# Bayesian Probabilistic Numerical Methods

Jon Cockayne* Chris Oates† Tim Sullivan‡ Mark Girolami§

February 14, 2017

The emergent field of probabilistic numerics has thus far lacked rigorous statistical principals. This paper establishes *Bayesian* probabilistic numerical methods as those which can be cast as solutions to certain Bayesian inverse problems, albeit problems that are non-standard. This allows us to establish general conditions under which Bayesian probabilistic numerical methods are well-defined, encompassing both non-linear and non-Gaussian models. For general computation, a numerical approximation scheme is developed and its asymptotic convergence is established. The theoretical development is then extended to pipelines of computation, wherein probabilistic numerical methods are composed to solve more challenging numerical tasks. The contribution highlights an important research frontier at the interface of numerical analysis and uncertainty quantification, with some illustrative applications presented.

## 1. Introduction

Numerical computation underpins almost all of modern scientific and industrial research and development. The impact of a finite computational budget is that problems which are infinite-dimensional, such as the solution of differential equations, must be both discretised and represented with finite precision arithmetic. The result is an approximation to the exact solution of the original problem. The declining rate of processor improvement, as physical limits are reached, is in contrast to the surge in complexity of modern inference problems and the error incurred by discretisation is attracting increased interest [e.g. Capistrán et al., 2013]. The situation is epitomised in modern climate models, where single-precision arithmetic has been suggested to permit finer temporal resolution. However, for these models a detailed time discretisation can *increase* total error, due to the increased total number of single precision computations, and in practice some form of *ad-hoc* trade-off is sought [Harvey and Verseghy, 2015]. Statistical considerations can permit more principled error control strategies for such models; we elaborate on this important research frontier next.

Numerical methods are designed to mitigate discretisation errors of all forms [Press et al., 2007]. Nonetheless, the introduction of error is unavoidable and it is the role of the numerical analyst to provide control of this error [Oberkampf and Roy, 2013]. The central theoretical

*University of Warwick, j.cockayne@warwick.ac.uk
†University of Technology Sydney, chris.oates@ncl.ac.uk
‡Free University of Berlin and Zuse Institute Berlin, sullivan@zib.de
§Imperial College London and Alan Turing Institute, m.girolami@imperial.ac.uk

results of numerical analysis have in general not been obtained through statistical considerations. However, the connection of discretisation error to statistics was noted as far back as Hull and Swenson [1966], who concluded that for a simple class of differential equations:

> ... the propagation of roundoff error, whether it is due to chopping or ordinary rounding, can be represented very well by the usual rather simple probabilistic models for these processes.

In their recent work on elliptic partial differential equations, Babuška and Söderlind [2016] provided examples of non-trivial numerical error propagation that do not admit such simple descriptions. To address more challenging problems, the field of probabilistic numerics has emerged with the aim to develop methods for quantification of the uncertainty introduced through discretisation error on the output of numerical methods.

The foundations of probabilistic numerics were laid in the 1970s and 1980s, where an important shift in emphasis occurred from descriptive statistical models of Hull and Swenson [1966] to the use of formal generative models that generalise across classes of numerical tasks [Larkin, 1972]. The role for statistics in this new outlook was captured in Kadane and Wasilkowski [1985]:

> Statistics can be thought of as a set of tools used in making decisions and inferences in the face of uncertainty. Algorithms typically operate in such an environment. Perhaps then, statisticians might join the teams of scholars addressing algorithmic issues.

The decade culminated in development of popular Bayesian optimisation methods [Mockus, 1989, Torn and Zilinskas, 1989], as well as the relation of smoothing splines to Bayesian estimation [Kimeldorf and Wahba, 1970b, Diaconis and Freedman, 1983].

The modern notion of a probabilistic numerical method (henceforth PNM) was described in Hennig et al. [2015]; these are algorithms whose output is a distribution over an unknown, deterministic quantity of interest, such as the integral of a function over its domain. Recent developments include PNMs for numerical linear algebra [Hennig, 2015, Bartels and Hennig, 2016], numerical solution of ordinary differential equations [ODEs; Schober et al., 2014, Kersting and Hennig, 2016, Schober et al., 2016, Conrad et al., 2016, Chkrebtii et al., 2016], numerical solution of partial differential equations [PDEs; Owhadi, 2015a, Cockayne et al., 2016, Conrad et al., 2016] and numerical integration [Ghahramani and Rasmussen, 2002, Briol et al., 2016].

**Open Problems** Despite numerous recent successes and achievements in the field, there remains a lack of statistical foundations for PNMs. This is due to the infinite-dimensional nature of the problems being solved. For instance, at present it is not clear under what conditions a PNM is well-defined, except for the standard conjugate Gaussian framework that is popular in machine learning applications [Rasmussen and Williams, 2006]. This limits the extent to which domain-specific knowledge, such as boundedness of an integrand or monotonicity of a solution to a differential equation, can be encoded in PNMs. In contrast, classical numerical methods often exploit such contextual information to achieve substantial reduction in discretisation error. For instance, finite element methods for solution of PDEs proceed based on a mesh that is designed to be more refined in areas of the domain where greater variation of the solution is anticipated [Strang and Fix, 1973].

Furthermore, although PNMs have been proposed for many standard numerical tasks (see Section 2.5.1 for a literature review), the lack of common theoretical foundations makes comparison of these methods difficult. To take PDEs as an example, Cockayne et al. [2016] placed

a prior measure on the unknown solution of the PDE, whereas Conrad et al. [2016] placed a prior on the unknown discretisation error of a numerical method. The uncertainty in each case is fundamentally different, and at present there is no framework in which to articulate the relationship between these PNMs. In particular there is no clear definition of a *Bayesian* PNM.

An even more profound consequence of the lack of common foundation occurs when we seek to compose PNMs. For example, in computational biology, models of the heart involve coupled systems of ODEs and PDEs which must each be solved numerically to produce an overall quantity of interest [Niederer et al., 2011]. The composition of successive discretisations leads to non-trivial error propagation and accumulation. However, in seeking to capture this error with composed PNMs, we observe that PNMs cannot be properly composed unless they share common statistical foundations that ensure coherence of the overall statistical output.

**Contributions** The main contribution of this paper is to establish rigorous foundations for PNMs. Foremost is a study of when PNMs are well-defined outside of the Gaussian context. This is achieved through an interpretation of PNMs as algorithms for Bayesian inversion [Stuart, 2010]. For exploration of non-linear, non-Gaussian models, a numerical approximation scheme is developed and shown to asymptotically approach the posterior distribution of interest. Our aim here is not to develop new or more efficient PNMs, but to understand when such development can be well-defined.

A second contribution is to clarify what it means for a PNM to be Bayesian. This illuminates subtle distinctions among existing methods and allows examination of the sense in which non-Bayesian methods are approximations to Bayesian PNMs.

The third contribution is to discuss pipelines of composed PNMs. This is a critical area of development for probabilistic numerics; in isolation, the error of a numerical method can more easily be studied and understood, but when composed into a pipeline the resulting error structure may be non-trivial. The real power of probabilistic numerics lies in its application to pipelines of numerical methods, where such understanding presents significant challenges. This paper introduces conditions under which a composition of PNMs can be considered to provide meaningful output.

**Structure of the Paper** In Section 2 we rigorously define what it means for a PNM to be Bayesian and establish results related to when such methods are well-defined. Section 3 establishes connections to other related fields, in particular with relation to evaluating the performance of PNMs. In Section 4 we develop numerical approximations to the output of Bayesian PNMs. Section 5 discusses the composition of multiple PNMs. Last, Section 6 presents applications of the techniques discussed in this paper.

All proofs of theorems presented herein can be found in either the Appendix or the Electronic Supplement.

## 2. Probabilistic Numerical Methods

The aim of this section is to provide rigorous statistical foundations for PNMs.

### 2.1. Notation

For a measurable space $(\mathcal{X}, \Sigma_{\mathcal{X}})$, the shorthand $\mathcal{P}_{\mathcal{X}}$ will be used to denote the set of all distributions on $(\mathcal{X}, \Sigma_{\mathcal{X}})$. For $\mu, \mu' \in \mathcal{P}_{\mathcal{X}}$ we write $\mu \ll \mu'$ when $\mu$ is absolutely continuous with respect to $\mu$. The notation $\delta(x)$ will be used to denote a Dirac measure on $x \in \mathcal{X}$, so that $\delta(x) \in \mathcal{P}_{\mathcal{X}}$. Let

$1[S]$ denote the indicator function of an event $S \in \Sigma_{\mathcal{X}}$. For a measurable function $f : \mathcal{X} \to \mathbb{R}$ and a distribution $\mu \in \mathcal{P}_{\mathcal{X}}$, we will on occasion use the notation $\mu(f) = \int f(x)\mu(\mathrm{d}x)$ and $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$. The point-wise product of two functions $f$ and $g$ is denoted $f \cdot g$. For a function or operator $T$, $T_\#$ denotes the associated push-forward operator that acts on measures on the domain of $T$. Let $\perp\!\!\!\perp$ denote conditional independence. The subset $\ell^p \subset \mathbb{R}^\infty$ is defined to consist of sequences $(u_i)$ for which $\sum_{i=1}^\infty |u_i|^p$ is convergent.

## 2.2. Definition of a PNM

To first build intuition, consider numerical approximation of the Lebesgue integral

$$\int x(t)\nu(\mathrm{d}t)$$

for some integrable function $x\colon D \to \mathbb{R}$, with respect to a measure $\nu$ on $D$. Here we may directly interrogate the integrand $x(t)$ at any $t \in D$, but unless $D$ is finite we cannot evaluate $x$ at all $t \in D$ with a finite computational budget. Nonetheless, there are many algorithms for approximation of this integral based on information $\{x(t_i)\}_{i=1}^n$ at some collection of locations $\{t_i\}_{i=1}^n$.

To see the abstract structure of this problem, assume the state variable $x$ exists in a measurable space $(\mathcal{X}, \Sigma_{\mathcal{X}})$. Information about $x$ is provided through the *information operator* $A\colon \mathcal{X} \to \mathcal{A}$ whose range is a measurable space $(\mathcal{A}, \Sigma_{\mathcal{A}})$. Thus, for the Lebesgue integration problem, the information operator is

$$A(x) = \begin{bmatrix} x(t_1) \\ \vdots \\ x(t_n) \end{bmatrix} = a \in \mathcal{A}. \tag{2.1}$$

The space $\mathcal{X}$, in this case a space of functions, can be high- or infinite-dimensional, but the space $\mathcal{A}$ of information is assumed to be finite-dimensional in accordance with our finite computational budget. In this paper we make explicit a quantity of interest (QoI) $Q(x)$, defined by a map $Q\colon \mathcal{X} \to \mathcal{Q}$ into a measurable space $(\mathcal{Q}, \Sigma_{\mathcal{Q}})$. This captures that $x$ itself may not be the object of interest for the numerical problem; for the Lebesgue integration illustration, the QoI is not $x$ itself but $Q(x) = \int x(t)\nu(\mathrm{d}t)$. Both $A$ and $Q$ will be assumed to be measurable functions.

The standard approach to such computational problems is to construct an algorithm which, when applied, produces some approximation $\hat{q}(a)$ of $Q(x)$ based on the information $a$, whose theoretical convergence order can be studied. A successful algorithm will often tailor the information operator $A$ to the QoI $Q$. For example, classical Gaussian cubature specifies *sigma points* $\{t_i^*\}_{i=1}^n$ at which the integrand must be evaluated, based on exact integration of certain polynomial test functions.

The probabilistic numerical approach, instead, begins with the introduction of a random variable $X$ on $(\mathcal{X}, \Sigma_{\mathcal{X}})$. The true state $X = x$ is fixed but unknown; the randomness is used an abstract device to represent epistemic uncertainty about $x$ prior to numerical computation [Hennig et al., 2015]. This is now formalised:

**Definition 2.1** (Belief Distribution)**.** An element $\mu \in \mathcal{P}_{\mathcal{X}}$ is a *belief distribution*[1] for $x$ if it carries the formal semantics of belief about the true, unknown state variable $x$.

[1]Two remarks are in order: First, we have avoided the use of "prior" as our framework encompasses both Bayesian and non-Bayesian PNMs (to be defined). Second, the use of "belief" differs to the set-valued *belief functions* in Dempster–Shafer theory, which do not require that $\mu(E) + \mu(E^{\mathrm{c}}) = 1$ [Shafer, 1976].

Thus we may consider $\mu$ to be the law of $X$. The construction of an appropriate belief distribution $\mu$ for a specific numerical task is not the focus of this research and has been considered in detail in previous work; see the Electronic Supplement for an overview of this material. Rather we consider the problem of how one *updates* the belief distribution $\mu$ in response to the information $A(x) = a$ obtained about the unknown $x$. Generic approaches to update belief distributions, which generalise Bayesian inference beyond the unique update demanded in Bayes theorem, were formalised in Bissiri et al. [2016], de Carvalho et al. [2017].

**Definition 2.2** (Probabilistic Numerical Method)**.** Let $(\mathcal{X}, \Sigma_{\mathcal{X}})$, $(\mathcal{A}, \Sigma_{\mathcal{A}})$ and $(\mathcal{Q}, \Sigma_{\mathcal{Q}})$ be measurable spaces and let $A\colon \mathcal{X} \to \mathcal{A}$, $Q\colon \mathcal{X} \to \mathcal{Q}$ and $B\colon \mathcal{P}_{\mathcal{X}} \times \mathcal{A} \to \mathcal{P}_{\mathcal{Q}}$ be measurable functions. The pair $M = (A, B)$ is called a *probabilistic numerical method* for estimation of a quantity of interest $Q$. The map $A$ is called an *information operator*, and the map $B$ is called a *belief update operator*.

The output of a PNM is a distribution $B(\mu, a) \in \mathcal{P}_{\mathcal{Q}}$. This holds the formal status of belief distribution for the value of $Q(x)$, based on both the initial belief $\mu$ about the value of $x$ and the information $a$ that are input to the PNM.

An objection sometimes raised to this construction is that $x$ itself is not random. We emphasise that this work does not propose that $x$ should be considered as such; the random variable $X$ is a formal statistical device used to represent epistemic uncertainty [Kadane, 2011, Lindley, 2014]. Thus, there is no distinction from mainstream statistics, in which $x$ represents a fixed but unknown parameter and $X$ encodes epistemic uncertainty about this parameter.

Before presenting specific instances of this general framework, we comment on the potential analogy between $A$ and the likelihood function, and between $B$ and Bayes' theorem. Whilst morally correct, the mathematical developments in this paper are not well-suited to these concepts; in Section 2.4 we show that Bayes formula is not well-defined due to a vanishing normalising constant.

To strengthen intuition we now give specific examples of established PNMs:

**Example 2.3** (Probabilistic Integration)**.** Consider the numerical integration problem earlier discussed. Diaconis [1988] and O'Hagan [1991] proposed a PNM for approximation of this integral, described next. Take $D \subseteq \mathbb{R}^d$, $\mathcal{X}$ a separable Banach space of real-valued functions on $D$, and $\Sigma_{\mathcal{X}}$ the Borel $\sigma$-algebra for $\mathcal{X}$. The space $(\mathcal{X}, \Sigma_{\mathcal{X}})$ is endowed with a Gaussian belief distribution $\mu \in \mathcal{P}_{\mathcal{X}}$. Given information $A(x) = a$, define $\mu^a$ to be the restriction of $\mu$ to those functions which interpolate $x$ at the points $\{t_i\}_{i=1}^n$; that $\mu^a$ is again Gaussian follows from linearity of the information operator [see Bogachev, 1998, for details]. The QoI $Q$ remains $Q(x) = \int x(t)\nu(\mathrm{d}t)$.

This problem was first considered by Larkin [1972]. The belief update operator proposed therein, and later considered in Diaconis [1988] and O'Hagan [1991], was $B(\mu, a) = Q_{\#}\mu^a$. Since Gaussian distributions are closed under linear projections, the PNM output $B(\mu, a)$ is a univariate Gaussian whose mean and variance can be expressed in closed-form for certain choices of Gaussian covariance function and reference measure $\nu$ on $D$. Specifically, if $\mu$ has mean function $m\colon \mathcal{X} \to \mathbb{R}$ and covariance function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, then

$$B(\mu, a) = \mathrm{N}(z^\top \mathrm{K}^{-1}(a - \bar{m}), z_0 - z^\top \mathrm{K}^{-1} z) \tag{2.2}$$

where $\bar{m}, z \in \mathbb{R}^n$ are defined as $\bar{m}_i = m(t_i)$, $z_i = \int k(t, t_i)\nu(\mathrm{d}t)$, $\mathrm{K} \in \mathbb{R}^{n \times n}$ is defined as $\mathrm{K}_{i,j} = k(t_i, t_j)$ and $z_0 = \iint k(t, t')(\nu \times \nu)(\mathrm{d}(t \times t')) \in \mathbb{R}$. Briol et al. [2016] extensively studied this method and provided a comprehensive list of $(\nu, k)$ combinations for which $z$ and $z_0$ possess a closed-form.

An interesting fact is that the mean of $B(\mu, a)$ coincides with classical cubature rules for different choices of $\mu$ and $A$ [Diaconis, 1988, Särkkä et al., 2015]. In Section 3 we will show that this is a typical feature of PNMs. The crucial distinction between PNMs and classical numerical methods is the distributional nature of $B(\mu, a)$, which carries the formal semantics of belief about the QoI. The full distribution $B(\mu, a)$ was examined in Briol et al. [2016], who established contraction to the exact value of the integral under certain smoothness conditions on the Gaussian covariance function and on the integrand.

**Example 2.4** (Probabilistic Meshless Method)**.** As a canonical example of a PDE, take the following elliptic problem with Dirichlet boundary conditions

$$\begin{aligned} -\nabla \cdot (\kappa \nabla x) &= f && \text{in } D \\ x &= b && \text{on } \partial D \end{aligned} \tag{2.3}$$

where we assume $D \subset \mathbb{R}^d$ and $\kappa \colon D \to \mathbb{R}^{d \times d}$ is a known coefficient. Let $\mathcal{X}$ be a separable Banach space of appropriately differentiable real-valued functions and take $\Sigma_{\mathcal{X}}$ to be the Borel $\sigma$-algebra for $\mathcal{X}$.

Such problems were considered in Cockayne et al. [2016] wherein $\mu$ was restricted to be a Gaussian distribution on $\mathcal{X}$. The information operator was constructed by choosing finite sets of locations $T_1 = \{t_{1,1}, \ldots, t_{1,n_1}\} \subset D$ and $T_2 = \{t_{2,1}, \ldots, t_{2,n_2}\} \subset \partial D$ at which the system defined in Eq. (2.3) was evaluated, so that

$$A(x) = \begin{bmatrix} -\nabla \cdot (\kappa(t_{1,1}) \nabla x(t_{1,1})) \\ \vdots \\ -\nabla \cdot (\kappa(t_{1,n_1}) \nabla x(t_{1,n_1})) \\ x(t_{2,1}) \\ \vdots \\ x(t_{2,n_2}) \end{bmatrix} \qquad a = \begin{bmatrix} f(t_{1,1}) \\ \vdots \\ f(t_{1,n_1}) \\ b(t_{2,1}) \\ \vdots \\ b(t_{2,n_2}) \end{bmatrix} .$$

The belief update operator was chosen to be $B(\mu, a) = \mu^a$, where $\mu^a$ is the restriction of $\mu$ to those functions for which $A(x) = a$ is satisfied. In contrast to the first illustration, the QoI here is $Q(x) = x$, as the goal is to make inferences about the solution of the PDE itself. In the setting of a linear system of PDEs such as that in Eq. (2.3), the distribution $B(\mu, a)$ is again Gaussian [Bogachev, 1998]. Full details are provided in Cockayne et al. [2016].

As in the previous example, we note that the mean of $B(\mu, a)$ coincides with the numerical solution to the PDE provided by a classical method [the *symmetric collocation* method; Fasshauer, 1999]. The full distribution $B(\mu, a)$ provides uncertainty quantification for the unknown exact solution and can again be shown to contract to the exact solution under certain smoothness conditions [Cockayne et al., 2016]. This method was further analysed for a specific choice of covariance operator in the belief distribution $\mu$, in an impressive contribution from Owhadi [2015b].

### 2.2.1. Classical Numerical Methods

Standard numerical methods fit into the above framework, as can be seen by taking

$$B(\mu, a) = \delta \circ c(a) \tag{2.4}$$

independent of the distribution $\mu$, where a function $c \colon \mathcal{A} \to \mathcal{Q}$ gives the output of some classical numerical method for solving the problem of interest. Here $\delta \colon \mathcal{Q} \to \mathcal{P}_{\mathcal{Q}}$ maps $c(a) \in \mathcal{Q}$ to a

Dirac measure centred on $c(a)$. Thus, information in $a \in \mathcal{A}$ is used to construct a point estimate $c(a) \in \mathcal{Q}$ for the QoI.

The formal language of probabilities is not used in classical numerical analysis to describe numerical error. However, in many cases the classical and probabilistic analyses are mathematically equivalent. For instance, Briol et al. [2016] highlighted an equivalence between the standard deviation of $B(\mu, a)$ for probabilistic integration and the worst-case error for numerical cubature rules from numerical analysis [Novak and Woźniakowski, 2010]. The explanation for this phenomenon will be given in Section 3.

## 2.3. Bayesian PNMs

Having defined a PNM, we now state the central definition of this paper, that is of *Bayesian* PNMs. Define $\mu^a$ to be the conditional distribution of the random variable $X$, given the event $A(X) = a$. For now we assume that this can be defined without ambiguity and reserve a more technical treatment of conditional probabilities for Section 2.4.

The main contribution of this work is to establish that the problem of determining $x$ in Eq. (2.1) can be cast as Bayesian inversion [Stuart, 2010], a methodology now popular in applied mathematics and uncertainty quantification research. However, there is a crucial technical distinction between problems of the form in Eq. (2.1) and more widely studied Bayesian inverse problems. Specifically, in a standard Bayesian inverse problem the observed quantity $a$ is assumed to be corrupted with noise, which is described by a "likelihood" function or "potential". This leads, under mild assumptions, to general versions of Bayes' theorem [see Stuart, 2010, Section 2.2]

For PNM, the information is *not* corrupted with measurement error. As a result, the support of the "likelihood" is a set of measure zero, making the standard approaches to such problems, including Bayes' theorem, ill-defined outside of a limited class of conjugate models. This necessitates a new definition:

**Definition 2.5** (Bayesian Probabilistic Numerical Method)**.** A probabilistic numerical method $M = (A, B)$ is said to be *Bayesian*[2] for a quantity of interest $Q$ if, for all $\mu \in \mathcal{P}_{\mathcal{X}}$, the output

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

That is, a PNM is Bayesian if the output of the PNM is the push-forward of the conditional distribution $\mu^a$ through $Q_{\#}$. This definition is familiar from the examples in Section 2.2, which are both examples of Bayesian PNMs.

For Bayesian PNMs we adopt the traditional terminology in which $\mu$ is the *prior* over $x$ and the output $Q_{\#}\mu^a$ the *posterior* over $Q(x)$. Note that, for fixed $A$ and $\mu$, the Bayesian choice for $B$ is uniquely defined, if it exists.

It is emphasised that the class of Bayesian PNMs is a subclass of all PNMs; examples of non-Bayesian PNMs are provided in Section 2.5.1. Our analysis is focussed on Bayesian PNMs due to their appealing Bayesian interpretation and ease of generalisation to pipelines of computation in Section 5. For non-Bayesian PNMs, careful definition and analysis of the belief update operator is necessary to enable proper interpretation of the uncertainty quantification being provided. In particular, the analysis of non-Bayesian PNMs may present considerable challenges in the context of computational pipelines, whereas for Bayesian PNMs this is shown in Section 5 to be straight-forward.

[2]The use of "Bayesian" contrasts with Bissiri et al. [2016], for whom all belief update operators represent Bayesian learning algorithms to some greater or lesser extent. An alternative term could be "lossless", since the information in $a$ is conditioned upon in $\mu^a$.

## 2.4. The Disintegration Theorem

The purpose of this section is to interpret $\mu^a$ and to determine conditions under which $\mu^a$ exists and is well-defined. From Definition 2.5, the object of interest is $B(\mu, a) = Q_{\#}\mu^a$. If $\mu^a$ exists, the pushforward $Q_{\#}\mu^a$ exists as $Q$ is assumed to be measurable; thus, in this section, we focus on a rigorous definition of $\mu^a$.

Unlike many problems of Bayesian inversion, proceeding by an analogue of Bayes' theorem is not possible. Let $\mathcal{X}^a = \{x \in \mathcal{X} : A(x) = a\}$. Then we observe that, if it is indeed measurable, $\mathcal{X}^a$ could be a set of zero measure under $\mu$. Standard techniques for infinite-dimensional Bayesian inversion rely on constructing a conditional distribution based on a non-vanishing normalisation constant [Stuart, 2010], so we must turn to other approaches to establish when a Bayesian PNM is well-defined.

Conditioning on null sets is technical and was formalised in the celebrated construction of measure-theoretic probability by Kolmogorov [1933]. The central challenge is to establish uniqueness of conditional probabilities. For this work we exploit the *disintegration theorem* to ensure our constructions are well-defined. The definition below is due to Dellacherie and Meyer [1978, p.78], and a statistical introduction to disintegration can be found in Chang and Pollard [1997].

**Definition 2.6** (Disintegration)**.** For $\mu \in \mathcal{P}_{\mathcal{X}}$, a collection $\{\mu^a\}_{a\in\mathcal{A}} \subset \mathcal{P}_{\mathcal{X}}$ is a *disintegration* of $\mu$ with respect to the (measurable) map $A\colon \mathcal{X} \to \mathcal{A}$ if:

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

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

2 (Measurability:) $a \mapsto \mu^a(f)$ is measurable;

3 (Conditioning:) $\mu(f) = \int \mu^a(f) A_{\#}\mu(\mathrm{d}a)$.

The concept of disintegration extends the usual concept of conditioning of random variables to the case where $\mathcal{X}^a$ is a null set. Existence of disintegrations is guaranteed under general (weak) conditions:

**Theorem 2.7** (Disintegration Theorem; Thm. 1 of Chang and Pollard [1997])**.** *Let $\mathcal{X}$ be a metric space, $\Sigma_{\mathcal{X}}$ be the Borel $\sigma$-algebra and let $\mu \in \mathcal{P}_{\mathcal{X}}$ be Radon. Let $\Sigma_{\mathcal{A}}$ be countably generated and contain all singletons $\{a\}$ for $a \in \mathcal{A}$. Then there exists a disintegration $\{\mu^a\}_{a\in\mathcal{A}}$ of $\mu$ with respect to $A$. Moreover, if $\{\nu^a\}_{a\in\mathcal{A}}$ is another such disintegration, then $\{a \in \mathcal{A} : \mu^a \neq \nu^a\}$ is a $A_{\#}\mu$ null set.*

The requirement that $\mu$ is Radon is weak and is implied when $\mathcal{X}$ is a *Radon space*, which encompasses, for example, separable complete metric spaces. The requirement that $\Sigma_{\mathcal{A}}$ is countably generated is also weak and includes the standard case where $\mathcal{A} = \mathbb{R}^n$ with the Borel $\sigma$-algebra. From Theorem 2.7 it follows that $\{\mu^a\}_{a\in\mathcal{A}}$ exists and is essentially unique for all of the examples considered in this paper. Thus, under mild conditions, we have established that Bayesian PNMs are well-defined, in that an essentially unique disintegration $\{\mu^a\}_{a\in\mathcal{A}}$ exists.

## 2.5. Prior Construction

The Gaussian distribution is popular as a prior in the PNM literature for its tractability, both in the fact that finite-dimensional distributions take a closed-form and that an explicit conditioning formula exists. More general priors, such as Besov priors [Dashti et al., 2012] and Cauchy priors

[Sullivan, 2016] are less easily accessed. In this section we summarise the common construction for all such prior distributions.

Let $\{\phi_i\}_{i=0}^{\infty}$ denote a Schauder basis for $\mathcal{X}$, assumed to be a separable Banach space in this section. Then any $x \in \mathcal{X}$ can be represented through an expansion

$$x = x_0 + \sum_{i=0}^{\infty} u_i \phi_i \tag{2.5}$$

for some fixed element $x_0 \in \mathcal{X}$ and a sequence $u \in \mathbb{R}^{\infty}$. Construction of measures on $\mathcal{X}$ is then reduced to construction of almost-surely convergent measures on $\mathbb{R}^{\infty}$ and studying the pushforward of such measures into $\mathcal{X}$.

To this end it is common to split $u$ into a stochastic and deterministic component; let $\xi \in \mathbb{R}^{\infty}$ represent an i.i.d sequence of random variables, and $\gamma \in \ell^p$ for some $p \in (1, \infty)$. Then with $u_i = \gamma_i \xi_i$, for the prior distribution to be well-posed we require that almost-surely $u \in \ell^1$. Different choices of $(\xi, \gamma)$ give rise to different distributions on $\mathcal{X}$. For instance, $\xi_i \sim \text{Uniform}(-1, 1)$, $\gamma \in \ell^1$ is termed a *uniform* prior and $\xi_i \sim \mathcal{N}(0, 1)$ gives a *Gaussian* prior, where $\gamma$ determines the regularity of the covariance operator $\mathcal{C}$ [Bogachev, 1998]. The choice of $\xi_i \sim \text{Cauchy}(0, 1)$ gives a *Cauchy* prior in the sense of Sullivan [2016]. Here we require $\gamma \in \ell^1 \cap \ell \log \ell$ for $\mathcal{X}$ a separable Banach space, or $\gamma \in \ell^2$ for when $\mathcal{X}$ is a Hilbert space.

A range of prior specifications will be explored in Section 6, including non-Gaussian prior distributions for numerical solution of nonlinear ODEs.

#### 2.5.1. Dichotomy of Existing PNMs

This section concludes with an overview of existing PNMs with respect to our proposed definition of Bayesian PNMs. This serves to clarify some subtle distinctions in existing literature, as well as to highlight the generality of our framework. To maintain brevity we have summarised our findings in Table 1.

## 3. Decision-Theoretic Treatment

In this section we assess the performance of PNMs from a decision-theoretic perspective [Berger, 1985] and explore connections to average-case analysis [Ritter, 2000]. Note that the treatment here is agnostic to whether the PNM in question is Bayesian, and also encompasses classical numerical methods. Throughout, the existence of a disintegration $\{\mu^a\}_{a \in \mathcal{A}}$ will be assumed.

### 3.1. Loss and Risk

Consider a generic loss function $L \colon \mathcal{Q} \times \mathcal{Q} \to \mathbb{R}$ where $L(q^\dagger, q)$ describes the loss incurred when the QoI $Q(x)$ is estimated with $q$, but its true value is $q^\dagger$. In this paper we restrict attention to symmetric loss functions, so that $L(q^\dagger, q) = L(q, q^\dagger)$, and integrability will be assumed. Further, we consider inner product spaces $(\mathcal{Q}, \langle \cdot, \cdot \rangle_{\mathcal{Q}})$ for the QoI. A common choice is the squared-error loss $L(q^\dagger, q) = \|q^\dagger - q\|_{\mathcal{Q}}^2$ where $\| \cdot \|_{\mathcal{Q}}$ is the norm induced by the inner product.

The belief update operator $B$ returns a distribution over $\mathcal{Q}$ which can be cast as a randomised decision rule for estimation of $q^\dagger$. For randomised decision rules, the *risk function* $r \colon \mathcal{Q} \times \mathcal{P}_{\mathcal{Q}} \to \mathbb{R}$ is defined as

$$r(q^\dagger, \beta) = \int L(q^\dagger, q) \beta(\mathrm{d}q) \, .$$

| **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] |
| | | $\nabla x$ + 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] | |

Notes

1 Here an extended definition of Bayesian PNM is used, that allows for auxiliary randomisation and adaptivity in the information operators; see the Discussion for details.

2 A prior was proposed for the floating point error, rather than for $x$ itself.

3 A prior was proposed for the discretisation error, rather than for $x$ itself.

4 When the system of ODEs is linear, these methods are Bayesian for more specific quantities of interest; $Q(x) = x(t_{\text{end}})$ but not for the full solution $x$ to the initial value problem (IVP).

5 A prior was placed on $\arg\min x(t)$ rather than on $x$ itself. However, only these marginal probabilities are actually required for Bayesian PNMs.

6 This literature focuses on point estimators that summarise the full distributional output; however, the full output constitutes a Bayesian PNM.

7 A finite difference approximation to the differential operator was used.

Table 1: Comparison of existing Probabilistic Numerical Methods (PNMs).

The *average risk* of the PNM $M = (A, B)$ with respect to $\mu \in \mathcal{P}_{\mathcal{X}}$ is defined as

$$R(\mu, M) = \int r(Q(x), B(\mu, A(x)))\mu(\mathrm{d}x). \tag{3.1}$$

We follow the convention of terming $R(\mu, M)$ the *Bayes risk* of the PNM, though the usual objection that a frequentist expectation enters into the definition of the Bayes risk could be raised.

**Definition 3.1** (Contraction)**.** A sequence $M^{(n)} = (A^{(n)}, B^{(n)})$ of PNMs is said to *contract* at a *rate* $r_n$ under a belief distribution $\mu$ and a loss function $L$ if $R(\mu, M^{(n)}) = O(r_n)$.

The above construction allows for comparison of both classical and probabilistic numerical methods in the same framework. In each case an important goal is to determine methods $M^{(n)}$ that contract as quickly as possible for a given distribution $\mu$ that defines the Bayes risk. This is the approach taken in average-case analysis [ACA; Ritter, 2000] and will be discussed in Section 3.4. For the earlier Examples 2.3 and 2.4 of Bayesian PNMs, Briol et al. [2016] and Cockayne et al. [2016] established rates of contraction for particular prior distributions $\mu$; we refer the reader to those papers for details.

## 3.2. Bayes Decision Rules

A (possibly randomised) decision rule is said to be a *Bayes rule* if it achieves the minimum Bayes risk among all decision rules. In the context of PNMs, let $M = (A, B)$ and let

$$\mathfrak{B}(A) = \left\{ B : R(\mu, (A, B)) = \inf_{B'} R(\mu, (A, B')) \right\}$$

that is, for fixed $A$, $\mathfrak{B}(A)$ is the set of all belief update operators that achieve minimum Bayes risk.

This raises the natural question of which belief update operators yield Bayes rules. Although the definition of a Bayes rule applies generically to both probabilistic and deterministic numerical methods, it can be shown[3] that if $\mathfrak{B}(A)$ is non-empty, then there exists a $B \in \mathfrak{B}(A)$ which takes the form of a classical numerical method, as expressed in Eq. (2.4). This may explain why neither classical statistical analysis nor numerical analysis have considered PNMs, as there exist deterministic numerical methods that are optimal in the sense of minimising Bayes risk.

In general, Bayesian PNMs do *not* constitute Bayes rules, as the extra uncertainty inflates the Bayes risk, so that such methods are not Bayes-optimal. Nonetheless, there is a natural connection between Bayesian PNMs and Bayes rules:

**Theorem 3.2.** *Let $M = (A, B)$ be a Bayesian probabilistic numerical method for the QoI $Q$. Let the loss function $L$ have the form $L(q^\dagger, q) = \|q^\dagger - q\|_Q^2$. Then the mean of $B(\mu, a)$ is a Bayes rule for estimation of $q^\dagger$.*

This well-known fact from Bayesian decision theory[4] is interesting in light of recent research in constructing PNMs whose mean functions correspond to classical numerical methods [Schober et al., 2014, Hennig, 2015, Särkkä et al., 2015, Teymur et al., 2016, Schober et al., 2016], and the famous observation of Diaconis [1988] that several classical numerical methods can be viewed as Bayes rules for specific choices of the prior distribution $\mu$. Theorem 3.2 explains the results in Examples 2.3 and 2.4, in which both instances of Bayesian PNMs were demonstrated to be centred on an established classical method.

[3] The proof is included in the Electronic Supplement.

[4] This is the fact that the Bayes act is the posterior mean under squared-error loss [Berger, 1985].

## 3.3. Optimal Information

The previous section considered selection of the belief update operator $B$, but not of the information operator $A$. The choice of $A$ determines the Bayes risk for a PNM, which leads to a problem of experimental design to minimise that risk.

The theoretical study of optimal information is the focus of the information complexity literature [Traub et al., 1988, Novak and Woźniakowski, 2010], while other fields such as quasi-Monte Carlo [QMC, Dick and Pillichshammer, 2010] attempt to develop asymptotically optimal information operators for specific numerical tasks, such as the choice of evaluation points for numerical approximation of integrals in the case of QMC. Here we characterise optimal information for Bayesian PNMs.

Consider the choice of $A$ from a fixed subset $\Lambda$ of the set of all possible information operators. To build intuition, for the task of numerical integration $\Lambda$ could represent all possible choices of information $\{x(t_i)\}_{i=1}^n$ where the locations $\{t_i\}_{i=1}^n$ are to be selected. For a Bayesian PNM $M = (A, B)$, one can ask for optimal information:

$$A_\mu \in \operatorname*{arg\,inf}_{A\in\Lambda} R(\mu, M)$$

where we have made explicit the fact that the optimal information depends on the choice of prior $\mu$. Once $A_\mu$ is determined, the corresponding belief operator $B_\mu$ is uniquely determined from the fact that $M_\mu$ is a Bayesian PNM. Next we characterise $A_\mu$, while an explicit example of optimal information for a Bayesian PNM is detailed in Example 3.4.

## 3.4. Connection to Average Case Analysis

The decision theoretic framework in Section 3.1 is closely related to ACA [Ritter, 2000]. In ACA the performance of a classical numerical method $c\colon \mathcal{A} \to \mathcal{Q}$, defined in Eq. (2.4), is studied in terms of the average error

$$e(\mu, A, c) = \int L(Q(x), c(A(x)))\mu(\mathrm{d}x). \tag{3.2}$$

which we recognise as the Bayes risk given in Eq. (3.1), for a deterministic PNM $B(\mu, a) = \delta \circ c(a)$ as in Eq. (2.4). ACA is concerned with the study of optimal information:

$$A^*_\mu \in \operatorname*{arg\,inf}_{A\in\Lambda} \inf_c e(\mu, A, c).$$

For Gaussian belief distributions $\mu$, a detailed theoretical analysis of existence of ACA optimal methods was provided in Tarieladze and Vakhania [2007]. Here we prove that optimal information for a Bayesian PNM is, under a constraint on the loss function $L$, the same as optimal information for ACA.

**Theorem 3.3.** *Let the loss function $L$ have the form $L(q^\dagger, q) = \|q^\dagger - q\|^2_{\mathcal{Q}}$. Then the optimal information $A_\mu$ for a Bayesian PNM and $A^*_\mu$ for ACA are identical.*

It is perhaps natural to expect that optimal information for PNM and ACA coincide in this context, since from Theorem 3.2 the mean of a Bayesian PNM are Bayes rules, and the Bayes risk is the object of interest in ACA.

This discussion connects Bayesian PNMs to the *design* of classical numerical methods, a perspective that was first clearly exposed in Larkin [1972] and Kadane and Wasilkowski [1985]. Thus we can extract results on optimal average case information from the ACA literature and use them to construct optimal Bayesian PNMs. An example is provided next.

**Example 3.4** (Optimal Information for Probabilistic Integration)**.** To illustrate optimal information for Bayesian PNMs, we observe historical precedent and revisit the first worked example of ACA, due to Sul′din [1959, 1960]. Set $\mathcal{X} = \{x \in C(0,1) : x(0) = 0\}$ and take the belief distribution $\mu$ to be induced from the Weiner process on $\mathcal{X}$, i.e. a Gaussian distribution with mean 0 and covariance function $k(t,t') = \min(t,t')$. Our QoI is $Q(x) = \int_0^1 x(t)\mathrm{d}t$ and the loss function is $L(q,q') = (q-q')^2$.

Consider standard information $A(x) = (x(t_1), \ldots, x(t_n))$ for $n$ fixed knots $0 \leq t_1 < \cdots < t_n \leq 1$. Our aim is to determine knots $t_i$ that represent optimal information for a Bayesian PNM with respect to $\mu$ and $L$.

Motivated by Theorem 3.3 we first solve the optimal information problem for ACA and then derive the associated PNM. It will be sufficient to restrict attention to linear methods $c(a) = \sum_{i=1}^n w_i x(t_i)$ with $w_i \in \mathbb{R}$. Substitution into Eq. (3.2) produces a closed-form expression for the average error:

$$e(\mu, A, c) = \frac{1}{3} - 2\sum_{i=1}^n w_i \left(t_i - \frac{1}{2}t_i^2\right) + \sum_{i,j=1}^n w_i w_j \min(t_i, t_j). \tag{3.3}$$

Standard calculus can be used to minimise Eq. (3.3) over both the weights $\{w_i\}_{i=1}^n$ and the locations $\{t_i\}_{i=1}^n$; the full calculation can be found in Chapter 2, Section 3.3 of Ritter [2000]. The result is an ACA optimal method

$$c_{\mu,A}(A(x)) = \frac{2}{2n+1}\sum_{i=1}^n x(t_i^*), \quad t_i^* = \frac{2i}{2n+1}$$

which is recognised as the trapezium rule with equally spaced knots. The associated contraction rate $r_n$ is $n^{-1}$ [Lee and Wasilkowski, 1986].

From Theorem 3.3 we have that ACA optimal information is also optimal information for the Bayesian PNM. Thus the optimal Bayesian PNM $M = (A, B)$ for the belief distribution $\mu$ is uniquely determined:

$$A(x) = \begin{bmatrix} x(t_1^*) \\ \vdots \\ x(t_n^*) \end{bmatrix}, \qquad B(\mu, a) = \mathrm{N}\left(\frac{2}{2n+1}\sum_{i=1}^n a_i \ , \ \frac{1}{3(2n+1)^2}\right).$$

Note how the PNM is centred on the ACA optimal method. However the PNM itself is not a Bayes rule; it in fact carries twice the Bayes risk as the ACA method.

This illustration can be generalised. It is known that for $\mu$ induced from the Weiner process on $\partial^s x$, $Q$ a linear functional and $\phi$ a loss function that is convex and symmetric, equi-spaced evaluation points are essentially optimal information, the Bayes rule is the natural spline of degree $2s+1$, and the contraction rate $r_n$ is essentially $n^{-(s+1)}$; see Lee and Wasilkowski [1986] for a complete treatment.

This completes our performance assessment for PNMs; next we turn to computational matters.

## 4. Numerical Disintegration

In this section we discuss algorithms to access the output from a Bayesian PNM. The approach considered in this paper uses standard statistical techniques to draw approximate samples from

$B(\mu, a)$, which can be achieved by drawing approximate samples $x' \sim \mu^a$ and evaluating $Q(x')$. The construction of a sampling scheme can exploit sophisticated Monte Carlo methods and allow probing $B(\mu, a)$ at a computational cost that is de-coupled from the potentially substantial cost of obtaining the information $a$ itself.

The construction of a sampling scheme is non-trivial on a technical level. As shown in Section 2.4, under weak conditions on the space $\mathcal{X}$ and the operator $A$, the disintegration $\mu^a$ is well-defined for $A_{\#}\mu$-almost all $a \in \mathcal{A}$. However, sampling from $\mu^a$ is a non-trivial problem that is equivalent to approximating derivatives of set functions. The approach considered in this work is based on sampling from an approximate distribution $\mu^a_\delta$ which converges in an appropriate sense to $\mu^a$ in the $\delta \downarrow 0$ limit. The use of a convergent approximation is reminiscent of Pfanzagl [1979], who established a weak convergence result similar to that we present in Section 4.1. However, his result relies upon the assumption that the topological convergence of the approximation is sufficiently regular, in that it forms a *differentiation basis* [in the terminology of Rao, 2005]. Differentiation bases have been shown to exist in all finite dimensional settings as discussed in Pfanzagl [1979]. However this result relies upon the Vitali covering lemma, which does not generally hold in infinite-dimensional settings, even such regular ones as Gaussian measures on Hilbert spaces [Tišer, 2003]. This motivates a novel theoretical approach which we next outline in detail.

## 4.1. Sequential Approximation of a Disintegration

Suppose that $\mathcal{A}$ is an open subset of $\mathbb{R}^n$ and that the distribution $A_{\#}\mu \in \mathcal{P}_{\mathcal{A}}$, admits a continuous and positive density $p_A$ with respect to Lebesgue measure on $\mathcal{A}$. Further endow $\mathcal{A}$ with the structure of a Hilbert space, with norm $\|\cdot\|_{\mathcal{A}}$.

Let $\phi\colon \mathbb{R}_+ \to \mathbb{R}_+$ denote a decreasing function, to be specified, that is continuous at 0, with $\phi(0) = 1$ and $\lim_{r\to\infty} \phi(r) = 0$. Then a "relaxed" version of the disintegration $\{\mu^a\}_{a\in\mathcal{A}}$ can be defined as

$$\mu^a_\delta(\mathrm{d}x) \coloneqq \frac{1}{Z^a_\delta}\phi\left(\frac{\|A(x) - a\|_{\mathcal{A}}}{\delta}\right)\mu(\mathrm{d}x)$$

where the normalisation constant

$$Z^a_\delta \coloneqq \int \phi\left(\frac{\|\tilde{a} - a\|_{\mathcal{A}}}{\delta}\right) p_A(\mathrm{d}\tilde{a})$$

is non-zero since $p_A$ is bounded away from 0 on a neighbourhood of $a \in \mathcal{A}$ and $\phi$ is bounded away from 0 on a sufficiently small interval $[0, \gamma]$. Our aim is to approximate $\mu^a$ with $\mu^a_\delta$ for small bandwidth parameter $\delta$. The construction, which can be considered a mathematical generalisation of approximate Bayesian computation [Del Moral et al., 2012] to the case of infinite-dimensional distributions, ensures that $\mu^a_\delta \ll \mu$. The role of $\phi$ is to admit states $x \in \mathcal{X}$ for which $A(x)$ is close to $a$ but not necessarily equal. It is assumed to be sufficiently regular:

**Assumption 4.1.** There exists $\alpha > 0$ such that $C^\alpha_\phi \coloneqq \int r^{\alpha+n-1}\phi(r)\mathrm{d}r < \infty$.

To discuss the convergence of $\mu^a_\delta$ to $\mu^a$ we must first specify a metric on $\mathcal{P}_{\mathcal{X}}$. Let $\mathcal{F}$ be a normed space of (measurable) functions $f\colon \mathcal{X} \to \mathbb{R}$ with norm $\|\cdot\|_{\mathcal{F}}$. For measures $\nu, \nu' \in \mathcal{P}_{\mathcal{X}}$, define an integral probability metric based on $\mathcal{F}$:

$$d_{\mathcal{F}}(\nu, \nu') = \sup_{\|f\|_{\mathcal{F}}\leq 1} |\nu(f) - \nu'(f)|$$

This formulation encompasses many common probability metrics such as the total variation distance and Wasserstein distance [Müller, 1997]. However, not all spaces of functions $\mathcal{F}$ lead

to useful theory. In particular the total variation distance between $\mu^a$ and $\mu^{a'}$ for $a \neq a'$ will be one by the definition of the disintegration. Sufficient conditions for weak convergence with respect to $\mathcal{F}$ are now established:

**Assumption 4.2.** The map $a \mapsto \mu^a$ is almost everywhere Hölder continuous in $d_{\mathcal{F}}$, i.e.

$$d_{\mathcal{F}}(\mu^a, \mu^{a'}) \leq C_\mu \|a - a'\|_{\mathcal{A}}^\alpha$$

for some constants $C_\mu > 0$, $\alpha$ and for $A_\# \mu$ almost all $a, a' \in \mathcal{A}$.

**Theorem 4.3.** *Let $C_\phi \coloneqq C_\phi^\alpha / C_\phi^0$. Then, for $\delta$ sufficiently small,*

$$d_{\mathcal{F}}(\mu_\delta^a, \mu^a) \leq C_\mu (1 + C_\phi) \delta^\alpha$$

*for $A_\# \mu$ almost all $a \in \mathcal{A}$.*

This result is stronger than that of Pfanzagl [1979] in that it holds for infinite-dimensional $\mathcal{X}$, though it also relies upon the stronger Hölder continuity assumption. The specific form for $\phi$ is not fundamental, but can impact upon rate constants. For the choice $\phi(r) = 1[r < 1]$ we have $C_\phi = \frac{n}{\alpha+n}$, which is well-behaved. On the other hand, for $\phi(r) = \exp(-\frac{1}{2}r^2)$ it can be shown that, for $\alpha \in \mathbb{N}$,

$$C_\phi = \frac{(\alpha + n - 1)!!}{(n-1)!!} \tag{4.1}$$

so that the constant $C_\phi$ might not be small. In general this necessitates effective Monte Carlo methods that are able to sample from the regime where $\delta$ can be extremely small, in order to control the overall approximation error.

## 4.2. Computation for Series Priors

The series representation of $\mu$ in Eq. (2.5) of Section 2.5 is infinite-dimensional and thus cannot, in general, be instantiated. To this end, define $\mathcal{X}_N = x_0 + \operatorname{span}\{\phi_0, \ldots, \phi_N\}$ and define the associated projection operator $P_N \colon \mathcal{X} \to \mathcal{X}_N$ as

$$P_N \left( x_0 + \sum_{i=0}^{\infty} u_i \phi_i \right) \coloneqq x_0 + \sum_{i=0}^{N} u_i \phi_i.$$

A natural approach is to compute with the modified information operator $A \circ P_N$ instead of $A$. This has the effect of updating the first $N+1$ coefficients and leaving the remainder unchanged, distributed according to the prior. Thus computation required in the Bayesian *update* step is finite-dimensional, whilst instantiation of the posterior itself remains infinite-dimensional. In the case of a Gaussian prior $\mu$ and nonlinear information operator $A$, this permits an algorithm for "exact approximation" since the remainder $\sum_{i=N+1}^{\infty} u_i \phi_i$ is also Gaussian and this can be sampled. Convergence of the output, denoted $\mu_{\delta,N}^a$, to $\mu_\delta^a$ in the limit $N \to \infty$ will be considered next.

In this section it is required that $\phi$ be everywhere continuous with $\phi > 0$. Let $\varphi = -\log \phi$, so that $\varphi$ is a continuous bijection of $\mathbb{R}_+$ to itself. The following are then also assumed:

**Assumption 4.4.** For each $R > 0$, it holds that $|\varphi(r) - \varphi(r')| \leq C_R |r - r'|$ for some constant $C_R$ and all $r, r' < R$.

**Assumption 4.5.** $\|A(x) - A \circ P_N(x)\|_{\mathcal{A}} \leq \exp(m(\|x\|_{\mathcal{X}}))\Psi(N)$ for all $x \in \mathcal{X}$, where $m$ is measurable and satisfies $\mathbb{E}_{X \sim \mu}[\exp(2m(\|X\|_{\mathcal{X}}))] < \infty$ and $\Psi(N)$ vanishes as $N$ is increased.

**Assumption 4.6.** $\sup_{x\in\mathcal{X}} \|A(x)\|_{\mathcal{A}} < \infty$.

**Assumption 4.7.** $\|f\|_\infty \leq C_{\mathcal{F}} \|f\|_{\mathcal{F}}$ for some constant $C_{\mathcal{F}}$ and all $f \in \mathcal{F}$.

Assumption 4.4 holds for the case $\varphi(r) = \frac{1}{2}r^2$ with constant $C_R = R$. Assumption 4.5 is standard in the inverse problem literature; for instance it is shown to hold for certain series priors in Theorem 3.4 of Cotter et al. [2010]. Assumption 4.6 is, in essence, a compactness assumption, in that for linear information operators $A$ it is implied by compactness of the state space $\mathcal{X}$. In this sense it is a strong assumption; however it can be enforced in our experiments, where $\mathcal{X}$ is unbounded, through a threshold map

$$\tilde{A}(x) \coloneqq \begin{cases} A(x) & \text{if } \|A(x)\|_{\mathcal{A}} \leq \lambda_{\max}, \\ \lambda_{\max} \frac{A(x)}{\|A(x)\|_{\mathcal{A}}} & \text{if } \|A(x)\|_{\mathcal{A}} > \lambda_{\max}, \end{cases}$$

where $\lambda_{\max}$ is a large pre-defined constant. Assumption 4.7 places a restriction on the integral probability metric $d_{\mathcal{F}}$ in which our result holds.

Next we establish a bound on the error due to prior truncation. This theorem has its proof in the Electronic Supplement:

**Theorem 4.8.** *For some constant $C_\delta$, dependent on $\delta$, it holds that $d_{\mathcal{F}}(\mu^a_{\delta,N}, \mu^a_\delta) \leq C_\delta \Psi(N)$.*

An immediate consequence of Theorems 4.3 and 4.8 is that the total approximation error can be bounded by applying the triangle inequality:

$$d_{\mathcal{F}}(\mu^a, \mu^a_{\delta,N}) \leq C_\mu(1 + C_\phi)\delta^\alpha + C_\delta \Psi(N).$$

In particular, we have convergence of $\mu^a_{\delta,N}$ to $\mu^a$ in the $\delta \downarrow 0$ limit provided that the number of basis functions satisfies $C_\delta \Psi(N) = o(1)$.

Recall that, although the Bayesian *update* step is finite-dimensional, the posterior $\mu^a_{\delta,N}$ itself remains infinite-dimensional in the form specified here. For the experiments in Section 6, in which both Gaussian and non-Gaussian priors $\mu$ are considered, we in addition employed a truncation of the series in Eq. (2.5) at level $N + 1$, with the resultant prior denoted $\mu_N$. The posterior $\mu^a_{\delta,N}$ was then entirely supported on the finite-dimensional subspace $\mathcal{X}_N$. It is emphasised that analysis of prior truncation, as opposed to modification of the information operator just reported, is known to be difficult. Indeed, while $\mu_N$ converges to $\mu$ weakly, it does not do so in total variation, and this deficiency generally transfers to the associated posteriors. In general the impact of prior perturbation is a subtle topic — see e.g. Owhadi et al. [2015] and the references therein — and we therefore defer theoretical analysis of this approximation to future work.

## 4.3. Monte Carlo Methods for Numerical Disintegration

The previous sections established a sequence of well-defined distributions $\mu^a_\delta$ (or $\mu^a_{\delta,N}$ for non-Gaussian models) which converge (in a specific weak sense) to the exact disintegration $\mu^a$. General Monte Carlo methods can then be used to sample from $\mu^a_\delta$ (or $\mu^a_{\delta,N}$). The construction of Monte Carlo methods is de-coupled from the core material in the main text and the main methodological considerations are well-documented [e.g. Girolami and Calderhead, 2011]. For the experiments reported in subsequent sections two approaches were explored; a Sequential Monte Carlo (SMC) method [Doucet et al., 2001] and a parallel tempering method [Geyer, 1991]. This provided a transparent sampling scheme, whose non-asymptotic approximation error can be theoretically understood [Del Moral, 2004]. Full details of the Monte-Carlo methods used for this work, along with associated theoretical analysis for the SMC method, are contained in Section C.1 of the Electronic Supplement.

# 5. Computational Pipelines and PNM

The last theoretical development in this paper concerns composition of several PNMs. Most analysis of numerical methods focuses on the error incurred by an individual method. However, real-world computational procedures typically rely on the composition of several numerical algorithms. The manner in which accumulated numerical error affects computational output may be highly non-trivial [Roy, 2010, Anderson, 2011]. An extreme example occurs when one of the numerical algorithms in a pipeline is charged with integration of a chaotic dynamical system [Strogatz, 2014].

In recent work, Chkrebtii et al. [2016], Conrad et al. [2016] and Cockayne et al. [2016] each used PNMs within a broader statistical procedure to estimate unknown parameters in systems of differential equations. The probabilistic description of error was incorporated into the data-likelihood, resulting in posterior distributions for parameters with inflated uncertainty to account for the unknown numerical error. However, beyond these limited works, no examination of the composition of PNMs has been performed. In particular, the question of which PNMs can be composed, and when the output of such a composition is meaningful, has not been addressed. This is important; for instance, if the output of a composition of PNMs is to be used for analysis of variance to elucidate the main sources of numerical error, then it is important that such output is meaningful.

This section defines a "pipeline" as an abstract graphical object within which a collection of PNMs can be "compatible". It is then proven that when compatible Bayesian PNMs are employed in the pipeline, the output of the probabilistic computation carries a Bayesian interpretation under an explicit condition on $\mu$.

To build intuition, for the simple case where two Bayesian PNMs are composed in series, our results provide conditions for when, informally, the distribution $B_2(B_1(\mu, a_1), a_2)$ corresponds to a single, coherent Bayesian procedure $B(\mu, (a_1, a_2))$. To reduce the notational and technical burden, in this section we will not provide rigorous measure theoretic details; however we note that those details broadly follow the same pattern as in Section 2.4.

## 5.1. Computational Pipelines

To analyse pipelines of PNMs, we consider $n$ such methods $M_1, \ldots, M_n$, where each method $M_i = (A_i, B_i)$ is defined on a common state space $\mathcal{X}$ and targets a QoI $Q_i \in \mathcal{Q}_i$. A pipeline will be represented as a directed graphical model, wherein the QoIs $Q_i$ from parent methods constitute information operators for child methods. It may be that a method will take as its input quantities from multiple parents. To allow for this, we suppose that the information operator $A_i \colon \mathcal{X} \to \mathcal{A}_i$ can be decomposed into components $A_{i,j} \colon \mathcal{X} \to \mathcal{A}_{i,j}$ such that $A_i = (A_{i,1}, \ldots, A_{i,m(i)})$ and $\mathcal{A}_i = \mathcal{A}_{i,1} \times \cdots \times \mathcal{A}_{i,m(i)}$. Thus, each component $A_{i,j}$ can be thought of as the QoI output by one of the parents of the method $M_i$.

Without loss of generality we designate the $n$th QoI $Q_n$ to be the *principal* QoI. That is, the purpose of the computational pipeline is to estimate $Q_n$. The case of multiple principal QoI is a simple extension not described herein. Nodes with no immediate children are called *terminal* nodes, while nodes with no immediate parents are called *source nodes*. We denote by $A$ the set of all source nodes.

**Definition 5.1** (Pipeline)**.** A *pipeline* $P$ is a directed acyclic graph defined as follows:

- Nodes are of two kinds: *Information* nodes are depicted by □, and *method* nodes are depicted by ■.

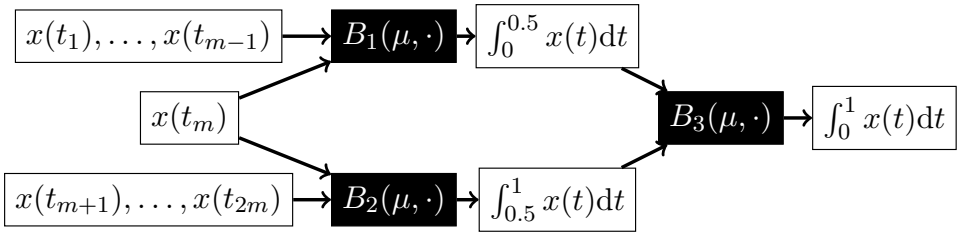



Figure 1: An intuitive representation of Example 5.2.

- The graph is bipartite, so that edges connect a method node to an information node or vice-versa. That is, edges are of the form $\square \to \blacksquare$ or $\blacksquare \to \square$.
- There are $n$ method nodes, each with a unique label in $\{1, \ldots, n\}$.
- The method node labelled $i$ has $m(i)$ parents and one child. Its in-edges are assigned a unique label in $\{1, \ldots, m(i)\}$.
- There is a unique terminal node and it is the child of method node $n$. This represents the principal QoI $Q_n$.

**Example 5.2** (Distributed Integration)**.** Recall the numerical integration problem of Example 3.4 and consider partitioning the domain of integration in order to distribute computation:

$$\underbrace{\int_0^1 x(t)\mathrm{d}t}_{\text{(c)}} = \underbrace{\int_0^{0.5} x(t)\mathrm{d}t}_{\text{(a)}} + \underbrace{\int_{0.5}^1 x(t)\mathrm{d}t}_{\text{(b)}} \tag{5.1}$$

To keep presentation simple we consider an integral over $[0, 1]$ with $2m + 1$ equidistant knots $t_i = i/2m$. Let $M_1$ be a Bayesian PNM for estimating $Q_1(x) =$ (a) and $M_2$ be a Bayesian PNM for estimating $Q_2(x) =$ (b).

To perform this computation we divide the information operator into four components; $A_{i,j}$, for $i, j \in \{1, 2\}$. $A_{1,1}$ and $A_{2,2}$ contain the information unique to $M_1$ and $M_2$. Specifically

$$A_{1,1}(x) = \begin{bmatrix} x(t_1) \\ \vdots \\ x(t_{m-1}) \end{bmatrix}, \qquad A_{2,2}(x) = \begin{bmatrix} x(t_{m+1}) \\ \vdots \\ x(t_{2m}) \end{bmatrix}.$$

$A_{1,2}$ and $A_{2,1}$ contain the information that is shared between the two methods; that is $A_{1,2} = A_{2,1} = \{x(t_m)\}$. To complete the computation we need a third PNM for estimation of $Q_3(x) =$ (c) which we denote $M_3$ and which combines the outputs of $M_1$ and $M_2$ by simply adding them together. Formally this has information operator $A_3(x) = (A_{3,1}(x), A_{3,2}(x))$ where $A_{3,1}(x) =$ (a) and $A_{3,2}(x) =$ (b). Its belief update operator is given by:

$$B_3(\mu, (a_{3,1}, a_{3,2})) = a_{3,1} + a_{3,2}$$

An intuitive graphical representation of this set-up is shown in Figure 1. The pipeline $P$ itself, which is identical to Figure 1 but with additional node and edge labels, is shown in Figure 2.

In general, the method node labelled $i$ is taken to represent the method $M_i$. The in-edge to this node labelled $j$ is taken to represent the information provided by the relationship $A_{i,j}(x_i) =$

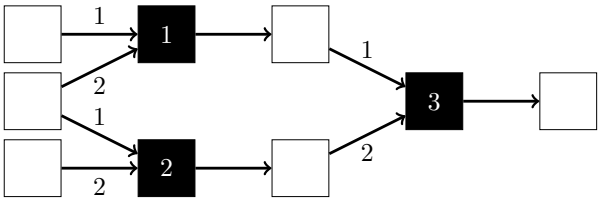
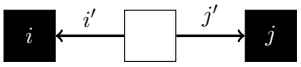



Figure 2: The pipeline $P$ corresponding to Figure 1.

$a_{i,j}$. Here $a_{i,j}$ can either be deterministic information provided to the pipeline, or statistical information derived from the output of another PNM. To make this formal and to "match the input-output spaces" we next define what it means for the collection of methods $M_i$ to be compatible with the pipeline $P$. Informally, this describes the conditions that must be satisfied for method nodes in a pipeline to be able to connect to each other.

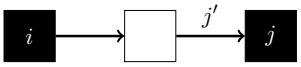
**Definition 5.3** (Compatible)**.** The collection $(M_1, \ldots, M_n)$ of PNMs is *compatible* with the pipeline $P$ if the following two requirements are satisfied:

(i) (Method nodes which share an information node must have consistent information spaces and information operators.) For a motif




we have that $A_{i,i'} = A_{j,j'}$ and $\mathcal{A}_{i,i'} = \mathcal{A}_{j,j'}$.

(ii) (The space $\mathcal{Q}_i$ for the output of a previous method must be consistent with the information space of the next method.) For a motif




we have that $\mathcal{Q}_i = \mathcal{A}_{j,j'}$.

Note that we do not require the converse of (i) at this stage; that is, the same information can be represented by more than one node in the pipeline. This permits redundancy in the pipeline, in that information is not recycled. It will be argued that pipelines with such redundancy are non-Bayesian.

The role of the pipeline $P$ is to specify the order in which information, either deterministic of statistical, is propagated through the collection of PNMs. This is illustrated next:

**Example 5.4** (Propagation of Information)**.** For the pipeline in Figure 2 the propagation of information proceeds as follows::

1. The source nodes, representing $A(x) = \{A_{1,1}(x), A_{1,2}(x) = A_{2,1}(x), A_{2,2}(x)\}$ are evaluated as $\{a_{1,1}, a_{1,2} = a_{2,1}, a_{2,2}\}$. This represents all the (deterministic) information that we are given on $x$ at the outset.

2. The distributions

$$\mu^{(1)} \coloneqq B_1(\mu, (a_{1,1}, a_{1,2}))$$
$$\mu^{(2)} \coloneqq B_2(\mu, (a_{2,1}, a_{2,2}))$$

are computed.

3. The push-forward distribution

$$\mu^{(3)} \coloneqq (B_3)_{\#}(\mu, \mu^{(1)} \times \mu^{(2)})$$

is computed.

Here $\mu^{(1)} \times \mu^{(2)}$ is defined on the Cartesian product $\Sigma_{\mathcal{A}_{3,1}} \times \Sigma_{\mathcal{A}_{3,2}}$ with independent components $\mu^{(1)}$ and $\mu^{(2)}$. The notation $(B_3)_{\#}$ refers to the push-forward of the function $B_3(\mu, \cdot)$ over its second argument. The distribution $\mu^{(3)}$ is the output and represents belief about the principal QoI $Q_3(x)$.

The procedure in Example 5.4 can be formalised, but to keep the presentation and notation succinct, we leave this implicit:

**Definition 5.5** (Computation)**.** For a collection $(M_1, \ldots, M_n)$ of PNMs that are compatible with a pipeline $P$, the *computation* $P(M_1, \ldots, M_n)$ is defined as the PNM with information operator $A$ and belief operator $B$ that takes $\mu$ and $A(x) = a$ as input and returns the distribution $\mu^{(n)}$ as its output $B(\mu, a)$, obtained through the procedure outlined in Example 5.4.

That is, the *computation* $P(M_1, \ldots, M_n)$ is a PNM for the principal QoI $Q_n$. Note that this definition includes a classical numerical work-flow just as a PNM encompasses a standard numerical method.

## 5.2. Bayesian Computational Pipelines

Noting that $P(M_1, \ldots, M_n)$ is itself a PNM, there is a natural definition for when such a computation can be called Bayesian:

**Definition 5.6** (Bayesian Computation)**.** Denote by $(A, B)$ the information and belief operators associated with the computation $P(M_1, \ldots, M_n)$ and let $\{\mu^a\}_{a \in \mathcal{A}}$ be a disintegration of $\mu$ with respect to the information operator $A$. The computation $P(M_1, \ldots, M_n)$ is said to be *Bayesian* for the QoI $Q_n$ if

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

This is clearly an appealing property; the output of a Bayesian computation can be interpreted as a posterior distribution over the QoI $Q_n(x)$ given the prior $\mu$ and the information $A(x)$. Or, more informally, the "pipeline is lossless with information". However, at face value it seems difficult to verify whether a given computation $P(M_1, \ldots, M_n)$ is Bayesian, since it depends on both the individual PNMs $M_i$ and the pipeline $P$ that combines them. Our next aim is to establish verifiable sufficient conditions, for which we require another definition:

**Definition 5.7** (Dependence Graph)**.** The *dependence graph* of a pipeline $P$ is the directed acyclic graph $G(P)$ obtained by taking the pipeline $P$, removing the method nodes and replacing all $\square \to \blacksquare \to \square$ motifs with direct edges $\square \to \square$.

The dependency graph for Example 5.2 is shown in Figure 3.

For a computation $P(M_1, \ldots, M_n)$, each of the $J$ distinct nodes in $G(P)$ can be associated with a random variable $Y_j$ where either $Y_j = A_{k,l}(X)$ for some $k, l$, when the node is a source, or otherwise $Y_j = Q_k(X)$, for some $k$. Randomness here is understood to be due to $X \sim \mu$, so that the distribution of the $\{Y_j\}_{j=1}^J$ is a function of $\mu$. The convention used here is that the $Y_j$ are indexed according to a topological ordering on $G(P)$, which has the properties that (i) the source nodes correspond to indices $1, \ldots, I$, and (ii) the final random variable is $Y_J = Q_n(X)$.

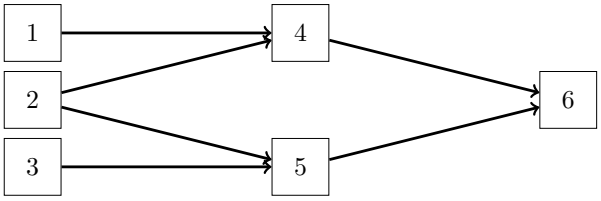



Figure 3: Dependence graph $G(P)$ corresponding to the pipeline $P$ in Figure 2. The nodes are indexed with a topological ordering (shown).

**Definition 5.8** (Coherence)**.** Consider a computation $P(M_1, \ldots, M_n)$. Denote by $\pi(j) \subseteq \{1, \ldots, j-1\}$ the parent set of node $j$ in the dependence graph $G(P)$. Then we say that $\mu \in \mathcal{P}_\mathcal{X}$ is *coherent* for the computation $P(M_1, \ldots, M_n)$ if the implied joint distribution of the random variables $Y_1, \ldots, Y_J$ satisfies:

$$Y_j \perp\!\!\!\perp Y_{\{1,\ldots,j-1\}\setminus\pi(j)} \mid Y_{\pi(j)}$$

for all $j = I+1, \ldots, J$.

Note that this is weaker than the Markov condition for directed acyclic graphs [see Lauritzen, 1991], since we do not insist that the variables represented by the source nodes are independent. It is emphasised that, for a given $\mu \in \mathcal{P}_\mathcal{X}$, the coherence condition can in general be checked and verified.

The following result provides sufficient and verifiable conditions which ensure that a computation composed of individual Bayesian PNMs is a Bayesian computation:

**Theorem 5.9.** *Let $M_1, \ldots, M_n$ be Bayesian PNMs and let $\mu \in \mathcal{P}_\mathcal{X}$ be coherent for the computation $P(M_1, \ldots, M_n)$. Then it holds that the computation $P(M_1, \ldots, M_n)$ is Bayesian for the QoI $Q_n$.*

Conversely, if non-Bayesian PNM are combined then the computation $P(M_1, \ldots, M_n)$ need not be Bayesian in general.

**Example 5.10** (Example 5.2, continued)**.** The random variables $Y_i$ in this example are:

$$Y_1 = \{X(t_i)\}_{i=1}^{m-1}, \quad Y_2 = X(t_m), \quad Y_3 = \{X(t_i)\}_{i=m+1}^{2m}, \quad Y_4 = \int_0^{0.5} X(t)\mathrm{d}t, \quad Y_5 = \int_{0.5}^{1} X(t)\mathrm{d}t.$$

From $G(P)$ in Figure 3, coherence condition in Definition 5.8 requires that the non-trivial conditional independences $Y_4 \perp\!\!\!\perp Y_3 \mid \{Y_1, Y_2\}$ and $Y_5 \perp\!\!\!\perp Y_1 \mid \{Y_2, Y_3\}$ hold. Thus the distribution $\mu$ is coherent for the computation $P(M_1, M_2, M_3)$ if and only if, for $X \sim \mu$, the associated information variables satisfy $\int_0^{0.5} X(t)\mathrm{d}t \perp\!\!\!\perp \{X(t_i)\}_{i=m+1}^{2m} | \{X(t_i)\}_{i=1}^{m}$ and $\int_{0.5}^{1} X(t)\mathrm{d}t \perp\!\!\!\perp \{X(t_i)\}_{i=1}^{m-1} | \{X(t_i)\}_{i=m}^{2m}$.

The distribution $\mu$ induced by the Weiner process on $x$ in Example 3.4 satisfies these conditions. Indeed, under $\mu$ the stochastic process $\{x(t) : t > t_m\}$ is conditionally independent of its history $\{x(t) : t < t_m\}$ given the current state $x(t_m)$. Thus for this choice of $\mu$, from Theorem 5.9 we have that $P(M_1, M_2, M_3)$ is Bayesian and parallel computation of $(a)$ and $(b)$ in Eq. (5.1) can be justified from a Bayesian statistical standpoint.

However, for the alternative of belief distributions induced by the Weiner process on $\partial^s x$, this condition is not satisfied and the computation $P(M_1, M_2, M_3)$ is not Bayesian. To turn this into a Bayesian procedure for these alternative belief distributions it would be required that $A_{1,2}(x)$ provides information about the derivatives $\partial^k x(t_m)$ for all orders $k \leq s$.

### 5.3. Monte Carlo Methods for Probabilistic Computation

The most direct approach to access $\mu^{(n)}$ is to sample from each Bayesian PNM and treat the output samples as inputs to subsequent PNM. This is sometimes known as *ancestral sampling* in the Bayesian network literature [e.g. Paige and Wood, 2016], and is illustrated in the following example:

**Example 5.11** (Ancestral Sampling for PNM)**.** For Example 5.2, ancestral sampling proceeds as follows:

1. Draw initial samples

$$q_1 \sim B_1(\mu, (a_{1,1}, a_{1,2}))$$
$$q_2 \sim B_2(\mu, (a_{2,1}, a_{2,2}))$$

2. Draw a final sample

$$q_3 \sim B_3(\mu, (q_1, q_2))$$

Then $q_3$ is a draw from $\mu^{(3)}$.

Ancestral sampling requires that PNM outputs can be sampled. Such sampling methods were discussed in Section 4.3. For a more general approach, sequential Monte Carlo methods can be used to propagate a collection of particles through the pipeline $P$, similar to work on SMC for general graphical models [Briers et al., 2005, Ihler and McAllester, 2009, Lienart et al., 2015, Lindsten et al., 2016, Paige and Wood, 2016]. For brevity we do not go into the details.

## 6. Numerical Experiments

In this final section of the paper we present two illustrative numerical experiments. The first is a nonlinear ODE and the second is a linear PDE. In each case we experiment with non-Gaussian belief distributions and, in doing so, go beyond previous work.

### 6.1. The Painlevé ODE

In this section numerical disintegration is applied to a nonlinear ODE based on Painlevé's first transcendental

$$x'' = x^2 - t$$

where $x = x(t)$ and $t \in \mathbb{R}^+$. This is equipped with the boundary conditions

$$x(0) = 0$$
$$t^{-1/2}x(t) \to 1 \text{ as } t \to \infty\,.$$

To permit computation, the right-boundary condition was relaxed by truncating the domain to $[0, 10]$ and using the modified condition $x(10) = \sqrt{10}$.

This system was recently presented in Farrell et al. [2015] as an example of a boundary value problem with multiple solutions. Two distinct solutions are known, and are illustrated in Figure 4. These model solutions were obtained using the deflation technique described in Farrell et al. [2015]. The spectrum plot represents the coefficients obtained when each solution is represented over a basis of normalised Chebyshev polynomials. As those polynomials are

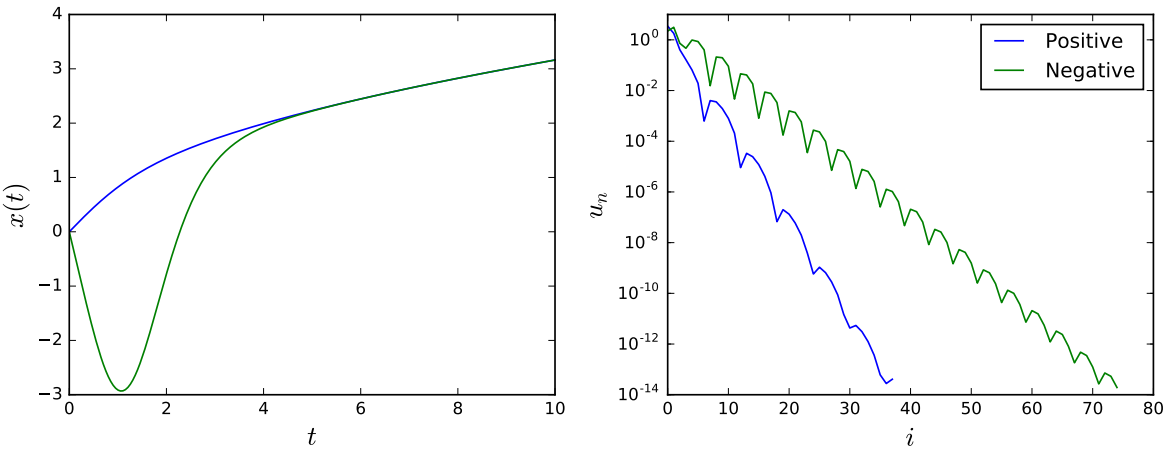



Figure 4: Two distinct solutions for the Painlevé ODE. The spectral plot on the right shows the true coefficients under the expression given in Eqn. (6.1), as determined by a model solver (`chebfun`).

orthonormal with respect to the $L_2$-inner-product, the slower decay for the negative solution compared to the positive solution is equivalent to the negative solution having a larger $L_2$-norm. This explains the preference that numerical solvers have for the positive solution in general, and also explains some of the results now presented.

Such systems for which multiple solutions exist have been studied before in the context of PNM, both in Chkrebtii et al. [2016] and in Cockayne et al. [2016]. It was noted in both papers that existence of multiple solutions can present a substantial challenge to classical numerical methods.

The Bayesian approach was taken, and a prior $\mu$ for this problem was defined by using a series expansion as detailed in Sections 2.5 and 4.2:

$$x^N(t) = x_0(t) + \sum_{i=0}^{\infty} u_i \phi_i(t). \tag{6.1}$$

The basis functions $\phi_i$ were normalised Chebyshev polynomials of the first kind $C_i(t')$. Recall that these are defined for $t' \in [-1, 1]$, necessitating the transformation $\phi_i(t) := C_i(\frac{1}{2}(t-5))$. Second derivatives for these polynomials are available in closed-form. For this illustration, both Gaussian and Cauchy priors were considered by taking $u_i := \gamma_i \xi_i$, where $\xi_i$ were taken to be either standard Gaussian or standard Cauchy. In each case $x_0(t) \equiv 0$. Full details of this construction are omitted and we refer the reader to Sullivan [2015] for further details.

In accordance with the exponential convergence rate for spectral methods when the solution to the system is a smooth function, the sequence of scale parameters was set to $\gamma_i = \alpha\beta^{-i}$, where $\alpha = 8$ and $\beta = 1.5$. These values were chosen by inspection of the true spectra (obtained with Matlab's "`chebfun`" package) in Figure 4 (right) to ensure that both solutions are in the support of the prior.

The information operator $A$ was defined by the choice of locations $\{t_j\}$, $j = 1, \ldots, N_t$, which determine the locations at which the posterior is constrained. Analysis for several values of $N_t$ was performed. In each case $t_1 = 0$, $t_{N_t} = 10$ and the remaining $t_j$ were equally spaced on

$[0, 10]$. To be explicit, the information operator was

$$A(x) = \begin{bmatrix} x''(t_1) - (x(t_1))^2 \\ \vdots \\ x''(t_{N_t}) - (x(t_{N_t}))^2 \\ x(0) \\ x(10) \end{bmatrix}$$

with the last two elements enforcing the boundary conditions. Thus our information was $a = [-t_1, \ldots, -t_{N_t}, 0, \sqrt{10}]$.

The Bayesian PNM output $B(\mu, a)$ was accessed via numerical disintegration (Section 4) with the first $N = 40$ terms of the series representation used. Specific details relating to the Monte Carlo method are reserved for the Electronic Supplement. Results for a selection of $\delta$, with $N_t = 15$, are shown in Figure 5. Note that a strong preference for the positive solution is expressed at the smallest $\delta$, with mass on both modes at larger $\delta$. For the Gaussian prior, some particles remain on the negative mode at the smallest $\delta$, while this was not so for the Cauchy prior. This reflects the fact that, in $n$ draws from a univariate Cauchy distribution, one element is likely to be significantly larger in magnitude than the others, which favours faster decay for the remaining elements.

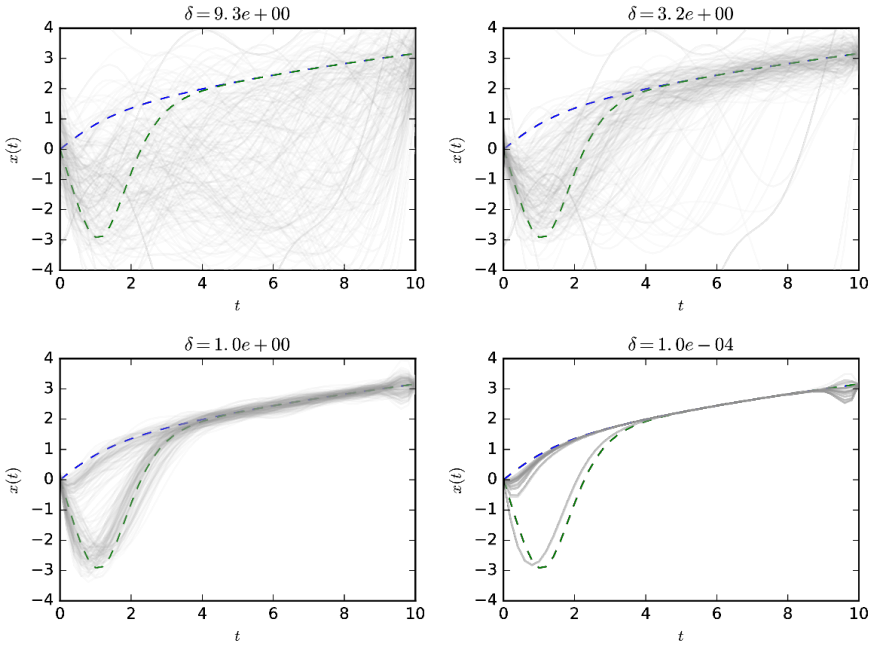
Furthermore the same sequence of scale parameters was used for both the Gaussian and Cauchy distribution, but the interpretation of this sequence differs. While for the Gaussian this can be thought of as the prior standard-deviation of each weight, the same interpretation is not valid for a Cauchy as the second moment is undefined. Since it is unclear how to make these parameters equivalent the same values have been used in each case, but this may further explain the difference.

Note that the particles near the negative mode in the Gaussian case appear to show significant bias away from the model negative solution. This is due to the fact that the amount of information supplied is relatively small.

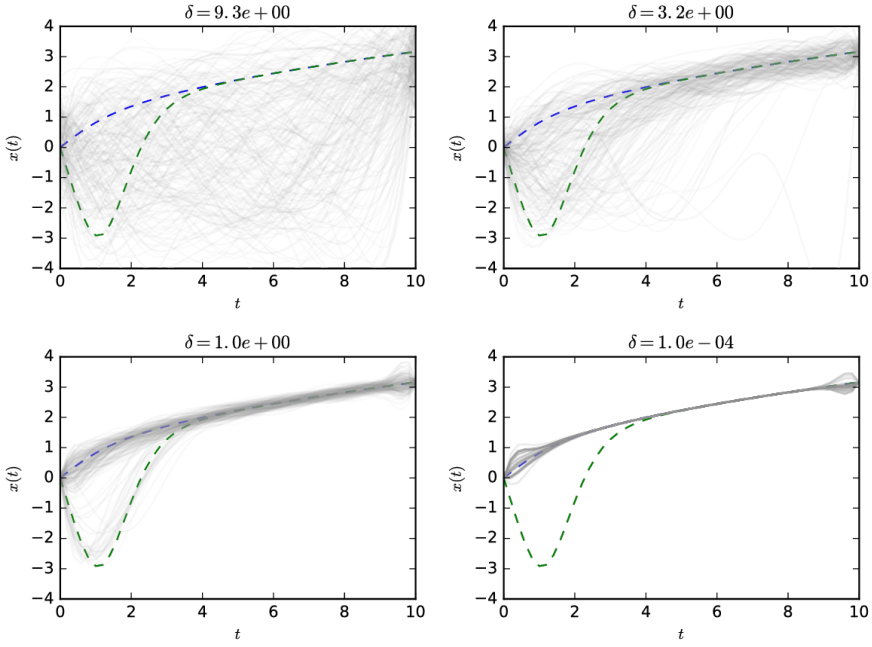
In Figure 6 the posterior distributions for first six coefficients $u_i$ at $N_t = 15$ and $\delta = 1$ are plotted. Strong multimodality is clear, as well as skewed correlation structure between the coefficients which verifies the practical importance of using effective Monte Carlo methods able to account for this structure, such as described in the Electronic Supplement. Illustration of such posteriors for smaller $\delta$ is difficult as the posteriors become extremely peaked.

Figure 7 displays convergence of the posterior distributions as $N_t$ is increased. Of particular interest is that for $N_t = 10$, the posterior distribution for the Gaussian prior becomes trimodal, while the Cauchy prior assigns no posterior mass to this third mode, again likely due to the aforementioned difference in interpretation of $\gamma$ for the Cauchy prior. Furthermore, for each prior the posterior mass settles on the positive solution to the system at $N_t = 20$. This is in accordance with the fact that this solution has smaller $L_2$-norm. This perhaps reflects the fact that, while in the limiting case both solutions should have an equal likelihood, the curvature of the likelihood at each mode may differ. Prior truncation may also be influential; in Figure 8 the log-likelihood of the negative mode increases at a slower rate than that of the positive mode. Thus, while in the setting of an infinite prior series neither solution should be preferred, in practice truncation might bias one solution over the other. Lastly, it is clear that the parameters $\alpha$ and $\beta$ may also have a significant effect on which solution is preferred.

Of further interest is how a preference for the negative solution could be encoded into a PNM. Owing to the flexible specification the information operator, there is considerable choice in this




(a) Gaussian Prior




(b) Cauchy Prior.

Figure 5: Posterior samples for the Painleve system at different values of $\delta$ for $N_t = 15$, with Gaussian and Cauchy priors. Blue and green dashed lines represent the positive and negative solutions determined by use of `chebfun`, while grey lines are posterior samples.

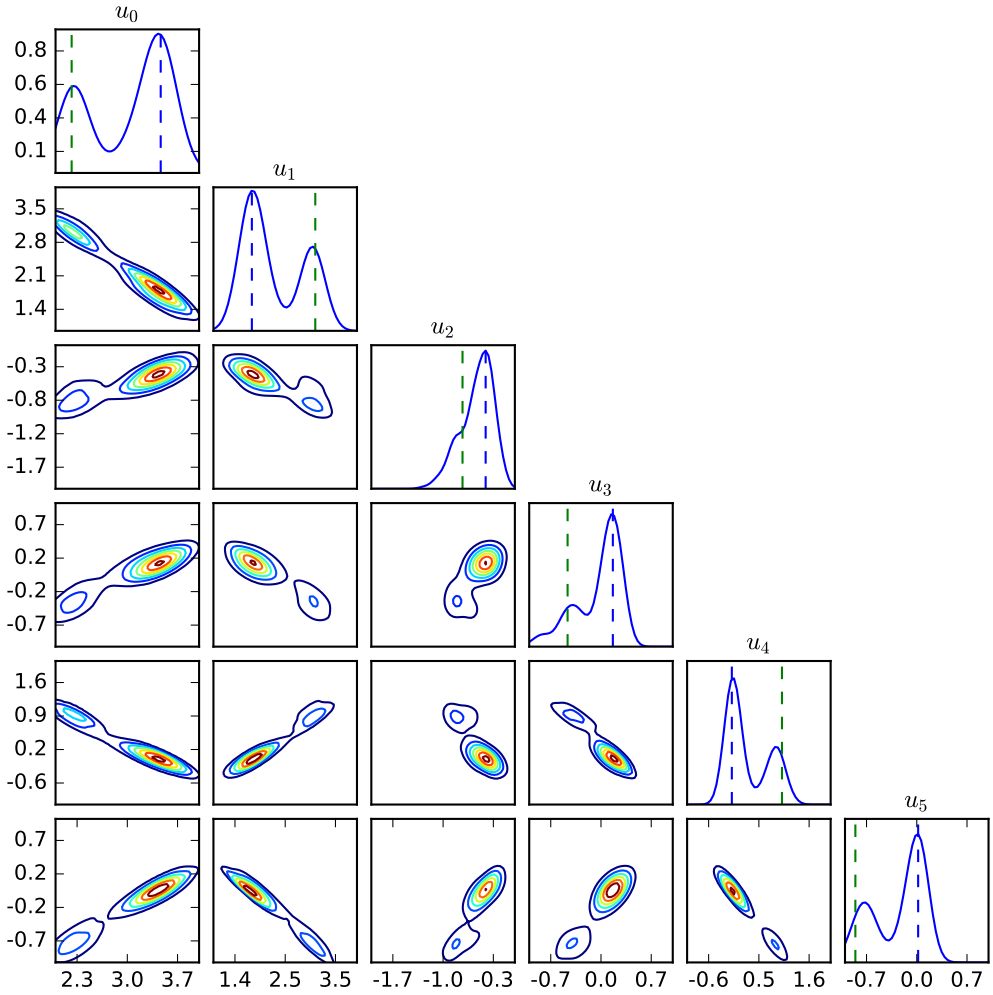



Figure 6: Posterior distributions for the first six coefficients of the expression for $x^N$ as given in Eqn. (6.1), at $\delta = 1$, $N_t = 15$. Vertical dashed lines on the diagonal plots indicate the value of the coefficients for the positive (blue) and negative (green) solutions generated by the model solver.

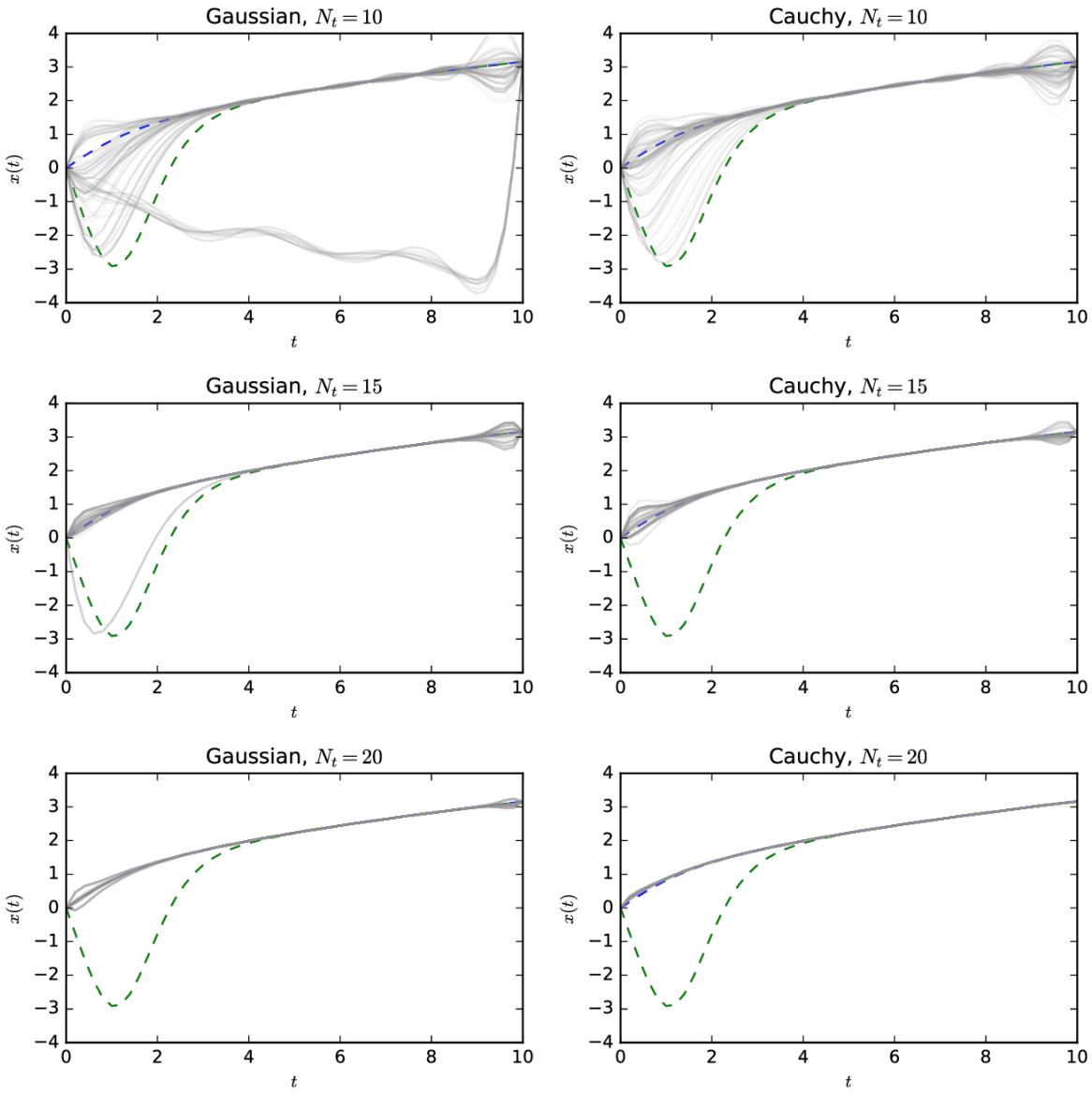



Figure 7: Convergence for the numerical disintegration scheme as $N_t$ is increased, for both Gaussian and Cauchy priors. In all cases $\delta = 10^{-4}$

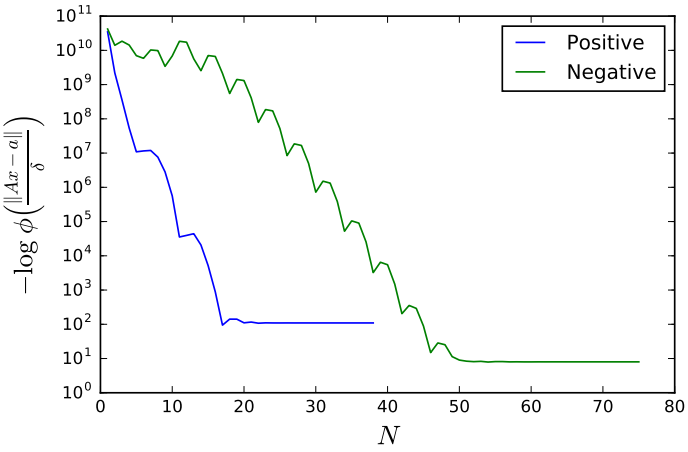



Figure 8: Negative-log-likelihoods for the point-estimates of coefficients for the postive and negative modes given by `chebfun`, as the truncation level is varied. The fact that the likelihood for the positive mode increases more rapidly than that of the negative mode suggests indicates that the sampler may have a preference for that mode over the other, though the level $N = 40$ has been selected in an attempt to minimise the impact.

matter. An elegant approach is the introduction of additional, inequality-based information

$$x'(0) \leq 0\,. \tag{6.2}$$

Such information can be difficult to incorporate in standard numerical algorithms, but is of interest in many physical problems [Kinderlehrer and Stampacchia, 2000]. For Bayesian PNM all that is required is an extension of the information operator to include $1[x'(0) \leq 0]$. Posterior distributions for the Gaussian prior at $N_t = 15$ are shown in Figure 9. Note that posterior mass has settled close to the negative solution. This highlights the simplicity with which Bayesian PNMs can encode a preference for a particular solution when a multiplicity of solutions exist.

## 6.2. Poisson Equation

Our second illustration is an instance of the Poisson equation, a linear PDE with mixed Dirichlet-Neumann boundary conditions:

$$-\nabla^2 x(\boldsymbol{t}) = 0 \qquad \boldsymbol{t} \in (0,1)^2 \tag{6.3}$$

$$x(\boldsymbol{t}) = t_1 \qquad t_1 \in [0,1] \qquad t_2 = 0 \tag{6.4}$$

$$x(\boldsymbol{t}) = 1 - t_1 \qquad t_1 \in [0,1] \qquad t_2 = 1 \tag{6.5}$$

$$\frac{\partial x}{\partial t_2} = 0 \qquad t_2 \in (0,1) \qquad t_1 = 0, 1 \tag{6.6}$$

This is a simple case of the models used for applications in electrical impedance tomography [Dunlop and Stuart, 2016]. A model solution to this system, generated with a finite-element method on a fine mesh, is shown in Figure 10.

As the spatial domain for this problem is two-dimensional, the basis used for specification of the belief distribution is more complex. Here tensor products of orthogonal polynomials have

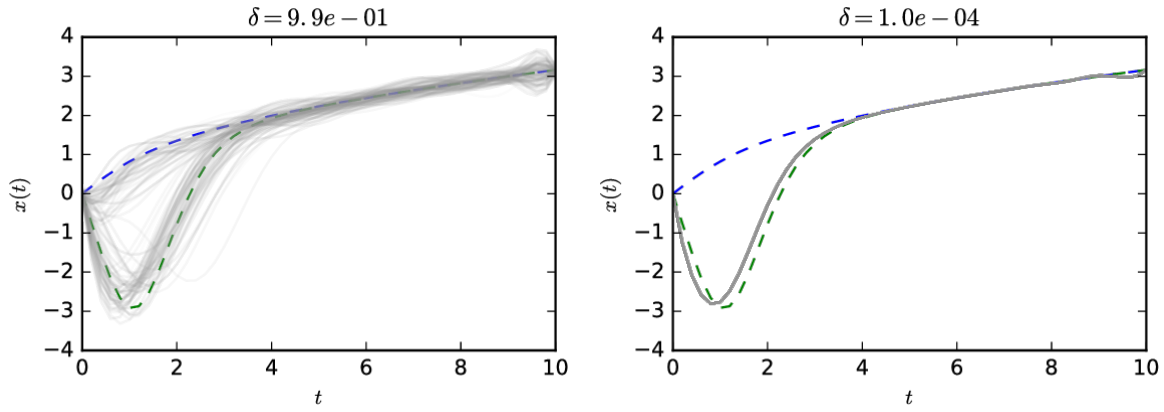



Figure 9: Posterior distribution at $N_t = 15$, based on a Gaussian prior, with the negative boundary condition given by Eqn. (6.2) enforced. On the right panel, the deviation from the true negative solution is explained by the relatively small number of design points used in calculating the solution.

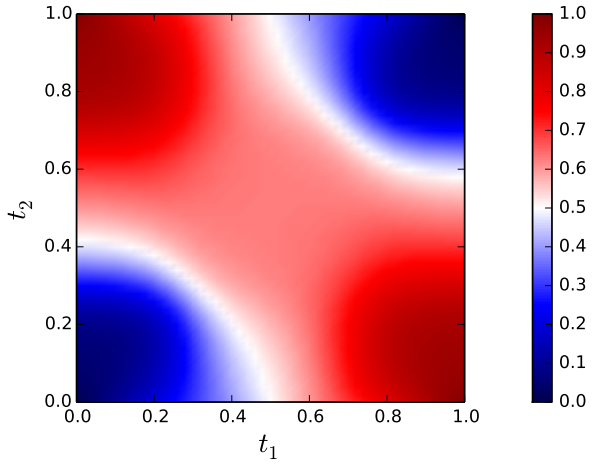



Figure 10: Model solution, generated by application of a finite element method based on a triangular mesh of $50 \times 50$ elements.

been used. As in the previous section, the underlying polynomials are chosen to be normalised Chebyshev polynomials of the first kind. To form the two-dimensional basis, independent bases $\mathbb{C}_1 = \{C_0(t_1'), \ldots, C_{N_C}(t_1')\}$ and $\mathbb{C}_2 = \{C_0(t_1'), \ldots, C_{N_C}(t_1')\}$ were constructed to some maximum polynomial order $N_C$. The basis used was then

$$\mathbb{C} \coloneqq \left\{C_i(t_1')C_j(t_2') : i + j \leq N_C, C_i \in \mathbb{C}_1, C_j \in \mathbb{C}_2\right\}$$

that is, the product of polynomials in $\mathbb{C}_1$ and $\mathbb{C}_2$ up to a maximum total polynomial order $N_C$. Once again, the underlying Chebyshev polynomials being defined for $t' \in [-1, 1]$ necessitates a transformation of the spatial domain. To this end, we set $\phi_i(\boldsymbol{t}) \coloneqq C_j(2t_1-1)C_k(2t_2-1)$ for some $j, k$. Prior specification then follows the formulation given in Section 2.5, where the remaining parameters are chosen to be $x_0 \equiv 1$, and $\gamma_i = \alpha(i+1)^{-2}$. The random variables $\xi$ were taken to be either Gaussian or Cauchy, and the polynomial basis was truncated to $N = 45$ terms, corresponding to a maximum polynomial degree of $N_C = 8$. For both priors the parameter $\alpha$ was set to $\alpha = 1$. Note that closed-form expressions are available for analysis under the Gaussian prior [Cockayne et al., 2016] but were not exploited, to keep the comparison as fair as possible.

The information operator was defined by a set of locations $\boldsymbol{t}_i \in [0, 1]^2$, $i = 0, \ldots, N_t$, where depending on the location either the interior condition or one of the boundary conditions was enforced. Denote by $\left\{\boldsymbol{t}_i^I\right\}$ the set of interior points, $\left\{\boldsymbol{t}_j^D\right\}$ the set of Dirichlet boundary points and $\left\{\boldsymbol{t}_k^N\right\}$ the set of Neumann boundary points, where $i = 1, \ldots, N_I$, $j = 1, \ldots, N_D$ and $k = 1, \ldots, N_N$, with $N_t = N_I + N_D + N_N$. Then, the information operator is given by the concatenation of the conditions defined above:

$$A(x) = [A^I(x)^\top, A^D(x)^\top, A^N(x)^\top]^\top$$

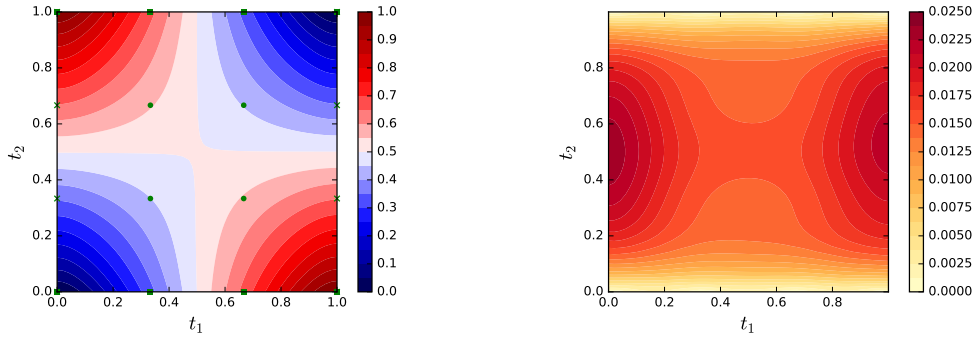
$$A^I(x) = \begin{bmatrix} -\nabla^2 x(\boldsymbol{t}_1^I) \\ \vdots \\ -\nabla^2 x(\boldsymbol{t}_{N_I}^I) \end{bmatrix} \quad A^D(x) = \begin{bmatrix} x(\boldsymbol{t}_1^D) \\ \vdots \\ x(\boldsymbol{t}_{N_D}^D) \end{bmatrix} \quad A^N(x) = \begin{bmatrix} \frac{\partial}{\partial t_1} x(\boldsymbol{t}_1^N) \\ \vdots \\ \frac{\partial}{\partial t_1} x(\boldsymbol{t}_{N_N}^N) \end{bmatrix}$$

while $a$ is formed as usual by evaluating the appropriate right-hand-sides of the system at the chosen locations.

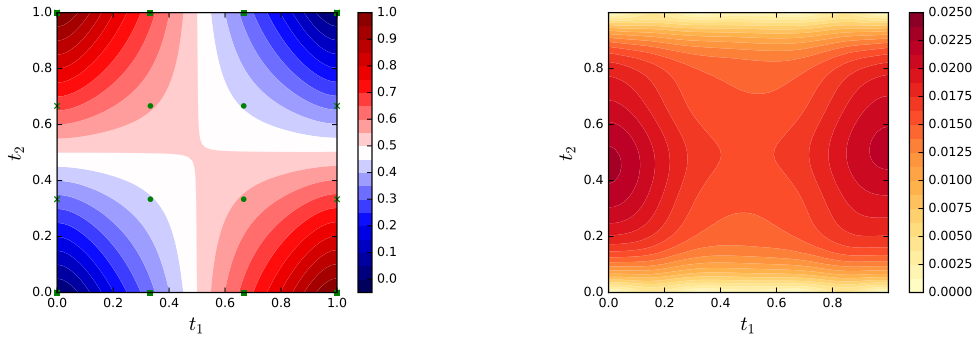
The Bayesian PNM output was accessed by numerical disintegration and details of the Monte Carlo method are reserved for the Electronic Supplement. In Figure 11 the mean and pointwise standard-deviations of the posterior distributions are plotted for Gaussian and Cauchy priors with $N_t = 16$. There is little qualitative difference between the posterior distributions for the Gaussian and Cauchy priors. The mean functions match closely to the mean function from the model solution, as given in Figure 10. The posterior variance is lowest near to the Dirichlet boundaries where the solution is known, and peaks where the Neumann condition is imposed. This is to be expected, as evaluations of the Neumann boundary condition provide little information about the solution itself.

Although no model spectral solution has been generated, analysis of the spectra is still pertinent. In the Gaussian case, conjugacy means that the posterior distribution over the coefficients is also Gaussian. In Figure 12 the posterior distributions over these coefficients are plotted and are seen to be Gaussian (up to Monte Carlo error), though the correlation structure between coefficients is non-trivial, as can be seen in the joint distribution between $u_0$ and $u_3$.

Lastly, in Figure 13 convergence of the posterior distribution is plotted as the number of design points is varied, for $N_t = 16, 25, 36$. In each case a Gaussian prior was used. As expected, the variance in the posterior distribution is seen to decrease as the number of design points is




(a) Gaussian prior




(b) Cauchy prior

Figure 11: Posterior distributions for the solution $x$ of the Poisson equation, with $N_t = 16$ and different choices of prior distribution. Design points for the interior, Dirichlet and Neumann boundary conditions are indicated by green dots, green squares and green crosses, respectively.

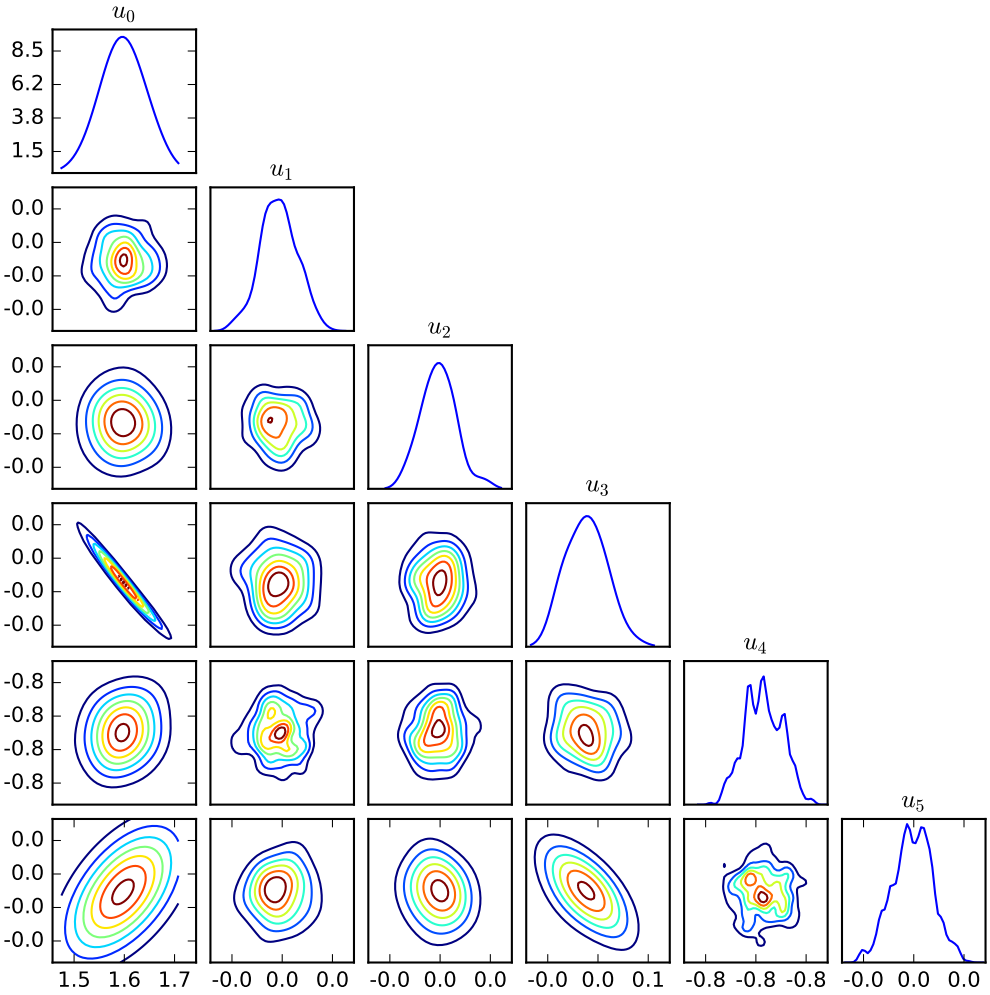
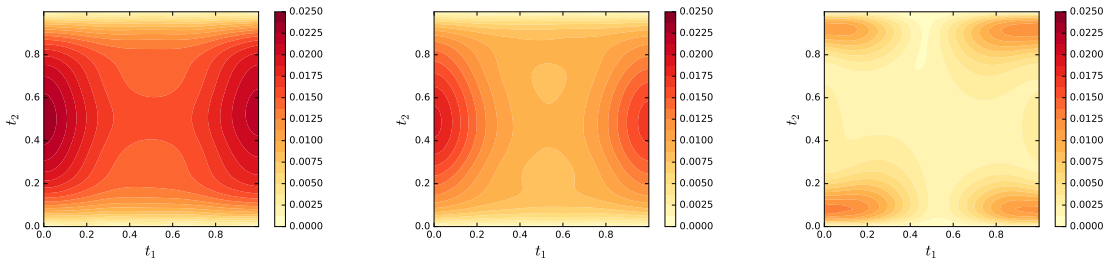



Figure 12: Posterior distributions for the first six coefficients of the spectrum for the posterior distribution on $x$ in the Poisson equation, obtained with Monte Carlo methods and numerical disintegration, based on $\delta = 0.0008$, $N_t = 16$.




(a) $N_t = 16$ (b) $N_t = 25$ (c) $N_t = 36$

Figure 13: Heat map of the point-wise variance for the solution $x$ to the Poisson equation as $N_t$ is varied. In each case a Gaussian prior has been used.

increased. At $N_t = 36$, the shape of the region of highest uncertainty changes markedly, with the most uncertain region lying between the Dirichlet boundary and the first evaluation points on the Neumann boundary. This is likely due to the fact that the number of evaluation points is approaching the size of the polynomial basis. When the number of points equals the size of the basis the system is completely determined for a linear model such as this, so that no uncertainty remains. Thus, as this numer of points is approached thus the number $N$ of basis functions ought also to be increased.

This completes our illustration of the properties and behaviour of Bayesian PNM.

# 7. Discussion

This paper has established statistical foundations for PNMs and investigated the Bayesian case in detail. Through connection to Bayesian inverse problems [Stuart, 2010], we have established when Bayesian PNM can be well-defined and when the output can be considered meaningful. The presentation touched on several important issues and a brief discussion of the most salient points is now provided.

**Bayesian vs Non-Bayesian PNMs** The decision to focus on Bayesian PNMs was motivated by the realisation that the output of a pipeline of PNMs can only be guaranteed to admit a valid Bayesian interpretation if the constituent PNMs are each Bayesian and the prior distribution is coherent. Indeed, Theorem 5.9 demonstrated that prior coherence can be established at a local level, essentially via a local Markov condition, so that Bayesian PNMs provide a extensible modelling framework as required to solve more challenging numerical tasks. These results support a research strategy that focuses on Bayesian PNMs, so that error can be propagated in a manner that is meaningful.

On the other hand, there are pragmatic reasons why either approximations to Bayesian PNMs, or indeed, non-Bayesian PNMs might be useful. The predominant reason would be to circumvent the off-line computational costs that are often associated with Bayesian PNMs, such as the use of numerical disintegration developed in this research. Recent research efforts, such as Schober et al. [2014, 2016] and Kersting and Hennig [2016] for the solution of ODEs, have aimed for computational costs that are competitive with classical methods, at the expense of fully Bayesian estimation for the solution of the ODE. Such methods are of interest as non-Bayesian PNMs, but their role in pipelines of PNMs is unclear. Our contribution serves to make this explicit.

**Computational Cost** The present research focused on the more fundamental cost of access to the information $A(x)$, rather than e.g. the additional CPU time required to process the PNM output. Indeed, numerical disintegration constituted the predominant computational cost in the applications that were reported. However, we stress that in many challenging applications gated by discretisation error, such as occur with climate models, the fundamental cost of the information $A(x)$ will be dominant. Furthermore, the Monte Carlo methods that were employed for numerical disintegration admit substantial improvements [e.g. in a similar vein to Botev and Kroese, 2012, Koskela et al., 2016]. The objective of this paper was to establish statistical foundations that will permit the development of more sophisticated and efficient Bayesian PNMs.

**Prior Elicitation** Throughout this work we assumed that a belief distribution $\mu$ was provided. The question of *whose* belief is represented in $\mu$ has been discussed by several authors and a chronology is included in the Electronic Supplement. Of these perspectives we mention in particular Hennig et al. [2015], wherein $\mu$ is the belief of an agent that "we get to design". This offers a connection to frequentist statistics, in that an agent can be designed to ensure favourable frequentist properties hold.

A robust statistics perspective is also relevant to prior elicitation. One such approach would be to consider a generalised Bayes risk (Eq. (3.1)) wherein the state variable $X$ used for assessment is assumed to be drawn from a distribution $\tilde{\mu}$ that differs from the $\mu$. This offers an opportunity to derive Bayesian PNMs that are robust to certain forms of prior mis-specification. This direction was not considered in the present paper, but has been pursued in the ACA literature, where parameters of a Gaussian distribution were considered as unknown and a numerical method was designed to perform well over a range of parameter values [see Chapter IV, Section 4 of Ritter, 2000].

In general, the specification of prior distributions for robust inference on an infinite-dimensional (non-parametric) state space can be problematic. The consistency and robustness of Bayesian inference procedures — particularly with respect to perturbations of the prior such as those arising from numerical approximations — in such settings is a subtle topic, with both positive [Castillo and Nickl, 2014, Doob, 1949, Kleijn and van der Vaart, 2012, Le Cam, 1953] and negative [Diaconis and Freedman, 1986, Freedman, 1963, Owhadi et al., 2015] results depending upon fine topological and geometric details.

In the context of computational pipelines, the challenge of eliciting a coherent prior is closely connected to the challenge of eliciting a single unified prior based on the conflicting input of multiple experts [French, 2011, Albert et al., 2012]. Future work will seek to deploy Bayesian computation to real-world numerical work-flows.

**Consistent Estimation** The present paper focused on foundations. Further methodological work will be required to establish sufficient conditions for when $B(\mu, A_n(x^\dagger))$ collapses to an atom on a single element $q^\dagger = Q(x^\dagger)$ representing the data-generating QoI in the limit as the amount of information, $n$, is increased. There are two questions here; (i) when is $q^\dagger$ identifiable from the given information, and (ii) at what rate does $B(\mu, A_n(x^\dagger))$ concentrate on $q^\dagger$.

**Generalisation and Extensions** Two directions are highlighted for extension of this work. First, note that in this paper the observation map $A : \mathcal{X} \to \mathcal{A}$ was treated as a deterministic object. However, in some applications there is auxiliary randomness in the acquisition of information. For our integration example, nodes $t_i$ might arise as random samples from a reference

distribution on $[0, 1]$. Or, observations $x(t_i)$ themselves might occur with measurement error, for example due to finite precision arithmetic. Then a more elaborate model $A\colon \mathcal{X} \times \Omega \to \mathcal{A}$ would be required, where $\Omega$ is a probability space that injects randomness into the information operator. This is the setting of, for instance, randomised quasi-Monte Carlo methods. Future work will extend the framework of PNMs to include randomised information operators of this kind.

As a second direction, recall that in an adaptive algorithm the choice of the information is made in an iterative procedure that is informed by the information observed up to that point. For the canonical illustration in Example 3.4 and its generalisations discussed there, it can be proven that adaptive algorithms do not out-perform non-adaptive algorithms in average case error [Lee and Wasilkowski, 1986]. However, outside this setting adaptation can be beneficial. One example of adaptation is the mesh refinement procedure used in Conrad et al. [2016]. Future work will extend the framework of PNMs to include adaptive information operators.

**Connection with Probabilistic Programming** The central goal of Probabilistic Programming (PP) is to automate statistical computation, through symbolic representation of statistical objects and operations on those objects. The formalism of pipelines as graphical models presented in this work can be compared to similar efforts to establish PP languages [Goodman et al., 2012]. For instance, a method node in a pipeline can be related to a *monad* aggregating several distributions into a single output distribution [Ścibior et al., 2015]. An important challenge in PP is the automation of the operation of conditioning, or disintegration. Numerical disintegration and extensions thereof might be of independent interest to this field [e.g. extending Wood et al., 2014].

**Acknowledgements** The research forms part of the SAMSI working group on Probabilistic Numerics co-led by CJO and TJS. CJO was supported by the Australian Research Council (ARC) Centre of Excellence for Mathematical and Statistical Frontiers. TJS was supported by the Excellence Initiative of the German Research Foundation (DFG) through the Free University of Berlin. MG was supported by the Engineering and Physical Sciences (EPSRC) grants EP/J016934/1, EP/K034154/1, an EPSRC Mathematical Sciences Established Career Research Fellowship and a Lloyds Register Foundation grant for Programme on Data-Centric Engineering.

The authors are grateful to Amazon for the provision of AWS credits which were used to generate the results of the experiments in Section 6, and to the authors of the Eigen and Eigency libraries which made implementation of code for generating those results immeasurably simpler.

# A. Proofs

*Proof of Theorem 3.3.* The following observation will be required; the joint density of $X$ and $A = A(X)$ can be expressed in two ways:

$$\delta(A(x))(\mathrm{d}a)\mu(\mathrm{d}x) = \mu^a(\mathrm{d}x)A_{\#}\mu(\mathrm{d}a) \tag{A.1}$$

which holds almost everywhere from the definition of a disintegration $\{\mu^a\}_{a\in\mathcal{A}}$. Here $\delta(a')(\cdot) \in \mathcal{P}_{\mathcal{A}}$ denotes an atomic distribution on $a' \in \mathcal{A}$. To see this, note that we can either first simulate $X$ and then extract $A(X)$ (left hand side) or we can first simulate $A$ and then simulate $X$ conditional on $A$ (right hand side).

Fix $a \in \mathcal{A}$ and denote the random variables $Q_A^a = Q(X) - c_{\mu,A}(a)$ that are induced according to $X \sim \mu^a$, indexed by the choice of information operator $A$. Denote by $\tilde{Q}_A^a$ an independent copy of $Q_A^a$ generated from $\tilde{X} \sim \mu^a$. Then we have

$$\begin{aligned} Q(X) - Q(\tilde{X}) &= (Q(X) - c_{\mu,A}(a)) - (Q(\tilde{X}) - c_{\mu,A}(a)) \\ &= Q_A^a - \tilde{Q}_A^a \end{aligned}$$

which will be used later. From Theorem 3.2 the posterior mean of $Q(X)$ is $c_{\mu,A}(a)$ and thus $\mathbb{E}[Q_A^a] = 0$.

The Bayes risk for a Bayesian PNM $M = (A, B)$ can be expressed as:

$$\begin{aligned} R(\mu, M) &= \int r(x, B(\mu, A(x)))\mu(\mathrm{d}x) \\ &= \iint \|Q(x) - q\|_{\mathcal{Q}}^2 Q_{\#}\mu^{A(x)}(\mathrm{d}q)\mu(\mathrm{d}x) \quad \text{(since } M \text{ Bayesian)} \\ &= \iiint \|Q(x) - q\|_{\mathcal{Q}}^2 Q_{\#}\mu^{a}(\mathrm{d}q)\delta(A(x))(\mathrm{d}a)\mu(\mathrm{d}x) \\ &= \iiint \|Q(x) - Q(x')\|_{\mathcal{Q}}^2 \mu^a(\mathrm{d}x')\mu^a(\mathrm{d}x)A_{\#}\mu(\mathrm{d}a) \quad \text{(from Eq. (A.1))} \\ &= \int \mathbb{E}[\|Q_A^a - \tilde{Q}_A^a\|_{\mathcal{Q}}^2]A_{\#}\mu(\mathrm{d}a) \\ &= \int \mathbb{E}[\|Q_A^a\|_{\mathcal{Q}}^2 - 2\langle Q_A^a, \tilde{Q}_A^a\rangle_{\mathcal{Q}} + \|\tilde{Q}_A^a\|_{\mathcal{Q}}^2]A_{\#}\mu(\mathrm{d}a) \\ &= 2\int \mathbb{E}[\|Q_A^a\|_{\mathcal{Q}}^2]A_{\#}\mu(\mathrm{d}a) \quad \text{(since } \mathbb{E}[Q_A^a] = 0 \text{ and } Q_A^a,\ \tilde{Q}_A^a \text{ are i.i.d.).} \end{aligned}$$

Note that our integrability assumption justifies the interchange of integrals from Fubini's theorem. On the other hand, the average error associated with the Bayes act can be expressed as:

$$\begin{aligned} e(\mu, A, c_{\mu,A}) &= \int \|Q(x) - c_{\mu,A}(A(x))\|_{\mathcal{Q}}^2 \mu(\mathrm{d}x) \\ &= \iint \|Q(x) - c_{\mu,A}(a)\|_{\mathcal{Q}}^2 \delta(A(x))(\mathrm{d}a)\mu(\mathrm{d}x) \\ &= \iint \|Q(x) - c_{\mu,A}(a)\|_{\mathcal{Q}}^2 \mu^a(\mathrm{d}x)A_{\#}\mu(\mathrm{d}a) \quad \text{(from Eq. (A.1))} \\ &= \int \mathbb{E}[\|Q_A^a\|_{\mathcal{Q}}^2]A_{\#}\mu(\mathrm{d}a) \end{aligned}$$

where we have again used integrability to exploit Fubini's theorem. Thus $R(\mu, M) = 2e(\mu.A.c_{\mu,A})$, which in turn implies that the optimal information $A_\mu$ for Bayesian PNM and $A_\mu^*$ for ACA are identical. □

*Proof of Theorem 4.3.* Fix $f \in \mathcal{F}$ and $a \in \mathcal{A}$. Then:

$$\begin{aligned}
\mu_\delta^a(f) &= \frac{1}{Z_\delta^a} \int f(x)\phi\left(\frac{\|A(x)-a\|_{\mathcal{A}}}{\delta}\right)\mu(\mathrm{d}x) \\
&= \frac{1}{Z_\delta^a} \iint f(x)\phi\left(\frac{\|\tilde{a}-a\|_{\mathcal{A}}}{\delta}\right)\mu^{\tilde{a}}(\mathrm{d}x)A_{\#}\mu(\mathrm{d}\tilde{a}) \quad \text{(from Eq. (A.1))} \\
&= \frac{1}{Z_\delta^a} \int \phi\left(\frac{\|\tilde{a}-a\|_{\mathcal{A}}}{\delta}\right)\mu^{\tilde{a}}(f)A_{\#}\mu(\mathrm{d}\tilde{a}) \\
&= \int \mu^{\tilde{a}}(f)A_{\#}\mu_\delta^a(\mathrm{d}\tilde{a}), \quad \text{since} \quad A_{\#}\mu_\delta^a(\mathrm{d}\tilde{a}) = \frac{1}{Z_\delta^a}\phi\left(\frac{\|\tilde{a}-a\|_{\mathcal{A}}}{\delta}\right)A_{\#}\mu(\mathrm{d}\tilde{a}).
\end{aligned}$$

Thus

$$\begin{aligned}
|\mu_\delta^a(f) - \mu^a(f)| &= \left|\int [\mu^{\tilde{a}}(f) - \mu^a(f)]A_{\#}\mu_\delta^a(\mathrm{d}\tilde{a})\right| \\
&\leq \left|C_\mu \|f\|_{\mathcal{F}} \int \|\tilde{a}-a\|_{\mathcal{A}}^\alpha A_{\#}\mu_\delta^a(\mathrm{d}\tilde{a})\right| \quad \text{(Assumption 4.2).} \qquad \text{(A.2)}
\end{aligned}$$

Now consider the random variable

$$R \coloneqq \frac{\|A(X)-a\|_{\mathcal{A}}}{\delta}$$

induced from $X \sim \mu$. The existence of a continuous and positive density $p_A$ implies that $R$ also admits a density on $[0,\infty)$, denoted $p_{R,\delta}$. The fact that $p_A$ is uniform on an infinitesimal neighbourhood of $a$ implies that

$$p_{R.\delta}(r) = \frac{2\pi^{n/2}}{\Gamma(\frac{n}{2})}(\delta r)^{n-1}(p_A(a) + o(1)) \qquad \text{(A.3)}$$

valid for $r \ll 1$, with $p_{R,\delta}(r)$ being proportional to the surface area of a hypersphere of radius $\delta r$ in $\mathcal{A}$.

The final integral in Eq. (A.2) can then be evaluated:

$$\begin{aligned}
\int \|\tilde{a}-a\|_{\mathcal{A}}^\alpha A_{\#}\mu_\delta^a(\mathrm{d}\tilde{a}) &= \frac{\int \|\tilde{a}-a\|_{\mathcal{A}}^\alpha \phi\left(\frac{\|\tilde{a}-a\|_{\mathcal{A}}}{\delta}\right)A_{\#}\mu(\mathrm{d}\tilde{a})}{\int \phi\left(\frac{\|\tilde{a}-a\|_{\mathcal{A}}}{\delta}\right)A_{\#}\mu(\mathrm{d}\tilde{a})} \\
&= \delta^\alpha \frac{\int r^\alpha \phi(r)p_{R,\delta}(r)\mathrm{d}r}{\int \phi(r)p_{R,\delta}(r)\mathrm{d}r} \quad \text{(change of variables).} \qquad \text{(A.4)}
\end{aligned}$$

The ratio of integrals in Eq. (A.4) can be simplified from Eq. (A.3):

$$\frac{\int r^\alpha \phi(r)p_{R,\delta}(r)\mathrm{d}r}{\int \phi(r)p_{R,\delta}(r)\mathrm{d}r} \xrightarrow[\delta\downarrow 0]{} \frac{\int r^{\alpha+n-1}\phi(r)\mathrm{d}r}{\int r^{n-1}\phi(r)\mathrm{d}r} = \frac{C_\phi^\alpha}{C_\phi^0} \quad (< \infty \text{ from Assumption 4.1}).$$

Thus, for $\delta$ sufficiently small, Eq. (A.4) can be bounded above by $\delta^\alpha(1+C_\phi)$ where $C_\phi \coloneqq C_\phi^\alpha/C_\phi^0$ and "1" is in this case an arbitrary positive constant. This establishes the upper bound

$$|\mu_\delta^a(f) - \mu^a(f)| \leq C_\mu(1+C_\phi)\|f\|_{\mathcal{F}}\,\delta^\alpha$$

for $\delta$ sufficiently small and completes the proof. □

*Proof of Theorem 5.9.* To reduce the notation, suppose that the random variables $Y_1, \dots, Y_J$ admit a joint density $p(y_1, \dots, y_J)$, However, we emphasise that existence of a density is not required for the proof to hold. To further reduce notation, denote $y_{a:b} = (y_a, \dots, y_b)$.

The output of the computation $P(M_1, \dots, M_n)$ was defined algorithmically in Definition 5.5 and illustrated in Example 5.4. Our aim is to show that this algorithmic output coincides with the distribution $(Q_n)_{\#}\mu^a$ on $\mathcal{Q}_n$, which is identified in the present notation with $p(y_J|y_{1:I})$.

For $j \in \{I+1, \dots, J\}$, the coherence condition on $Y_1, \dots, Y_J$ translates into the present notation as $p(y_j|y_{1:j-1}) = p(y_j|y_{\pi(j)})$. This allows us to deduce that:

$$\begin{aligned} p(y_J|y_{1:I}) &= \int \cdots \int p(y_{I+1:J}|y_{1:I}) \mathrm{d}y_{I+1:J-1} \\ &= \int \cdots \int \prod_{j=I+1}^{J} p(y_j|y_{1:j-1}) \mathrm{d}y_{I+1:J-1} \\ &= \int \cdots \int \prod_{j=I+1}^{J} p(y_j|y_{\pi(j)}) \mathrm{d}y_{I+1:J-1}. \end{aligned}$$

The right hand side is recognised as the output of the computation $P(M_1, \dots, M_n)$, as defined in Definition 5.5. This completes the proof. □

*University of Warwick, j.cockayne@warwick.ac.uk
†University of Technology Sydney, chris.oates@ncl.ac.uk
‡Free University of Berlin and Zuse Institute Berlin, sullivan@zib.de
§Imperial College London and Alan Turing Institute, m.girolami@imperial.ac.uk

# Electronic Supplement

to the paper *Bayesian Probabilistic Numerical Methods*

## A. Philosophical Status of the Belief Distribution

The aim of this section is to discuss in detail the semantic status of the belief distribution $\mu$ in a probabilistic numerical method (PNM). In Section A.1 we survey historical work on this topic, while in Section A.2 more recent literature is covered. Then in Section A.3 we highlight some philosophical objections and their counter-arguments.

### A.1. Historical Precedent

The use of probabilistic and statistical methods to model a deterministic mathematical object can be traced back to Poincaré [1912], who used a stochastic model to construct interpolation formulae. In brief, Poincaré formulated a polynomial

$$f(x) = a_0 + a_1 x + \cdots + a_m x^m$$

whose coefficients $a_i$ were modelled as independent Gaussian random variables. Thus Poincaré in effect constructed a Gaussian measure over the Hilbert space with basis $\{1, x, \ldots, x^m\}$. This pre-empted the later extensions of Kimeldorf and Wahba [1970a,b] and others, which associated spline interpolation formulae to the means of Gaussian measures over Hilbert spaces.

The first explicit statistical model for numerical error (of which we are aware) was in the literature on rounding error in the numerical solution of ordinary differential equations (ODE), as summarised in Hull and Swenson [1966]. Therein it was supposed that rounding, by which we mean representation of a real number

$$x = 0.a_1 a_2 a_3 a_4 \ldots \quad \in [0, 1]$$

in a truncated form

$$\hat{x} = 0.a_1 a_2 a_3 a_4 \ldots a_n,$$

is such that the error $e = x - \hat{x}$ can be reasonably modelled by a uniform random variable on $[-5 \times 10^{-(n+1)}, 5 \times 10^{-(n+1)}]$. This implies a distribution $\mu$ over the unknown value of $x$ given $\hat{x}$. The contribution of Hull and Swenson [1966] and others was to replace the last digit $a_n$, in each stored number that arises in the numerical solution of an ODE, with a uniformly chosen element of $\{0, \ldots, 9\}$. This performs approximate propagation of the *numerical uncertainty* due to rounding error through further computation and, in their case, induces a distribution over the solution space of the ODE. Note that this work focused on rounding error, rather than the (time) discretisation error that is intrinsic to numerical ODE solvers; this could reflect the limited precision arithmetic that was available from the computer hardware of the period.

Larkin [1972] was an important historical paper for PNMs, being the first to set out the modern statistical agenda for PNMs:

> In any particular problem situation we are given certain specific properties of the solution, e.g. a finite number of ordinate or derivative values at fixed abscissae. If we can assume no more than this basic information we can conclude only that our required solution is a member of that class of functions which possesses the given properties - a tautology which is unlikely to appeal to an experimental scientist! Clearly, we need to be given, or to assume, extra information in order to make more definite statements about the required function.
>
> Typically, we shall assume general properties, such as continuity or non-negativity of the solution and/or its derivatives, and use the given specific properties in order to assist in making a selection from the class $K$ of all functions possessing the assumed general properties. We shall choose $K$ either to be a Hilbert space or to be simply related to one.

This description defines a set $K$ of permissible functions, rather than an explicit distribution over $K$, but it is clear that Larkin envisaged numerical analysis as an instance of statistical estimation:

> In the present approach, an *a priori* localisation is achieved effectively by making an assumption about the relative likelihoods of elements of the Hilbert space of possible candidates for the solution to the original problem. Among other things, this permits, at least in principle, the derivation of joint probability density functions for 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) without any extra information about the norm of the function in question.

Later, Diaconis [1988] re-iterated this argument for the construction of $K$ more explicitly, considering numerical integration of the function

$$f(x) = \exp\left\{\cosh\left(\frac{x + x^2 + \cos(x)}{3 + \sin(x^3)}\right)\right\} .$$

over the unit interval. In particular, Diaconis asked:

> "What does it mean to 'know' a function?" The formula says some things (e.g. $f$ is smooth, positive and bounded by 20 on $[0, 1]$) but there are many other facts about $f$ that we don't know (e.g. is $f$ monotone, unimodal or convex?)

This argument was provided as justification for belief distributions that encode certain basic features, such as the smoothness of the integrand. The belief distributions that were then considered in Diaconis' paper were Gaussian distributions on $K$. Diaconis famously observed that some classical numerical methods are Bayes rules for particular Gaussian prior distributions.

The arguments of these papers are intrinsic to modern PNMs. However, the associated theoretical analysis of computation under finite information has proceeded outside of statistics, in the applied mathematical literature, where it is usually presented without a statistical context. That research is reviewed next.

## A.2. Contemporary Outlook

The mathematical foundations of computation based on finite information are established in the field of *information-based complexity* (IBC). The monograph of Traub et al. [1988] presents the foundations of IBC. In brief, the starting point for IBC is the mantra that

> To compute fast you need to compute with partial information ($\sim$ Houman Owhadi, SIAM UQ 2016)

This motivates the search for optimal approximations based on finite information, in either the worst-case or average-case sense of optimal. The particular development of PNMs that we presented in the main text is naturally aligned to *average-case analysis* (ACA) and we focus on that literature in what follows.

Among the earliest work on ACA, Sul′din [1959, 1960] studied numerical integration and $L_2$ function approximation in the average case setting where $\mu$ was induced from the Weiner process, with a focus on optimal linear methods. Later, Sacks and Ylvisaker [1970] moved from analysis with fixed $\mu$ to analysis over a class of $\mu$ defined by the smoothness properties of their covariance kernels. At the same time Kimeldorf and Wahba [1970a,b] established optimality properties of splines in reproducing kernel Hilbert spaces in the ACA context. Kadane and Wasilkowski [1985], Diaconis [1988] discussed the connection between ACA and Bayesian statistics. A general framework for ACA was formalised in the IBC monograph of Traub et al. [1988], while Ritter [2000] provides a more recent account.

Game theoretic arguments have recently been explored in Owhadi [2015a], who argued that the optimal prior for probabilistic meshless methods [Cockayne et al., 2016] is a particular Gaussian measure under a game theoretic framework where the energy norm is the loss function. This provides one route to the specification of default or *objective* priors for PNMs which deserves further exploration in general.

The question of "whose" belief is captured in $\mu$ was addressed in Hennig et al. [2015], where it was argued that the prior information in $\mu$ represents that of a hypothetical agent (numerical analyst) which

> [...] we are allowed to design ($\sim$ Michael Osborne, personal correspondence, 2016).

This draws connections with frequentist statistics, where the properties of a statistical methods are assessed in a repeated experiment sense rather than conditional on the data at hand. This mirrors our decision theoretic assessment of PNMs presented in the main text.

## A.3. Paradise Lost?

Typical numerical algorithms contain several different sources of discretisation error. Consider the solution of the wave equation: A standard finite element method involves both spatial and temporal discretisations, a series of numerical quadrature problems, as well as the use of finite precision arithmetic for all numerical calculations. Yet, decades of numerical analysis have led to highly optimised computer codes such that these methods can be routinely used. To develop PNM for solution of the wave equation, which accounts for each separate source of discretisation error, is it required to unpick and reconstruct such established numerical algorithms? This would be an unattractive prospect that would detract from further research into PNMs.

Our view is that there is a choice for which numerical errors to model. In practice the PNMs implemented for this work were run on floating point precision machines, yet we did not model rounding error in their output. This was because, in our examples, floating point error can be shown to be insignificant compared to discretisation error and so we chose not to model it. Indeed, an implicit assumption in this paper was that, given the inputs $\mu$ and $a$, the output $B(\mu, a)$ can be obtained without further approximation error. This assumption was necessary to circumvent an infinite regress in which the PNMs cannot themselves be implemented and must be analysed with further PNMs. However, since for many PNMs the exact output $B(\mu, a)$

is unavailable in a closed form, we in practice adopted the weaker requirement that the output $B(\mu, a)$ can be sampled from, which allows us to explore the PNM output through sampling schemes whose computational cost and theoretical analysis can both be decoupled from the problem of interest.

## B. Existence of Non-Randomised Bayes Rule

In this section we recall an argument for the general existence of non-randomised Bayes rules, that was stated without proof in the main text. Sufficient conditions for Fubini's theorem to hold are assumed in what follows.

**Proposition B.1.** *Let $\mathfrak{B}(A)$ be non-empty. Then $\mathfrak{B}(A)$ contains a classical numerical method of the form $B(\mu, a) = \delta \circ c(a)$ for some $c\colon \mathcal{A} \to \mathcal{Q}$.*

*Proof.* Let $\mathfrak{C}$ be the set of belief update operators of the classical form $B(\mu, a) = \delta \circ c(a)$. Without loss of generality, suppose there exists a belief update operator $B^* \in \mathfrak{B}(A)$ but that $B^* \notin \mathfrak{C}$. The Bayes risk satisfies $R(\mu, (A, B^*)) \leq R(\mu, (A, \delta \circ c))$ for all $\delta \circ c \in \mathfrak{C}$ since $B^* \in \mathfrak{B}(A)$. On the other hand, $B^*$ can be characterised as a non-atomic distribution $\pi$ over the elements of $\mathfrak{C}$. Its risk can be computed as:

$$\begin{aligned} R(\mu, (A, B^*)) &= \int r(Q(x), B^*(\mu, A(x)))\mu(\mathrm{d}x) \\ &= \iint L(Q(x), c(A(x)))\pi(\mathrm{d}c)\mu(\mathrm{d}x) \\ &= \iint L(Q(x), c(A(x)))\mu(\mathrm{d}x)\pi(\mathrm{d}c) \quad \text{(Fubini)} \\ &= \int R(\mu, (A, \delta \circ c))\pi(\mathrm{d}c). \end{aligned}$$

Thus if we had $R(\mu, (A, B^*)) < R(\mu, (A, \delta \circ c))$ for all $\delta \circ c \in \mathfrak{C}$ we would have a contradiction. It follows that $\mathfrak{B}(A) \cap \mathfrak{C}$ is non-empty. This completes the proof. □

## C. Monte Carlo Methods for Numerical Disintegration

In this section, Monte Carlo methods for sampling from the distribution $\mu_\delta^a$ (or $\mu_{\delta,N}^a$; the $N$ subscript will be suppressed to reduce notation in the sequel) are considered. The Monte Carlo approximation of $\mu_\delta^a$ is, in effect, a problem in rare event simulation as most of the mass of $\mu_\delta^a$ will be confined to a set $S$ such that $\mu(S)$ is small. Rare events pose some difficulties for classical Monte Carlo, as an enormous number of draws can be required to study the rare event of interest.

In the literature there are two major solutions proposed. *Importance sampling* [Robert and Casella, 2013] samples from a modified process, under which the event of interest is more likely, then re-weights these samples to compensate for the adjustment. Conversely, in *splitting* [Botev and Kroese, 2012] trajectories of the process are constructed in a genetic fashion, by retaining and duplicating those which approach the events of interest and discarding others. Splitting is closely related to SMC [Cérou et al., 2012] and Feynman–Kac models [Del Moral, 2004].

The splitting approach is described in the following section, while in Section C.3 a parallel tempering (PT) algorithm is described. In spirit these approaches are similar in that they employ a tempering approach to ease sampling the relaxed posterior distribution for a small

value of $\delta$. The SMC method employs a particle approximation to accomplish this, while the PT algorithm uses coupled Markov chains.

## C.1. Sequential Monte Carlo Algorithms for Numerical Disintegration

Let $\{\delta_i\}_{i=1}^n$ be such that $\delta_0 = \infty$ and $\delta_i > \delta_{i+1} > 0$ for all $i < n-1$. Furthermore let $\{K_i\}_{i=1}^n$ be some set of Markov transition kernels that leave $\mu^a_{\delta_i}$ invariant, for which $K_i(\cdot, S)$ is measurable for all $S \in \Sigma_{\mathcal{X}}$ and $K_i(x, \cdot)$ is an element of $\mathcal{P}_{\mathcal{X}}$ for all $x \in \mathcal{X}$. Then our SMC for numerical disintegration (SMC-ND) algorithm, based on $P$ particles, is given in Algorithm 1. Here we have used Discrete$(\{x_j\}_{j=1}^P; \{w_j\}_{j=1}^P)$ to denote the discrete distribution which puts mass $w_j / \sum_k w_k$ on the state $x_j \in \mathcal{X}$.

Sample $x_j^0 \sim \mu$ for $j = 1, \ldots, P$ [Initialise]
**for** $i = 1, \ldots, n$ **do**
  Sample $x_j^{i-1} \sim K_i(x_j^{i-1}, \cdot)$ for $j = 1, \ldots, P$ [Move]
  Set $w_j^i \leftarrow \frac{\phi\left(\delta_i^{-1} \|A(x_j^{i-1}) - a\|_{\mathcal{A}}\right)}{\phi\left(\delta_{i-1}^{-1} \|A(x_j^{i-1}) - a\|_{\mathcal{A}}\right)}$ for $j = 1, \ldots, P$ [Re-weight]
  Sample $x_j^i \sim$ Discrete$(\{x_j^{i-1}\}_{j=1}^P; \{w_j^i\}_{j=1}^P)$ for $j = 1, \ldots, P$ [Re-sample]
**end**

**Algorithm 1:** Sequential Monte Carlo for Numerical Disintegration (SMC-ND).

The output of the SMC-ND algorithm is an empirical approximation

$$\mu^a_{\delta_n, P} = \frac{1}{P} \sum_{j=1}^{P} \delta(x_j^n)$$

to $\mu^a_{\delta_n}$ based on a population of $P$ particles $\{x_j^n\}_{j=1}^P$. There is substantial room to extend and improve the SMC-ND algorithm based on the wide body of literature available on this subject [e.g. Doucet et al., 2001, Del Moral et al., 2006, Beskos et al., 2016, Ellam et al., 2016], but we defer all such improvements for future work. Our aim in the remainder is to establish the approximation properties of the SMC-ND output. This will be based on theoretical results in Del Moral et al. [2006].

**Assumption C.1.** $\phi > 0$ on $\mathbb{R}_+$.

**Assumption C.2.** For all $i = 0, \ldots, n-1$ and all $x, y \in \mathcal{X}$, it holds that $K_{i+1}(x, \cdot) \ll K_{i+1}(y, \cdot)$. Furthermore there exist constants $\epsilon_i > 0$ such that the Radon–Nikodým derivative

$$\frac{\mathrm{d}K_{i+1}(x, \cdot)}{\mathrm{d}K_{i+1}(y, \cdot)} \geq \epsilon_i.$$

Assumption C.1 ensures that Algorithm 1 is well-defined, else it can happen that all particles are assigned zero weight and re-sampling will fail. However, the result that we obtain in Theorem C.3 below can also be established in the special case of an indicator function $\phi(r) = 1[r < 1]$. The details for this variation of the results on Theorem C.3 are also included in the sequel.

The interpretation of Assumption C.2 is that, for fixed $i$, transition kernels do not allocate arbitrarily large or small amounts of mass to different areas of the state space, as a function of their first argument. This poses a constraint on the choice of Markov kernels for the SMC-ND algorithm.

**Theorem C.3.** *For all $\delta \in \{\delta_i\}_{i=0}^n$ and fixed $p \geq 1$ it holds that*

$$\mathbb{E}\left(\left[\mu_{\delta,P}^a(f) - \mu_\delta^a(f)\right]^p\right)^{\frac{1}{p}} \leq \frac{C_{p,\delta}\,\|f\|_{\mathcal{F}}}{\sqrt{P}}$$

*for some constant $C_{p,\delta}$ independent of $P$ but dependent on $\{\delta_i\}_{i=0}^n$, $p$ and $\{\epsilon_i\}_{i=0}^{n-1}$.*

The proof of Theorem C.3 is presented next. Note that the established bound is independent of $\delta \in \{\delta_i\}_{i=0}^n$; this is therefore a uniform convergence result. The assumptions and the conclusion of Theorem C.3 can be weakened in several directions, as discussed in detail in [Del Moral et al., 2006]. Development of SMC methods in the context of high-dimensional and infinite-dimensional state spaces has also been considered in Beskos et al. [2014, 2015].

## C.2. Proof of Theorem C.3

In this section we establish the uniform convergence of the SMC-ND algorithm as claimed in Theorem C.3. This relies on a powerful technical result from Del Moral [2004], whose context is now established.

### C.2.1. Feynman–Kac Models

Let $(E_i, \mathcal{E}_i)$ for $i = 0, \ldots, n$ be a collection of measurable spaces. Let $\eta_0$ be a measure on $E_0$ and let $\Gamma_i$ index a collection of Markov transition kernels from $E_{i-1}$ to $E_i$. Let $G_i \colon E_i \to (0, 1]$ be a collection of functions, which are referred to as *potentials*. The triplet $(\eta_0, G_i, \Gamma_i)$ is associated with *Feynman–Kac* measures $\eta_i$ on $E_i$ defined as, for bounded and measurable functions $f_i$ on $E_i$;

$$\eta_i(f_i) = \frac{\gamma_i(f_i)}{\gamma_i(1)}$$
$$\gamma_i(f_i) = \mathbb{E}_{\eta_0}\left[f_i(X^i)\prod_{j=0}^{i-1} G_j(X^j)\right]$$

where the expectation is taken with respect to the Markov process $X^i$ defined by $X^0 \sim \eta_0$ and $X^i | X^{i-1} \sim \Gamma_i(X^{i-1}, \cdot)$.

The Feynman–Kac measures can be associated with a (non-unique) *McKean interpretation* of the form $\eta_{i+1} = \eta_i \Lambda_{i+1,\eta_i}$ where the $\Lambda_{i+1,\eta}$ are a collection of Markov transitions for which the following compatibility condition holds:

$$\eta\Lambda_{i+1,\eta} = \frac{G_i}{\eta(G_i)}\eta\Gamma_{i+1}$$

Then the $\eta_i$ can be interpreted as the $i$th step marginal distribution of the non-homogeneous Markov chain defined by $X^0 \sim \eta_0$ and $X^{i+1} | X^i \sim \Lambda_{i+1,\eta_i}(X^i, \cdot)$. The corresponding $P$-particle model is defined on $E_i^P = E_i \times \cdots \times E_i$ and has

$$\mathbf{X}^0 \sim \eta_0^P$$
$$\mathbb{P}(\mathbf{X}^i \in \mathrm{d}\mathbf{x}^i | \mathbf{X}^i) = \prod_{j=1}^{P} \Lambda_{i,\eta_{i-1}^P}(X_j^{i-1}, \mathrm{d}x_j^i)$$

where $\eta_i^P = \frac{1}{P}\sum_{j=1}^P \delta(X_j^i)$ is an empirical (random) measure on $E_i$. The SMC-ND algorithm can be cast as an instance of such a $P$-particle model, as is made clear later.

The result that we require from Del Moral [2004] is given next. Denote by $\mathrm{Osc}_1(E_i)$ the set of measurable functions $f_i$ on $E_i$ for which $\sup\{|f_i(x^i) - f_i(y^i)| \ : \ x^i, y^i \in E_i\} \leq 1$.

**Theorem** (Theorem 7.4.4 in Del Moral [2004])**.** *Suppose that:*

$(G)$ *There exist* $\epsilon_i^G \in (0,1]$ *such that* $G_i(x^i) \geq \epsilon_i^G G_i(y^i) > 0$ *for all* $x^i, y^i \in E_i$.

$(M_1)$ *There exist* $\epsilon_i^\Gamma \in (0,1)$ *such that* $\Gamma_{i+1}(x^i, \cdot) \geq \epsilon_i^\Gamma \Gamma_{i+1}(y^i, \cdot)$ *for all* $x^i, y^i \in E_i$.

*Then for* $p \geq 1$ *and any valid McKean interpretation* $\Lambda_{i,\eta}$*, the associated* $P$*-particle model* $\eta_i^P$ *satisfies the uniform (in* $i$*) bound*

$$\sup_{0\leq i\leq n} \sup_{f_i \in Osc_1(E_i)} \sqrt{P}\mathbb{E}[|\eta_i^P(f_i) - \eta_i(f_i)|^p]^{1/p} \leq C_p$$

*for some constant* $C_p$ *independent of* $P$ *but dependent on* $\{\epsilon_i^G\}_{i=0}^n$ *and* $\{\epsilon_i^\Gamma\}_{i=0}^{n-1}$.

The actual statement in Del Moral [2004] contains a more general version of $(M_1)$ and a more explicit decomposition of the constant $C_p$; however the simpler version presented here is sufficient for the purposes of the present paper.

### C.2.2. Case A: Positive Function $\phi(r) > 0$

First we prove Theorem C.3 as it is stated. Later the assumption of $\phi > 0$ will be relaxed.

**SMC-ND as a Feynman–Kac Model** The aim here is to demonstrate that the SMC-ND algorithm fits into the framework of Section C.2.1 for a specific McKean interpretation. This connection will then be used to establish uniform convergence for the SMC-ND algorithm as a consequence of Theorem 7.4.4 in Del Moral [2004].

For the state spaces we associate each $E_i = \mathcal{X}$ and $\mathcal{E}_i = \Sigma_\mathcal{X}$. For the potentials we associate

$$G_i(x^i) = \frac{\phi\left(\frac{1}{\delta_{i+1}}\|A(x^i) - a\|_\mathcal{A}\right)}{\phi\left(\frac{1}{\delta_i}\|A(x^i) - a\|_\mathcal{A}\right)}$$

which clearly does not vanish and takes values in $(0,1]$ since $\delta_i > \delta_{i+1}$ and $\phi$ is decreasing. For the Markov transitions we associate $\Gamma_{i+1}$ with $K_{i+1}$.

The Feynman–Kac measures associated with the SMC-ND algorithm can be cast as a non-homogeneous Markov chain with transitions $\Lambda_{i+1,\eta}$. Here $\Lambda_{i+1,\eta_i}$ acts on the current measure $\eta_i$ on $\mathcal{X}$ by first propagating as $\eta_i K_{i+1}$ and then "warping" this measure with the potential $G_i$; i.e.

$$\eta\Lambda_{i+1,\eta} = \frac{G_i}{\eta(G_i)}\eta\Gamma_{i+1}.$$

This demonstrates that the SMC-ND algorithm is the $P$-particle model corresponding to the McKean interpretation $\Lambda_{i+1,\eta}$ of the Feynman–Kac triplet $(\eta_0, G_i, \Gamma_i)$. Thus the SMC-ND algorithm can be studied in the context of Section C.2.1, which we report next.

Note that it is common in applications of SMC to perform the "Re-sample" step before the "Move" step - our choice of order was required for the McKean framework that is the basis of the theoretical results in Del Moral et al. [2006]. It is known in the SMC "folk lore" that the order of these steps can be interchanged.

**Proof of Uniform Convergence Result for SMC-ND** It remains to verify the hypotheses of Theorem 7.4.4 in Del Moral [2004]. Condition $(G)$ is satisfied if and only if

$$\phi\left(\frac{1}{\delta_{i+1}}\|A(x^i) - a\|_{\mathcal{A}}\right)$$

is bounded below, since

$$\phi\left(\frac{1}{\delta_i}\|A(x^i) - a\|_{\mathcal{A}}\right)$$

is bounded above by 1. Since $\phi$ is continuous, decreasing and satisfies $\phi > 0$ (Assumption C.1), it suffices to show that its argument $\frac{1}{\delta_{i+1}}\|A(x) - a\|_{\mathcal{A}}$ is upper-bounded. This is the content of Assumption 4.6 in the main text, which shows that

$$\begin{aligned}\frac{1}{\delta_i}\|A(x) - a\|_{\mathcal{A}} &\leq \frac{1}{\delta_i}\sup_{x\in\mathcal{X}}\|A(x)\|_{\mathcal{A}} + \|a\|_{\mathcal{A}} \\ &=: \frac{1}{\epsilon_i^G} \;<\; \infty.\end{aligned}$$

Condition $(M_1)$ requires that

$$\Gamma_{i+1}(x^i, S) \geq \epsilon_i^{\Gamma}\Gamma_{i+1}(y^i, S)$$

for all $x^i, y^i \in E_i$ and $S \in \mathcal{E}_{i+1}$. From construction this is equivalent to

$$K_{i+1}(x^i, S) \geq \epsilon_i^{\Gamma}K_{i+1}(y^i, S)$$

for all $x^i, y^i \in \mathcal{X}$ and $S \in \Sigma_{\mathcal{X}}$. This is the content of Assumption C.2.

Thus we have established the hypotheses of Theorem 7.4.4 in Del Moral [2004] for the SMC-ND algorithm. Theorem C.3 is a re-statement of this result. For the statement of the result we used the $\|f\|_{\mathcal{F}}$ norm, based on the fact that (from Assumption 4.7) $\|f_i\|_{\mathrm{Osc}(E_i)} \leq 2\|f\|_{\infty} \leq 2C_{\mathcal{F}}\|f\|_{\mathcal{F}}$.

### C.2.3. Case B: Indicator Function $\phi(r) = 1[r < 1]$

The previous analysis required that $\phi > 0$ on $\mathbb{R}_+$. However, the most basic choice for $\phi$ is the indicator function $\phi(r) = 1[r < 1]$ which does not satisfy this condition. The case of an indicator function demands special attention, since Algorithm 1 can fail in this case if all particles are assigned zero weight. If this occurs, then we just define $\mu^a_{\delta,P}(f) = 0$. To be specific, the SMC-ND algorithm associated to the indicator function $\phi$ for approximation of the integral $\mu^a_{\delta}(f)$ is stated as Algorithm 2 next.

Let $\mathcal{X}^a_{\delta} = \{x \in \mathcal{X} : \|A(x) - a\|_{\mathcal{A}} < \delta\}$. If there is some iteration $i$ at which, after applying the kernel $K_i$ to each particle, no particle lies within $\mathcal{X}^a_{\delta_i}$, the algorithm fails. As a result it is critical to ensure that the distance between successive $\delta_i$ is small so that the probability of failure is controlled. This requirement is made formal next. To establish the approximation properties of the random measure $\mu^a_{\delta_n,P}$, two assumptions are required. These are intended to replace Assumptions C.1, C.2 and Assumption 4.6 from the main text:

**Assumption C.4.** For all $i = 0, \ldots, n-1$ and all $x^i \in \mathcal{X}^a_{\delta_i}$, it holds that $K_{i+1}(x^i, \mathcal{X}^a_{\delta_{i+1}}) > 0$.

**Assumption C.5.** For all $i = 0, \ldots, n-1$ and all $x^i, y^i \in \mathcal{X}^a_{\delta_i}$, $K_{i+1}(x^i, \cdot) \ll K_{i+1}(y^i, \cdot)$. Furthermore there exist constants $\epsilon_i > 0$ such that the Radon–Nikodým derivative

$$\frac{\mathrm{d}K_{i+1}(x^i, \cdot)}{\mathrm{d}K_{i+1}(y^i, \cdot)} \geq \epsilon_i.$$

Sample $x_j^0 \sim \mu$ for $j = 1, \ldots, P$ [Initialise]
**for** $i = 1, \ldots, n$ **do**
  Sample $x_j^i \sim K_i(x_j^{i-1}, \cdot)$ for $j = 1, \ldots, P$ [Sample]
  $E_i \leftarrow \{x_j^i : x_j^i \in \mathcal{X}_{\delta_i}^a\}$
  **if** $E_i = \emptyset$ **then**
    Return $\mu_{\delta,P}^a(f) \leftarrow 0$
  **end**
  **for** $j = 1, \ldots, P$ **do**
    **if** $x_j^i \notin E_i$ **then**
      $x_j^i \sim \text{Uniform}(E_i)$ [Re-sample]
    **end**
  **end**
**end**
Return $\mu_{\delta,P}^a(f) \leftarrow \frac{1}{P}\sum_{j=1}^P f(x_j^n)$.

**Algorithm 2:** Sequential Monte Carlo for Numerical Disintegration (SMC-ND), for the case where $\phi(r) = 1[r < 1]$.

Assumption C.4 requires that the probability of reaching $\mathcal{X}_{\delta_{i+1}}^a$ when starting in $\mathcal{X}_{\delta_i}^a$ and applying the transition kernel $K_{i+1}$, is bounded away from zero. Assumption C.5 ensures that, for fixed $i$, transition kernels do not allocate arbitrarily large or small amounts of mass to different areas of the state space, as a function of their first argument.

**Theorem C.6.** *For the alternative situation of an indicator function, it holds that for all* $\delta \in \{\delta_i\}_{i=0}^n$ *and fixed* $p \geq 1$*,*

$$\mathbb{E}\left(\left[\mu_{\delta,P}^a(f) - \mu_\delta^a(f)\right]^p\right)^{\frac{1}{p}} \leq \frac{C_p \|f\|_{\mathcal{F}}}{\sqrt{P}}$$

*for some constant* $C_p$ *independent of* $P$ *but dependent on* $p$ *and* $\{\epsilon_i\}_{i=0}^{n-1}$*.*

Cérou et al. [2012] proposed an algorithm similar to the one herein but focussed on approximation of the *probability* of a rare event rather than sampling from the rare event itself. In particular the theoretical results provided are in terms of these probabilities rather than how well the measure restricted to the rare event is approximated. Furthermore, many of the results therein focused upon an idealised version of the problem, in which it was assumed that the intermediate restricted measures can be sampled directly; this avoids the issues with vanishing potentials indicated in Del Moral [2004]. A similar algorithm was discussed in Ścibior et al. [2015] but was not shown to be theoretically sound.

The remainder of this Section establishes Theorem C.6.

**SMC-ND as a Feynman–Kac Model** The aim here is to demonstrate that Algorithm 2 fits into the framework of Section C.2.1 for a specific McKean interpretation. This is analogous to the proof of Theorem C.3.

A technical complication is that the potentials $G_i$ must take values in $(0, 1]$, which precludes the "obvious" choice of $E_i = \mathcal{X}$ and $G_i(x^i)$ as indicator functions for the sets $\mathcal{X}_{\delta_i}^a$. Instead, we associate $E_i = \mathcal{X}_{\delta_i}^a$ and $\mathcal{E}_i$ with the corresponding restriction of $\Sigma_{\mathcal{X}}$. For the potentials we then take $G_i(x^i) = 1$ for all $x_i \in E_i$, which clearly does not vanish and takes values in $(0, 1]$. For the

Markov transitions $\Gamma_{i+1}$ from $E_i$ to $E_{i+1}$ we consider

$$\Gamma_{i+1}(x^i, \mathrm{d}x^{i+1}) \propto K_{i+1}(x^i, x^{i+1})$$

which is the restriction of $K_{i+1}$ to $E_{i+1}$. For the latter to be well-defined it is required that the normalisation constant

$$\int_{E_{i+1}} K_{i+1}(x^i, x^{i+1})\mathrm{d}x^{i+1} > 0$$

for all $x^i \in E_i$, so that there is a positive probability of reaching $E_{i+1}$ from $E_i$. This is the content of Assumption C.4.

The Feynman–Kac measures associated with Algorithm 2 can be cast as a non-homogeneous Markov chain with transitions $\Lambda_{i+1,\eta}$. Here $\Lambda_{i+1,\eta_i}$ acts on the current measure $\eta_i$ on $E_i$ by first propagating as $\eta_i K_{i+1}$ and then restricting this measure to $E_{i+1}$. This procedure is seen to be identical to the Markov transition $\Gamma_{i+1}$ defined above and, since the potentials $G_i \equiv 1$, it follows that

$$\begin{aligned}\eta\Lambda_{i+1,\eta} &= \eta\Gamma_{i+1} \\ &= \frac{G_i}{\eta(G_i)}\eta\Gamma_{i+1}.\end{aligned}$$

This demonstrates that Algorithm 2 is the $P$-particle model corresponding to the McKean interpretation $\Lambda_{i+1,\eta}$ of the Feynman–Kac triplet $(\eta_0, G_i, \Gamma_i)$. Thus the SMC-ND algorithm can be studied in the context of Section C.2.1, which we report next.

**Proof of Uniform Convergence Result for SMC-ND** It remains to verify the hypotheses of Theorem 7.4.4 in Del Moral [2004]. Condition $(G)$ is satisfied with no further assumption, since $G_i \equiv 1$ and we can take $\epsilon_i^G = 1$. Condition $(M_1)$ requires that

$$\Gamma_{i+1}(x^i, S) \geq \epsilon_i^\Gamma \Gamma_{i+1}(y^i, S)$$

for all $x^i, y^i \in E_i$ and $S \in \mathcal{E}_{i+1}$. From construction this is equivalent to

$$K_{i+1}(x^i, S) \geq \epsilon_i^\Gamma K_{i+1}(y^i, S)$$

for all $x^i, y^i \in E_i$ and $S \in \mathcal{E}_{i+1}$. This is the content of Assumption C.5.

Thus we have established the hypotheses of Theorem 7.4.4 in Del Moral [2004] for Algorithm 2 and in doing so have established Theorem C.6.

## C.3. Parallel Tempering for Numerical Disintegration

Let $K_i$, $\{\delta_i\}_{i=1}^n$ be as in Section C.1. The PT algorithm [Geyer, 1991] for sampling from $\mu_{\delta_n}^a$ runs $n$ Markov chains in parallel, one for each temperature, by alternately applying $K_i$, then randomly proposing to "swap" the current state of two of the chains. Commonly only swaps of adjacent chains are considered; to this end suppose at iteration $j$ an integer index $q \in [0, n-1]$ has been selected. Denote by $x_j^q$ the state of the chain with $\mu_{\delta_q}^a$ as its invariant measure. Then to ensure the correct invariant distribution of all chains is maintained, the swap of state $x_j^q$ and $x_j^{q+1}$ is accepted with probability

$$\alpha(x_j^q, x_j^{q+1}) = \frac{\pi_q(x_j^{q+1})\pi_{q+1}(x_j^q)}{\pi_q(x_j^q)\pi_{q+1}(x_j^{q+1})} \tag{C.1}$$

where $\pi_q$ denotes the density of the target distribution $\mu^a_{\delta_q}$ with respect to a suitable reference measure. The density notation can be justified since in our experiments the sampler was applied to the finite-dimensional distributions $\mu^a_{\delta_q,N}$ and so the reference measure can be taken to be the Lebesgue measure on $\mathbb{R}^N$.

The PT algorithm for numerical disintegration is described in Algorithm 3. The samples $\{x^n_j\}_{j=1}^{N_{\text{iter}}}$ are approximate draws from the distribution $\mu^a_{\delta_n}$.

Given some initial $x^i_0$ for $i = 1, \ldots, n$ [Initialise]
**for** $j = 1, \ldots, N_{iter}$ **do**
  Sample $\hat{x}^i_j \sim K_i(x^i_{j-1}, \cdot)$ for $i = 1, \ldots, n$ [Move]
  Sample $q \sim \text{Uniform}(0, i-1)$
  **if** $U(0,1) < \alpha(x^q_j, x^{q+1}_j)$ **then**
    Set $x^q_j = \hat{x}^{q+1}_j$ and $x^{q+1}_j = \hat{x}^q_j$ [Accept Swap]
  **else**
    Set $x^q_j = \hat{x}^q_j$ and $x^{q+1}_j = \hat{x}^{q+1}_j$ [Reject Swap]
  For $i \neq q, q+1$, set $x^i_j = \hat{x}^i_j$ [Update]
**end**

**Algorithm 3:** Parallel Tempering for Numerical Disintegration

Algorithms 1 and 3 are each valid for sampling from a target measure $\mu^a_\delta$. The choice of which algorithm to use is problem dependent, and each algorithm has been applied in the experiments in Section 6.

## C.4. Monte Carlo Details for Painlevé Transcendental

Sampling of the posterior was performed for a temperature schedule of 1600 $\delta_i$, equally spaced in log-space from 10 to $10^{-4}$, for an ensemble of 200 particles.

Specification of appropriate transition kernels $K_i$ for this problem was challenging due both to the high dimension and the empirical observation that, for small $\delta$, mixing of the chains tends to be poor. This is likely due to the nonlinearity of the information operator which leads to highly a complex posterior structure. For this reason, a gradient-based sampler was used to construct the transition kernel, the Metropolis-adjusted Langevin algorithm (MALA) [Girolami and Calderhead, 2011].

Denote by $\boldsymbol{u}^k$ the coefficients $[u^k_j]_{j=1}^N$ at iteration $k$ of MALA. Then, recall that MALA has proposals given by

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \tau_i P \nabla \log \pi_i(\boldsymbol{u}^k) + \sqrt{2\tau_i P} W$$

where $W$ is a standard Gaussian distribution and $P \in \mathbb{R}^{N\times N}$ is a positive definite preconditioning matrix. The $\tau_i$ were taken to be fixed for each kernel $K_i$ to a value found empirically to provide a reasonable acceptance rate. $\pi_i$ denotes the unnormalised target distribution for $K_i$, here given by

$$\pi_i(\boldsymbol{u}^k) = \phi\left(\frac{\left\|Ax^N - a\right\|}{\delta_i}\right) q^N(\boldsymbol{u}^k)$$

where $x^N$ is as given in Eq. (6.1) and $q^N(\cdot)$ denotes the density of the truncated prior distribution.

To ensure proposals are scaled appropriately to the decay of the prior distribution for the coefficients, we take $P = \text{diag}(\gamma)$, the diagonal matrix which has the coefficients $\gamma_i$ on its

diagonal. Even with such a transition kernel, mixing is generally poor. To compensate $k$ is taken to be large; for $N_t = 10, 15$ we take $k = 10,000$, while for $N_t = 20$ we take $k = 40,000$. We note that such a large number of temperature levels and transitions makes computation prohibitively expensive, highlighting the importance of future work toward methods for approximating the Bayesian posterior in a more computationally efficient manner.

### C.5. Monte Carlo Details for Poisson Equation

The posterior distribution was obtained by use of the PT algorithm, for 20 temperatures equally spaced in log-space between $10^{-2}$ and $10^{-4}$. For all kinds of prior, the transition kernels $K_i$ were given by 10 iterations of a MALA sampler, with preconditioner as described earlier and parameter $\tau$ chosen to achieve a good acceptance rate. The number of swaps $N_{\text{iter}}$ was taken to be $10^6$ when $N_t = 15$ and $10^7$ when $N_t = 25$ or $N_t = 36$.

## D. Truncation of the Prior Distribution (Proof of Theorem 4.8)

In this section we present the proof for Theorem 4.8 in the main text. We use a general result on the well-posedness of Bayesian inverse problems, of which there are many in the literature, with many variations in terms of the complexity of the function spaces and priors that are considered; see e.g. Stuart [2010], Dashti and Stuart [2013], and Sullivan [2016].

**Theorem D.1** (Theorem 4.6 in Sullivan [2016])**.** *Let $\mathcal{X}$ and $\mathcal{A}$ be separable quasi-Banach spaces over $\mathbb{R}$. Suppose that*

$$\frac{\mathrm{d}\mu_\delta^a}{\mathrm{d}\mu} = \frac{\exp(-\Phi_\delta(x;a))}{Z_\delta^a} \tag{D.1}$$

*where the potential function $\Phi_\delta$ satisfies:*

*S0 $\Phi_\delta(x;\cdot)$ is continuous for each $x \in \mathcal{X}$, $\Phi_\delta(\cdot;a)$ is measurable for each $a \in \mathcal{A}$, and for every $r > 0$, there exists $M_{0,r,\delta} \in \mathbb{R}$ such that, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$ with $\|x\|_{\mathcal{X}} < r$ and $\|a\|_{\mathcal{A}} < r$,*

$$|\Phi_\delta(x;a)| \leq M_{0,r,\delta}.$$

*S1 For every $r > 0$, there exists a measurable $M_{1,r,\delta} \colon \mathbb{R}_+ \to \mathbb{R}$ such that, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$ with $\|a\|_{\mathcal{A}} < r$,*

$$\Phi_\delta(x;a) \geq M_{1,r,\delta}\big(\|x\|_{\mathcal{X}}\big).$$

*S2 For every $r > 0$, there exists a measurable $M_{2,r,\delta} \colon \mathbb{R}_+ \to \mathbb{R}_+$ such that, for all $(x,a,\tilde{a}) \in \mathcal{X} \times \mathcal{A} \times \mathcal{A}$ with $\|a\|_{\mathcal{A}} < r$, $\|\tilde{a}\|_{\mathcal{A}} < r$,*

$$|\Phi_\delta(x;a) - \Phi_\delta(x;\tilde{a})| \leq \exp\big(M_{2,r,\delta}\big(\|x\|_{\mathcal{X}}\big)\big)\|a - \tilde{a}\|_{\mathcal{A}}.$$

*Let $\Phi_{\delta,N}$ be an approximation to $\Phi_\delta$ that satisfies (S1-S3) with $M_{i,r,\delta}$ independent of $N$, and such that*

*S3 $\Psi \colon \mathbb{N} \to \mathbb{R}_+$ is such that, for every $r > 0$, there exists a measurable $M_{3,r,\delta} \colon \mathbb{R}_+ \to \mathbb{R}_+$, such that, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$ with $\|a\|_{\mathcal{A}} < r$,*

$$|\Phi_{\delta,N}(x;a) - \Phi_\delta(x;a)| \leq \exp\big(M_{3,r,\delta}\big(\|x\|_{\mathcal{X}}\big)\big)\Psi(N).$$

*S4 For some $r > 0$,*

$$\mathbb{E}_{X\sim\mu}\big[\exp(2M_{3,r,\delta}(\|X\|_{\mathcal{X}}) - M_{1,r,\delta}(\|X\|_{\mathcal{X}}))\big] < \infty. \tag{D.2}$$

*Let $d_H$ denote the Hellinger distance on $\mathcal{P}_{\mathcal{X}}$. Then there exists a constant $C_\delta$, independent of $N$, such that*

$$d_H\big(\mu^a_{\delta,N}, \mu^a_\delta\big) \le C_\delta \Psi(N)$$

*where $\mu^a_{\delta,N}$ is the posterior distribution based on the potential function $\Phi_{\delta,N}$ instead of $\Phi_\delta$.*

This allows us to establish conditions on $A$ and $\mu$ that guarantee stability under truncation of the prior:

*Proof of Theorem 4.8.* Let $\varphi$ be as in Section 4.1, and let

$$\Phi_\delta(x;a) = \varphi\left(\frac{\|A(x)-a\|_{\mathcal{A}}}{\delta}\right)$$
$$\Phi_{\delta,N}(x;a) = \varphi\left(\frac{\|A\circ P_N(x)-a\|_{\mathcal{A}}}{\delta}\right).$$

Our task is to check the conditions of Theorem D.1 hold for $\Phi_\delta$ and $\Phi_{\delta,N}$.

S0 First, note that $\Phi_\delta(x;\cdot)$ is continuous (since $\varphi$ is continuous from Assumption 4.1 and $\Phi_\delta(x;\cdot)$ is a composition of continuous functions) and that $\Phi_\delta(\cdot;a)$ is measurable (since $\phi$ is measurable and $\Phi_\delta(\cdot;a)$ is a composition of measurable functions). Second, note that $\varphi$ is a continuous bijection from $(0,\infty)$ to itself with $\varphi(0) = 0$. Thus $\varphi^{-1}$ exists and we can consider

$$\begin{aligned}\delta\varphi^{-1}\sup\{|\Phi_\delta(x;a)| : \|x\|_{\mathcal{X}}, \|a\|_{\mathcal{A}} < r\} &= \sup\{\|A(x)-a\|_{\mathcal{A}} : \|x\|_{\mathcal{X}}, \|a\|_{\mathcal{A}} < r\}\\ &\le \sup_{x\in\mathcal{X}}\|A(x)\|_{\mathcal{A}} + r\\ &\le \infty \quad \text{(Assumption 4.6).}\end{aligned}$$

Thus we can take $M_{0,r,\delta} = \varphi(\frac{1}{\delta}\sup_{x\in\mathcal{X}}\|A(x)\|_{\mathcal{A}} + \frac{r}{\delta})$.

S1 Since $\Phi_\delta(x;a) \ge 0$ we can take $M_{1,r,\delta} = 0$.

S2 Given $r > 0$ let $R = \frac{1}{\delta}\sup_{x\in\mathcal{X}}\|A(x)\|_{\mathcal{A}} + \frac{r}{\delta}$, which is finite by Assumption 4.6. The upper bound

$$\begin{aligned}|\Phi_\delta(x;a) - \Phi_\delta(x;\tilde{a})| &= \left|\varphi\left(\frac{\|A(x)-a\|_{\mathcal{A}}}{\delta}\right) - \varphi\left(\frac{\|A(x)-\tilde{a}\|_{\mathcal{A}}}{\delta}\right)\right|\\ &\le C_R\left|\frac{\|A(x)-a\|_{\mathcal{A}}}{\delta} - \frac{\|A(x)-\tilde{a}\|_{\mathcal{A}}}{\delta}\right| \quad \text{(Assumption 4.4)}\\ &\le \frac{C_R}{\delta}\|a-\tilde{a}\|_{\mathcal{A}} \quad \text{(reverse triangle inequality)}\end{aligned}$$

demonstrates that we can take $M_{2,r,\delta} = \max\{0, \log(\frac{C_R}{\delta})\}$.

Minor variation on the above arguments show that S1-3 also hold for $\Phi_{\delta,N}$ with the same constants $M_{i,r,\delta}$.

S3 Let $C_R$ be defined as in S2. The upper bound

$$\begin{aligned}
|\Phi_{\delta,N}(x;a) - \Phi_\delta(x;a)| &= \left|\varphi\left(\frac{\|A\circ P_N(x)-a\|_{\mathcal{A}}}{\delta}\right) - \varphi\left(\frac{\|A(x)-a\|_{\mathcal{A}}}{\delta}\right)\right| \\
&\leq C_R\left|\frac{\|A\circ P_N(x)-a\|_{\mathcal{A}}}{\delta} - \frac{\|A(x)-a\|_{\mathcal{A}}}{\delta}\right| \quad \text{(Assumption 4.4)} \\
&\leq \frac{C_R}{\delta}\|A\circ P_N(x) - A(x)\|_{\mathcal{A}} \quad \text{(reverse triangle inequality)} \\
&\leq \frac{C_R}{\delta}\exp(m(\|x\|_{\mathcal{X}}))\Psi(N) \quad \text{(Assumption 4.5)}
\end{aligned}$$

demonstrates that we can take $M_{3,r,\delta}(\|x\|_{\mathcal{X}}) = \max\{0, \log(\frac{C_R}{\delta}) + m(\|x\|_{\mathcal{X}})\}$.

S4 Let $C_R$ be defined as in S2. The upper bound

$$\begin{aligned}
\mathbb{E}_{X\sim\mu}[\exp(2M_{3,r,\delta}(\|X\|_{\mathcal{X}}) - M_{1,r,\delta}(\|X\|_{\mathcal{X}}))] &= \mathbb{E}_{X\sim\mu}[\exp(2\max\{0, \log(C_R/\delta) + m(\|X\|_{\mathcal{X}})\})] \\
&\leq 1 + \frac{C_R}{\delta}\mathbb{E}_{X\sim\mu}[\exp(2m(\|X\|_{\mathcal{X}}))] \\
&< \infty \quad \text{(Assumption 4.5)}
\end{aligned}$$

establishes the last of the conditions for Theorem D.1 to hold.

Thus from Theorem D.1, $d_{\mathrm{H}}\big(\mu^a_{\delta,N}, \mu^a_\delta\big) \leq C_\delta\Psi(N)$. The proof is completed since Assumption 4.7 implies that $d_{\mathcal{F}} \leq C_{\mathcal{F}}^{-1} d_{\mathrm{TV}}$ where $d_{\mathrm{TV}}$ is the total variation distance based on $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$; in turn it is a standard fact that $d_{\mathrm{TV}} \leq \sqrt{2} d_{\mathrm{H}}$. □

## References


I. Albert, S. Donnet, C. Guihenneuc-Jouyaux, S. Low-Choy, K. Mengersen, and J. Rousseau. Combining expert opinions in prior elicitation. *Bayesian Anal.*, 7(3):503–531, 2012. URL `http://dx.doi.org/10.1214/12-BA717`.

T. V. Anderson. Efficient, accurate, and non-gaussian error propagation through nonlinear, closed-form, analytical system models. Master's thesis, Department of Mechanical Engineering, Brigham Young University, 2011.

I. Babuška and G. Söderlind. On round-off error growth in elliptic problems, 2016. In preparation.

S. Bartels and P. Hennig. Probabilistic approximate least-squares. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, 2016.

J. O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag New York, 1985.

A. Beskos, D. Crisan, and A. Jasra. On the stability of sequential Monte Carlo methods in high dimensions. *Ann. Appl. Probab.*, 24(4):1396–1445, 2014. URL `http://dx.doi.org/10.1214/13-AAP951`.

A. Beskos, A. Jasra, E. A. Muzaffer, and A. M. Stuart. Sequential Monte Carlo methods for Bayesian elliptic inverse problems. *Stat. Comput.*, 25(4):727–737, 2015. URL `http://dx.doi.org/10.1007/s11222-015-9556-7`.

A. Beskos, M. Girolami, S. Lan, P. E. Farrell, and A. M. Stuart. Geometric MCMC for infinite-dimensional inverse problems. *Journal of Computational Physics*, 2016. Advanced access.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B*, 78(5):1103–1130, 2016.

V. I. Bogachev. *Gaussian Measures*, volume 62 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1998. URL `http://dx.doi.org/10.1090/surv/062`.

Z. I. Botev and D. P. Kroese. Efficient Monte Carlo simulation via the generalized splitting method. *Statistics and Computing*, 22(1):1–16, 2012.

M. Briers, A. Doucet, and S. S. Singh. Sequential auxiliary particle belief propagation. In *International Conference on Information Fusion*, 2005.

F.-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic. Probabilistic integration: A role for statisticians in numerical analysis?, 2016. arXiv:1512.00933v4.

M. Capistrán, J. A. Christen, and S. Donnet. Bayesian analysis of ODE's: Solver optimal accuracy and Bayes factors. 2013. arXiv:1311.2281.

I. Castillo and R. Nickl. On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.*, 42(5):1941–1969, 2014. URL `http://dx.doi.org/10.1214/14-AOS1246`.

F. Cérou, P. Del Moral, T. Furon, and A. Guyader. Sequential Monte Carlo for rare event estimation. *Stat. Comput.*, 22(3):795–808, 2012. URL `http://dx.doi.org/10.1007/s11222-011-9231-6`.

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

O. A. Chkrebtii, D. A. Campbell, B. Calderhead, and M. A. Girolami. Bayesian solution uncertainty quantification for differential equations. *Bayesian Analysis*, 11(4):1239–1267, 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.

S. L. Cotter, M. Dashti, and A. M. Stuart. Approximation of Bayesian inverse problems for PDEs. *SIAM J. Numer. Anal.*, 48(1):322–345, 2010. URL `http://dx.doi.org/10.1137/090770734`.

M. Dashti and A. M. Stuart. The Bayesian approach to inverse problems, 2013. arXiv:1302.6989.

M. Dashti, S. Harris, and A. Stuart. Besov priors for Bayesian inverse problems. *Inverse Probl. Imaging*, 6(2):183–200, 2012. URL `http://dx.doi.org/10.3934/ipi.2012.6.183`.

M. de Carvalho, G. L. Page, and B. J. Barney. On the geometry of Bayesian inference. 2017. arXiv:1701.08994.

P. Del Moral. *Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Probability and its Applications (New York). Springer-Verlag, New York, 2004. URL http://dx.doi.org/10.1007/978-1-4684-9393-1.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(3):411–436, 2006. URL http://dx.doi.org/10.1111/j.1467-9868.2006.00553.x.

P. Del Moral, A. Doucet, and A. Jasra. An adaptive sequential Monte Carlo method for approximate bayesian computation. *Statistics and Computing*, 22(5):1009–1020, 2012.

C. Dellacherie and P.-A. Meyer. *Probabilities and Potential*. North-Holland Publishing Co., Amsterdam-New York, 1978.

P. Diaconis. Bayesian numerical analysis. *Statistical Decision Theory and Related Topics IV*, 1:163–175, 1988.

P. Diaconis and D. Freedman. Frequency properties of Bayes rules. In *Scientific inference, data analysis, and robustness (Madison, Wis., 1981)*, volume 48 of *Publ. Math. Res. Center Univ. Wisconsin*, pages 105–115. Academic Press, Orlando, FL, 1983.

P. Diaconis and D. A. Freedman. On the consistency of Bayes estimates. *Ann. Statist.*, 14(1): 1–67, 1986. URL http://dx.doi.org/10.1214/aos/1176349830. With a discussion and a rejoinder by the authors.

J. Dick and F. Pillichshammer. *Digital Nets and Sequences: Discrepancy Theory and Quasi–Monte Carlo Integration*. Cambridge University Press, 2010.

J. L. Doob. Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications*, Colloques Internationaux du Centre National de la Recherche Scientifique, no. 13, pages 23–27. Centre National de la Recherche Scientifique, Paris, 1949.

A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science. Springer-Verlag, New York, 2001. URL http://dx.doi.org/10.1007/978-1-4757-3437-9.

M. M. Dunlop and A. M. Stuart. The Bayesian formulation of EIT: Analysis and algorithms. *Inverse Problems and Imaging*, 10(4):1007–1036, 2016.

L. Ellam, N. Zabaras, and M. Girolami. A Bayesian approach to multiscale inverse problems with on-the-fly scale determination. *J. Comput. Phys.*, 326:115–140, 2016. URL http://dx.doi.org/10.1016/j.jcp.2016.08.031.

P. E. Farrell, A. Birkisson, and S. W. Funke. Deflation techniques for finding distinct solutions of nonlinear partial differential equations. *SIAM J. Sci. Comput.*, 37(4):A2026–A2045, 2015. URL http://dx.doi.org/10.1137/140984798.

G. E. Fasshauer. Solving differential equations with radial basis functions: Multilevel methods and smoothing. *Advances in Computal Mathathematics*, 11(2-3):139–159, 1999. Radial basis functions and their applications.

D. A. Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.*, 34:1386–1403, 1963.

S. French. Aggregating expert judgement. *Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Math. RACSAM*, 105(1):181–206, 2011. URL `http://dx.doi.org/10.1007/s13398-011-0018-6`.

C. J. Geyer. Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface*, 1991.

Z. Ghahramani and C. E. Rasmussen. Bayesian Monte Carlo. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 489–496, 2002.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(2):123–214, 2011. URL `http://dx.doi.org/10.1111/j.1467-9868.2010.00765.x`. With discussion and a reply by the authors.

N. Goodman, V. Mansinghka, D. M. Roy, K. Bonawitz, and J. B. Tenenbaum. Church: a language for generative models. 2012. arXiv:1206.3255.

T. Gunter, M. A. Osborne, R. Garnett, P. Hennig, and S. J. Roberts. Sampling for inference in probabilistic models with fast Bayesian quadrature. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 2789–2797, 2014.

R. Harvey and D. Verseghy. The reliability of single precision computations in the simulation of deep soil heat diffusion in a land surface model. *Climate Dynamics*, 46(3865), 2015.

P. Hennig. Probabilistic interpretation of linear solvers. *SIAM J. Optim.*, 25(1):234–260, 2015. URL `http://dx.doi.org/10.1137/140955501`.

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

P. Hennig, M. A. Osborne, and M. Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A*, 471(2179):20150142, 2015.

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.

A. T. Ihler and D. A. McAllester. Particle belief propagation. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, 2009.

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

J. B. Kadane. *Principles of Uncertainty*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL, 2011. URL `http://dx.doi.org/10.1201/b11322`.

J. B. Kadane and G. W. Wasilkowski. *Bayesian Statistics*, chapter Average Case $\epsilon$-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.

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

G. S. Kimeldorf and G. Wahba. Spline functions and stochastic processes. *Sankhyā Ser. A*, 32: 173–180, 1970b.

D. Kinderlehrer and G. Stampacchia. An Introduction to Variational Inequalities and their Applications, 2000. URL `http://dx.doi.org/10.1137/1.9780898719451`. Reprint of the 1980 original.

B. J. K. Kleijn and A. W. van der Vaart. The Bernstein–Von-Mises theorem under misspecification. *Electron. J. Stat.*, 6:354–381, 2012. URL `http://dx.doi.org/10.1214/12-EJS675`.

A. N. Kolmogorov. *Foundations of Probability*. 1933.

A. Kong, P. McCullagh, X.-L. Meng, D. Nicolae, and Z. Tan. A theory of statistical models for Monte Carlo integration. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 65(3):585–618, 2003. URL `http://dx.doi.org/10.1111/1467-9868.00404`. With discussion and a reply by the authors.

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. Koskela, D. Spano, and P. A. Jenkins. Inference and rare event simulation for stopped Markov processes via reverse-time sequential Monte Carlo. 2016. arXiv:1603.02834.

J. T. N. Krebs. Consistency and asymptotic normality of stochastic Euler schemes for ordinary differential equations. 2016. arXiv:1609.06880.

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

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

L. Le Cam. On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. California Publ. Statist.*, 1:277–329, 1953.

D. Lee and G. W. Wasilkowski. Approximation of linear functionals on a Banach space with a Gaussian measure. *J. Complexity*, 2(1):12–43, 1986. URL `http://dx.doi.org/10.1016/0885-064X(86)90021-X`.

T. Lienart, Y. W. Teh, and A. Doucet. Expectation particle belief propagation. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2015.

D. V. Lindley. *Understanding Uncertainty*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, revised edition, 2014. URL `http://dx.doi.org/10.1002/9781118650158.indsp2`.

F. Lindsten, A. M. Johansen, C. A. Naesseth, B. Kirkpatrick, T. B. Schön, J. Aston, and A. Bouchard-Côté. Divide-and-conquer with sequential Monte Carlo. *Journal of Computational and Graphical Statistics*, 2016. Advanced access.

M. Mahsereci and P. Hennig. Probabilistic line searches for stochastic optimization. In *Proceedings of Advances In Neural Information Processing Systems (NIPS)*, 2015.

J. Mockus. *Bayesian Approach to Global Optimization: Theory and Applications*. 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*, 57(7):1157–1167, 2009.

A. Müller. Integral probability metrics and their generating classes of functions. *Adv. in Appl. Probab.*, 29(2):429–443, 1997. URL `http://dx.doi.org/10.2307/1428011`.

S. Niederer, L. Mitchell, N. Smith, and G. Plank. Simulating human cardiac electrophysiology on clinical time-scales. *Frontiers in Physiology*, 2:14, 2011.

E. Novak and H. Woźniakowski. *Tractability of Multivariate Problems: Standard Information for Functionals*. European Mathematical Society, 2010.

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

W. L. Oberkampf and C. J. Roy. *Verification and Validation in Scientific Computing*. Cambridge University Press, Cambridge, 2013.

A. O'Hagan. Bayes–Hermite quadrature. *Journal of Statistical Planning and Inference*, 29(3): 245–260, 1991.

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 *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2012a.

M. A. Osborne, R. Garnett, S. J. Roberts, C. Hart, S. Aigrain, N. Gibson, and S. Aigrain. Bayesian quadrature for ratios. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, 2012b.

H. Owhadi. Bayesian numerical homogenization. *Multiscale Model. Simul.*, 13(3):812–828, 2015a. URL `http://dx.doi.org/10.1137/140974596`.

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

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

B. Paige and F. Wood. Inference networks for sequential Monte Carlo in graphical models. 2016. arXiv:1602.06701.

J. Pfanzagl. Conditional distributions as derivatives. *Ann. Probab.*, 7(6):1046–1050, 1979.

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

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, Cambridge, third edition, 2007.

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

M. M. Rao. *Conditional Measures and Applications*, volume 271 of *Pure and Applied Mathematics (Boca Raton)*. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2005. URL `http://dx.doi.org/10.1201/9781420027433`.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

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

C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Science & Business Media., 2013.

C. Roy. Review of discretization error estimators in scientific computing. In *Proceedings of AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition*, 2010.

J. Sacks and D. Ylvisaker. Statistical designs and integral approximation. In *Proc. Twelfth Biennial Sem. Canad. Math. Congr. on Time Series and Stochastic Processes; Convexity and Combinatorics (Vancouver, B.C., 1969)*, pages 115–136. Canad. Math. Congr., Montreal, Que., 1970.

S. Särkkä, J. Hartikainen, L. Svensson, and F. Sandblom. On the relation between Gaussian process quadratures and sigma-point methods. 2015. arXiv:1504.05994.

M. Schober, D. K. Duvenaud, and P. Hennig. Probabilistic ODE solvers with Runge–Kutta means. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2014.

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

A. Ścibior, Z. Ghahramani, and A. D. Gordon. Practical probabilistic programming with monads. *SIGPLAN Notices*, 50(12):165–176, 2015.

G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J., 1976.

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

G. Strang and G. Fix. *An analysis of the finite element method*. Englewood Cliffs, NJ: Prentice-Hall., 1973.

S. H. Strogatz. *Nonlinear Dynamics and Chaos*. Westview Press, 2014.

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

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

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

T. J. Sullivan. *Introduction to Uncertainty Quantification*, volume 63 of *Texts in Applied Mathematics*. Springer, Cham, 2015. URL `http://dx.doi.org/10.1007/978-3-319-23395-6`.

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.

V. Tarieladze and N. Vakhania. Disintegration of Gaussian measures and average-case optimal algorithms. *J. Complexity*, 23(4-6):851–866, 2007. URL `http://dx.doi.org/10.1016/j.jco.2007.04.005`.

O. Teymur, K. Zygalakis, and B. Calderhead. Probabilistic linear multistep methods. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2016.

J. Tišer. Vitali covering theorem in Hilbert space. *Trans. Amer. Math. Soc.*, 355(8):3277–3289, 2003. URL `http://dx.doi.org/10.1090/S0002-9947-03-03296-3`.

A. Torn and A. Zilinskas. *Global Optimization*. Springer-Verlag New York, Inc., 1989.

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

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.

F. Wood, J.-W. van de Meent, and V. Mansinghka. A new approach to probabilistic programming inference. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, 2014.