Bayesian Probabilistic Numerical Methods (Part I)

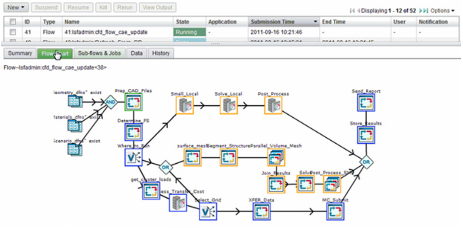
Chris. J. Oates
Newcastle University
Alan Turing Institute

Joint work with: Jon Cockayne, Tim Sullivan and Mark Girolami

June 2017 @ ICERM

Motivation: Computational Pipelines

Numerical analysis for the "drag and drop" era of computational pipelines:



[Fig: IBM High Performance Computation]

The sophistication and scale of modern computer models creates an urgent need to better understand the propagation and accumulation of numerical error within arbitrary - often large - pipelines of computation, so that "numerical risk" to end-users can be controlled.

Motivation: Solution of Poisson's Equation

Consider numerical solution for $x \in \mathcal{X}$ of the Poisson equation

$$\begin{aligned} -\Delta x &= f && \text{in } D \\ x &= g && \text{on } \partial D \end{aligned}$$

based on (noiseless) information of the form

$$A(x) = \begin{bmatrix} -\Delta x(t_1) \\ \vdots \\ -\Delta x(t_m) \\ x(t_{m+1}) \\ \vdots \\ x(t_n) \end{bmatrix} = \begin{bmatrix} f(t_1) \\ \vdots \\ f(t_m) \\ g(t_{m+1}) \\ \vdots \\ g(t_n) \end{bmatrix}, \qquad \{t_i\}_{i=1}^m \in D, \quad \{t_i\}_{i=m+1}^d \in \partial D.$$

This is an ill-posed inverse problem and must be regularised.

The onus is on us to establish principled statistical foundations that are general.

Motivation: Solution of Poisson's Equation

Consider numerical solution for $x \in \mathcal{X}$ of the Poisson equation

$$\begin{aligned} -\Delta x &= f && \text{in } D \\ x &= g && \text{on } \partial D \end{aligned}$$

based on (noiseless) information of the form

$$A(x) = \begin{bmatrix} -\Delta x(t_1) \\ \vdots \\ -\Delta x(t_m) \\ x(t_{m+1}) \\ \vdots \\ x(t_n) \end{bmatrix} = \begin{bmatrix} f(t_1) \\ \vdots \\ f(t_m) \\ g(t_{m+1}) \\ \vdots \\ g(t_n) \end{bmatrix}, \qquad \{t_i\}_{i=1}^{m} \in D, \quad \{t_i\}_{i=m+1}^{d} \in \partial D.$$

This is an ill-posed inverse problem and must be regularised.

The onus is on us to establish principled statistical foundations that are general.

Motivation: Solution of Poisson's Equation

Consider numerical solution for $x \in \mathcal{X}$ of the Poisson equation

$$\begin{aligned} -\Delta x &= f && \text{in } D \\ x &= g && \text{on } \partial D \end{aligned}$$

based on (noiseless) information of the form

$$A(x) = \begin{bmatrix} -\Delta x(t_1) \\ \vdots \\ -\Delta x(t_m) \\ x(t_{m+1}) \\ \vdots \\ x(t_n) \end{bmatrix} = \begin{bmatrix} f(t_1) \\ \vdots \\ f(t_m) \\ g(t_{m+1}) \\ \vdots \\ g(t_n) \end{bmatrix}, \qquad \{t_i\}_{i=1}^m \in D, \quad \{t_i\}_{i=m+1}^d \in \partial D.$$

This is an ill-posed inverse problem and must be regularised.

The onus is on us to establish principled statistical foundations that are general.

The *Bayesian* approach, popularised in Stuart (2010), can be used:

- a *prior* measure P_x is placed on $\mathcal{X}$
- a *posterior* measure $P_{x|a}$ is defined as the "restriction of P_x to those functions $x \in \mathcal{X}$ for which

$$A(x) = a \qquad\qquad \text{e.g.} \quad A(x) = \begin{bmatrix} -\Delta x(t_1) \\ \vdots \\ -\Delta x(t_n) \end{bmatrix} = a$$

is satisfied" (to be formalised).

$\implies$ Principled and general uncertainty quantification for numerical methods.

Insight: Numerical Analysis as Bayesian Inversion

The *Bayesian* approach, popularised in Stuart (2010), can be used:

- a *prior* measure P_x is placed on $\mathcal{X}$
- a *posterior* measure $P_{x|a}$ is defined as the "restriction of P_x to those functions $x \in \mathcal{X}$ for which

$$A(x) = a \qquad \text{e.g.} \quad A(x) = \begin{bmatrix} -\Delta x(t_1) \\ \vdots \\ -\Delta x(t_n) \end{bmatrix} = a$$

is satisfied" (to be formalised).

$\Longrightarrow$ Principled and general uncertainty quantification for numerical methods.

The Research Agenda

Part I

1. First Job: Elicit the Abstract Structure
2. Second Job: Check Well-Defined, Existence and Uniqueness
3. Third Job: Characterise Optimal Information

Part II

4. Fourth Job: Algorithms to Access $P_{x|a}$
5. Fifth Job: Extend to Pipelines of Computation

First Job: Elicit the Abstract Structure

Abstract Structure

Abstractly, consider an unobserved state variable $x \in \mathcal{X}$ together with:

- A *quantity of interest*, denoted $Q(x) \in \mathcal{Q}$
- An *information operator*, denoted $x \mapsto A(x) \in \mathcal{A}$.

Examples:

Task	$Q(x)$	$A(x)$
Integration	$\int x(t)\nu(\mathrm{d}t)$	$\{x(t_i)\}_{i=1}^n$
Optimisation	$\arg\max x(t)$	$\{x(t_i)\}_{i=1}^n$
Solution of Poisson Eqn	$x(\cdot)$	$\{-\Delta x(t_i)\}_{i=1}^m \cup \{x(t_i)\}_{i=m+1}^n$

Abstract Structure

Abstractly, consider an unobserved state variable $x \in \mathcal{X}$ together with:

- A *quantity of interest*, denoted $Q(x) \in \mathcal{Q}$
- An *information operator*, denoted $x \mapsto A(x) \in \mathcal{A}$.

Examples:

Task	$Q(x)$	$A(x)$
Integration	$\int x(t)\nu(\mathrm{d}t)$	$\{x(t_i)\}_{i=1}^n$
Optimisation	$\arg\max x(t)$	$\{x(t_i)\}_{i=1}^n$
Solution of Poisson Eqn	$x(\cdot)$	$\{-\Delta x(t_i)\}_{i=1}^m \cup \{x(t_i)\}_{i=m+1}^n$

Let $\mathcal{P}_\bullet$ denote the set of distributions on $\bullet$.

Let $M_\# \mu$ denote the "pushforward" measure, st $(M_\# \mu)(S) = \mu(M^{-1}(S))$.

		Classical Numerical Method	Probabilistic Numerical Method
Inputs	Assumed	e.g. smoothness	$P_x \in \mathcal{P}_\mathcal{X}$
	Information	$a \in \mathcal{A}$	$a \in \mathcal{A}$
Output		$b(a) \in \mathcal{Q}$	$B(P_x, a) \in \mathcal{P}_\mathcal{Q}$

A Probabilistic Numerical Method is Bayesian iff $B(P_x, a) = Q_\# P_{x|a}$.

Let $\mathcal{P}_\bullet$ denote the set of distributions on $\bullet$.

Let $M_\# \mu$ denote the "pushforward" measure, st $(M_\# \mu)(S) = \mu(M^{-1}(S))$.

		Classical Numerical Method	Probabilistic Numerical Method
Inputs	Assumed	e.g. smoothness	$P_x \in \mathcal{P}_\mathcal{X}$
	Information	$a \in \mathcal{A}$	$a \in \mathcal{A}$
Output		$b(a) \in \mathcal{Q}$	$B(P_x, a) \in \mathcal{P}_\mathcal{Q}$

A Probabilistic Numerical Method is Bayesian iff $B(P_x, a) = Q_\# P_{x|a}$.

Abstract Structure

Let $\mathcal{P}_\bullet$ denote the set of distributions on $\bullet$.

Let $M_\# \mu$ denote the "pushforward" measure, st $(M_\# \mu)(S) = \mu(M^{-1}(S))$.

		Classical Numerical Method	Probabilistic Numerical Method
Inputs	Assumed	e.g. smoothness	$P_x \in \mathcal{P}_\mathcal{X}$
	Information	$a \in \mathcal{A}$	$a \in \mathcal{A}$
Output		$b(a) \in \mathcal{Q}$	$B(P_x, a) \in \mathcal{P}_\mathcal{Q}$

A Probabilistic Numerical Method is Bayesian iff $B(P_x, a) = Q_\# P_{x|a}$.

Abstract Structure

Let $\mathcal{P}_\bullet$ denote the set of distributions on $\bullet$.

Let $M_\# \mu$ denote the "pushforward" measure, st $(M_\# \mu)(S) = \mu(M^{-1}(S))$.

		Classical Numerical Method	Probabilistic Numerical Method
Inputs	Assumed	e.g. smoothness	$P_x \in \mathcal{P}_\mathcal{X}$
	Information	$a \in \mathcal{A}$	$a \in \mathcal{A}$
Output		$b(a) \in \mathcal{Q}$	$B(P_x, a) \in \mathcal{P}_\mathcal{Q}$

A Probabilistic Numerical Method is Bayesian iff $B(P_x, a) = Q_\# P_{x|a}$.

Abstract Structure

Let $\mathcal{P}_\bullet$ denote the set of distributions on $\bullet$.

Let $M_\# \mu$ denote the "pushforward" measure, st $(M_\# \mu)(S) = \mu(M^{-1}(S))$.

		Classical Numerical Method	Probabilistic Numerical Method
Inputs	Assumed	e.g. smoothness	$P_x \in \mathcal{P}_\mathcal{X}$
	Information	$a \in \mathcal{A}$	$a \in \mathcal{A}$
Output		$b(a) \in \mathcal{Q}$	$B(P_x, a) \in \mathcal{P}_\mathcal{Q}$

A Probabilistic Numerical Method is Bayesian iff $B(P_x, a) = Q_\# P_{x|a}$.

Abstract Structure

Let $\mathcal{P}_\bullet$ denote the set of distributions on $\bullet$.

Let $M_\# \mu$ denote the "pushforward" measure, st $(M_\# \mu)(S) = \mu(M^{-1}(S))$.

		Classical Numerical Method	Probabilistic Numerical Method
Inputs	Assumed	e.g. smoothness	$P_x \in \mathcal{P}_\mathcal{X}$
	Information	$a \in \mathcal{A}$	$a \in \mathcal{A}$
Output		$b(a) \in \mathcal{Q}$	$B(P_x, a) \in \mathcal{P}_\mathcal{Q}$

A Probabilistic Numerical Method is Bayesian iff $B(P_x, a) = Q_\# P_{x|a}$.

Dichotomy of Probabilistic Numerical Methods

Method	**QoI** $Q(x)$	**Information** $A(x)$	**Non-Bayesian PNMs**	**Bayesian PNMs**
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]
	$\int f(t)x(\mathrm{d}t)$	$\{t_i\}_{i=1}^n$ 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]
		$\{\nabla x(t_i)\}_{i=1}^n$		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]
		$\{\mathbb{I}[t_{\min} < t_i] + \text{error}\}_{i=1}^n$	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] Finite Difference Methods [John and Wu, 2017]	Skilling [1992]
		∇x + rounding error	Hull and Swenson [1966], Mosbach and Turner [2009]	
	$x(t_{\mathrm{end}})$	$\{\nabla x(t_i)\}_{i=1}^n$	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]	

Second Job: Check Well-Defined, Existence and Uniqueness

Well-Defined?

Limitations of existing Bayesian probabilistic numerical methods:

- Restriction to Gaussian prior distributions $P_x \in \mathcal{P}_\mathcal{X}$
- Often focused just on linear information operator $x \mapsto A(x)$

Outside of this context even existence of Bayesian probabilistic numerical methods is non-trivial:

$$p(x|a) = \frac{p(a|x)p(x)}{p(a)}$$

No Lebesgue measure $\implies$ work instead with Radon-Nikodym derivatives:

$$\frac{\mathrm{d}P_{x|a}}{\mathrm{d}P_x} = \frac{p(a|x)}{p(a)}$$

But when "$p(a|x) = \delta(a - A(x))$", the posterior $P_{x|a}$ will not be absolutely continuous wrt the prior P_x, so no Radon-Nikodym theorem!

Well-Defined?

Limitations of existing Bayesian probabilistic numerical methods:

- Restriction to Gaussian prior distributions $P_x \in \mathcal{P}_\mathcal{X}$
- Often focused just on linear information operator $x \mapsto A(x)$

Outside of this context even existence of Bayesian probabilistic numerical methods is non-trivial:

$$p(x|a) = \frac{p(a|x)p(x)}{p(a)}$$

No Lebesgue measure $\implies$ work instead with Radon-Nikodym derivatives:

$$\frac{\mathrm{d}P_{x|a}}{\mathrm{d}P_x} = \frac{p(a|x)}{p(a)}$$

But when "$p(a|x) = \delta(a - A(x))$", the posterior $P_{x|a}$ will not be absolutely continuous wrt the prior P_x, so no Radon-Nikodym theorem!

Well-Defined?

Limitations of existing Bayesian probabilistic numerical methods:

- Restriction to Gaussian prior distributions $P_x \in \mathcal{P}_\mathcal{X}$
- Often focused just on linear information operator $x \mapsto A(x)$

Outside of this context even existence of Bayesian probabilistic numerical methods is non-trivial:

$$p(x|a) = \frac{p(a|x)p(x)}{p(a)}$$

No Lebesgue measure $\implies$ work instead with Radon-Nikodym derivatives:

$$\frac{\mathrm{d}P_{x|a}}{\mathrm{d}P_x} = \frac{p(a|x)}{p(a)}$$

But when "$p(a|x) = \delta(a - A(x))$", the posterior $P_{x|a}$ will not be absolutely continuous wrt the prior P_x, so no Radon-Nikodym theorem!

Well-Defined?

Limitations of existing Bayesian probabilistic numerical methods:

- Restriction to Gaussian prior distributions $P_x \in \mathcal{P}_\mathcal{X}$
- Often focused just on linear information operator $x \mapsto A(x)$

Outside of this context even existence of Bayesian probabilistic numerical methods is non-trivial:

$$p(x|a) = \frac{p(a|x)p(x)}{p(a)}$$

No Lebesgue measure $\implies$ work instead with Radon-Nikodym derivatives:

$$\frac{\mathrm{d}P_{x|a}}{\mathrm{d}P_x} = \frac{p(a|x)}{p(a)}$$

But when "$p(a|x) = \delta(a - A(x))$", the posterior $P_{x|a}$ will not be absolutely continuous wrt the prior P_x, so no Radon-Nikodym theorem!

Well-Defined?

Limitations of existing Bayesian probabilistic numerical methods:

- Restriction to Gaussian prior distributions $P_x \in \mathcal{P}_\mathcal{X}$
- Often focused just on linear information operator $x \mapsto A(x)$

Outside of this context even existence of Bayesian probabilistic numerical methods is non-trivial:

$$p(x|a) = \frac{p(a|x)p(x)}{p(a)}$$

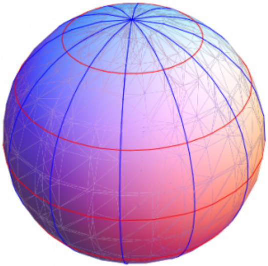
No Lebesgue measure $\implies$ work instead with Radon-Nikodym derivatives:

$$\frac{\mathrm{d}P_{x|a}}{\mathrm{d}P_x} = \frac{p(a|x)}{p(a)}$$

But when "$p(a|x) = \delta(a - A(x))$", the posterior $P_{x|a}$ will not be absolutely continuous wrt the prior P_x, so no Radon-Nikodym theorem!

Well-Defined?

Borel-Kolmogorov paradox[1]:



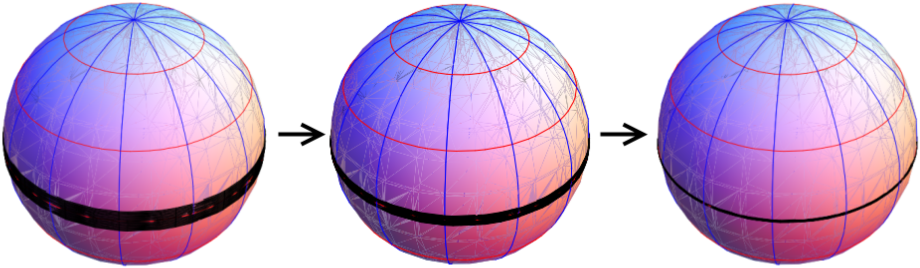
(latitude = red, longitude = blue)

To make progress it is required to introduce measure-theoretic detail.

[1] Figures from Greg Gandenberger's blog post

Well-Defined?

Borel-Kolmogorov paradox[1]:

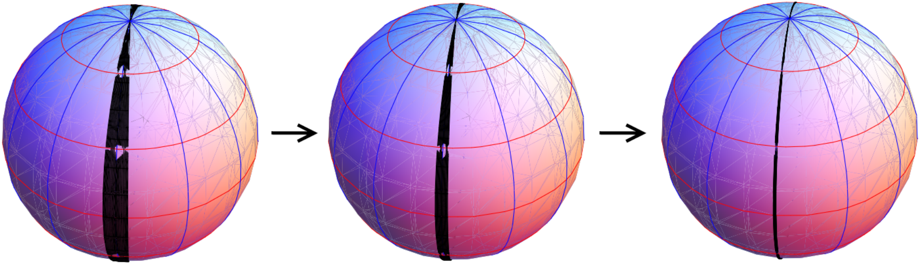


To make progress it is required to introduce measure-theoretic detail.

[1] Figures from Greg Gandenberger's blog post

Well-Defined?

Borel-Kolmogorov paradox[1]:



To make progress it is required to introduce measure-theoretic detail.

[1] Figures from Greg Gandenberger's blog post

Disintegration

High-level idea: Additional structure on $\mathcal{X}$, $\mathcal{A}$ and $A : \mathcal{X} \to \mathcal{A}$ is needed:

Let $(\mathcal{X}, \Sigma_{\mathcal{X}})$, $(\mathcal{A}, \Sigma_{\mathcal{A}})$ and $(\mathcal{Q}, \Sigma_{\mathcal{Q}})$ be measurable spaces and A, Q be measurable.

Due to Dellacherie and Meyer [1978, p.78]:

For $P_x \in \mathcal{P}_{\mathcal{X}}$, a collection $\{P_{x|a}\}_{a \in \mathcal{A}} \subset \mathcal{P}_{\mathcal{X}}$ is a disintegration of P_x with respect to the map $A : \mathcal{X} \to \mathcal{A}$ if:

1. (Concentration:) $P_{x|a}(\mathcal{X} \setminus \{x \in \mathcal{X} : A(x) = a\}) = 0$ for $A_{\#} P_x$-almost all $a \in \mathcal{A}$;

and for each measurable $f : \mathcal{X} \to [0, \infty)$ it holds that

2. (Measurability:) $a \mapsto P_{x|a}(f)$ is measurable;
3. (Conditioning:) $P_x(f) = \int P_{x|a}(f) A_{\#} P_x(\mathrm{d}a)$.

Disintegration

High-level idea: Additional structure on $\mathcal{X}$, $\mathcal{A}$ and $A : \mathcal{X} \to \mathcal{A}$ is needed:

Let $(\mathcal{X}, \Sigma_{\mathcal{X}})$, $(\mathcal{A}, \Sigma_{\mathcal{A}})$ and $(\mathcal{Q}, \Sigma_{\mathcal{Q}})$ be measurable spaces and A, Q be measurable.

Due to Dellacherie and Meyer [1978, p.78]:

For $P_x \in \mathcal{P}_{\mathcal{X}}$, a collection $\{P_{x|a}\}_{a \in \mathcal{A}} \subset \mathcal{P}_{\mathcal{X}}$ is a disintegration of P_x with respect to the map $A : \mathcal{X} \to \mathcal{A}$ if:

1. (Concentration:) $P_{x|a}(\mathcal{X} \setminus \{x \in \mathcal{X} : A(x) = a\}) = 0$ for $A_{\#}P_x$-almost all $a \in \mathcal{A}$;

and for each measurable $f : \mathcal{X} \to [0, \infty)$ it holds that

2. (Measurability:) $a \mapsto P_{x|a}(f)$ is measurable;
3. (Conditioning:) $P_x(f) = \int P_{x|a}(f) A_{\#}P_x(\mathrm{d}a)$.

Disintegration

High-level idea: Additional structure on $\mathcal{X}$, $\mathcal{A}$ and $A : \mathcal{X} \to \mathcal{A}$ is needed:

Let $(\mathcal{X}, \Sigma_{\mathcal{X}})$, $(\mathcal{A}, \Sigma_{\mathcal{A}})$ and $(\mathcal{Q}, \Sigma_{\mathcal{Q}})$ be measurable spaces and A, Q be measurable.

Due to Dellacherie and Meyer [1978, p.78]:

For $P_x \in \mathcal{P}_{\mathcal{X}}$, a collection $\{P_{x|a}\}_{a \in \mathcal{A}} \subset \mathcal{P}_{\mathcal{X}}$ is a disintegration of P_x with respect to the map $A : \mathcal{X} \to \mathcal{A}$ if:

1. (Concentration:) $P_{x|a}(\mathcal{X} \setminus \{x \in \mathcal{X} : A(x) = a\}) = 0$ for $A_{\#}P_x$-almost all $a \in \mathcal{A}$;

and for each measurable $f : \mathcal{X} \to [0, \infty)$ it holds that

2. (Measurability:) $a \mapsto P_{x|a}(f)$ is measurable;
3. (Conditioning:) $P_x(f) = \int P_{x|a}(f) A_{\#} P_x(\mathrm{d}a)$.

Existence and Uniqueness

Disintegration Theorem; statement from Thm. 1 of Chang and Pollard [1997]:

- Let $\mathcal{X}$ be a metric space, $\Sigma_{\mathcal{X}}$ be the Borel σ-algebra.
- Let $P_x \in \mathcal{P}_{\mathcal{X}}$ be Radon.
- Let $\Sigma_{\mathcal{A}}$ be a countably generated σ-algebra that contains singletons $\{a\}$ for $a \in \mathcal{A}$.

Then there exists an (essentially) unique disintegration $\{P_{x|a}\}_{a \in \mathcal{A}}$ of P_x with respect to A.

Thus Bayesian probabilistic numerical methods $B(P_x, a) = Q_{\#} P_{x|a}$ are well-defined under quite general conditions.

In particular, $Q_{\#} P_{x|a}$ exists and is unique for $A_{\#} P_x$ almost all $a \in \mathcal{A}$.

Disintegration Theorem; statement from Thm. 1 of Chang and Pollard [1997]:

- Let $\mathcal{X}$ be a metric space, $\Sigma_\mathcal{X}$ be the Borel σ-algebra.
- Let $P_x \in \mathcal{P}_\mathcal{X}$ be Radon.
- Let $\Sigma_\mathcal{A}$ be a countably generated σ-algebra that contains singletons $\{a\}$ for $a \in \mathcal{A}$.

Then there exists an (essentially) unique disintegration $\{P_{x|a}\}_{a\in\mathcal{A}}$ of P_x with respect to A.

Thus Bayesian probabilistic numerical methods $B(P_x, a) = Q_\# P_{x|a}$ are <u>well-defined</u> under quite general conditions.

In particular, $Q_\# P_{x|a}$ exists and is unique for $A_\# P_x$ almost all $a \in \mathcal{A}$.

Example: Solution of a Non-linear ODE

Consider Painlevé's first transcendental:

$$\begin{aligned} x''(t) &= x(t)^2 - t, \quad t \in \mathbb{R}_+ \\ x(0) &= 0 \\ t^{-1/2}x(t) &\to 1 \text{ as } t \to \infty \end{aligned}$$

The information operator is

$$A(x) = \begin{bmatrix} x''(t_1) - x(t_1)^2 \\ \vdots \\ x''(t_n) - x(t_n)^2 \\ x(0) \\ \lim_{t\to\infty} t^{-1/2}x(t) \end{bmatrix} = \begin{bmatrix} t_1 \\ \vdots \\ t_n \\ 0 \\ 1 \end{bmatrix}.$$

Construct an infinite-dimensional prior $P_x \in \mathcal{P}_\mathcal{X}$ as

$$x(t) = \sum_{i=0}^{\infty} u_i \gamma_i \phi_i(t)$$

with u_i i.i.d. std. Cauchy coefficients, weights $\gamma_i = (i+1)^{-2}$ and $\phi_i(t)$ (normalized) Chebyshev polynomials of the first kind. [See Sullivan, 2016, for mathematical details.]

Example: Solution of a Non-linear ODE

Consider Painlevé's first transcendental:

$$\begin{aligned} x''(t) &= x(t)^2 - t, \quad t \in \mathbb{R}_+ \\ x(0) &= 0 \\ t^{-1/2}x(t) &\to 1 \text{ as } t \to \infty \end{aligned}$$

The information operator is

$$A(x) = \begin{bmatrix} x''(t_1) - x(t_1)^2 \\ \vdots \\ x''(t_n) - x(t_n)^2 \\ x(0) \\ \lim_{t\to\infty} t^{-1/2}x(t) \end{bmatrix} = \begin{bmatrix} t_1 \\ \vdots \\ t_n \\ 0 \\ 1 \end{bmatrix}.$$

Construct an infinite-dimensional prior $P_x \in \mathcal{P}_\mathcal{X}$ as

$$x(t) = \sum_{i=0}^{\infty} u_i \gamma_i \phi_i(t)$$

with u_i i.i.d. std. Cauchy coefficients, weights $\gamma_i = (i+1)^{-2}$ and $\phi_i(t)$ (normalized) Chebyshev polynomials of the first kind. [See Sullivan, 2016, for mathematical details.]

Consider Painlevé's first transcendental:

$$\begin{aligned} x''(t) &= x(t)^2 - t, \quad t \in \mathbb{R}_+ \\ x(0) &= 0 \\ t^{-1/2}x(t) &\to 1 \text{ as } t \to \infty \end{aligned}$$

The information operator is

$$A(x) = \begin{bmatrix} x''(t_1) - x(t_1)^2 \\ \vdots \\ x''(t_n) - x(t_n)^2 \\ x(0) \\ \lim_{t\to\infty} t^{-1/2}x(t) \end{bmatrix} = \begin{bmatrix} t_1 \\ \vdots \\ t_n \\ 0 \\ 1 \end{bmatrix}.$$

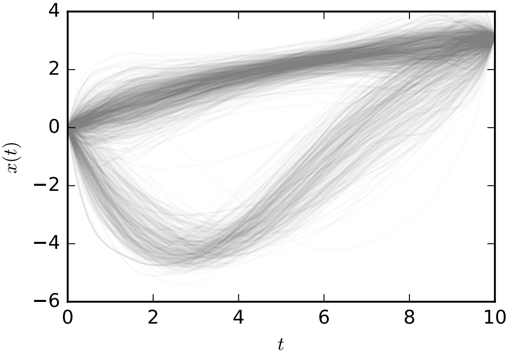
Construct an infinite-dimensional prior $P_x \in \mathcal{P}_\mathcal{X}$ as

$$x(t) = \sum_{i=0}^{\infty} u_i \gamma_i \phi_i(t)$$

with u_i i.i.d. std. Cauchy coefficients, weights $\gamma_i = (i+1)^{-2}$ and $\phi_i(t)$ (normalized) Chebyshev polynomials of the first kind. [See Sullivan, 2016, for mathematical details.]

Example: Solution of a Non-linear ODE

For this illustration the information, $n = 10$, is fixed.



[samples via *Numerical Disintegration* algorithm; see Part II]

Third Job: Characterise Optimal Information

Optimal Information

Recall the contribution of Kadane and Wasilkowski [1985]:

Consider a classical numerical method (A, b) with information operator $A : \mathcal{X} \to \mathcal{A}$, such that $A \in \Lambda$ for some set Λ, and estimator $b : \mathcal{A} \to \mathcal{Q}$. Let $L : \mathcal{Q} \times \mathcal{Q} \to \mathbb{R}$ be a loss function that is pre-specified. Then consider the minimal average case error

$$\inf_{A \in \Lambda, b} \int L(b(A(x)), Q(x)) \mathrm{d}P_x.$$

The minimiser $b(\cdot)$ is a non-randomised Bayes rule and the minimiser A is "optimal information" over Λ, or optimal experimental design for this numerical task.

Generalisation of optimal information to probabilistic numerical methods?

Optimal Information

Recall the contribution of Kadane and Wasilkowski [1985]:

Consider a classical numerical method (A, b) with information operator $A : \mathcal{X} \to \mathcal{A}$, such that $A \in \Lambda$ for some set Λ, and estimator $b : \mathcal{A} \to \mathcal{Q}$. Let $L : \mathcal{Q} \times \mathcal{Q} \to \mathbb{R}$ be a loss function that is pre-specified. Then consider the minimal average case error

$$\inf_{A \in \Lambda, b} \int L(b(A(x)), Q(x)) \mathrm{d}P_x.$$

The minimiser $b(\cdot)$ is a non-randomised Bayes rule and the minimiser A is "optimal information" over Λ, or optimal experimental design for this numerical task.

Generalisation of optimal information to probabilistic numerical methods?

Optimal Information

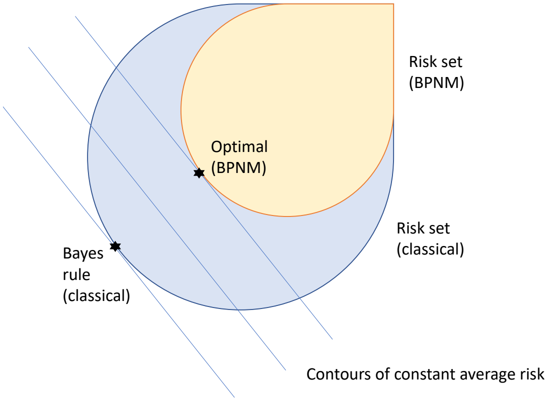
For Bayesian probabilistic numerical methods $B(P_x, a) = Q_{\#}P_{x|a}$, optimal information is defined as

$$\arg\inf_{A \in \Lambda} \int \int L(Q_{\#}P_{x|A(x)}(\omega), Q(x)) \mathrm{d}P_x \, \mathrm{d}\omega.$$

Important point: The Bayesian probabilistic numerical method output $Q_{\#}P_{x|a}$ will not in general be supported on the set of Bayes acts. This presents a non-trivial constraint on the risk set...

Optimal Information

Average Case Analysis $\overset{1985}{\leftrightarrow}$ Bayesian Decision Theory $\overset{?}{\leftrightarrow}$ Bayesian Probabilistic Numerical Methods



Optimal Information

We have established the following (new) result:

Let $(\mathcal{Q}, \langle \cdot, \cdot \rangle_{\mathcal{Q}})$ be an inner-product space with associated norm $\| \cdot \|_{\mathcal{Q}}$ and consider the canonical loss $L(q, q') = \|q - q'\|_{\mathcal{Q}}^2$. Then optimal information for Bayesian probabilistic numerical methods coincides with average-case optimal information.

The assumption is non-trivial:

Consider the following counter-example:

- $\mathcal{X} = \{b, c, d, e\}$,
- $Q(x) = 1[x = b]$,
- P_x uniform,
- $A(x) = 1[x \in S]$, where we are allowed either $S = \{b, c\}$ or $\{b, c, d\}$,
- $L(q, q') = 1[q \neq q']$.

Then average-case optimal information can be either $S = \{b, c\}$ or $\{b, c, d\}$. On the other hand, optimal information in the Bayesian probabilistic numerical context is just $S = \{b, c\}$.

Optimal Information

We have established the following (new) result:

Let $(\mathcal{Q}, \langle \cdot, \cdot \rangle_{\mathcal{Q}})$ be an inner-product space with associated norm $\| \cdot \|_{\mathcal{Q}}$ and consider the canonical loss $L(q, q') = \|q - q'\|_{\mathcal{Q}}^2$. Then optimal information for Bayesian probabilistic numerical methods coincides with average-case optimal information.

The assumption is non-trivial:

Consider the following counter-example:

- $\mathcal{X} = \{b, c, d, e\}$,
- $Q(x) = 1[x = b]$,
- P_x uniform,
- $A(x) = 1[x \in S]$, where we are allowed either $S = \{b, c\}$ or $\{b, c, d\}$,
- $L(q, q') = 1[q \neq q']$.

Then average-case optimal information can be either $S = \{b, c\}$ or $\{b, c, d\}$. On the other hand, optimal information in the Bayesian probabilistic numerical context is just $S = \{b, c\}$.

Optimal Information

We have established the following (new) result:

Let $(\mathcal{Q}, \langle \cdot, \cdot \rangle_{\mathcal{Q}})$ be an inner-product space with associated norm $\| \cdot \|_{\mathcal{Q}}$ and consider the canonical loss $L(q, q') = \|q - q'\|_{\mathcal{Q}}^2$. Then optimal information for Bayesian probabilistic numerical methods coincides with average-case optimal information.

The assumption is non-trivial:

Consider the following counter-example:

- $\mathcal{X} = \{b, c, d, e\}$,
- $Q(x) = 1[x = b]$,
- P_x uniform,
- $A(x) = 1[x \in S]$, where we are allowed either $S = \{b, c\}$ or $\{b, c, d\}$,
- $L(q, q') = 1[q \neq q']$.

Then average-case optimal information can be either $S = \{b, c\}$ or $\{b, c, d\}$. On the other hand, optimal information in the Bayesian probabilistic numerical context is just $S = \{b, c\}$.

Conclusion

Conclusion

In Part I it has been argued that:

- The onus is on us to establish principled statistical foundations that are general.
- The Bayesian approach to inverse problems, popularised in Stuart [2010], provides such a framework.
- Bayesian probabilistic numerical methods (BPNM) are well-defined under weak conditions ($\mathcal{X}$ metric space, P_x radon, $\Sigma_{\mathcal{A}}$ countably generated).
- Optimal information for BPNM is not always equivalent to optimal information in Average Case Analysis.

Full details (Parts I and II) can be found in the preprint:

Cockayne *et al.* (2017) "Bayesian Probabilistic Numerical Methods" (on arXiv).

Thank you for your attention!

Conclusion

In Part I it has been argued that:

- The onus is on us to establish principled statistical foundations that are general.
- The Bayesian approach to inverse problems, popularised in Stuart [2010], provides such a framework.
- Bayesian probabilistic numerical methods (BPNM) are well-defined under weak conditions ($\mathcal{X}$ metric space, P_x radon, $\Sigma_{\mathcal{A}}$ countably generated).
- Optimal information for BPNM is not always equivalent to optimal information in Average Case Analysis.

Full details (Parts I and II) can be found in the preprint:

Cockayne *et al.* (2017) "Bayesian Probabilistic Numerical Methods" (on arXiv).

Thank you for your attention!

Conclusion

In Part I it has been argued that:

- The onus is on us to establish principled statistical foundations that are general.
- The Bayesian approach to inverse problems, popularised in Stuart [2010], provides such a framework.
- Bayesian probabilistic numerical methods (BPNM) are well-defined under weak conditions ($\mathcal{X}$ metric space, P_x radon, $\Sigma_{\mathcal{A}}$ countably generated).
- Optimal information for BPNM is not always equivalent to optimal information in Average Case Analysis.

Full details (Parts I and II) can be found in the preprint:

Cockayne *et al.* (2017) "Bayesian Probabilistic Numerical Methods" (on arXiv).

Thank you for your attention!

Conclusion

In Part I it has been argued that:

- The onus is on us to establish principled statistical foundations that are general.
- The Bayesian approach to inverse problems, popularised in Stuart [2010], provides such a framework.
- Bayesian probabilistic numerical methods (BPNM) are well-defined under weak conditions ($\mathcal{X}$ metric space, P_x radon, $\Sigma_{\mathcal{A}}$ countably generated).
- Optimal information for BPNM is not always equivalent to optimal information in Average Case Analysis.

Full details (Parts I and II) can be found in the preprint:

Cockayne *et al.* (2017) "Bayesian Probabilistic Numerical Methods" (on arXiv).

Thank you for your attention!

Conclusion

In Part I it has been argued that:

- The onus is on us to establish principled statistical foundations that are general.
- The Bayesian approach to inverse problems, popularised in Stuart [2010], provides such a framework.
- Bayesian probabilistic numerical methods (BPNM) are well-defined under weak conditions ($\mathcal{X}$ metric space, P_x radon, $\Sigma_{\mathcal{A}}$ countably generated).
- Optimal information for BPNM is not always equivalent to optimal information in Average Case Analysis.

Full details (Parts I and II) can be found in the preprint:

Cockayne *et al.* (2017) "Bayesian Probabilistic Numerical Methods" (on arXiv).

Thank you for your attention!

Conclusion

In Part I it has been argued that:

- The onus is on us to establish principled statistical foundations that are general.
- The Bayesian approach to inverse problems, popularised in Stuart [2010], provides such a framework.
- Bayesian probabilistic numerical methods (BPNM) are well-defined under weak conditions ($\mathcal{X}$ metric space, P_x radon, $\Sigma_{\mathcal{A}}$ countably generated).
- Optimal information for BPNM is not always equivalent to optimal information in Average Case Analysis.

Full details (Parts I and II) can be found in the preprint:

Cockayne *et al.* (2017) "Bayesian Probabilistic Numerical Methods" (on arXiv).

Thank you for your attention!

Conclusion

In Part I it has been argued that:

- The onus is on us to establish principled statistical foundations that are general.
- The Bayesian approach to inverse problems, popularised in Stuart [2010], provides such a framework.
- Bayesian probabilistic numerical methods (BPNM) are well-defined under weak conditions ($\mathcal{X}$ metric space, P_x radon, $\Sigma_{\mathcal{A}}$ countably generated).
- Optimal information for BPNM is not always equivalent to optimal information in Average Case Analysis.

Full details (Parts I and II) can be found in the preprint:

Cockayne *et al.* (2017) "Bayesian Probabilistic Numerical Methods" (on arXiv).

Thank you for your attention!

References I

S. Bartels and P. Hennig. Probabilistic approximate least-squares. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics AISTATS*, 51:676–684, 2016.

J. Chang and D. Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317, 1997.

O. A. Chkrebtii, D. A. Campbell, B. Calderhead, and M. A. Girolami. Bayesian solution uncertainty quantification for differential equations. *Bayesian Analysis*, 2016.

J. Cockayne, C. Oates, T. J. Sullivan, and M. Girolami. Probabilistic meshless methods for partial differential equations and Bayesian inverse problems, 2016. arXiv:1605.07811v1.

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. *Statistics and Computing*, 2016. doi: 10.1007/s11222-016-9671-0.

C. Dellacherie and P. Meyer. *Probabilities and Potential*. North-Holland, Amsterdam, 1978.

P. Diaconis. Bayesian numerical analysis. *Statistical decision theory and related topics IV*, 1:163–175, 1988.

T. Gunter, M. A. Osborne, R. Garnett, P. Hennig, and S. J. Roberts. Sampling for inference in probabilistic models with fast bayesian quadrature. In *Advances in neural information processing systems*, pages 2789–2797, 2014.

P. Hennig. Probabilistic interpretation of linear solvers. *SIAM Journal on Optimization*, 25(1):234–260, jan 2015.

P. Hennig and M. Kiefel. Quasi-newton method: A new direction. *Journal of Machine Learning Research*, 14(Mar):843–865, 2013.

M. Horstein. Sequential transmission using noiseless feedback. *IEEE Transactions on Information Theory*, 9(3):136–143, 1963.

T. E. Hull and J. R. Swenson. Tests of probabilistic models for propagation of roundoff errors. *Communications of the ACM*, 9(2):108–113, 1966.

M. John and Y. Wu. Confidence intervals for finite difference solutions. *arXiv:1701.05609*, 2017.

J. B. Kadane and G. W. Wasilkowski. *Bayesian Statistics*, chapter Average Case ϵ-Complexity in Computer Science: A Bayesian View, pages 361–374. Elsevier, North-Holland, 1985.

H. Kersting and P. Hennig. Active uncertainty calibration in Bayesian ODE solvers, 2016. arXiv:1605.03364.

A. Kong, P. McCullagh, X.-L. Meng, D. Nicolae, and Z. Tan. A theory of statistical models for monte carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):585–604, 2003.

A. Kong, P. McCullagh, X.-L. Meng, and D. L. Nicolae. Further explorations of likelihood theory for monte carlo integration. In *Advances In Statistical Modeling And Inference: Essays in Honor of Kjell A Doksum*, pages 563–592. World Scientific, 2007.

J. T. Krebs. Consistency and asymptotic normality of stochastic euler schemes for ordinary differential equations. *arXiv preprint arXiv:1609.06880*, 2016.

M. Mahsereci and P. Hennig. Probabilistic line searches for stochastic optimization. In *Advances In Neural Information Processing Systems*, pages 181–189, 2015.

References II

J. Mockus. *Bayesian Approach to Global Optimization: Theory and Applications*, volume 37. Springer Science & Business Media, 1989.

S. Mosbach and A. G. Turner. A quantitative probabilistic investigation into the accumulation of rounding errors in numerical ODE solution. *Computers & Mathematics with Applications*, 2009.

C. Oates, F.-X. Briol, M. Girolami, et al. Probabilistic integration and intractable distributions. *arXiv preprint arXiv:1606.06841*, 2016.

A. O'Hagan. Bayes–Hermite quadrature. *Journal of Statistical Planning and Inference*, 29(3):245–260, nov 1991. doi: 10.1016/0378-3758(91)90002-v. URL `http://dx.doi.org/10.1016/0378-3758(91)90002-V`.

M. Osborne, R. Garnett, Z. Ghahramani, D. K. Duvenaud, S. J. Roberts, and C. E. Rasmussen. Active learning of model evidence using bayesian quadrature. In *Advances in neural information processing systems*, pages 46–54, 2012a.

M. A. Osborne, R. Garnett, S. J. Roberts, C. Hart, S. Aigrain, N. Gibson, and S. Aigrain. Bayesian quadrature for ratios. In *AISTATS*, pages 832–840, 2012b.

H. Owhadi. Bayesian numerical homogenization. *Multiscale Modeling & Simulation*, 13(3):812–828, 2015a.

H. Owhadi. Multigrid with rough coefficients and multiresolution operator decomposition from hierarchical information games. *arXiv preprint arXiv:1503.03467*, 2015b.

M. Raissi, P. Perdikaris, and G. E. Karniadakis. Inferring solutions of differential equations using noisy multi-fidelity data. *arXiv preprint arXiv:1607.04805*, 2016.

M. Schober, D. K. Duvenaud, and P. Hennig. Probabilistic ODE solvers with Runge–Kutta means. *Advances in Neural Information Processing Systems 27*, pages 739–747, 2014.

M. Schober, S. Särkkä, and P. Hennig. A probabilistic model for the numerical solution of initial value problems, 2016. arXiv:1610.05261v1.

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

A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, May 2010.

T. J. Sullivan. Well-posed Bayesian inverse problems and heavy-tailed stable quasi-Banach space priors, 2016. arXiv:1605.05898.

Z. Tan. On a likelihood approach for monte carlo integration. *Journal of the American Statistical Association*, 99(468):1027–1036, 2004.

O. Teymur, K. Zygalakis, and B. Calderhead. Probabilistic linear multistep methods. In *Advances in Neural Information Processing Systems*, pages 4314–4321, 2016.

R. Waeber, P. I. Frazier, and S. G. Henderson. Bisection search with noisy responses. *SIAM Journal on Control and Optimization*, 51(3):2261–2279, 2013.