Statistics and Computing (2019) 29:1335–1351
https://doi.org/10.1007/s11222-019-09902-z

A modern retrospective on probabilistic numerics

C. J. Oates[1,2] · **T. J. Sullivan**[3,4]

Published online: 4 October 2019

Abstract
This article attempts to place the emergence of probabilistic numerics as a mathematical–statistical research field within its historical context and to explore how its gradual development can be related both to applications and to a modern formal treatment. We highlight in particular the parallel contributions of Sul′din and Larkin in the 1960s and how their pioneering early ideas have reached a degree of maturity in the intervening period, mediated by paradigms such as average-case analysis and information-based complexity. We provide a subjective assessment of the state of research in probabilistic numerics and highlight some difficulties to be addressed by future works.

Keywords Probabilistic numerics · Scientific computation · Reasoning under uncertainty · Uncertainty quantification

Mathematics Subject Classification 62-03 · 65-03 · 01A60 · 01A65 · 01A67

1 Introduction

The field of probabilistic numerics (PN), loosely speaking, attempts to provide a *statistical* treatment of the errors and/or approximations that are made en route to the output of a deterministic numerical method, e.g. the approximation of an integral by quadrature, or the discretised solution of an ordinary or partial differential equation. This decade has seen a surge of activity in this field. In comparison with historical developments that can be traced back over more than a hundred years, the most recent developments are particularly interesting because they have been characterised by simultaneous input from multiple scientific disciplines: mathematics, statistics, machine learning, and computer science. The field has, therefore, advanced on a broad front, with contributions ranging from the building of over-arching general theory to practical implementations in specific problems of interest. Over the same period of time, and because of increased interaction among researchers coming from different communities, the extent to which these developments were—or were not—presaged by twentieth-century researchers has also come to be better appreciated.

Thus, the time appears to be ripe for an update of the 2014 *Tübingen Manifesto* on probabilistic numerics (Hennig 2014; Osborne 2014a, b, c, d) and the position paper of Hennig et al. (2015) to take account of the developments between 2014 and 2019, an improved awareness of the history of this field, and a clearer sense of its future directions and potential.

In this article, we aim to summarise some of the history of probabilistic perspectives on numerics (Sect. 2), to place more recent developments into context (Sect. 3), and to articulate a vision for future research in, and use of, probabilistic numerics (Sect. 4).

The authors are grateful to the participants of *Prob Num 2018*, 11–13 April 2018, at the Alan Turing Institute, UK—and in particular the panel discussants Oksana Chkrebtii, Philipp Hennig, Youssef Marzouk, Mike Osborne, and Houman Owhadi—for many stimulating discussions on these topics. However, except where otherwise indicated, the views that we present here are our own, and if we have misquoted or misrepresented the views of others, then the fault is entirely ours.

✉ C. J. Oates
chris.oates@ncl.ac.uk; coates@turing.ac.uk

T. J. Sullivan
t.j.sullivan@fu-berlin.de; sullivan@zib.de

[1] Newcastle University, Herschel Building, Newcastle upon Tyne NE1 7RU, UK

[2] Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, UK

[3] Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany

[4] Zuse Institute Berlin, Takustraße 7, 14195 Berlin, Germany

2 Historical developments

The first aim of this article is to reflect on the gradual emergence of probabilistic numerics as a research field. The account in this section is not intended to be comprehensive in terms of the literature that is cited. Rather, our aim is to provide an account of how the philosophical status of probabilistic approaches to numerical tasks has evolved, and in particular to highlight the parallel, pioneering, but often-overlooked contributions of Sul′din in the USSR and Larkin in the UK and Canada.

2.1 Prehistory (–1959)

The origins of PN can be traced to a discussion of probabilistic approaches to polynomial interpolation by Poincaré in his *Calcul des Probabilités* (Poincaré 1896, Ch. 21;Poincaré 1912, Ch. 25). Poincaré considered what, in modern terms, would be a particular case of a Gaussian infinite product measure prior on a function f, expressing it as a power series

$$f(x) = \sum_{k=0}^{\infty} A_k x^k$$

with independent normally-distributed coefficients A_k; one is then given n pointwise observations of the values of f and seeks the probable values of $f(x)$ for another (not yet observed) value of x.

> "Je suppose que l'on sache a priori que la fonction f(x) est développable, dans une certain domaine, suivant les puissances croissantes des x,
>
> $$f(x) = A_0 + A_1 x + \ldots.$$
>
> Nous ne savons rien sur les A, sauf que la probabilité pour que l'un d'eux, A_i, *soit compris entre certaines limites, y et* $y + \mathrm{d}y$, est
>
> $$\sqrt{\frac{h_i}{\pi}} e^{-h_i y^2}\, \mathrm{d}y.$$
>
> Nous connaissons par n observations
>
> $$\begin{aligned} f(a_1) &= B_1, \\ f(a_2) &= B_2, \\ &\ldots\ldots\ldots\ldots \\ f(a_n) &= B_n. \end{aligned}$$
>
> Nous cherchons la valeur probable de $f(x)$ pour une autre valeur de x." (Poincaré 1912, p. 292)

Note that, in using a Gaussian prior, Poincaré was departing from the Laplacian principle of indifference (Laplace 1812), which would have mandated a uniform prior.[1]

Poincaré's analytical treatment predates the first digital multipurpose computers by decades, yet it clearly illustrates a non-trivial probabilistic perspective on a classic numerical task, namely function approximation by interpolation, a hybrid approach that is entirely in keeping with Poincaré's reputation as one of the last universalist mathematicians (Ginoux and Gerini 2013).

However, our focus here is on the development of probabilistic numerical methods for use on a computer. The limited nature of the earliest computers led authors to focus initially on the phenomenon of *round-off error* (Henrici 1962; Hull and Swenson 1966; von Neumann and Goldstine 1947), whether of fixed-point or floating-point type, without any particular statistical *inferential* motivation; more recent contributions to the statistical study of round-off error include those of Barlow and Bareiss (1985), Chatelin and Brunet (1990), and Tienari (1970). According to von Neumann and Goldstine, writing in 1947,

> "[round-off errors] are strictly very complicated but uniquely defined number theoretical functions [of the inputs], yet our ignorance of their true nature is such that we best treat them as random variables." (von Neumann and Goldstine 1947, p. 1027).

Thus, von Neumann and Goldstine seem to have held a utilitarian view that probabilistic models in computation are useful shortcuts, simply easier to work with than the unwieldy deterministic truth.[2]

Concerning the numerical solution of ordinary differential equations (ODEs), Henrici (1962, 1963) studied classical finite difference methods and derived expected values and covariance matrices for accumulated round-off error, under an assumption that individual round-off errors can be modelled as independent random variables. In particular, given posited means and covariance matrices of the individual errors, Henrici demonstrated how these moments can be propagated through the computation of a finite difference method. In contrast with more modern treatments, Henrici was concerned with the *analysis* of an established numeri-

[1] Indeed, while an improper uniform prior distribution on $\mathbb{R}$ makes sense for each A_k individually, no such countably additive uniform measure (an "infinite-dimensional Lebesgue measure") can exist on $\mathbb{R}^\infty$ for $(A_k)_{k=0}^\infty$ (Sudakov 1959). That said, Poincaré does not impose any summability constraints on the h_i either, so the covariance operator associated to his Gaussian prior may fail to be trace class.

[2] Decades later, the discovery of chaotic dynamical systems would yield a similar conundrum: after long enough time, one may as well assume that the system's state is randomly distributed according to its invariant measure, if it possesses one.

cal method and did not attempt to statistically *motivate* the numerical method itself.

2.2 The parallel contributions of Larkin and Sul′din (1959–1980)

One of the earliest attempts to motivate a numerical algorithm from a statistical perspective was due to Al′bert Valentinovich Sul′din (1924–1996) (Fig. 1), working at Kazan State University in the USSR (now Kazan Federal University in the Russian Federation) (Norden et al. 1978; Zabotin et al. 1996). After first making contributions to the study of Lie algebras, towards the end of the 1950s Sul′din turned his attention to computational and applied mathematics, and in particular to probabilistic and statistical methodology. His work in this direction led to the establishment of the Faculty of Computational Mathematics and Cybernetics (now Institute of Computational Mathematics and Information Technologies) in Kazan, of which he was the founding Dean.

Sul′din began by considering the problem of quadrature. Suppose that we wish to approximate the definite integral $\int_a^b u(t)\,\mathrm{d}t$ of a function $u \in \mathcal{U} := C^0([a,b];\mathbb{R})$, the space of continuous real-valued functions on $[a,b]$, under a statistical assumption that $(u(t)-u(a))_{t\in[a,b]}$ follows a standard Brownian motion (Wiener measure, μ_{W}). For this task, we receive pointwise data about the integrand u in the form of the values of u at $J \in \mathbb{N}$ arbitrarily located nodes $t_1, \ldots, t_J \in [a,b]$, although for convenience we assume that

$$a = t_1 < t_2 < \cdots < t_J = b.$$

In more statistical language, anticipating the terminology of Sect. 3.2, our *observed data* or *information* concerning the integrand u is $y := (t_j, u(t_j))_{j=1}^J$, which takes values in the space $\mathcal{Y} := ([a,b]\times\mathbb{R})^J$.

Since μ_{W} is a Gaussian measure and both the integral and pointwise evaluations of u are linear functions of u, Sul′din (1959, 1960, 1963b) showed by direct calculation that the quadrature rule $B\colon \mathcal{Y} \to \mathbb{R}$ that minimises the mean squared error

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

is the classical trapezoidal rule[3]

$$B_{\mathrm{tr}}\big((t_j,z_j)_{j=1}^J\big) := \frac{1}{2}\sum_{j=1}^{J-1}(z_{j+1}+z_j)(t_{j+1}-t_j) \tag{2}$$

$$= z_1\frac{t_2-t_1}{2} + \sum_{j=2}^{J-1} z_j \frac{t_{j+1}-t_{j-1}}{2} + z_J\frac{t_J - t_{J-1}}{2}, \tag{3}$$

i.e. the definite integral of the piecewise linear interpolant of the observed data. This result was a precursor to a sub-field of numerical analysis that became known as *average-case analysis*; see Sect. 2.3.

Sul′din was aware of the connection between his methods and statistical regression (Sul′din 1963a) and conditional probability (Sul′din 1963c), although it is difficult to know whether he considered his work to be an expression of *statistical inference* as such. Indeed, since Sul′din's methods were grounded in Hilbert space theory (Sul′din 1968, 1969), the underlying mathematics (the linear conditioning of Gaussian measures on Hilbert spaces) is linear algebra which can be motivated without recourse to a probabilistic framework.

In any case, Sul′din's contributions were something entirely novel. Up to this point, the role of statistics in numerical analysis was limited to providing insight into the *performance* of a traditional numerical method. The 1960s brought forth a new perspective, namely the statistically motivated *design* of numerical methods. Indeed,

> "A.V. Sul′din's 1969 habilitation thesis concerned the development of probabilistic methods for the solution of problems in computational mathematics. His synthesis of two branches of mathematics turned out to be quite fruitful, and deep connections were discovered between the robustness of approximation formulae and their precision. Building on the general concept of an enveloping Hilbert space, A.V.Sul′din proved a projection theorem that enabled the solution of a number of approximation-theoretic problems." (Zabotin et al. 1996)

Fig. 1 Al′bert Valentinovich Sul′din (1924–1996). (Sul′din 2018, reproduced with permission)

[3] Note that formulation (2) of B_{tr} emphasises the trapezoidal geometry being used to approximate the integral, whereas formulation (3) emphasises that the integrand need only be evaluated J and not $2J-2$ times.

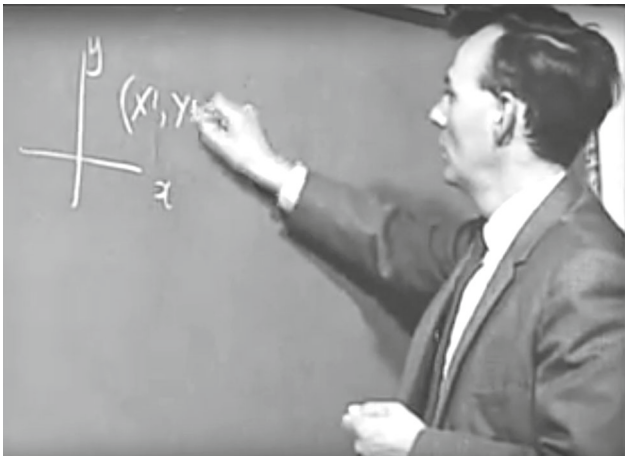


Fig. 2 Frederick Michael Larkin (1936–1982). (Larkin et al. 1967, reproduced with permission)

However, Sul′din was not alone in arriving at this point of view. On the other side of the Iron Curtain, between 1957 and 1969, Frederick Michael ("Mike") Larkin (1936–1982) (Fig. 2) worked for the UK Atomic Energy Authority in its laboratories at Harwell and Culham (the latter as part of the Computing and Applied Mathematics Group), as well as working for two years at Rolls Royce, England. Following a parallel path to that of Sul′din, over the next decade Larkin would further blend numerical analysis and statistical thinking (Kuelbs et al. 1972; Larkin 1969, 1972, 1974, 1979a, b, c), arguably laying the foundations on which PN would be developed. At Culham, Larkin worked on building some of the first graphical calculators, the GHOST graphical output system and the accompanying GHOUL graphical output language. It can be speculated that an intimate familiarity with the computational limitations of GHOST and GHOUL may have motivated Larkin to seek a richer description of the numerical error associated to their output.

The perspective developed by Larkin was fundamentally statistical and, in modern terminology, the probabilistic numerical methods he developed would be described as *Bayesian*,[4] which we discuss further in Sect. 3.2. Nevertheless, the pioneering nature of this research motivated Larkin to focus on specific numerical tasks, as opposed to establishing a unified framework. In particular, he considered in detail the problems of approximating a non-negative function (Larkin 1969), quadrature (Larkin 1972, 1974), and estimating the zeros of a complex function (Larkin 1979a, b). In the context of the earlier numerical integration example of Sul′din, the alternative proposal of Larkin was to consider the Wiener measure as a prior, the information $(t_j, u(t_j))_{j=1}^J$ as (noiseless) data, and to output the posterior marginal for the integral $\int_a^b u(t)\,\mathrm{d}t$. That is, Larkin took the fundamental step of considering a distribution over the solution space of the numerical task to be the output of a computation—this is what we would now recognise as the defining property of a *probabilistic numerical method*:

> "Among other things, this permits, at least in principle, the derivation of joint probability density functions for [both observed and unobserved] functionals on the space and also allows us to evaluate confidence limits on the estimate of a required functional (in terms of given values of other functionals)." (Larkin 1972)[5]

Thus, in contrast to Sul′din's description of the trapezoidal rule B_{tr} from (2) as a frequentist point estimator obtained from minimising (1), which just happens to produce an unbiased estimator with variance $\frac{1}{12}\sum_{j=1}^{J-1}(t_{j+1}-t_j)^3$, the Larkin viewpoint is to see the normal distribution

$$\mathcal{N}\left(B_{\mathrm{tr}}\big((t_j, z_j)_{j=1}^J\big), \frac{1}{12}\sum_{j=1}^{J-1}(t_{j+1}-t_j)^3\right) \tag{4}$$

on $\mathbb{R}$ as the measure-valued output of a probabilistic quadrature rule, of which $B_{\mathrm{tr}}\big((t_j, z_j)_{j=1}^J$ is a convenient point summary. Note also that the technical development in this pioneering work made fundamental contributions to the study of Gaussian measures on Hilbert spaces (Kuelbs et al. 1972; Larkin 1972).

Larkin moved to Canada in 1969 to start work as a Consultant in Numerical Methods and Applied Mathematics within the Computing Centre and, subsequently in 1974, as Associate Professor in the Department of Computing and Information Science (now the School of Computing) at Queen's University in Kingston, Ontario. He received tenure in 1977 and was promoted to full professor in 1980.

> "He worked in isolation at Queen's in that few graduate students and fewer faculty members were aware of the nature of his research contributions to the field.

[4] Larkin used the term *relative likelihood* for what we would recognise as a Bayesian prior (Larkin 1972, Section 3.3). We may speculate, but cannot be sure, that such terminological differences are largely accidents of history. Larkin was educated and did his early work exactly when the frequentist paradigm was starting to lose its dominance and Bayesian methods were starting to come back into fashion, driven by Cox's logical justification of the Bayesian paradigm (Cox 1946, 1961) and the development of theory, hardware, and software for methods like Markov chain Monte Carlo. See Dale (1999) for a comprehensive history of this area of statistics.

[5] In this passage "the estimate" refers to the posterior mean in a linear-Gaussian setup and "confidence limit" refers to what we would now call a highest-posterior-density credible interval. We suspect that the cultural dominance of frequentist statistics, in which estimators are reported alongside confidence intervals, led Larkin to adopt a similar presentation of the posterior—though we emphasise that Larkin was fundamentally providing a Bayesian treatment.

> […] Michael pioneered the idea of using a probabilistic approach to give an alternative local approximation technique. In some cases this leads to the classical methods, but in many others leads to new algorithms that appear to have practical advantages over more classical methods. This work has finally begun to attract attention and I expect that the importance of his contribution will grow in time." (Queen's University at Kingston, 11 Feb. 1982)

From our perspective, writing in 2019, it seems that Sul′din and Larkin were working in parallel but were ahead of their time. Their probabilistic perspectives on approximation theory were similar, but limited to a Gaussian measure context. Naturally, given the linguistic barriers and nearly disjoint publication cultures of their time, it would not have been easy for Larkin and Sul′din to be conversant with each other's work, though these barriers were not always as great as is sometimes thought (Hollings 2016). At least by 1972 (Larkin 1972), Larkin was aware of and cited Sul′din's work on minimal variance estimators for the values of linear functionals on Wiener space (Sul′din 1959, 1960), but apparently did not know of Sul′din's 1969 habilitation thesis, which laid out a broader agenda for the role of probability in numerics. Conversely, Soviet authors writing in 1978 were aware of Sul′din's influence on, e.g. Ulf Grenander and Walter Freiberger at Brown University, but make no mention of Larkin (Norden et al. 1978). Sul′din, for his part, at least as judged by his publication record, seems to have turned his attention to topics such as industrial mathematics [perhaps an "easier sell" in the production-oriented USSR (Hollings 2016)], mathematical biology, and of course the pressing concerns of faculty administration.

Finally, concerning the practicality of Sul′din and Larkin's ideas, one has to bear in mind the limited computational resources available at even cutting-edge facilities in the 1960s:[6] probabilistic numerics was an idea ahead of its time, and the computational power needed to make it a reality simply did not exist.

2.3 Optimal numerical methods are Bayes rules (1980–1990)

In the main, research contributions until 1990 continued to focus on deriving insight into traditional numerical methods through probabilistic analyses. In particular, the *average-case analysis* (ACA) of numerical methods received interest and built on the work of Kolmogorov (1936) and Sard (1963). In ACA, the performance of a numerical method is assessed in terms of its *average error* over an ensemble of numerical problems, with the ensemble being represented by a probability measure over the problem set; a prime example is univariate quadrature with the average quadratic loss (1) given earlier. Root-finding, optimisation, etc. can all be considered similarly, and we defer to, e.g. Ritter (2000) and Traub et al. (1983) for comprehensive treatments of this broad topic.

A traditional (deterministic) numerical method can also be regarded as a decision rule and the probability measure used in ACA can be used to instantiate the Bayesian decision-theoretic framework (Berger 1985). The average error is then recognised as the *expected loss*, also called the *risk*. The fact that ACA is mathematically equivalent to Bayesian decision theory (albeit limited to the case of an experiment that produces a deterministic dataset) was noted by Kimeldorf and Wahba (1970a, b), and Parzen (1970)—and also by Larkin (1970).

Armed with an optimality criterion for a numerical method, it is natural to ask about the existence and performance of method(s) that minimise it. Such methods are called *average-case optimal* in ACA and are recognised as *Bayes rules* or *Bayes acts* in the decision-theoretic context. A key result in this area is the insight of Kadane and Wasilkowski (1985) that ACA-optimal methods coincide with (non-randomised) Bayes rules when the measure used to define the average error is the Bayesian prior; for a further discussion of the relationships among these optimality criteria, including the Bayesian probabilistic numerical methods of Sect. 3.2, see Cockayne et al. (2019a) and Oates et al. (2019b).

Many numerical methods come in parametric families, being parametrised by, e.g. the number of quadrature nodes, a mesh size, or a convergence tolerance. For any "sensible" method, the error can be driven to zero by sending the parameter to infinity or zero as appropriate. If one is prepared to pay an infinite computational cost, then essentially any method can be optimal! Thus, when asking about the optimality of a numerical method, it is natural to consider the optimality of methods of a given computational cost or complexity.

With such concerns in mind, the field of *information-based complexity* (IBC) (Novak 1988; Traub et al. 1983; Traub and Woźniakowsi 1980) developed simultaneously with ACA, with the aim of relating the computational complexity and optimality properties of algorithms to the available information on the unknowns, e.g. the partial nature of the information and any associated observational costs and errors. For example, Smale (1985, Theorem D) compared the accuracies (with respect to mean absolute error) for a given cost of the Riemann sum, trapezoidal, and Simpson

[6] To first approximation, a single modern laptop has a hundred times the computing power of all five then-cutting-edge IBM System/360 Model 75J mainframe computers used for the ground support of the Apollo missions (Manber and Norvig 2012).

quadrature rules;[7] in the same paper, Smale also considered root-finding, optimisation via linear programming, and the solution of systems of linear equations.

The example of Bayesian quadrature was again discussed in detail by Diaconis (1988), who repeated Sul′din's observation that the posterior mean for $\int_a^b u(t)\,\mathrm{d}t$ under the Wiener measure prior is the trapezoidal method (2), which is an ACA-optimal numerical method. However, Diaconis posed a further question: can other classical numerical integration methods, or numerical methods for other tasks, be similarly recovered as Bayes rules in a decision-theoretic framework? For linear cubature methods, a positive and constructive answer was recently provided by Karvonen et al. (2018), but the question remains open in general.

2.4 Probabilistic numerical methods (1991–2009)

After a period in which probabilistic numerical methods were all but forgotten, research interest was again triggered by various contributions on numerical integration (Minka 2000; O'Hagan 1991; Rasmussen and Ghahramani 2003), each to a greater or lesser extent a rediscovery of earlier work due to Larkin (1972). In each case, the output of computation was considered to be a probability distribution over the quantity of interest.

The 1990s saw an expansion in the PN agenda, first with early work on an area that was to become *Bayesian optimisation* (Močkus 1975, 1977, 1989) and then with an entirely novel contribution on the numerical solution of ODEs by Skilling (1992). Skilling presented a Bayesian[8] perspective on the numerical solution of initial value problems of the form

$$\begin{aligned} u'(t) \equiv \frac{\mathrm{d}u}{\mathrm{d}t} &= f(t, u(t)) \quad t \in [0, T], \\ u(0) &= u_0, \end{aligned} \tag{5}$$

and considered, for example, how regularity assumptions on f should be reflected in correlation functions and the hypothesis space, how to choose a prior and likelihood, and potential sampling strategies. Despite this work's then-new explicit emphasis on its Bayesian statistical character, Skilling himself considered his contributions to be quite natural:

> "This paper arose from long exposure to Laplace/Cox/Jaynes probabilistic reasoning, combined with the University of Cambridge's desire that the author teach some (traditional) numerical analysis. The rest is common sense. […] Simply, Bayesian ideas are 'in the air'." (Skilling 1992)

2.5 Modern perspective (2010–)

The last two decades have seen an explosion of interest in *uncertainty quantification* (UQ) for complex systems, with a great deal of research taking place in this area at the meeting point of applied mathematics, statistics, computational science, and application domains (Le Maître and Knio 2010; Smith 2014; Sullivan 2015):

> "UQ studies all sources of error and uncertainty, including the following: systematic and stochastic measurement error; ignorance; limitations of theoretical models; limitations of numerical representations of those models; limitations of the accuracy and reliability of computations, approximations, and algorithms; and human error. A more precise definition is UQ is the end-to-end study of the reliability of scientific inferences." (U.S. Department of Energy 2009, p. 135)

Since 2010, perhaps stimulated by this activity in the UQ community, a perspective on PN has emerged that sees PN part of UQ (broadly understood) and should be performed with a view to propagating uncertainty in computational pipelines. This is discussed further in Sects. 3.1 and 3.2.

A notable feature of PN research since 2010 is the way that it has advanced on a broad front. The topic of quadrature/cubature, in the tradition of Sul′din and Larkin, continues to be well represented: see, e.g. Briol et al. (2019); Gunter et al. (2014); Karvonen et al. (2018); Oates et al. (2017); Osborne et al. (2012a, b); Särkkä et al. (2016), and Xi et al. (2018), as well as Ehler et al. (2019); Jagadeeswaran and Hickernell (2019); Karvonen et al. (2019a), and Karvonen et al. (2019b) in this special issue. The Bayesian approach to global optimisation continues to be widely used (Chen et al. 2018; Snoek et al. 2012), whilst probabilistic perspectives on quasi-Newton methods (Hennig and Kiefel 2013) and line search methods (Mahsereci and Hennig 2015) have been put forward. In the context of numerical linear algebra, Bartels and Hennig (2016); Cockayne et al. (2019b), and Hennig (2015), as well as Bartels et al. (2019) in this special issue, have approached the solution of a large linear system of equations as a statistical learning task and developed probabilistic alternatives to the classical conjugate gradient method.

Research has been particularly active in the development and analysis of statistical methods for the solution of ordinary and partial differential equations (ODEs and PDEs). One line of research has sought to cast the solution of ODEs in the context of Bayesian filtering theory by building a Gaussian process (GP) regression model for the solution u of the initial value problem of the form (5). The observational data

[7] On page 95 of the same paper, Smale highlighted Larkin (1972) as an *"important earlier paper in this area"*.

[8] To be pedantic, the method of Skilling (1992) does not satisfy the definition of a Bayesian PNM as given in Sect. 3.2. However, the method can be motivated as exact Bayesian inference under an approximate likelihood; see Wang et al. (2018).

consist of the evaluations of the vector field f, interpreted as imperfect observations of the true time derivative u', since one evaluates f at the "wrong" points in space. In this context, the key result is the Bayesian optimality of evaluating f according to the classical Runge–Kutta (RK) scheme, so that the RK methods can be seen as point estimators of GP filtering schemes (Kersting and Hennig 2016; Schober et al. 2014, 2018); see also Tronarp et al. (2019) in this special issue. Related iterative probabilistic numerical methods for ODEs include those of Abdulle and Garegnani (2018); Chkrebtii et al. (2016); Conrad et al. (2017); Kersting et al. (2018); Teymur et al. (2016, 2018). The increased participation of mathematicians in the field has led to correspondingly deeper local and global convergence analysis of these methods in the sense of conventional numerical analysis, as performed by Conrad et al. (2017); Kersting et al. (2018); Schober et al. (2018), and Teymur et al. (2018), as well as Lie et al. (2019) in this special issue; statistical principles for time step adaptivity have also been discussed, e.g. by Chkrebtii and Campbell (2019) in this special issue.

For PDEs, resent research includes Chkrebtii et al. (2016); Cockayne et al. (2016, 2017), and Owhadi (2015), with these contributions making substantial use of reproducing kernel Hilbert space (RKHS) structure and Gaussian processes. Unsurprisingly, given the deep connections between linear algebra and numerical methods for PDEs, the probabilistically motivated theory of *gamblets* for PDEs (Owhadi 2017; Owhadi and Scovel 2017a; Owhadi and Zhang 2017) has gone hand-in-hand with the development of fast solvers for structured matrix inversion and approximation problems (Schäfer et al. 2017); see also Yoo and Owhadi (2019) in this special issue.

Returning to the point made at the beginning of this section, however, motivation for the development of probabilistic numerical methods has become closely linked to the traditional motivations of UQ (e.g. accurate and honest estimation of parameters of a so-called *forward model*), with a role for PN due to the need to employ numerical methods to simulate from a forward model. The idea to substitute a probability distribution in place of the (in general erroneous) output of a traditional numerical method can be used to prevent undue bias and over-confidence in the UQ task and is analogous to *robust likelihood* methods in statistics (Bissiri et al. 2016; Greco et al. 2008). This motivation is already present in Conrad et al. (2017) and forms a major theme of Cockayne et al. (2019a); Oates et al. (2019a). Analysis of the impact of probabilistic numerical methods in simulation of the forward model within the context of Bayesian inversion has been provided by Lie et al. (2018) and Stuart and Teckentrup (2018).

2.6 Related fields and their development

The field of PN did not emerge in isolation and the research cited above was undoubtedly influenced by parallel developments in mathematical statistics, some of which are now discussed.

First, the mathematical theory of *optimal approximation using splines* was applied by Schoenberg (1965, 1966) and Karlin (1969, 1971, 1972, 1976) in the late 1960s and early 1970s to the linear problem of quadrature. Indeed, Larkin (1974) cites Karlin (1969). However, the works cited above were not concerned with randomness and equivalent probabilistic interpretations were not discussed; in contrast, the Bayesian interpretation of spline approximation was highlighted by Kimeldorf and Wahba (1970a).

Second, the *experimental design* literature of the late 1960s and early 1970s, including a sequence of contributions from Sacks and Ylvisaker (1966, 1968, 1970a, b), considered optimal selection of a design $0 \leq t_1 < t_2 < \cdots < t_J \leq 1$ to minimise the covariance of the best linear estimator of β given discrete observations of stochastic process

$$Y(t) = \sum_{i=1}^{m} \beta_i \phi_i(t) + Z(t),$$

where Z is a stochastic process with $\mathbb{E}[Z(t)] = 0$ and $\mathbb{E}[Z(t)^2] < \infty$, based on the data $\{(t_j, Y(t_j))\}_{j=1}^{J}$. As such, the mathematical content of these works concerns optimal approximation in RKHSs, e.g. Sacks and Ylvisaker (1970a, p. 2064, Theorem 1); we note that Larkin (1970) simultaneously considered optimal approximation in RKHSs. However, the extent to which probability enters these works is limited to the measurement error process Z that is entertained.

Third, the literature on *emulation of black-box functions* that emerged in the late 1970s and 1980s, with contributions from, e.g. O'Hagan (1978) and Sacks et al. (1989), provided Bayesian and frequentist statistical perspectives (respectively) on interpolation of a black-box function based on a finite number of function evaluations. This literature did not present interpolation as an exemplar of other more challenging numerical tasks, such as the solution of differential equations, which could be similarly addressed but rather focused on the specific problem of black-box interpolation in and of itself. Sacks et al. (1989) were aware of the work of Sul′din but Larkin's work was not cited. The challenges of proposing a suitable stochastic process model for a deterministic function were raised in the accompanying discussion of Sacks et al. (1989) and were further discussed by Currin et al. (1991).

2.7 Conceptual evolution—a summary

To conclude and summarise this section, we perceive the following evolution of the concepts used in, and interpretation applied to, probability in numerical analysis:

1. In the traditional setting of numerical analysis, as seen circa 1950, all objects and operations are seen as being strictly deterministic. Even at that time, however, it was accepted by some that these deterministic objects are sometimes exceedingly complicated, to the extent that they may be treated as being stochastic, à la von Neumann and Goldstine.
2. Sard and Sul′din considered the questions of optimal performance of a numerical method in, respectively, the worst-case and the average-case context. Though it is a fact that some of the average-case performance measures amount to variances of point estimators, they were not *viewed* as such and in the early 1960s these probabilistic aspects were not a motivating factor.
3. Larkin's innovation, in the late 1960s and early 1970s, was to formulate numerical tasks in terms of a joint distribution over latent quantities and quantities of interest, so that the quantity-of-interest output can be seen as a stochastic object. However, perhaps due to the then-prevailing statistical culture, Larkin summarised his posterior distributions using a point estimator accompanied by a credible interval.
4. The fully modern viewpoint, circa 2019, is to explicitly think of the output as a probability measure to be realised, sampled, and possibly summarised.

3 Probabilistic numerical methods come into focus

In this section, we wish to emphasise how some of the recent developments mentioned in the previous section have brought greater clarity to the philosophical status of probabilistic numerics, clearing up some old points of disagreement or providing some standardised frameworks for the comparison of tasks and methods.

3.1 A means to an end, or an end in themselves?

One aspect that has become clearer over the last few years, stimulated to some extent by disagreements between statisticians and numerical analysts over the role of probability in numerics, is that there are (at least) two distinct use cases or paradigms:

– (P1) a probability-based *analysis* of the performance of a (possibly classical) numerical method;
– (P2) a numerical method whose output carries the *formal semantics* of some statistical inferential paradigm (e.g. the Bayesian paradigm; cf. Sect. 3.2).

Representatives of the first class of methods include Abdulle and Garegnani (2018) and Conrad et al. (2017), which consider stochastic perturbations to explicit numerical integrators for ODEs in order to generate an ensemble of plausible trajectories for the unknown solution of the ODE. In some sense, this can be viewed as a probabilistic sensitivity/stability analysis of a classical numerical method. This first paradigm is also, clearly, closely related to ACA.

The second class of methods is exemplified by the Bayesian probabilistic numerical methods, discussed in Cockayne et al. (2019a) and Sect. 3.2. We can further enlarge the second class to include those methods that only *approximately* carry the appropriate semantics, e.g. because they are only approximately Bayesian, or only Bayesian for a particular quantity of interest or up to a finite time horizon, e.g. the filtering-based solvers for ODEs (Kersting and Hennig 2016; Kersting et al. 2018; Schober et al. 2014, 2018).

Note that the second class of methods can also be pragmatically motivated, in the sense that formal statistical semantics enable techniques such as ANOVA to be brought to bear on the design and optimisation of a computational pipeline (to target the aspect of the computation that contributes most to uncertainty in the computational output) (Hennig et al. 2015). In this respect, statistical techniques can in principle supplement the expertise that is typically provided by a numerical analyst.

We note that paradigm (P1), with its close relationship to the longer-established field of ACA, tends to be more palatable to the classical numerical analysis community. The typical, rather than worst-case, performance of a numerical method is of obvious practical interest (Trefethen 2008). Statisticians, especially practitioners of Bayesian and fiducial inference, are habitually more comfortable with paradigm (P2) than numerical analysts are. As we remark in Sect. 4.5, this difference stems in part from a difference of opinion in which quantities are / can be regarded as "random" by the two communities; this difference of opinion affects (P2) much more strongly than (P1).

3.2 Bayesian probabilistic numerical methods

A recent research direction, which provides formal foundations for the approach pioneered by Larkin, is to interpret both traditional numerical methods and probabilistic numerical methods as particular solutions to an *ill-posed inverse problem* (Cockayne et al. 2019a). Given that the latent quantities involved in numerical tasks are frequently functions,

this development is in accordance with recent years' interest in non-parametric inversion in infinite-dimensional function spaces (Stuart 2010; Sullivan 2015).

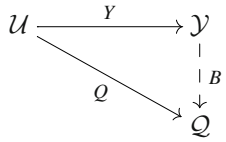
From the point of view of Cockayne et al. (2019a), which echoes IBC, the common structure of numerical tasks such as quadrature, optimisation, and the solution of an ODE or PDE, is the following:

- two known spaces: $\mathcal{U}$, where the unknown latent variable lives, and $\mathcal{Q}$, where the quantity of interest lives;
- and a known function $Q : \mathcal{U} \to \mathcal{Q}$, a *quantity-of-interest* function;

and the traditional role of the numerical analyst is to select/design

- a space $\mathcal{Y}$, where data about the latent variable live;
- and two functions: $Y : \mathcal{U} \to \mathcal{Y}$, an *information operator* that acts on the latent variable to yield information, and $B : \mathcal{Y} \to \mathcal{Q}$ such that $B \circ Y \approx Q$ in some sense to be determined.

With respect to this final point, Larkin (1970) observed that there are many senses in which $B \circ Y \approx Q$. One might ask, as Gaussian quadrature does, that the residual operator $R := B \circ Y - Q$ vanish on a large enough finite-dimensional subspace of $\mathcal{U}$; one might ask, as worst-case analysis does, that R be small in the supremum norm (Sard 1949); one might ask, as ACA does, that R be small in some integral norm against a probability measure on $\mathcal{U}$. In the chosen sense, numerical methods aim to make the following diagram approximately commute[9]:



(6)

A statistician might say that a deterministic numerical method $B : \mathcal{Y} \to \mathcal{U}$ as described above uses observed data $y := Y(u)$ to give a *point estimator* $B(y) \in \mathcal{Q}$ for a quantity of interest $Q(u) \in \mathcal{Q}$ derived from a latent variable $u \in \mathcal{U}$.

Example 1 The general structure is exemplified by univariate quadrature, in which $\mathcal{U} := C^0([a, b]; \mathbb{R})$, the information operator

$$Y(u) := (t_j, u(t_j))_{j=1}^J \in \mathcal{Y} := ([a, b] \times \mathbb{R})^J,$$

corresponds to pointwise evaluation of the integrand at J given nodes $a \leq t_1 < \cdots < t_J \leq b$, and the quantity of interest is

$$Q(u) := \int_a^b u(t) \, \mathrm{d}t \in \mathcal{Q} := \mathbb{R}.$$

Thus, we are interested in the definite integral of u, and we estimate it using only the information $Y(u)$, which does not completely specify u. Notice that *some but not all* quadrature methods $B : \mathcal{Y} \to \mathcal{Q}$ construct an estimate of u and then exactly integrate this estimate; Gaussian quadrature does this by polynomially interpolating the observed data $Y(u)$; by way of contrast, vanilla Monte Carlo builds no such functional estimate of u, since its estimate for the quantity of interest,

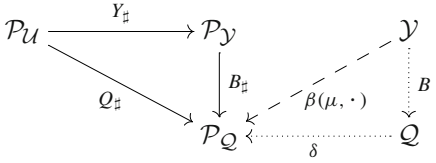
$$B_{\mathrm{MC}}\left((t_j, z_j)_{j=1}^J\right) = \frac{1}{J} \sum_{j=1}^J z_j, \tag{7}$$

forgets the locations t_j at which the integrand u was evaluated and uses only the values $z_j := u(t_j)$ of u. (Of course, the *accuracy* of B_{MC} is based on the assumption that the nodes t_j are uniformly distributed in $[a, b]$.)

This formal framework enables a precise definition of a probabilistic numerical method (PNM) to be stated (Cockayne et al. 2019a, Section 2). Assume that $\mathcal{U}$, $\mathcal{Y}$, and $\mathcal{Q}$ are measurable spaces, that Y and Q are measurable maps, and let $\mathcal{P}_{\mathcal{U}}$ etc. denote the corresponding sets of probability distributions on these spaces. Let $Q_\sharp : \mathcal{P}_{\mathcal{U}} \to \mathcal{P}_{\mathcal{Q}}$ denote the push-forward[10] of the map Q, and define $Y_\sharp$ etc. similarly.

Definition 1 A *probabilistic numerical method* for the estimation of a quantity of interest Q consists of an information operator $Y : \mathcal{U} \to \mathcal{Y}$ and a map $\beta : \mathcal{P}_{\mathcal{U}} \times \mathcal{Y} \to \mathcal{P}_{\mathcal{Q}}$, the latter being termed a *belief update operator*.

That is, given a belief μ about u, $\beta(\mu, \cdot)$ converts observed data $y \in \mathcal{Y}$ about u into a belief $\beta(\mu, y) \in \mathcal{P}_{\mathcal{Q}}$ about $Q(u)$, as illustrated by the dashed arrow in the following (not necessarily commutative) diagram:



(8)

As shown by the dotted arrows in (8), this perspective is general enough to contain classical numerical methods $B : \mathcal{Y} \to \mathcal{Q}$ as the special case $\beta(\mu, y) = \delta_{B(y)}$, where $\delta_q \in \mathcal{P}_{\mathcal{Q}}$ is the unit Dirac measure at $q \in \mathcal{Q}$.

[9] Recall that a diagram such as (6) or (9) is called *commutative* if all routes that follow the arrows (functions) from any starting point to any endpoint yield the same result. Thus, commutativity of (6) means exactly that $B(Y(u)) = Q(u)$ for all $u \in \mathcal{U}$.

[10] I.e. $Q_\sharp \mu(S) = \mu(Q^{-1}(S))$ for all measurable $S \subseteq \mathcal{Q}$

One desideratum for a PNM β is that its point estimators (e.g. mean, median, or mode) should be closely related to standard deterministic numerical methods B. This aspect is present in works such as Schober et al. (2014), which considers probabilistic ODE solvers with Runge–Kutta schemes as their posterior means, and Cockayne et al. (2016, 2017), which consider PDE solvers with the symmetric collocation method as the posterior mean. However, this aspect is by no means universally stressed.

A second, natural, desideratum for a PNM β is that the spread (e.g. the variance) of the distributional output should provide a fair reflection of the accuracy to which the quantity of interest is being approximated. In the statistics literature, this amounts to a deside for credible intervals to be *well calibrated* (Robins and van der Vaart 2006). In particular, one might desire that the distribution β contract to the true value of $Q(u)$ at an appropriate rate as the data dimension (e.g. the number of quadrature nodes) is increased.[11]

Diagram (6), when it commutes, characterises the "ideal" classical numerical method B; there is, as yet, no closed loop in diagram (8) involving β, which we would need in order to describe an "ideal" PNM β. This missing map in (8) is intimately related to the notion of a *Bayesian* PNM as defined by Cockayne et al. (2019a).

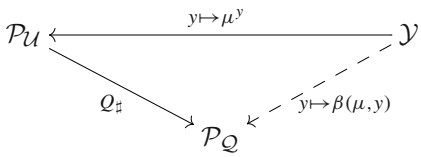
The key insight is that, given a prior belief expressed as a probability distribution $\mu \in \mathcal{P}_{\mathcal{U}}$ and the information operator $Y : \mathcal{U} \to \mathcal{Y}$, a Bayesian practitioner has a privileged map from $\mathcal{Y}$ into $\mathcal{P}_{\mathcal{U}}$ to add to diagram (8), namely the conditioning operator that maps any possible value $y \in \mathcal{Y}$ of the observed data to the corresponding conditional distribution $\mu^y \in \mathcal{P}_{\mathcal{U}}$ for u given y. In this situation, in contrast to the freedom[12] enjoyed by the designer of an arbitrary PNM, a Bayesian has no choice in her/his belief $\beta(\mu, y)$ about $Q(u)$: it must be nothing other than the image under Q of μ^y.

Definition 2 A probabilistic numerical method is said to be *Bayesian* for $\mu \in \mathcal{P}_{\mathcal{U}}$ if,

$$\beta(\mu, y) = Q_\sharp \mu^y \text{ for } Y_\sharp \mu\text{-almost all } y \in \mathcal{Y}.$$

In this situation μ is called a *prior* (for u) and $\beta(\mu, y)$ a *posterior* (for $Q(u)$).

In other words, being Bayesian means that the following diagram commutes:



(9)

Note that Definition 2 does not insist that a Bayesian PNM actually calculates μ^y and then computes the push-forward; only that the output of the PNM is equal to $Q_\sharp \mu^y$. Thus, whether or not a PNM is Bayesian is specific to the quantity of interest Q. Note also that a PNM $\beta(\mu, \cdot)$ can be Bayesian for some priors μ yet be non-Bayesian for other choices of μ; for details see Cockayne et al. (2019a, Sec. 5.2).

To be more formal for a moment, in Definition 2 the conditioning operation $y \mapsto \mu^y$ is interpreted in the sense of a *disintegration*, as advocated by Chang and Pollard (1997). This level of technicality is needed in order to make rigorous sense of the operation of conditioning on the μ-negligible event that $Y(u) = y$. Thus,

- for each $y \in \mathcal{Y}$, $\mu^y \in \mathcal{P}_{\mathcal{U}}$ is supported only on those values of u compatible with the observation $Y(u) = y$, i.e. $\mu^y(\{u \in \mathcal{U} \mid Y(u) \neq y\}) = 0$;
- for any measurable set $E \subseteq \mathcal{U}$, $y \mapsto \mu^y(E)$ is a measurable function from $\mathcal{Y}$ into $[0, 1]$ satisfying the *reconstruction property*, or *law of total probability*,

$$\mu(E) = \int_{\mathcal{Y}} \mu^y(E)\, (Y_\sharp \mu)(\mathrm{d}y).$$

Under mild conditions[13] such a disintegration always exists, and is unique up to modification on $Y_\sharp \mu$-null sets.

Observe that the fundamental difference between ACA (i.e. the probabilistic assessment of classical numerical methods) and Bayesianity of PNMs is that the former concerns the commutativity of diagram (6) in the average (i.e. the left-hand half of diagram (8)), whereas the latter concerns the commutativity of diagram (9).

The prime example of a Bayesian PNM is the following example of *kernel quadrature*, due to Larkin (1972):

Example 2 Recall the setup of Example 1. Take a Gaussian distribution μ on $C^0([a, b]; \mathbb{R})$, with mean function $m : [a, b] \to \mathbb{R}$ and covariance function $k : [a, b]^2 \to \mathbb{R}$. Then, given the data

$$y = (t_j, z_j)_{j=1}^J \equiv (t_j, u(t_j))_{j=1}^J,$$

[11] Here we abuse notation slightly: strictly speaking, we should refer not to one PNM β with input data y of varying dimension but to a *one-parameter family* of PNMs β_J parametrised by the data dimension J.

[12] The large and rapidly growing canon of PNMs, only some of which are cited in this article, is strong evidence of just how great this freedom is!

[13] Sufficient conditions are, e.g. that $\mathcal{U}$ be a complete and separable metric space with its Borel σ-algebra (so that every $\mu \in \mathcal{P}_{\mathcal{U}}$ is a Radon measure) and that the σ-algebra on $\mathcal{Y}$ be countably generated and contain all singletons.

the disintegration μ^y is again a Gaussian on $C^0([a,b];\mathbb{R})$ with mean and covariance functions

$$m^y(t) = m(t) + k_T(t)^\top k_{TT}^{-1}(z_T - m_T), \tag{10}$$
$$k^y(t,t') = k(t,t') - k_T(t)^\top k_{TT}^{-1} k_T(t'), \tag{11}$$

where $k_T : [a,b] \to \mathbb{R}^J$, $k_{TT} \in \mathbb{R}^{J\times J}$, $z_T \in \mathbb{R}^J$, and $m_T \in \mathbb{R}^J$ are given by

$$[k_T(t)]_j := k(t,t_j), \qquad [k_{TT}]_{i,j} := k(t_i,t_j),$$
$$[z_T]_j := z_j \equiv u(t_j), \qquad [m_T]_j := m(t_j).$$

The Bayesian PNM output $\beta(\mu, y)$, i.e. the push-forward $Q_\sharp \mu^y$, is a Gaussian on $\mathbb{R}$ with mean $\overline{m}^y$ and variance $(\overline{\sigma}^y)^2$ given by integrating (10) and (11) respectively, i.e.

$$\overline{m}^y = \int_a^b m(t)\,\mathrm{d}t + \left[\int_a^b k_T(t)\,\mathrm{d}t\right]^\top k_{TT}^{-1}(z_T - m_T),$$
$$(\overline{\sigma}^y)^2 = \int_a^b \int_a^b k(t,t')\,\mathrm{d}t\,\mathrm{d}t' - \left[\int_a^b k_T(t)\,\mathrm{d}t\right]^\top k_{TT}^{-1}\left[\int_a^b k_T(t')\,\mathrm{d}t'\right].$$

From a practical perspective, k is typically taken to have a parametric form k_θ and the parameters θ are adjusted in a data-dependent manner, for example to maximise the marginal likelihood of the information y under the Gaussian model.

One may also seek point sets that minimise the posterior variance $(\overline{\sigma}^y)^2$ of the estimate of the integral. For the Brownian covariance kernel $k(t,t') = \min(t,t')$, the posterior $Q_\sharp\mu = \mathcal{N}(\overline{m}^y, (\overline{\sigma}^y)^2)$ for $\int_a^b u(t)\,\mathrm{d}t$ is given by (4), the variance of which is clearly minimised by an equally spaced point set $\{t_j\}_{j=1}^J$. For more general kernels k, an early reference for selecting the point set $\{t_j\}_{j=1}^J$ to minimise $(\overline{\sigma}^y)^2$ is O'Hagan (1991).

This perspective, in which the Bayesian update is singled out from other possible belief updates, is reminiscent of foundational discussions such as those of Bissiri et al. (2016) and Zellner (1988). Interestingly, about half of the papers published on PN can be viewed as being (at least approximately) Bayesian; see the survey in the supplement of Cockayne et al. (2019a). This includes the work of Larkin, though, as previously mentioned, Larkin himself did not use the terminology of the Bayesian framework. Quite aside from questions of computational cost, non-Bayesian methods come into consideration because the requirement to be fully Bayesian can impose non-trivial constraints on the design of a practical numerical method, particularly for problems with a causal aspect or "time's arrow"; this point was discussed in detail for the numerical solution of ODEs by Wang et al. (2018).

As well as providing a clear formal benchmark, Cockayne et al. (2019a, Section 5) argue that a key advantage of Bayesian probabilistic numerical methods is that they are *closed under composition*, so that the output of a computational pipeline composed of Bayesian probabilistic numerical methods will inherit Bayesian semantics itself. This is analogous to the Markov condition that underpins directed acyclic graphical models (Lauritzen 1996) and may be an advantageous property in the context of large and/or distributed computational codes—an area where performing a classical numerical analysis can often be difficult. For non-Bayesian PNMs, it is unclear how these can/should be combined, but we note an analogous discussion of statistical "models made of modules" in the recent work of Jacob et al. (2017) [who observe, like Owhadi et al. (2015)], that strictly Bayesian models can be brittle under model misspecification, whereas non-Bayesianity confers additional robustness) and also the numerical analysis of probabilistic forward models in Bayesian inverse problems by Lie et al. (2018).

4 Discussion and outlook

> "Det er vanskeligt at spaa, især naar det gælder Fremtiden." [Danish proverb]

As it stands in 2019, our view is that there is much to be excited about. An intermittent stream of ad hoc observations and proposals, which can be traced back to the pioneering work of Larkin and Sul′din, has been unified under the banner of probabilistic numerics (Hennig et al. 2015) and solid statistical foundations have now been established (Cockayne et al. 2019a). In this section, we comment on some of the most important aspects of research that remain to be addressed.

4.1 Killer apps

The most successful area of research to date has been on the development of Bayesian methods for global optimisation (Snoek et al. 2012), which have become standard to the point of being embedded into commercial software (The MathWorks Inc. 2018) and deployed in realistic (Acerbi 2018; Paul et al. 2018) and indeed high-profile (Chen et al. 2018) applications. Other numerical tasks have yet to experience the same level of practical interest, though we note applications of probabilistic methods for cubature in computer graphics (Marques et al. 2013) and tracking (Prüher et al. 2018), as well as applications of probabilistic numerical methods in medical tractography (Hauberg et al. 2015) and nonlinear state estimation (Oates et al. 2019a) in an industrial context.

It has been suggested that probabilistic numerics is likely to experience the most success in addressing numerical tasks that are fundamentally difficult (Owen 2019). One area that

we highlight, in particular, in this regard is the solution of high-dimensional PDEs. There is considerable current interest in the deployment of neural networks as a substitute for more traditional numerical methods in this context, e.g. Sirignano and Spiliopoulos (2018), and the absence of interpretable error indicators for neural networks is a strong motivation for the development of more formal probabilistic numerical methods for this task. We note also that nonlinear PDEs in particular are prone to non-uniqueness of solutions. For some problems, physical reasoning may be used to choose among the various solutions, from the probabilistic or statistical perspective lack of uniqueness presents no fundamental philosophical issues: the multiple solutions are simply multiple maxima of a likelihood, and the prior is used to select among them, as in e.g. the treatment of Painlevé's transcendents by Cockayne et al. (2019a).

It has also been noted that the probabilistic approach provides a promising paradigm for the analysis of rounding error in mixed-precision calculations, where classical bounds "do not provide good estimates of the size of the error, and in particular […] overestimate the error growth, that is, the asymptotic dependence of the error on the problem size" (Higham and Mary 2018).

4.2 Adaptive Bayesian methods

The presentation of a PNM in Sect. 3.2 did not permit *adaptation*. It has been rigorously established that for *linear* problems adaptive methods (e.g. in quadrature, sequential selection of the notes t_j) do not outperform non-adaptive methods according to certain performance metrics such as worst-case error (Woźniakowski 1985, Section 3.2). However, adaptation is known to be advantageous in general for *nonlinear* problems (Woźniakowski 1985, Section 3.8). At a practical level, adaptation is usually an essential component in the development of stopping rules that enable a numerical method to terminate after an error indicator falls below a certain user-specified level. An analysis of adaptive PNMs would constitute a non-trivial generalisation of the framework of Cockayne et al. (2019a), who limited attention to static directed acyclic graph representation of conditional dependence structure. The generalisation to adaptive PNM necessitates the use of graphical models with a natural filtration, as exemplified by a dynamical Bayesian network (Murphy 2002).

It has been suggested that numerical analysis is a natural use case for *empirical Bayes methods* (Carlin and Louis 2000; Casella 1985), as opposed to related—but usually more computationally intensive—approaches such as hierarchical modelling and cross-validation. Empirical Bayes methods can be characterised as a specific instance of adaptation in which the observed data are used not only for inference but also to form a point estimator for the prior. For example, in a quadrature setting, the practitioner is in the fortunate position of being able to use evaluations of the integrand u both to estimate the regularity of u and the value of the integral. Empirical Bayesian methods are explored by Schober et al. (2018) and by Jagadeeswaran and Hickernell (2019) in this special issue.

4.3 Design of probabilistic numerical methods

Paradigmatic questions in the IBC literature are those of (i) an optimal information operator Y for a given task, and (ii) the optimal numerical method B for a given task, given information of a known type (Traub et al. 1983). In the statistical literature, there is also a long history of Bayesian optimal experimental design, in parametric and non-parametric contexts (Lindley 1956; Piironen 2005). The extent to which these principles can be used to design optimal numerical methods automatically (rather than by inspired guesswork on the mathematician's part, à la Larkin) remains a major open question, analogous to the automation of statistical reasoning envisioned by Wald and subsequent commentators on his work (Owhadi and Scovel 2017b).

4.4 Probabilistic programming

The theoretical foundations of probabilistic numerics have now been laid, but at present a library of compatible code has not been developed. In part, this is due to the amount of work needed in order to make a numerical implementation reliable and efficient, and in this respect PN lies far behind classical numerical analysis at present. Nevertheless, we anticipate that such efforts will be undertaken in coming years, and will lead to the wider adoption of probabilistic numerical methods. In particular, we are excited at the prospect of integrating probabilistic numerical methods into a probabilistic programming language, e.g. Carpenter et al. (2017), where tools from functional programming and category theory can be exploited in order to automatically compile codes built from probabilistic numerical methods (Ścibior et al. 2015).

4.5 Bridging the numerics–statistics gap

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

A major challenge faced by researchers in this area is the interdisciplinary gap between numerical analysts on the one hand and statisticians on the other. Though there are

some counterexamples, as a first approximation it is true to say that classically trained numerical analysts lack deep knowledge of probability or statistics, and classically trained statisticians are not well versed in numerical topics such as convergence and stability analysis. Indeed, not only do these two communities take interest in different questions, they often fail to even see the point of the other group's expertise and approaches to their common problems.

A caricature of this mutual incomprehension is the following: A numerical analyst will quite rightly point out that almost all problems have numerical errors that are provably non-Gaussian, not least because s/he can exhibit a rigorous a-priori or a-posteriori error bound. Therefore, to the numerical analyst it seems wholly inappropriate to resort to Gaussian models for any purpose at all; these are often the statistician's first models of choice, though they should not be the last. This non-paradox was explained in detail by Larkin (1974). (As a side note, it seems to us from our discussions that numerical analysts are happier to discuss the modelling of *errors* than the *latent quantities* which they regard as fixed, whereas statisticians seems to have the opposite preference; this is a difference in views that echoes the famous frequentist–subjectivist split in statistics.) The numerical analyst also wonders why, in the presence of an under-resolved integral, the practitioner does not simply apply an adaptive quadrature scheme and run it until an a posteriori global error indicator falls below a pre-set tolerance.

We believe that these difficulties are not fundamental and can be overcome by a more careful statement of the approach being taken to address the numerical task. In particular, the meeting ground for the numerical analysts and statisticians, and the critical arena of application for PN, consists of problems that *cannot* be run to convergence more cheaply than quantifying the uncertainties of the coarse solution—or, at least, where there is an interesting cost-v.-accuracy tradeoff to be had, which is a central enabling factor for multilevel methods (Giles 2015).

More generally, we are encouraged to see that epistemic uncertainty is being used once again and an analytical device in numerical analysis in the sense originally described by von Neumann and Goldstine (1947); see e.g. Higham and Mary (2018).

4.6 Summary

The first aim of this article was to better understand probabilistic numerics through its historical development. Aside from the pioneering work of Larkin, it was only in the 1990s that probabilistic numerical methods—i.e. algorithms returning a probability distribution as their output—were properly developed. A unified vision of probabilistic computation was powerfully presented by Hennig et al. (2015) and subsequently formalised by Cockayne et al. (2019a).

The second aim of this article was to draw a distinction between PN as a means to an end, as a form of probabilistic sensitivity / stability analysis, and PN as an end in itself. In particular, we highlighted the Bayesian subclass of PNMs as being closed under composition, a property that makes these particularly well suited for use in UQ; we also remarked that many problems—for reasons of problem structure, computational cost, or robustness to model misspecification—call for methods that are not formally Bayesian.

Finally, we highlighted areas for further development, which we believe will be essential if the full potential of probabilistic numerics highlighted by Hennig et al. (2015) is to be realised. From our perspective, the coming to fruition of this vision will require demonstrable success on problems that were intractable with the computational resources of previous decades and a wider acceptance of Larkin's observation quoted above, with which we wholeheartedly agree: numerical analysts and statisticians are indeed in the same business and do have much to offer one other!

Acknowledgements The authors wish to express their sincere thanks to Paul Constantine for highlighting the contribution of von Neumann and Goldstine (1947); to Matthew Larkin, Graham Larkin, Cherrilyn Yalin and Selim Akl for assisting with the historical content in Sect. 2.2 and giving permission to reproduce Fig. 2; to Sergey Mosin and Kazan Federal University for permission to reproduce Fig. 1; to Milena Kremakova for help with translating Norden et al. (1978) and Zabotin et al. (1996) from Russian into English and for assisting in communications with Kazan Federal University; and to Ilse Ipsen and four anonymous reviewers for their thoughtful comments. CJO was supported by the Lloyd's Register Foundation programme on Data-Centric Engineering at the Alan Turing Institute, London, UK. TJS was supported by the Freie Universität Berlin within the Excellence Initiative of the German Research Foundation (DFG), and by the DFG Collaborative Research Centre 1114 "Scaling Cascades in Complex Systems". This material was developed, in part, at the *Prob Num 2018* workshop hosted by the Lloyd's Register Foundation programme on Data-Centric Engineering at the Alan Turing Institute, UK, and supported by the National Science Foundation, USA, under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the above-named funding bodies and research institutions.

References

Abdulle, A., Garegnani, G.: Random time step probabilistic methods for uncertainty quantification in chaotic and geometric numerical integration (2018). arXiv:1801.01340

Acerbi, L.: Variational Bayesian Monte Carlo. In: 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)

(2018). https://papers.nips.cc/paper/8043-variational-bayesian-monte-carlo
Barlow, J.L., Bareiss, E.H.: Probabilistic error analysis of Gaussian elimination in floating point and logarithmic arithmetic. Computing **34**(4), 349–364 (1985). https://doi.org/10.1007/BF02251834
Bartels, S., Hennig, P.: Probabilistic approximate least-squares. In: Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, volume 51 of Proceedings of Machine Learning Research, pp. 676–684 (2016). http://proceedings.mlr.press/v51/bartels16.pdf
Bartels, S., Cockayne, J., Ipsen, I.C.F., Hennig, P.: Probabilistic linear solvers: a unifying view. Stat. Comput. (2019). https://doi.org/10.1007/s11222-019-09897-7
Berger, J.O.: Statistical Decision Theory and Bayesian Analysis. Springer Series in Statistics, 2nd edn. Springer, New York (1985). https://doi.org/10.1007/978-1-4757-4286-2
Bissiri, P.G., Holmes, C.C., Walker, S.G.: A general framework for updating belief distributions. J. R. Stat. Soc. Ser. B **78**(5), 1103–1130 (2016). https://doi.org/10.1111/rssb.12158
Briol, F.-X., Oates, C.J., Girolami, M., Osborne, M.A., Sejdinovic, D.: Probabilistic integration: A role in statistical computation? (with discussion and rejoinder). Stat. Sci. **34**(1), 1–22 (2019). https://doi.org/10.1214/18-STS660
Carlin, B.P., Louis, T.A.: Empirical Bayes: past, present and future. J. Am. Stat. Assoc. **95**(452), 1286–1289 (2000). https://doi.org/10.2307/2669771
Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: a probabilistic programming language. J. Stat. Softw. (2017). https://doi.org/10.18637/jss.v076.i01
Casella, G.: An introduction to empirical Bayes data analysis. Am. Stat. **39**(2), 83–87 (1985). https://doi.org/10.2307/2682801
Chang, J.T., Pollard, D.: Conditioning as disintegration. Stat. Neerl. **51**(3), 287–317 (1997). https://doi.org/10.1111/1467-9574.00056
Chatelin, F., Brunet, M.-C.: A probabilistic round-off error propagation model. Application to the eigenvalue problem. In: Reliable Numerical Computation, pp. 139–160. Oxford Sci. Publ., Oxford Univ. Press, New York (1990)
Chen, Y., Huang, A., Wang, Z., Antonoglou, I., Schrittwieser, J., Silver, D., de Freitas, N.: Bayesian optimization in AlphaGo (2018). arXiv:1812.06855
Chkrebtii, O.A., Campbell, D.A.: Adaptive step-size selection for state-space based probabilistic differential equation solvers. Stat. Comput. (2019). https://doi.org/10.1007/s11222-019-09899-5
Chkrebtii, O.A., Campbell, D.A., Calderhead, B., Girolami, M.A.: Bayesian solution uncertainty quantification for differential equations. Bayesian Anal. **11**(4), 1239–1267 (2016). https://doi.org/10.1214/16-BA1017
Cockayne, J., Oates, C., Sullivan, T.J., Girolami, M.: Probabilistic meshless methods for partial differential equations and Bayesian inverse problems (2016). arXiv:1605.07811
Cockayne, J., Oates, C., Sullivan, T.J., Girolami, M.: Probabilistic numerical methods for PDE-constrained Bayesian inverse problems. In: Verdoolaege, G. (ed.) Proceedings of the 36th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, vol. 1853 of AIP Conference Proceedings, pp. 060001–1–060001–8 (2017). https://doi.org/10.1063/1.4985359
Cockayne, J., Oates, C., Sullivan, T.J., Girolami, M.: Bayesian probabilistic numerical methods. SIAM Rev. (2019a). To appear. arXiv:1702.03673
Cockayne, J., Oates, C.J., Ipsen, I.C.F., Girolami, M.: A Bayesian conjugate gradient method. Bayesian Anal. (2019b). https://doi.org/10.1214/19-BA1145
Conrad, P.R., Girolami, M., Särkkä, S., Stuart, A., Zygalakis, K.: Statistical analysis of differential equations: introducing probability measures on numerical solutions. Stat. Comput. **27**(4), 1065–1082 (2017). https://doi.org/10.1007/s11222-016-9671-0
Cox, R.T.: Probability, frequency and reasonable expectation. Am. J. Phys. **14**(1), 1–13 (1946). https://doi.org/10.1119/1.1990764
Cox, R.T.: The Algebra of Probable Inference. The Johns Hopkins Press, Baltimore (1961)
Currin, C., Mitchell, T., Morris, M., Ylvisaker, D.: Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. J. Am. Stat. Assoc. **86**(416), 953–963 (1991). https://doi.org/10.1080/01621459.1991.10475138
Dale, A.I.: A History of Inverse Probability: From Thomas Bayes to Karl Pearson. Sources and Studies in the History of Mathematics and Physical Sciences, 2nd edn. Springer, New York (1999). https://doi.org/10.1007/978-1-4419-8652-8
Diaconis, P.: Bayesian numerical analysis. In: Statistical Decision Theory and Related Topics, IV, Vol. 1 (West Lafayette, Ind., 1986), pp. 163–175. Springer, New York (1988). https://doi.org/10.1007/978-1-4613-8768-8_20
Ehler, M., Gräf, M., Oates, C.J.: Optimal Monte Carlo integration on closed manifolds. Stat. Comput. (2019). https://doi.org/10.1007/s11222-019-09894-w
Giles, M.B.: Multilevel Monte Carlo methods. Acta Numer. **24**, 259–328 (2015). https://doi.org/10.1017/S096249291500001X
Ginoux, J.M., Gerini, C.: Henri Poincaré: A Biography Through the Daily Papers. World Scientific, Singapore (2013). https://doi.org/10.1142/8956
Greco, L., Racugno, W., Ventura, L.: Robust likelihood functions in Bayesian inference. J. Stat. Plann. Inference **138**(5), 1258–1270 (2008). https://doi.org/10.1016/j.jspi.2007.05.001
Gunter, T., Osborne, M.A., Garnett, R., Hennig, P., Roberts, S.J.: Sampling for inference in probabilistic models with fast Bayesian quadrature. In: Advances in Neural Information Processing Systems 27, pp. 2789–2797 (2014). URL https://papers.nips.cc/paper/5483-sampling-for-inference-in-probabilistic-models-with-fast-bayesian-quadrature
Hauberg, S., Schober, M., Liptrot, M., Hennig, P., Feragen, A.: A random Riemannian metric for probabilistic shortest-path tractography. vol. 9349 of Lecture Notes in Computer Science, pp. 597–604 (2015). https://doi.org/10.1007/978-3-319-24553-9_73
Hennig, P.: Roundtable in Tübingen (2014). http://www.probnum.org/2014/08/22/Roundtable-2014-in-Tuebingen/
Hennig, P.: Probabilistic interpretation of linear solvers. SIAM J. Optim. **25**(1), 234–260 (2015). https://doi.org/10.1137/140955501
Hennig, P., Kiefel, M.: Quasi-Newton methods: a new direction. J. Mach. Learn. Res. **14**(Mar), 843–865 (2013)
Hennig, P., Osborne, M.A., Girolami, M.: Probabilistic numerics and uncertainty in computations. Proc. R. Soc. A **471**(2179), 20150142 (2015). https://doi.org/10.1098/rspa.2015.0142
Henrici, P.: Discrete Variable Methods in Ordinary Differential Equations. Wiley, New York-London (1962)
Henrici, P.: Error Propagation for Difference Method. Wiley, New York-London (1963)
Higham, N.J., Mary, T.: A new approach to probabilistic rounding error analysis. Technical report, University of Manchester (2018). http://eprints.maths.manchester.ac.uk/2673/1/paper.pdf
Hollings, C.D.: Scientific Communication Across the Iron Curtain. Springer Briefs in History of Science and Technology. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-25346-6
Hull, T.E., Swenson, J.R.: Tests of probabilistic models for the propagation of roundoff errors. Commun. ACM **9**, 108–113 (1966). https://doi.org/10.1145/365170.365212
Jacob, P.E., Murray, L.M., Holmes, C.C., Robert, C.P.: Better together? Statistical learning in models made of modules (2017). arXiv:1708:08719

Jagadeeswaran, R., Hickernell, F.J.: Fast automatic Bayesian cubature using lattice sampling. Stat. Comput. (2019). https://doi.org/10.1007/s11222-019-09895-9

Kadane, J.B., Wasilkowski, G.W.: Average case ε-complexity in computer science. A Bayesian view. In: Bayesian Statistics, 2 (Valencia, 1983), pp. 361–374. North-Holland, Amsterdam (1985)

Karlin, S.: Best quadrature formulas and interpolation by splines satisfying boundary conditions. In: Approximations with Special Emphasis on Spline Functions (Proc. Sympos. Univ. of Wisconsin, Madison, Wis., 1969), pp. 447–466. Academic Press, New York (1969)

Karlin, S.: Best quadrature formulas and splines. J. Approx. Theory **4**, 59–90 (1971). https://doi.org/10.1016/0021-9045(71)90040-2

Karlin, S.: On a class of best nonlinear approximation problems. Bull. Am. Math. Soc. **78**, 43–49 (1972). https://doi.org/10.1090/S0002-9904-1972-12842-8

Karlin, S.: Studies in Spline Functions and Approximation Theory, Chapter on a Class of Best Nonlinear Approximation Problems and Extended Monosplines, pp. 19–66. Academic Press, New York (1976)

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

Karvonen, T., Kanagawa, M., Särkkä, S.: On the positivity and magnitudes of Bayesian quadrature weights. Stat. Comput. (2019a). https://doi.org/10.1007/s11222-019-09901-0

Karvonen, T., Särkkä, S., Oates, C.J.: Symmetry exploits for Bayesian cubature methods. Stat. Comput. (2019b). https://doi.org/10.1007/s11222-019-09896-8

Kazan Federal University (2018). https://kpfu.ru/portal/docs/F_261937733/suldin2.jpg. Accessed December 2018

Kersting, H., Hennig, P.: Active uncertainty calibration in Bayesian ODE solvers. In: Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016), pp. 309–318 (2016). http://www.auai.org/uai2016/proceedings/papers/163.pdf

Kersting, H., Sullivan, T.J., Hennig, P.: Convergence rates of Gaussian ODE filters (2018). arXiv:1807.09737

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

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

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

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

Laplace, P.S.: Théorie Analytique des Probabilités. Courcier, Paris (1812)

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

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

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

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

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

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

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

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

Lauritzen, S.L.: Graphical Models. Oxford Statistical Science Series, 17th edn. Oxford University Press, New York (1996)

Le Maître, O.P., Knio, O.M.: Spectral Methods for Uncertainty Quantification. Scientific Computation. Springer, New York (2010). https://doi.org/10.1007/978-90-481-3520-2

Lie, H.C., Sullivan, T.J., Teckentrup, A.L.: Random forward models and log-likelihoods in Bayesian inverse problems. SIAM/ASA J. Uncertain. Quantif. **6**(4), 1600–1629 (2018). https://doi.org/10.1137/18M1166523

Lie, H.C., Stuart, A.M., Sullivan, T.J.: Strong convergence rates of probabilistic integrators for ordinary differential equations. Stat. Comput. (2019). https://doi.org/10.1007/s11222-019-09898-6

Lindley, D.V.: On a measure of the information provided by an experiment. Ann. Math. Stat. **27**, 986–1005 (1956). https://doi.org/10.1214/aoms/1177728069

Mahsereci, M., Hennig, P.: Probabilistic line searches for stochastic optimization. In: Advances in Neural Information Processing Systems 28, pp. 181–189 (2015). https://papers.nips.cc/paper/5753-probabilistic-line-searches-for-stochastic-optimization

Manber, U., Norvig, P.: The power of the Apollo missions in a single Google search (2012). https://search.googleblog.com/2012/08/the-power-of-apollo-missions-in-single.html

Marques, R., Bouville, C., Ribardiere, M., Santos, L.P., Bouatouch, K.: A spherical Gaussian framework for Bayesian Monte Carlo rendering of glossy surfaces. IEEE Trans. Vis. Comput. Graph. **19**(10), 1619–1632 (2013). https://doi.org/10.1109/TVCG.2013.79

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

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

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

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

Murphy, K.P.: Dynamic Bayesian networks: representation, inference and learning. PhD thesis, University of California, Berkeley (2002)

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

Novak, E.: Deterministic and Stochastic Error Bounds in Numerical Analysis. Lecture Notes in Mathematics, vol. 1349. Springer, Berlin (1988). https://doi.org/10.1007/BFb0079792

Oates, C., Niederer, S., Lee, A., Briol, F.-X., Girolami, M.: Probabilistic models for integration error in the assessment of functional cardiac models. In: Advances in Neural Information Process-

ing Systems 30, pp. 110–118 (2017). http://papers.nips.cc/paper/6616-probabilistic-models-for-integration-error-in-the-assessment-of-functional-cardiac-models

Oates, C.J., Cockayne, J., Aykroyd, R.G., Girolami, M.: Bayesian probabilistic numerical methods in time-dependent state estimation for industrial hydrocyclone equipment. J. Am. Stat. Assoc. (2019a). https://doi.org/10.1080/01621459.2019.1574583

Oates, C.J., Cockayne, J., Prangle, D., Sullivan, T.J., Girolami, M.: Optimality criteria for probabilistic numerical methods. In: Multivariate Algorithms and Information-Based Complexity, Linz, 2018 (2019b). arXiv:1901.04326

O'Hagan, A.: Curve fitting and optimal design for prediction. J. R. Stat. Soc. Ser. B **40**(1), 1–42 (1978). https://doi.org/10.1111/j.2517-6161.1978.tb01643.x

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

Osborne, M.: Tübingen manifesto: uncertainty (2014a). URL http://probabilistic-numerics.org/2014/08/27/Roundtable-Uncertainty/

Osborne, M.: Tübingen manifesto: probabilistic numerics and probabilistic programming (2014b). http://probabilistic-numerics.org/2014/09/01/Roundtable-ProbNum-ProbProg/

Osborne, M.: Tübingen manifesto: priors and prior work (2014c). http://probabilistic-numerics.org/2014/08/27/Roundtable-Uncertainty/

Osborne, M.: Tübingen manifesto: community (2014d). http://probabilistic-numerics.org/2014/09/05/Roundtable-Community/

Osborne, M., Garnett, R., Ghahramani, Z., Duvenaud, D.K., Roberts, S.J., Rasmussen, C.E.: Active learning of model evidence using Bayesian quadrature. In: Advances in Neural Information Processing Systems 25, pp. 46–54 (2012a). https://papers.nips.cc/paper/4657-active-learning-of-model-evidence-using-bayesian-quadrature

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

Owen, A.: Unreasonable effectiveness of Monte Carlo. Stat. Sci. **34**, 29–33 (2019). https://doi.org/10.1214/18-STS676

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

Owhadi, H.: Multigrid with rough coefficients and multiresolution operator decomposition from hierarchical information games. SIAM Rev. **59**(1), 99–149 (2017). https://doi.org/10.1137/15M1013894

Owhadi, H., Scovel, C.: Universal scalable robust solvers from computational information games and fast eigenspace adapted multiresolution analysis (2017a). arXiv:1703.10761

Owhadi, H., Scovel, C.: Toward Machine Wald. In: Handbook of Uncertainty Quantification, pp. 157–191. Springer (2017b). https://doi.org/10.1007/978-3-319-12385-1_3

Owhadi, H., Zhang, L.: Gamblets for opening the complexity-bottleneck of implicit schemes for hyperbolic and parabolic ODEs/PDEs with rough coefficients. J. Comput. Phys. **347**, 99–128 (2017). https://doi.org/10.1016/j.jcp.2017.06.037

Owhadi, H., Scovel, C., Sullivan, T.J.: Brittleness of Bayesian inference under finite information in a continuous world. Electron. J. Stat. **9**(1), 1–79 (2015). https://doi.org/10.1214/15-EJS989

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

Paul, S., Chatzilygeroudis, K., Ciosek, K., Mouret, J.-B., Osborne, M.A., Whiteson, S.: Alternating optimisation and quadrature for robust control. In: The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18) (2018)

Piiroinen, P.: Statistical Measurements, Experiments and Applications. PhD thesis, University of Helsinki (2005)

Poincaré, H.: Calcul des Probabilités. Georges Carré (1896)

Poincaré, H.: Calcul des Probabilités, 2nd edn. Gauthier-Villars, Paris (1912)

Prüher, J., Karvonen, T., Oates, C.J., Straka, O., Särkkä, S.: Improved calibration of numerical integration error in sigma-point filters (2018). arXiv:1811.11474

Queen's University at Kingston. Frederick Michael Larkin (1936–1982) (1982). https://grahamlarkin.files.wordpress.com/2018/12/fmlarkin_obit.pdf

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

Ritter, K.: Average-Case Analysis of Numerical Problems, vol. 1733 of Lecture Notes in Mathematics. Springer, Berlin (2000). https://doi.org/10.1007/BFb0103934

Robins, J., van der Vaart, A.: Adaptive nonparametric confidence sets. Ann. Stat. **34**(1), 229–253 (2006). https://doi.org/10.1214/009053605000000877

Sacks, J., Ylvisaker, N.D.: Designs for regression problems with correlated errors. Ann. Math. Stat. **37**, 66–89 (1966). https://doi.org/10.1214/aoms/1177699599

Sacks, J., Ylvisaker, D.: Designs for regression problems with correlated errors; many parameters. Ann. Math. Stat. **39**, 49–69 (1968). https://doi.org/10.1214/aoms/1177698504

Sacks, J., Ylvisaker, D.: Designs for regression problems with correlated errors. III. Ann. Math. Stat. **41**, 2057–2074 (1970a). https://doi.org/10.1214/aoms/1177696705

Sacks, J., Ylvisaker, D.: 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), pp. 115–136. Canad. Math. Congr., Montreal, Que. (1970b)

Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P.: Design and analysis of computer experiments. Stat. Sci. **4**(4), 409–435 (1989). https://doi.org/10.1214/ss/1177012413

Sard, A.: Best approximate integration formulas; best approximation formulas. Am. J. Math. **71**, 80–91 (1949). https://doi.org/10.2307/2372095

Sard, A.: Linear Approximation Number 9 in Mathematical Surveys. American Mathematical Society, Providence (1963). https://doi.org/10.1090/surv/009

Särkkä, S., Hartikainen, J., Svensson, L., Sandblom, F.: On the relation between Gaussian process quadratures and sigma-point methods. J. Adv. Inf. Fusion **11**(1), 31–46 (2016)

Schäfer, F., Sullivan, T.J., Owhadi, H.: Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity (2017). arXiv:1706.02205

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

Schober, M., Särkkä, S., Hennig, P.: A probabilistic model for the numerical solution of initial value problems. Stat. Comput. **29**(1), 99–122 (2018). https://doi.org/10.1007/s11222-017-9798-7

Schoenberg, I.J.: On monosplines of least deviation and best quadrature formulae. J. Soc. Ind. Appl. Math. Ser. B Numer. Anal. **2**(1), 144–170 (1965). https://doi.org/10.1137/0702012

Schoenberg, I.J.: On monosplines of least square deviation and best quadrature formulae. II. SIAM J. Numer. Anal. **3**(2), 321–328 (1966). https://doi.org/10.1137/0703025

Ścibior, A., Ghahramani, Z., Gordon, A.: Practical probabilistic programming with monads. ACM SIGPLAN Not. **50**(12), 165–176 (2015). https://doi.org/10.1145/2804302.2804317

Sirignano, J., Spiliopoulos, K.: DGM: A deep learning algorithm for solving partial differential equations. J. Comput. Phys. **375**, 1339–1364 (2018). https://doi.org/10.1016/j.jcp.2018.08.029

Skilling, J.: Bayesian solution of ordinary differential equations. In: Maximum Entropy and Bayesian Methods, pp. 23–37. Springer (1992). URL https://doi.org/10.1007/978-94-017-2219-3

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

Smith, R. C.: Uncertainty Quantification: Theory, Implementation, and Applications, vol. 12 of Computational Science and Engineering. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2014)

Snoek, J., Larochelle, H., Adams, R.P.: Practical Bayesian optimization of machine learning algorithms. In: Advances in Neural Information Processing Systems, pp. 2951–2959 (2012). https://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms

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

Stuart, A.M., Teckentrup, A.L.: Posterior consistency for Gaussian process approximations of Bayesian posterior distributions. Math. Comput. **87**(310), 721–753 (2018). https://doi.org/10.1090/mcom/3244

Sudakov, V.N.: Linear sets with quasi-invariant measure. Dokl. Akad. Nauk SSSR **127**, 524–525 (1959)

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

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

Sul′din, A.V.: The method of regression in the theory of approximation. Kazan. Gos. Univ. Uchen. Zap. **123**(hn. 6), 3–35 (1963a)

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

Sul′din, A.V.: The solution of equations by the method of conditional mean values. In: Kazan State Univ. Sci. Survey Conf. 1962 (Russian), pp. 85–87. Izdat. Kazan. Univ., Kazan (1963c)

Sul′din, A.V.: Curves and operators in a Hilbert space. Kazan. Gos. Univ. Uchen. Zap. **128**(2), 15–47 (1968)

Sul′din, A.V., Zabotin, V.I., Semenihina, N.P.: Certain operators in Hilbert space. Kazan. Gos. Univ. Uchen. Zap. **129**(4), 90–95 (1969)

Sullivan, T.J.: Introduction to Uncertainty Quantification, vol. 63 of Texts in Applied Mathematics. Springer, Berlin (2015). https://doi.org/10.1007/978-3-319-23395-6

Teymur, O., Zygalakis, K., Calderhead, B.: Probabilistic linear multistep methods. In: Advances in Neural Information Processing Systems 29 (2016). https://papers.nips.cc/paper/6356-probabilistic-linear-multistep-methods

Teymur, O., Lie, H.C., Sullivan, T.J., Calderhead, B.: Implicit probabilistic integrators for ODEs. In: 32nd Conference on Neural Information Processing Systems (NeurIPS 2018) (2018). http://papers.nips.cc/paper/7955-implicit-probabilistic-integrators-for-odes

The MathWorks Inc. Bayesian optimization algorithm (2018). https://uk.mathworks.com/help/stats/bayesian-optimization-algorithm.html. Accessed Dec 2018

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

Traub, J.F., Woźniakowsi, H.: A General Theory of Optimal Algorithms. ACM Monograph Series. Academic Press, Inc., New York-London (1980)

Traub, J.F., Wasilkowski, G.W., Woźniakowski, H.: Information, Uncertainty Complexity. Advanced Book Program. Addison-Wesley Publishing Company, Reading (1983)

Trefethen, L.N.: Is Gauss quadrature better than Clenshaw–Curtis? SIAM Rev. **50**(1), 67–87 (2008). https://doi.org/10.1137/060659831

Tronarp, F., Kersting, H., Särkkä, S., Hennig, P.: Probabilistic solutions to ordinary differential equations as non-linear Bayesian filtering: a new perspective (2019). https://doi.org/10.1007/s11222-019-09900-1

U.S. Department of Energy: Scientific Grand Challenges for National Security: The Role of Computing at the Extreme Scale (2009)

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

Wang, J., Cockayne, J., Oates, C.: On the Bayesian solution of differential equations. In: Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2018) (2018)

Woźniakowski, H.: A survey of information-based complexity. J. Complex. **1**(1), 11–44 (1985). https://doi.org/10.1016/0885-064X(85)90020-2

Xi, X., Briol, F.-X., Girolami, M.: Bayesian quadrature for multiple related integrals. In: Proceedings of the 35th International Conference on Machine Learning, vol. 80, pp. 5373–5382 (2018). http://proceedings.mlr.press/v80/xi18a/xi18a.pdf

Yoo, G.R., Owhadi, H.: De-noising by thresholding operator adapted wavelets. Stat. Comput. (2019). https://doi.org/10.1007/s11222-019-09893-x

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

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