integral $\int_a^b u(t)\,\mathrm{d}t$ of $u \in \mathcal{U} := C^0([a,b];\mathbb{R})$ under a statistical assumption that $(u(t) - u(a))_{t\in[a,b]}$ follows a standard Brownian motion (Wiener measure, μ_{W}).
- We receive pointwise data about u in the form of the values of u at $J \in \mathbb{N}$ nodes $a = t_1 < t_2 < \cdots < t_J = b$.
- In more statistical language, anticipating the terminology of Cockayne et al. (2019):
  - we have a **latent quantity** (integrand) u living in a space $\mathcal{U}$,
  - our **observed data** or **information** concerning u is $y := (t_j, u(t_j))_{j=1}^J$, living in the space $\mathcal{Y} := ([a,b] \times \mathbb{R})^J$,
  - and we care about the **quantity of interest** $\mathsf{Q}(u) := \int_a^b u(t)\,\mathrm{d}t$, living in $\mathcal{Q} := \mathbb{R}$.

- Sul′din (1959, 1960, 1963) showed by direct calculation that the quadrature rule $\mathsf{B}\colon \mathcal{Y} \to \mathbb{R}$ that minimises the mean squared error

$$\int_{\mathcal{U}} \left| \int_a^b u(t)\,\mathrm{d}t - \mathsf{B}\big((t_j, u(t_j))_{j=1}^J\big) \right|^2 \mu_{\mathrm{W}}(\mathrm{d}u)$$

  is the classical trapezoidal rule

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

  i.e. the definite integral of the piecewise linear interpolant of the observed data.

- Thus, Sul′din describes the trapezoidal rule B_{tr} as a frequentist point estimator obtained from minimising the mean square error, which "just happens" to produce an unbiased estimator with variance $\frac{1}{12}\sum_{j=1}^{J-1}(t_{j+1}-t_j)^3$.

- However, Larkin sees the normal distribution

$$\mathcal{N}\left(\mathsf{B}_{\mathrm{tr}}\left((t_j, z_j)_{j=1}^{J}\right), \frac{1}{12}\sum_{j=1}^{J-1}(t_{j+1} - t_j)^3\right)$$

  on $\mathbb{R}$ as the measure-valued output of a probabilistic quadrature rule, of which $\mathsf{B}_{\mathrm{tr}}\left((t_j, z_j)_{j=1}^{J}\right.$ is a convenient point summary. The technical development in this pioneering work made fundamental contributions to the study of Gaussian measures on Hilbert spaces (Kuelbs et al., 1972; Larkin, 1972).
- However, neither Larkin nor Sul′din would have had access to the computing resources needed to realise their more general vision except in special cases.

- The average-case analysis (ACA) of numerical methods received interest and built on the work of Kolmogorov (1936) and Sard (1963).
- In ACA the performance of a numerical method is assessed in terms of its *average error* with respect to a probability measure over the problem set; a prime example is univariate quadrature with the average quadratic loss given earlier.
- A traditional (deterministic) NM can also be regarded as a decision rule and the probability measure used in ACA can be used to instantiate the Bayesian decision-theoretic framework (Berger, 1985). The average error is then recognised as the *expected loss*, also called the *risk*. ACA is mathematically equivalent to Bayesian decision theory — restricted to the case of an experiment that produces a deterministic dataset (Kimeldorf and Wahba, 1970a,b; Parzen, 1970; Larkin, 1970).

- *ACA optimal* methods are *Bayes rules* or *Bayes acts* in the decision-theoretic context. Kadane and Wasilkowski (1985) had the insight that ACA-optimal methods coincide with (non-randomised) Bayes rules when the measure used to define the MSE is the Bayesian prior. Recently it has become clear that ACA and Bayesian optimality differ in general (Cockayne et al., 2019; Oates et al., 2019b).

- Information-based complexity (IBC) Novak (1988); Traub et al. (1983); Traub and Woźniakowsi (1980) developed simultaneously with ACA, with the aim of relating the computational complexity and optimality properties of algorithms to the available information on the unknowns.
- For example, (Smale, 1985, Theorem D) compared the accuracies (with respect to mean absolute error) for a given cost of the Riemann sum, trapezoidal, and Simpson quadrature rules; in the same paper, Smale also considered root-finding, optimisation via linear programming, and the solution of systems of linear equations.
- Bayesian quadrature was again discussed in detail by Diaconis (1988), who repeated Sul′din's observation that the posterior mean for $\int_a^b u(t)\,\mathrm{d}t$ under the Wiener measure prior is the trapezoidal method, which is a ACA-optimal.
- Diaconis posed a further question: can other numerical methods for other tasks be similarly recovered as Bayes rules in a decision-theoretic framework? For linear cubature methods, a positive and constructive answer was recently provided by Karvonen et al. (2018), but the general question remains open.

- Research interest in PN was revived by contributions from on quadrature (Minka, 2000; O'Hagan, 1991; Rasmussen and Ghahramani, 2003), each to a greater or lesser extent a rediscovery of earlier work due to Larkin (1972). In each case the algorithmic output was considered to be a probability distribution over the quantity of interest.
- The 1990s saw an expansion in the PN agenda, first with early work on an area that would become Bayesian optimisation (Močkus, 1975, 1977, 1989).
- Skilling (1992) presented a novel (partially) Bayesian perspective on the numerical solution of ODE initial value problems of the form

$$\begin{aligned} u'(t) \equiv \frac{\mathrm{d}u}{\mathrm{d}t} &= f(t, u(t)) & t \in [0, T], \\ u(0) &= u_0. \end{aligned}$$

- Skilling (1992) considered, e.g. the role of regularity assumptions on f, prior and likelihood choice, and sampling strategies.
- Skilling himself considered his then-new explicit emphasis on a Bayesian statistical approach to be quite natural:

  "This paper arose from long exposure to Laplace/Cox/Jaynes probabilistic reasoning, combined with the University of Cambridge's desire that the author teach some (traditional) numerical analysis. The rest is common sense. [...] *Simply, Bayesian ideas are 'in the air'."*

  — Skilling (1992)

- The machine learning community took up the ODE theme again $\approx$ 5 years ago (Schober et al., 2014), provoking further mathematical analysis (Conrad et al., 2016) and then an explosion of more general studies.[1]

[1] I have a marvellous literature list, but this slide is too small to contain it...

Conceptual evolution — A summary

1. In the traditional setting of numerical analysis, c. 1950, all objects and operations are seen as being strictly deterministic. These deterministic objects are sometimes exceedingly complicated, to the extent that they may be treated as being stochastic.
2. Sard and Sul′din consider the questions of optimal performance of a numerical method in, respectively, the worst-case and the average-case context. Some of the average-case performance measures amount to variances of point estimators but are not *viewed* as such; probabilistic aspects are not a motivating factor.
3. Larkin's innovation, 1960s–1970s, is to formulate numerical tasks in terms of a joint distribution over latent quantities and quantities of interest; the quantity of interest is a stochastic object. Larkin summarises his posterior distributions using a point estimator accompanied by a credible interval.
4. The fully modern viewpoint, circa 2019, is to explicitly think of the output as a probability measure to be realised, sampled, and possibly summarised.

An inference perspective on numerical tasks

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 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 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 $\mathsf{Q}(u) \approx \mathsf{B}(\mathsf{Y}(u))$.

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 $\mathsf{Q}(u) \approx \mathsf{B}(\mathsf{Y}(u))$.

Example (Quadrature)

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

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

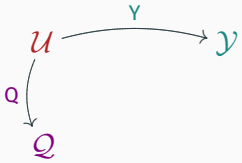
- Conventional numerical methods are cleverly-designed functions $\mathsf{B}\colon \mathcal{Y} \to \mathcal{Q}$: such a method **"believes"** that $\mathsf{Q}(u) \approx \mathsf{B}(\mathsf{Y}(u))$.
- N.B. *Some* methods try to invert Y, form an estimate of u, then apply Q, but not all do!
  - E.g. the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

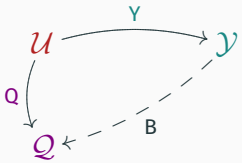
$$\mathsf{B}_{\text{trap}}\left((t_j,z_j)_{j=1}^J\right) := \sum_{j=1}^{J-1} \frac{z_{j+1}+z_j}{2}(t_{j+1}-t_j) = z_1\frac{t_2-t_1}{2} + \sum_{j=2}^{J-1} z_j \frac{t_{j+1}-t_{j-1}}{2} + z_J\frac{t_J - t_{J-1}}{2}.$$

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

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

- 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): residual $\mathsf{B} \circ \mathsf{Y} - \mathsf{Q}$ 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?


Y
$\mathcal{U}$
$\mathcal{Y}$
Q
$\mathcal{Q}$



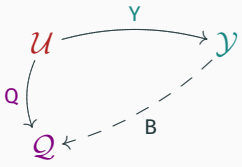
$\mathsf{B}\colon \mathcal{Y} \to \mathcal{Q}$

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, $\mathsf{Q}(u)$.

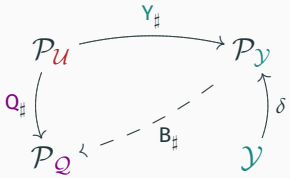


$\mathsf{B}\colon \mathcal{Y} \to \mathcal{Q}$

Go Probabilistic!

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

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



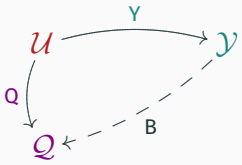
average-case performance of B?

Example (Quadrature)

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

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

A deterministic numerical method uses only the spaces and data to produce a point estimate of the integral, $\mathsf{Q}(u)$.

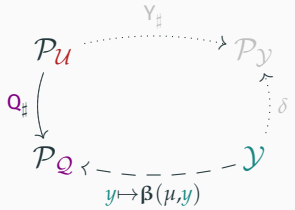


$\mathsf{B}\colon \mathcal{Y} \to \mathcal{Q}$

Go Probabilistic!

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

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



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

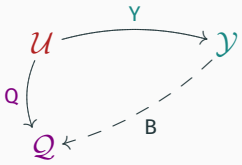
Example (Quadrature)

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

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

A deterministic numerical method uses only the spaces and data to produce a point estimate of the integral, $\mathsf{Q}(u)$.

A probabilistic numerical method converts an additional belief $\mu \in \mathcal{P}_{\mathcal{U}}$ about u into a belief $\beta(\mu, \mathsf{Y}(u)) \in \mathcal{P}_{\mathcal{Q}}$ about $\mathsf{Q}(u)$.

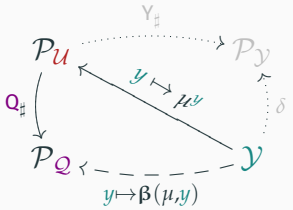


$\mathsf{B}\colon \mathcal{Y} \to \mathcal{Q}$

Go Probabilistic!

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

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



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

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:

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

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:

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

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 $\leftrightarrow$ 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	$\arg\min x(t)$	$\{x(t_i)\}_{i=1}^n$		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] 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 operators: the Worst, the Average, and the Bayesian

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
$$e_{\mathrm{WC}}(\mathsf{Y}, \mathsf{B}) := \sup_{u \in \mathcal{U}} L\big(\mathsf{B}(\mathsf{Y}(u)), \mathsf{Q}(u)\big).$$
- The **average-case loss** under a probability measure $\mu \in \mathcal{P}_{\mathcal{U}}$ is
$$e_{\mathrm{AC}}(\mathsf{Y}, \mathsf{B}) := \int_{\mathcal{U}} L\big(\mathsf{B}(\mathsf{Y}(u)), \mathsf{Q}(u)\big)\, \mu(\mathrm{d}u),$$
$$e_{\mathrm{AC}}(\mathsf{Y}, \beta) := \int_{\mathcal{U}} \left[\int_{\mathcal{Q}} L\big(q, \mathsf{Q}(u)\big)\, \beta(\mu, \mathsf{Y}(u))(\mathrm{d}q) \right] \mu(\mathrm{d}u).$$
- 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**
$$e_{\mathrm{BPN}}(\mathsf{Y}) := \int_{\mathcal{U}} \left[\int_{\mathcal{Q}} L\big(q, \mathsf{Q}(u)\big)\, (\mathsf{Q}_\sharp \mu^{\mathsf{Y}(u)})(\mathrm{d}q) \right] \mu(\mathrm{d}u).$$

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 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 $\neq$ 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.

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).

$$\mathcal{U} = \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\} \qquad \mu = \mathrm{Unif}_{\mathcal{U}} \qquad \mathcal{Q} = \{0, 1\} \subset \mathbb{R}$$
$$\mathcal{Y}_1 = \{0, 1\} \qquad \mathsf{Y}_1(u) = \mathbb{1}[u \in \{\diamondsuit, \heartsuit\}] \qquad \mathsf{Q}(u) = \mathbb{1}[u = \diamondsuit]$$
$$\mathcal{Y}_2 = \{0, 1\} \qquad \mathsf{Y}_2(u) = \mathbb{1}[u \in \{\diamondsuit, \heartsuit, \spadesuit\}] \qquad L(q, q') = \mathbb{1}[q \neq q']$$

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!

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

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

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

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

Optimal information: AC $\neq$ BPN!

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

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

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

$u =$	$\clubsuit$		$\diamondsuit$		$\heartsuit$		$\spadesuit$	
$e_{\text{AC}}(\mathsf{Y}_1, \mathsf{B}) = \frac{1}{4}($	$L(\mathsf{B}(\blacksquare), 0)$	$+$	$L(\mathsf{B}(\blacksquare), 1)$	$+$	$L(\mathsf{B}(\blacksquare), 0)$	$+$	$L(\mathsf{B}(\blacksquare), 0)$	$)$
$e_{\text{AC}}(\mathsf{Y}_1, 0) = \frac{1}{4}($	0	$+$	1	$+$	0	$+$	0	$) = \frac{1}{4}$
$e_{\text{AC}}(\mathsf{Y}_1, \text{id}) = \frac{1}{4}($	0	$+$	0	$+$	1	$+$	0	$) = \frac{1}{4}$

Optimal information: AC $\neq$ BPN!

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

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

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

$u =$	$\clubsuit$		$\diamondsuit$		$\heartsuit$		$\spadesuit$		
$e_{\mathrm{AC}}(\mathsf{Y}_1, \mathsf{B}) = \frac{1}{4}($	$L(\mathsf{B}(\blacksquare), 0)$	$+$	$L(\mathsf{B}(\blacksquare), 1)$	$+$	$L(\mathsf{B}(\blacksquare), 0)$	$+$	$L(\mathsf{B}(\blacksquare), 0)$	$)$	
$e_{\mathrm{AC}}(\mathsf{Y}_1, 0) = \frac{1}{4}($	0	$+$	1	$+$	0	$+$	0	$) = \frac{1}{4}$	
$e_{\mathrm{AC}}(\mathsf{Y}_1, \mathrm{id}) = \frac{1}{4}($	0	$+$	0	$+$	1	$+$	0	$) = \frac{1}{4}$	
$e_{\mathrm{AC}}(\mathsf{Y}_2, \mathsf{B}) = \frac{1}{4}($	$L(\mathsf{B}(\clubsuit), 0)$	$+$	$L(\mathsf{B}(\neg\clubsuit), 1)$	$+$	$L(\mathsf{B}(\neg\clubsuit), 0)$	$+$	$L(\mathsf{B}(\neg\clubsuit), 0)$	$)$	
$e_{\mathrm{AC}}(\mathsf{Y}_2, 0) = \frac{1}{4}($	0	$+$	1	$+$	0	$+$	0	$) = \frac{1}{4}$	

Optimal information: AC $\neq$ BPN!

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

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

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

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

Optimal information: AC $\neq$ BPN!

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

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

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

$u =$	$\clubsuit$	$\diamondsuit$	$\heartsuit$	$\spadesuit$	
$e_{\mathrm{AC}}(\mathsf{Y}_1, \mathsf{B}) = \frac{1}{4}($	$L(\mathsf{B}(\blacksquare), 0)$ +	$L(\mathsf{B}(\blacksquare), 1)$ +	$L(\mathsf{B}(\blacksquare), 0)$ +	$L(\mathsf{B}(\blacksquare), 0)$ $)$	
$e_{\mathrm{AC}}(\mathsf{Y}_1, 0) = \frac{1}{4}($	0 +	1 +	0 +	0 $)$	$= \frac{1}{4}$
$e_{\mathrm{AC}}(\mathsf{Y}_1, \mathrm{id}) = \frac{1}{4}($	0 +	0 +	1 +	0 $)$	$= \frac{1}{4}$
$e_{\mathrm{AC}}(\mathsf{Y}_2, \mathsf{B}) = \frac{1}{4}($	$L(\mathsf{B}(\clubsuit), 0)$ +	$L(\mathsf{B}(\neg\clubsuit), 1)$ +	$L(\mathsf{B}(\neg\clubsuit), 0)$ +	$L(\mathsf{B}(\neg\clubsuit), 0)$ $)$	
$e_{\mathrm{AC}}(\mathsf{Y}_2, 0) = \frac{1}{4}($	0 +	1 +	0 +	0 $)$	$= \frac{1}{4}$
$e_{\mathrm{BPN}}(\mathsf{Y}_1) = \frac{1}{4}($	$\mathbb{E}_{\mathsf{Q}_\sharp \mu(\cdot\mid\blacksquare)} L(\cdot, 0)$ +	$\mathbb{E}_{\mathsf{Q}_\sharp \mu(\cdot\mid\blacksquare)} L(\cdot, 1)$ +	$\mathbb{E}_{\mathsf{Q}_\sharp \mu(\cdot\mid\blacksquare)} L(\cdot, 0)$ +	$\mathbb{E}_{\mathsf{Q}_\sharp \mu(\cdot\mid\blacksquare)} L(\cdot, 0)$ $)$	
$= \frac{1}{4}($	$(\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0)$ +	$(\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 1)$ +	$(\frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 0)$ +	$(\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0)$ $)$	$= \frac{1}{4}$
$e_{\mathrm{BPN}}(\mathsf{Y}_2) = \frac{1}{4}($	$\mathbb{E}_{\mathsf{Q}_\sharp \mu(\cdot\mid\clubsuit)} L(\cdot, 0)$ +	$\mathbb{E}_{\mathsf{Q}_\sharp \mu(\cdot\mid\neg\clubsuit)} L(\cdot, 1)$ +	$\mathbb{E}_{\mathsf{Q}_\sharp \mu(\cdot\mid\neg\clubsuit)} L(\cdot, 0)$ +	$\mathbb{E}_{\mathsf{Q}_\sharp \mu(\cdot\mid\neg\clubsuit)} L(\cdot, 0)$ $)$	
$= \frac{1}{4}($	$(1\cdot 0)$	$+ (\frac{1}{3}\cdot 0 + \frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 1)$	$+ (\frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 0 + \frac{1}{3}\cdot 0)$	$+ (\frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 0 + \frac{1}{3}\cdot 0)$ $)$	$= \frac{1}{3}$

Disintegration: Exact and numerical

- 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**.

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.)

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.)

- The familiar conditional densities for a probability density on $\mathbb{R}^n$ conditioned on a "nice" subset such as a lower-dimensional submanifold $M \subset \mathbb{R}^n$ are disintegrations.
- In particular, the familiar Woodbury formula for the conditioning of Gaussian measures subject to linear constraints is a disintegration (Owhadi and Scovel, 2015).
- But, in general, disintegrations cannot be computed exactly — we have to work approximately.

- 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 $\mu^y_\delta(\mathrm{d}u) \propto \phi(\|\mathsf{Y}(u) - y\|_\mathcal{Y}/\delta)\,\mu(\mathrm{d}u)$ 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

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

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

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

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 $\mu_\delta^y \approx \mu^y$ as a function of δ. That is,

$$d_{\mathcal{F}}(\mu^y, \mu^{y'}) \le C \cdot \|y - y'\|^\gamma \qquad \text{for } y, y' \in \mathcal{Y}$$
$$\implies \quad d_{\mathcal{F}}(\mu^y, \mu_\delta^y) \le C \cdot C_\phi \cdot \delta^\gamma \qquad \text{for } \mathsf{Y}_\sharp\mu\text{-almost all } y \in \mathcal{Y}.$$

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.)

A simple but multivalent boundary value problem:

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

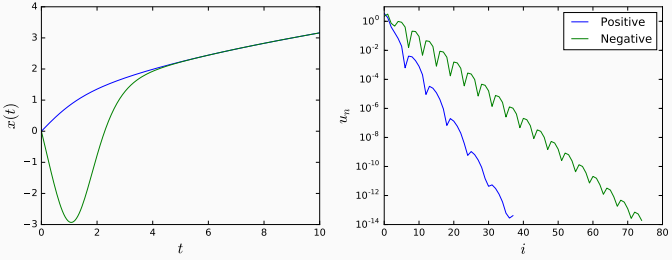


Figure 2: The two solutions of Painlevé's first transcendental and their spectra in the orthonormal Chebyshev polynomial basis over $[0, 10]$.

A simple but multivalent boundary value problem:

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

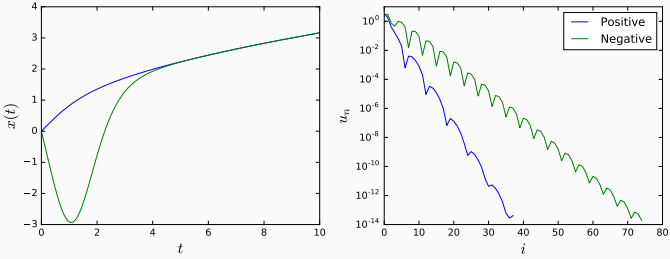
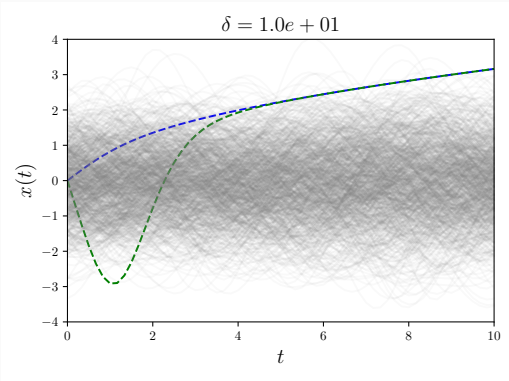


Figure 2: The two solutions of Painlevé's first transcendental and their spectra in the orthonormal Chebyshev polynomial basis over $[0, 10]$.

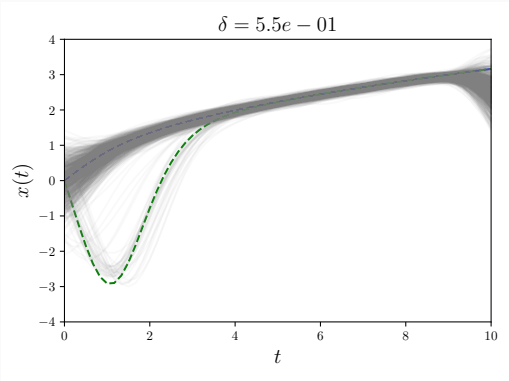
Example: Painlevé's first transcendental II

- 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... ✗



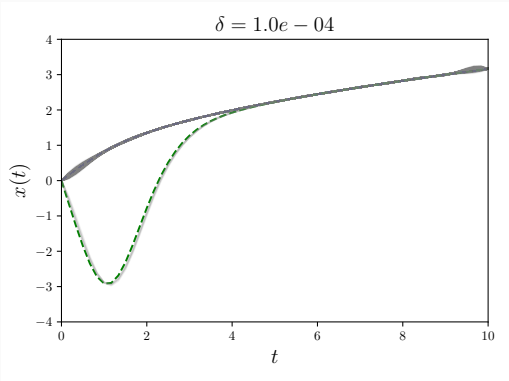
Example: Painlevé's first transcendental II

- 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... ✗

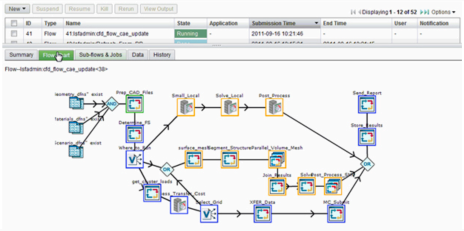


Example: Painlevé's first transcendental II

- 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... ✗



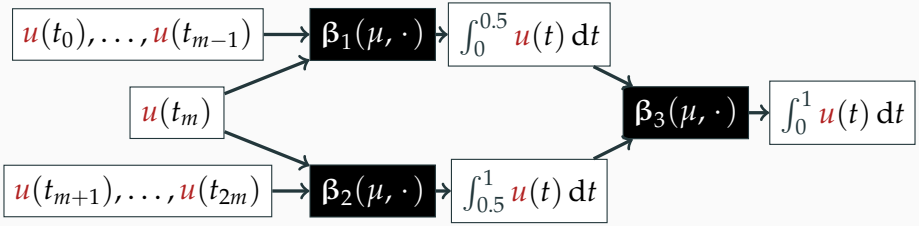
Coherent pipelines of PNMs, and Bayesian inverse problems



- 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:

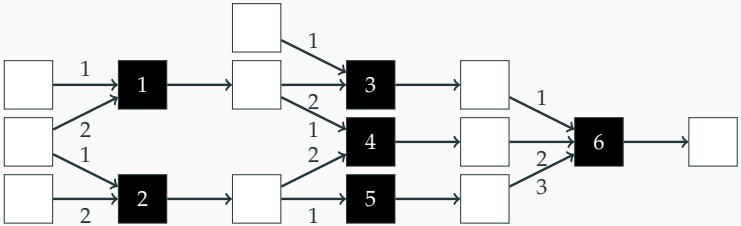
  Make the forward and inverse problem
  speak the same statistical language!

- We can compose PNMs in series, e.g. $\boldsymbol{\beta}_2(\boldsymbol{\beta}_1(\mu, y_1), y_2)$ is formally $\boldsymbol{\beta}(\mu, (y_1, y_2))$... although figuring out what the spaces $\mathcal{U}_i$, $\mathcal{Y}_i$ and operators Y_i etc. are is a headache!

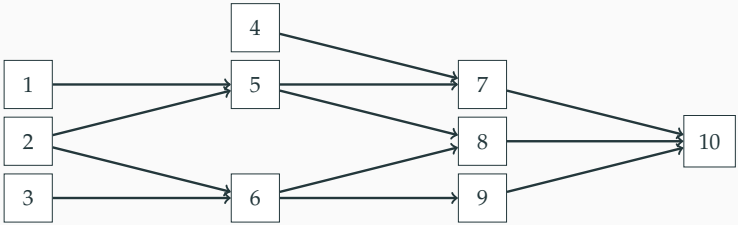


- Integrate a function over $[0, 1]$ in two steps using nodes $0 \leq t_0 < \cdots < t_{m-1} < 0.5$, $t_m = 0.5$, and $t_{m+1} < \cdots < t_{2m} \leq 1$.
- For example, the two nodal sets are very large, and so two are handled by two different processors with non-shared memory.
- A third processor handles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

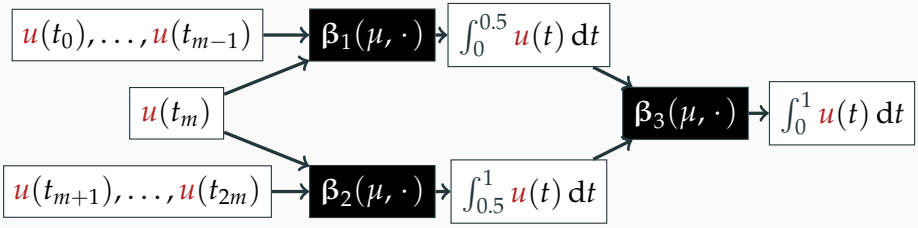
- We compose PNMs in a graphical way by allowing input information nodes (□) to feed into method nodes (■), which in turn output new information.
- N.B. one should at first think of having *deterministic* data at the left-most □ nodes, then *random variables* as outputs, *realisations* of which get fed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



- We 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.

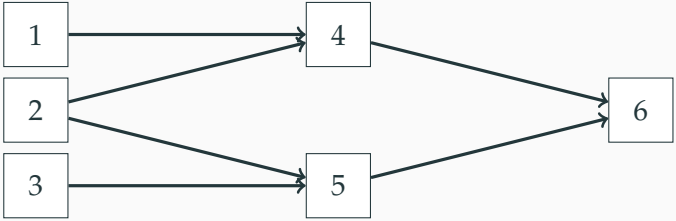


Definition

A prior μ is **coherent** for the dependency graph if — when the "leaf" input nodes are $Y_\sharp\mu$-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:

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

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.

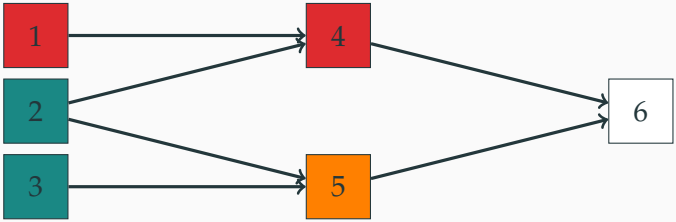


Definition

A prior μ is **coherent** for the dependency graph if — when the "leaf" input nodes are $Y_\sharp\mu$-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:

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

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.



Definition

A prior μ is **coherent** for the dependency graph if — when the "leaf" input nodes are $Y_\sharp\mu$-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:

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

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.

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 

.

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

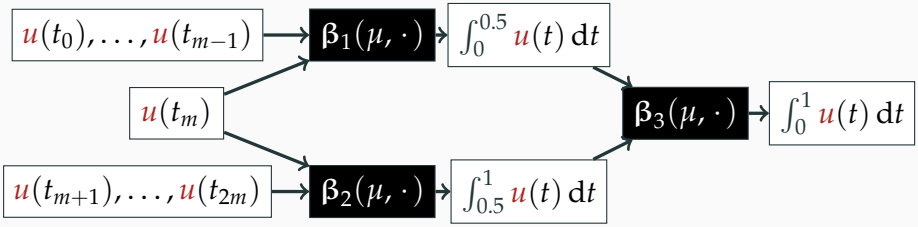
If a pipeline of PNMs is such that

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

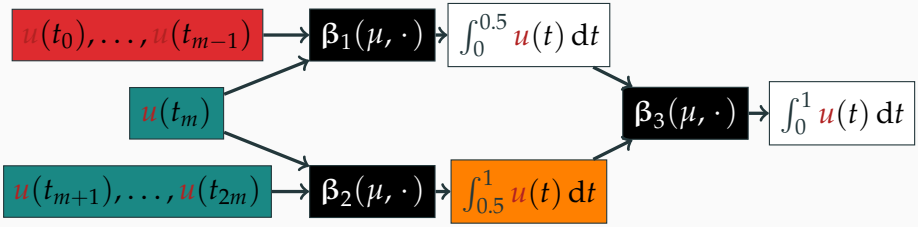
then the pipeline is the Bayesian pipeline 

.

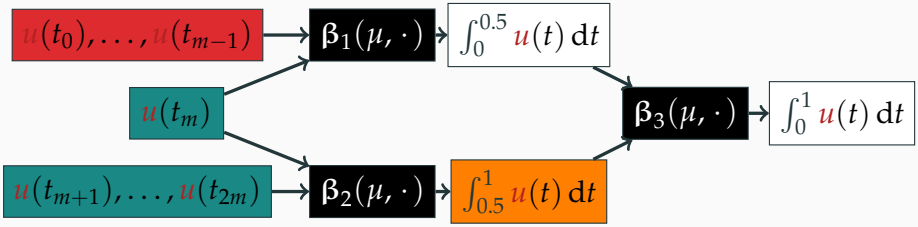
- Redundant structure in the pipeline (recycled information) will break coherence, and hence Bayesianity of the pipeline.
- In principle, coherence and hence being Bayesian depend upon the prior.
- This should not be surprising — as a loose analogy, one doesn't expect the trapezoidal rule to be a good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



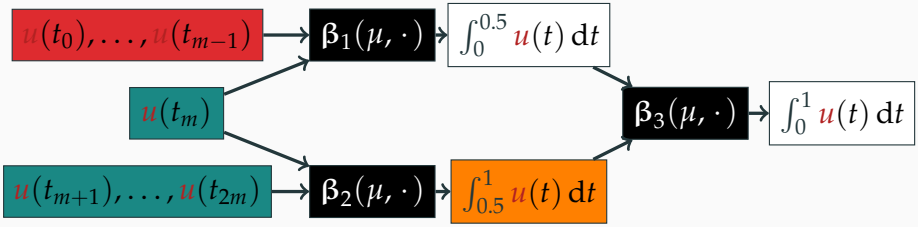
- Integrate a function over $[0, 1]$ in two steps using nodes $0 \leq t_0 < \cdots < t_{m-1} < 0.5$, $t_m = 0.5$, and $t_{m+1} < \cdots < t_{2m} \leq 1$.



- 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 ■ $(u(t_m), \ldots, u(t_{2m}))$?

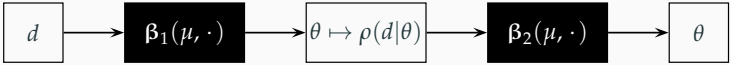


- Integrate a function over $[0, 1]$ in two steps using nodes $0 \leq t_0 < \cdots < t_{m-1} < 0.5$, $t_m = 0.5$, and $t_{m+1} < \cdots < t_{2m} \leq 1$.
- Is ■ $(\int_{0.5}^1 u(t)\,\mathrm{d}t)$ independent of ■ $(u(t_0), \ldots, u(t_{m-1}))$ given ■ $(u(t_m), \ldots, u(t_{2m}))$?
- For a Brownian motion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



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

- 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

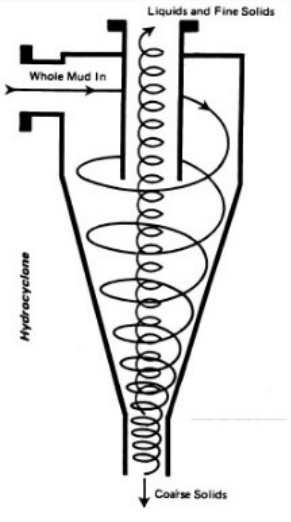


- β_1 converts data d into the likelihood function for parameters θ, and hence incorporates any forward model such as an O/PDE solver.
- β_2 converts the prior on θ and the likelihood into a joint distribution for (θ, d), then conditions upon the actual observation — it returns something in $\mathcal{P}_\Theta$.
- 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).

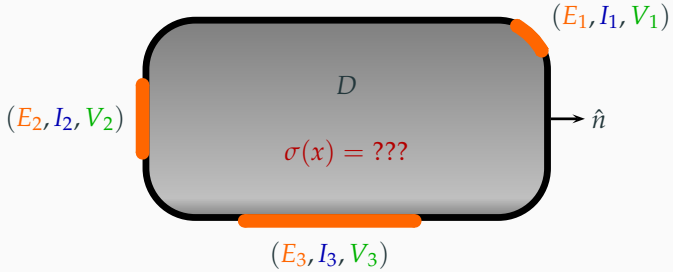


The interior conductivity field σ and electrical potential field v and the applied boundary currents I_i, measured voltages V_i, and known contact impedances ζ_i are related by

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

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

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



- 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).

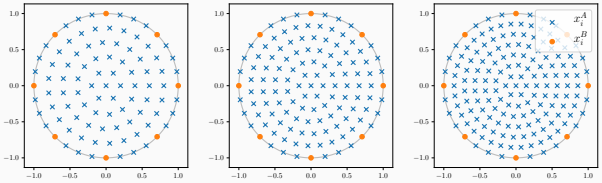


Figure 3: Like collocation, PMM imposes the PDE relation at $n_{\mathcal{A}}$ interior nodes and boundary conditions at $n_{\mathcal{B}}$ boundary nodes.

- 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.

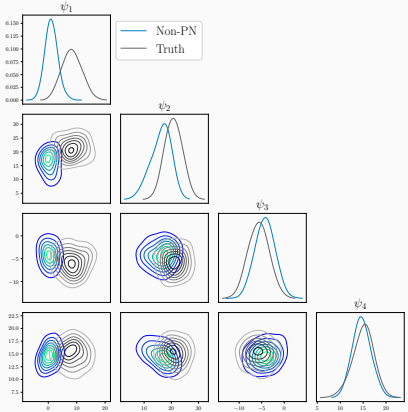


Figure 4: A small number $n_{\mathcal{A}} + n_{\mathcal{B}} = 71$ of collocation points was used to discretise the PDE, but the uncertainty due to discretisation was not modelled. The reference posterior distribution over the coefficients ψ_k is plotted (grey) and compared to the approximation to the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

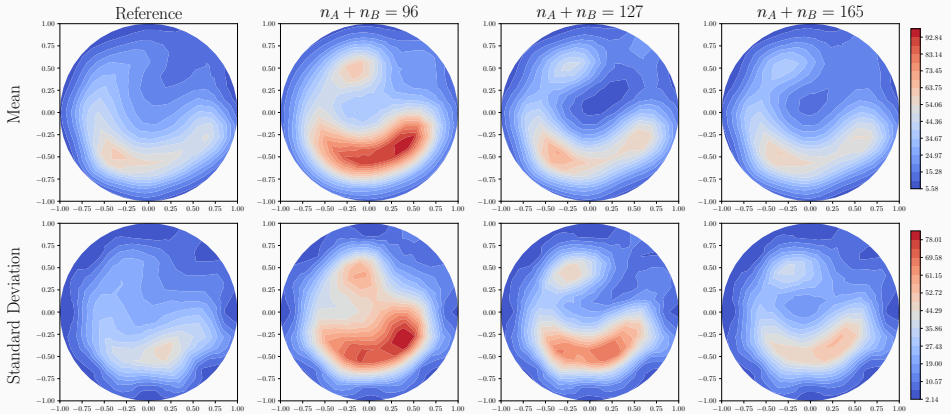


Figure 5: Posterior means and standard-deviations for the recovered conductivity field at $t = 14$. The first column shows the reference solution, obtained using symmetric collocation with a large number of collocation points. The remaining columns show the recovered field when PMM is used with $n_{\mathcal{A}}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

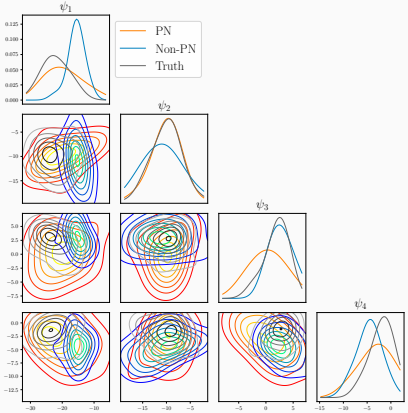


Figure 6: Posterior distribution over the coefficients ψ_k at the final time. A small number $n_{\mathcal{A}} + n_{\mathcal{B}} = 71$ of collocation points was used to discretise the PDE. The reference posterior distribution over the coefficients ψ_k is plotted (grey) and compared to the approximation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

- 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 1

J. L. Barlow and E. H. Bareiss. Probabilistic error analysis of Gaussian elimination in floating point and logarithmic arithmetic. *Computing*, 34(4):349–364, 1985. doi:10.1007/BF02251834.

J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 1985. doi:10.1007/978-1-4757-4286-2.

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.

F. Chatelin and M.-C. Brunet. A probabilistic round-off error propagation model. Application to the eigenvalue problem. In *Reliable numerical computation*, Oxford Sci. Publ., pages 139–160. Oxford Univ. Press, New York, 1990.

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.

References II

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.*, 61(4): 756–789, 2019. doi:10.1137/17M1139357.

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. 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. Henrici. *Discrete Variable Methods in Ordinary Differential Equations*. John Wiley & Sons, Inc., New York-London, 1962.

T. E. Hull and J. R. Swenson. Tests of probabilistic models for the propagation of roundoff errors. *Comm. ACM*, 9:108–113, 1966. doi:10.1145/365170.365212.

References III

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.

T. Karvonen, C. J. Oates, and S. Särkkä. A Bayes–Sard cubature method. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, 2018. papers.nips.cc/paper/7829-a-bayes-sard-cubature-method.

Kazan Federal University. kpfu.ru/portal/docs/F_261937733/suldin2.jpg. Accessed December 2018.

G. S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Stat.*, 41:495–502, 1970a. doi:10.1214/aoms/1177697089.

G. S. Kimeldorf and G. Wahba. Spline functions and stochastic processes. *Sankhyā Ser. A*, 32:173–180, 1970b. www.jstor.org/stable/25049652.

A. N. Kolmogorov. Über die beste Annäherung von Funktionen einer gegebenen Funktionenklasse. *Ann. of Math. (2)*, 37 (1):107–110, 1936. doi:10.2307/1968691.

J. Kuelbs, F. M. Larkin, and J. A. Williamson. Weak probability distributions on reproducing kernel Hilbert spaces. *Rocky Mountain J. Math.*, 2(3):369–378, 1972. doi:10.1216/RMJ-1972-2-3-369.

F. M. Larkin. Estimation of a non-negative function. *BIT Num. Math.*, 9(1):30–52, 1969. doi:10.1007/BF01933537.

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

References iv

F. M. Larkin. Gaussian measure in Hilbert space and applications in numerical analysis. *Rocky Mountain J. Math.*, 2(3): 379–421, 1972. doi:10.1216/RMJ-1972-2-3-379.

F. M. Larkin. Probabilistic error estimates in spline interpolation and quadrature. In *Information Processing 74 (Proc. IFIP Congress, Stockholm, 1974)*, pages 605–609, Amsterdam, 1974. North-Holland.

F. M. Larkin. A modification of the secant rule derived from a maximum likelihood principle. *BIT*, 19(2):214–222, 1979a. doi:10.1007/BF01930851.

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

F. M. Larkin. Bayesian estimation of zeros of analytic functions. Technical report, Queen's University of Kingston. Department of Computing and Information Science., 1979c.

F. M. Larkin, C. E. Brown, K. W. Morton, and P. Bond. Worth a thousand words, 1967. www.amara.org/en/videos/7De21CeNlz8b/info/worth-a-thousand-words-1967/.

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. Comp.*, 29(6):1265–1283, 2019. doi:10.1007/s11222-019-09898-6.

T. Minka. Deriving quadrature rules from Gaussian processes, 2000. www.microsoft.com/en-us/research/publication/deriving-quadrature-rules-gaussian-processes/.

J. Močkus. On Bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974. Optimization Techniques 1974*, volume 27 of *Lecture Notes in Computer Science*, pages 400–404. Springer, Berlin, Heidelberg, 1975. doi:10.1007/3-540-07165-2_55.

J. Močkus. On Bayesian methods for seeking the extremum and their application. In *Information Processing 77 (Proc. IFIP Congr., Toronto, Ont., 1977)*, pages 195–200. IFIP Congr. Ser., Vol. 7. North-Holland, Amsterdam, 1977.

J. Močkus. *Bayesian approach to global optimization*, volume 37 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht, 1989. doi:10.1007/978-94-009-0909-0.

A. P. Norden, Y. I. Zabotin, L. D. Èskin, S. V. Grigor′ev, and E. A. Begovatov. Al′bert Valentinovich Sul′din (on the occasion of his fiftieth birthday). *Izv. Vysš. Učebn. Zaved. Mat.*, 12:3–5, 1978.

E. Novak. *Deterministic and Stochastic Error Bounds in Numerical Analysis*, volume 1349 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1988. doi:10.1007/BFb0079792.

C. J. Oates and T. J. Sullivan. A modern retrospective on probabilistic numerics. *Stat. Comp.*, 29(6):1335–1351, 2019. doi:10.1007/s11222-019-09902-z.

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.

References vi

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

A. O'Hagan. Bayes–Hermite quadrature. *J. Stat. Plann. Inference*, 29(3):245–260, 1991. doi:10.1016/0378-3758(91)90002-V.

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

E. Parzen. Statistical inference on time series by RKHS methods. Technical report, Stanford University of California, Department of Statistics, 1970.

H. Poincaré. *Calcul des Probabilités*. Gauthier-Villars, second edition, 1912.

C. E. Rasmussen and Z. Ghahramani. Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems 16*, pages 505–512, 2003. papers.nips.cc/paper/2150-bayesian-monte-carlo.

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.

A. Sard. *Linear Approximation*. Number 9 in Mathematical Surveys. American Mathematical Society, Providence, RI, 1963. doi:10.1090/surv/009.

M. Schober, D. K. Duvenaud, and P. Hennig. Probabilistic ODE solvers with Runge–Kutta means. In *Advances in Neural Information Processing Systems 27*, 2014. papers.nips.cc/paper/5451-probabilistic-ode-solvers-with-runge-kutta-means.

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.

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.

S. Smale. On the efficiency of algorithms of analysis. *Bull. Amer. Math. Soc. (N.S.)*, 13(2):87–121, 1985. doi:10.1090/S0273-0979-1985-15391-1.

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

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

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

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

A. V. Sul′din. On the distribution of the functional $\int_0^1 x^2(t)\,\mathrm{d}t$ where $x(t)$ represents a certain Gaussian process. In *Kazan State Univ. Sci. Survey Conf. 1962 (Russian)*, pages 80–82. Izdat. Kazan. Univ., Kazan, 1963.

M. Tienari. A statistical model of roundoff error for varying length floating-point arithmetic. *Nordisk Tidskr. Informationsbehandling (BIT)*, 10:355–365, 1970. doi:10.1007/BF01934204.

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

J. F. Traub and H. Woźniakowsi. *A General Theory of Optimal Algorithms*. ACM Monograph Series. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1980.

J. F. Traub, G. W. Wasilkowski, and H. Woźniakowski. *Information, Uncertainty, Complexity*. Addison-Wesley Publishing Company, Advanced Book Program, Reading, MA, 1983.

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

J. von Neumann and H. H. Goldstine. Numerical inverting of matrices of high order. *Bull. Amer. Math. Soc.*, 53:1021–1099, 1947. doi:10.1090/S0002-9904-1947-08909-6.

Y. I. Zabotin, N. K. Zamov, L. A. Aksent'ev, and T. N. Zemtseva. Al'bert Valentinovich Sul'din (obituary). *Izv. Vyssˇ. Učebn. Zaved. Mat.*, 2(84), 1996.