Probability and numerics: A modern retrospective

T. J. Sullivan[1]

Warwick Mathematics Institute Colloquium
University of Warwick, UK, 22 January 2021

[1]Mathematics Institute and School of Engineering, University of Warwick, UK

Introduction

A role for probability in numerics?

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

Thanks to my co-conspirators since 2015: Ben Calderhead, Jon Cockayne, Mark Girolami, Philipp Hennig, Ilse Ipsen, Hans Kersting, Han Cheng Lie, Chris Oates, Art Owen, Dennis Prangle, Houman Owhadi, Florian Schäfer, Andrew Stuart, Aretha Teckentrup, Onur Teymur

Suppose that we have the ability to interrogate/evaluate a function $u\colon [0,1] \to \mathbb{R}$ pointwise (at finitely many points $t_1, \ldots, t_J$ in finite time), possibly with evaluation errors.

Tasks that we might be interested in performing — or quantities of interest — include:

- quadrature: find $q := \int_0^1 u(x)\,\mathrm{d}x$;
- interpolation: find $q\colon [0,1] \to \mathbb{R}$ such that $q(t_j) = u(t_j)$ for each $j = 1, \ldots, J$;
- approximation: find $q\colon [0,1] \to \mathbb{R}$ such that $q(t_j) \approx u(t_j)$ for each $j = 1, \ldots, J$, e.g. the closest such q to u in some norm;
- optimisation: find $q \in [0,1]$ such that $u(q) \leq u(x)$ for all $x \in [0,1]$;
- solution of an ODE with $u\colon [0,1] \times \mathbb{R} \to \mathbb{R}$ and $x_0 \in \mathbb{R}$: find $q\colon [0,1] \to \mathbb{R}$ such that $q'(t) = u(t, q(t))$ and $q(0) = x_0$.

These are all deterministic problems!

- 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 numerical method (Hennig et al., 2015; Oates and Sullivan, 2019).
- The history of such approaches goes back at least a century.
- The last decade has seen a surge of activity here, with simultaneous input from multiple scientific disciplines: mathematics, statistics, machine learning, and computer science.
- There have been advances on a broad front, with contributions ranging from general theory-building to practical implementations in real-world problems of interest.
- Over the same period, 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.

Historical developments

216. Je suppose que l'on sache *a priori* que la fonction $f(x)$ est développable, dans un certain domaine, suivant les puissances croissantes de x,

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

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

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

Nous connaissons par n observations

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

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

- In modern terms, Poincaré (1912, Ch. 25), in his *Calcul des Probabilités*, considered a (formal) Gaussian prior distribution on a function f, i.e. a randomised power series

$$f(x) = \sum_{k=0}^{\infty} A_k x^k, \quad A_k \sim \mathcal{N}\left(0, \frac{1}{\sqrt{2h_k}}\right).$$

  Given n pointwise observations of the values of f, one seeks the probable values of $f(x)$ for another (not yet observed) value of x.
- This analytical treatment predates the first digital multipurpose computers and rigorous Gaussian measure theory by decades, yet it clearly illustrates a non-trivial probabilistic perspective on interpolation, a hybrid approach that is entirely in keeping with Poincaré's reputation as one of the last universalist mathematicians (Ginoux and Gerini, 2013).

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

  "[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)

- Concerning the numerical solution of 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. (Side note: There are actually non-IEEE computing paradigms in which is is true by design!)
- 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 numerical method and did not attempt to statistically motivate the numerical method itself.

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

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

- 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 — every probabilist's first-choice for "a random continuous function").
- 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 Cockayne et al. (2019a), 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 $\mathsf{B}\colon \mathcal{Y} \to \mathbb{R}$ that minimises the mean squared error (MSE)

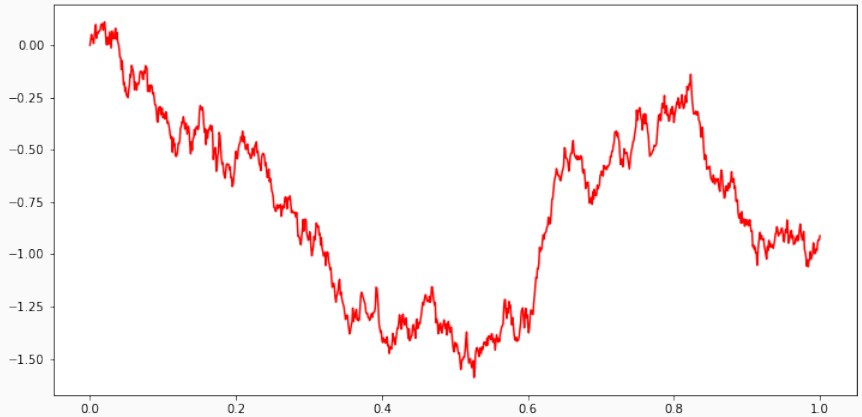
$$\int_{\mathcal{U}} \left| \int_a^b u(t)\,\mathrm{d}t - \mathsf{B}\big((t_j, u(t_j))_{j=1}^J\big) \right|^2 \mu_W(\mathrm{d}u) \tag{1}$$

is the classical trapezoidal rule

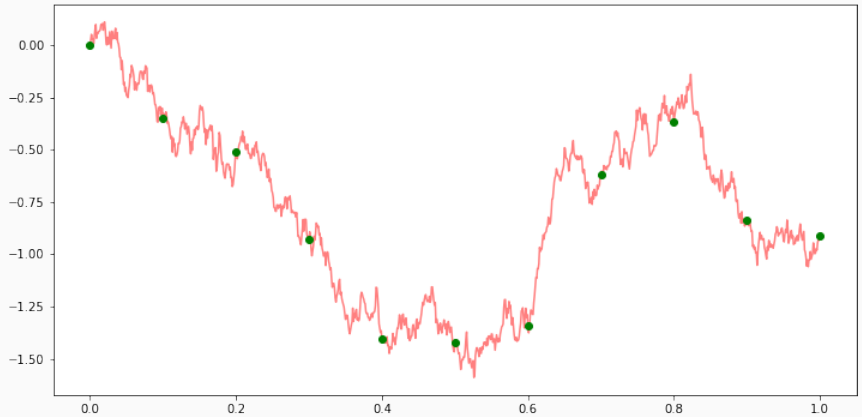
$$\mathsf{B}_{\mathrm{tr}}\big((t_j, z_j)_{j=1}^J\big) := \frac{1}{2}\sum_{j=1}^{J-1}(z_{j+1}+z_j)(t_{j+1}-t_j) = z_1\frac{t_2-t_1}{2} + \sum_{j=2}^{J-1} z_j \frac{t_{j+1}-t_{j-1}}{2} + z_J\frac{t_J - t_{J-1}}{2}, \tag{2}$$

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

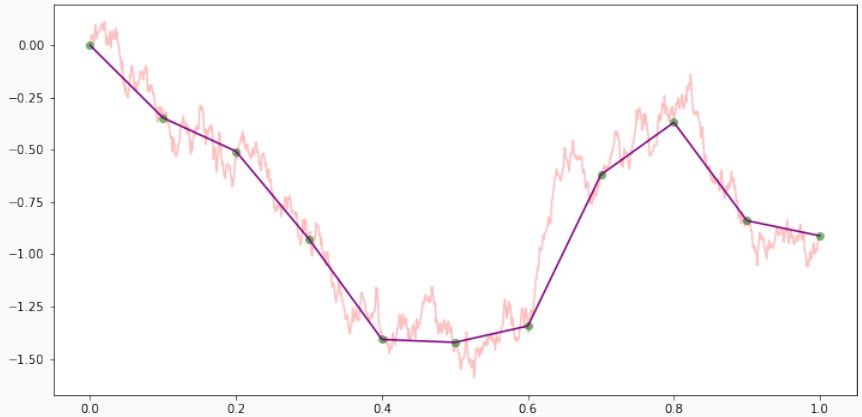
(One derivative smoother $\rightsquigarrow$ cubic spline quadrature.)



The estimator of $\int_a^b u(t)\,\mathrm{d}t$ with minimal MSE under a Brownian motion prior on u given pointwise evaluations is the trapezoidal rule, the definite integral of the piecewise linear interpolant of the observed data.



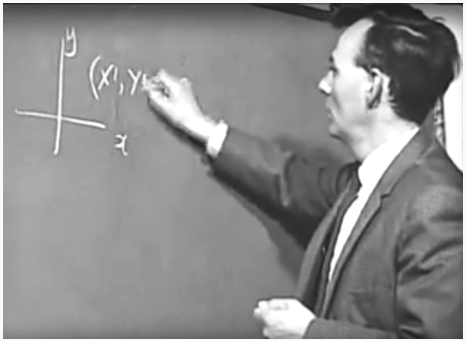
The estimator of $\int_a^b u(t)\,\mathrm{d}t$ with minimal MSE under a Brownian motion prior on u given pointwise evaluations is the trapezoidal rule, the definite integral of the piecewise linear interpolant of the observed data.



The estimator of $\int_a^b u(t)\,\mathrm{d}t$ with minimal MSE under a Brownian motion prior on u given pointwise evaluations is the trapezoidal rule, the definite integral of the piecewise linear interpolant of the observed data.

- Sul'din saw the connection between his methods and statistical regression (Sul'din, 1963a) and conditional probability (Sul'din, 1963c) — but did he consider his work to be an expression of statistical inference?
- Sul'din's methods were grounded in Hilbert space theory (Sul'din, 1968; Sul'din et al., 1969), so the underlying mathematics (the linear conditioning of Gaussian measures on Hilbert spaces) is linear algebra that can be motivated without recourse to probability.
- Sul'din *did* contribute something 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, as laid out in his 1969 habilitation thesis.

- On the other side of the Iron Curtain, between 1957 and 1969, Frederick Michael (“Mike”) Larkin (1936–1982) worked for the UK Atomic Energy Authority in its laboratories at Harwell and Culham, as well as working for two years at Rolls Royce.
- 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, 1979b,a,c), arguably laying the foundations on which modern PN would be developed.



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

- At Culham, Larkin worked on building some of the first graphical calculators, called GHOST (short for graphical output system) and GHOUL (graphical output language) — perhaps a motivation for seeking a richer description of numerical error.
- The perspective developed by Larkin was fundamentally statistical and, in modern terminology, the probabilistic numerical methods he developed would be described as *Bayesian* — though Larkin used the term *relative likelihood* for the prior.
- Larkin's perspective on quadrature: consider the Wiener measure as a prior, the information $(t_j, u(t_j))_{j=1}^{J}$ as (noiseless) data, and output the posterior marginal for $\int_a^b u(t)\,\mathrm{d}t$ — what we would now recognise as a probabilistic numerical method:
  "Among other things, this permits, at least in principle, the derivation of joint probability density functions for [both observed and unobserved] functionals on the space and also allows us to evaluate confidence limits on the estimate of a required functional (in terms of given values of other functionals)." — Larkin (1972)

- Sul'din describes the trapezoidal rule B_{tr} as a frequentist point estimator obtained from minimising the MSE (1), which "just happens" to produce an unbiased estimator with variance $\frac{1}{12}\sum_{j=1}^{J-1}(t_{j+1}-t_j)^3$.
- (Hence, the statistically optimal set of quadrature nodes is evenly spaced.)
- Larkin sees the normal distribution

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

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

- Larkin moved to Canada in 1969 to 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. [...] 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)

Sul'din ↔ Larkin?

- Sul'din and Larkin seem to have been working in parallel, with similar probabilistic perspectives on numerics, but limited to a Gaussian setting. It would not have been *easy* for Larkin and Sul'din to be conversant with each other's work (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.
- Soviet authors knew of Sul'din's influence on e.g. U. Grenander and W. Freiberger at Brown University, but make no mention of Larkin (Norden et al., 1978).
- Their ideas were ahead of their time: given the limited computational resources available at even cutting-edge facilities in the 1960s, the computational power needed to make PN a reality simply did not exist.[1]

[1] 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).

- The average-case analysis (ACA) of numerical methods built on the work of Kolmogorov (1936) and Sard (1963).
- In ACA the performance of a numerical method is assessed in terms of its *average error* with respect to a probability measure over the problem set; a prime example is univariate quadrature with the average quadratic loss (1) given earlier.
- A traditional (deterministic) NM can also be regarded as a statistical 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 the *expected loss* a.k.a. the *risk*. ACA is mathematically equivalent to Bayesian decision theory — restricted to the case of an experiment that produces a deterministic dataset (Kimeldorf and Wahba, 1970a,b; Parzen, 1970; Larkin, 1970).

- What about optimality of numerical methods, i.e. designs for data acquisition that yield minimal MSE among all possible designs of a given class, e.g. maximum number of function evaluations?
- Average-case optimal methods are **Bayes rules** or **Bayes acts** in the decision-theoretic context. Kadane and Wasilkowski (1985) showed that ACA-optimal methods coincide with (non-randomised) Bayes rules when the probability measure used to define the MSE is the Bayesian prior.
- Recently it has become clear that ACA optimality and the optimality of Bayesian *inferential* methods differ in general (Cockayne et al., 2019a; Oates et al., 2020). For example, if you're asked to guess whether or not a randomly-drawn card from a standard 52-card deck is ♦, out of the four equiprobable suits {♣, ♦, ♥, ♠}, ACA regards the information/questions "red or not?" and "♣ or not?" as equally optimal, whereas Bayesianity strictly prefers the 50-50 question.

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

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

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

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

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

  — Skilling (1992)

The last two decades have seen an explosion of interest in *uncertainty quantification* (UQ) for complex systems (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.

- **Quadrature**: 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); Xi et al. (2018); Ehler et al. (2019); Jagadeeswaran and Hickernell (2019); Karvonen et al. (2019a,b).
- **Optimisation**: Chen et al. (2018); Snoek et al. (2012), including probabilistic perspectives on quasi-Newton methods (Hennig and Kiefel, 2013) and line search methods (Mahsereci and Hennig, 2015).
- **Numerical linear algebra**: Bartels and Hennig (2016); Cockayne et al. (2019b); Hennig (2015); Bartels et al. (2019) 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.

- **ODEs**: approaches based on Gaussian filtering (Kersting and Hennig, 2016; Schober et al., 2014, 2018; Tronarp et al., 2019) and perturbation of dynamics and time steps (Abdulle and Garegnani, 2020; Chkrebtii et al., 2016; Conrad et al., 2017; Kersting et al., 2020; Teymur et al., 2018, 2016).

  A key result is the Bayesian optimality of evaluating f according to the classical Runge–Kutta scheme, and numerical-analysis-style convergence guarantees are being supplied (Conrad et al., 2017; Schober et al., 2018; Teymur et al., 2018; Lie et al., 2019; Kersting et al., 2020).

  Many applications here arise from machine learning algorithms whose *ideal* learning dynamics are ODEs in extremely high dimension (in both unknown parameters and training data) — the ideal vector field is subject to severe approximation, e.g. through random mini-batching, and fine time steps are impractical.

- **PDEs**: resent research includes (Owhadi, 2015; Chkrebtii et al., 2016; Cockayne et al., 2016, 2017), with these contributions making substantial use of RKHS structure and Gaussian processes.

  There are also recent statistical interpretations of finite element methods (Duffin et al., 2021; Girolami et al., 2021), in which non-conforming elements and "variational crimes" (Strang, 1972) correspond to misspecification of the statistical model.

  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., 2021; Yoo and Owhadi, 2019) — inversion of a dense $n \times n$ matrix in $O(n \operatorname{polylog} n)$ complexity.

- 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, and Larkin (1974) cites Karlin (1969) on this point.
- 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).

- The experimental design literature of the late 1960s and early 1970s, including a sequence of contributions from Sacks and Ylvisaker (1968, 1970a,b, 1966), 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 observed data $\{(t_j, Y(t_j))\}_{j=1}^{J}$.
- 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.

- The emulation of black-box functions, in the late 1970s and 1980s (O'Hagan, 1978; 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) cite Sul'din but not Larkin. The challenges of proposing a suitable stochastic process model for a deterministic function are discussed by Sacks et al. (1989) and Currin et al. (1991).

1. In the traditional setting of numerical analysis, c. 1950, all objects and operations are seen as being strictly deterministic. These deterministic objects are sometimes exceedingly complicated, to the extent that they may be treated as being stochastic.
2. Sard and Sul'din consider the questions of optimal performance of a numerical method in, respectively, the worst-case and the average-case context. Some of the average-case performance measures amount to variances of point estimators but are not *viewed* as such; probabilistic aspects are not a motivating factor.
3. Larkin's innovation, 1960s–1970s, is to formulate numerical tasks in terms of a joint distribution over latent quantities and quantities of interest; the quantity of interest is a stochastic object 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.

Bayesian probabilistic numerical methods come into focus

- 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 Bayesian inversion in infinite-dimensional spaces (Stuart, 2010).

From the point of view of Cockayne et al. (2019a), which echoes IBC and inverse problem theory (Stuart, 2010), the common structure of numerical tasks such as quadrature, optimisation, and the solution of an ODE or PDE, is the following:

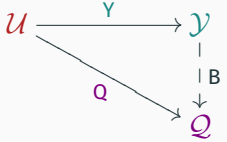
- two known spaces: $\mathcal{U}$, where the unknown latent variable lives, and $\mathcal{Q}$, where the quantity of interest lives;
- and a known function $\mathsf{Q} \colon \mathcal{U} \to \mathcal{Q}$, a quantity-of-interest function;

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: $\mathsf{Y} \colon \mathcal{U} \to \mathcal{Y}$, an information operator that acts on the latent variable to yield information, and a numerical method $\mathsf{B} \colon \mathcal{Y} \to \mathcal{Q}$ such that $\mathsf{B} \circ \mathsf{Y} \approx \mathsf{Q}$.

- E.g. Gaussian quadrature asks that the residual operator $R \coloneqq \mathsf{B} \circ \mathsf{Y} - \mathsf{Q}$ vanish on a large enough finite-dimensional subspace of $\mathcal{U}$.
- Worst-case analysis asks that R be small in the supremum norm (Sard, 1949).
- ACA asks that R be small in some integral norm against a probability measure on $\mathcal{U}$.

In the chosen sense, "good" NMs make the following diagram approximately commute:



(4)

A statistician might say that a deterministic NM $\mathsf{B}\colon \mathcal{Y} \to \mathcal{U}$ uses observed data $y \coloneqq \mathsf{Y}(u)$ to give a point estimator $\mathsf{B}(y) \in \mathcal{Q}$ for a quantity of interest $\mathsf{Q}(u) \in \mathcal{Q}$ derived from a latent variable $u \in \mathcal{U}$.

Consider, given nodes $a \leq t_1 < \cdots < t_J \leq b$,

$$\begin{aligned} \mathcal{U} &\coloneqq C^0([a,b];\mathbb{R}), \\ \mathsf{Q}(u) &\coloneqq \int_a^b u(t)\,\mathrm{d}t \in \mathcal{Q} \coloneqq \mathbb{R}, \\ \mathsf{Y}(u) &\coloneqq (t_j, u(t_j))_{j=1}^J \in \mathcal{Y} \coloneqq ([a,b]\times\mathbb{R})^J. \end{aligned}$$

Some but not all quadrature methods $\mathsf{B}\colon \mathcal{Y} \to \mathcal{Q}$ construct an estimate of u and then exactly integrate this estimate; Gaussian quadrature does this by polynomially interpolating the observed data $\mathsf{Y}(u)$; the vanilla Monte Carlo estimate,

$$\mathsf{B}_{\mathrm{MC}}\big((t_j,z_j)_{j=1}^J\big) = \frac{1}{J}\sum_{j=1}^J z_j,$$

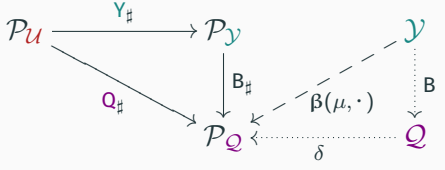
forgets the evaluation locations t_j and uses only the values $z_j \coloneqq u(t_j)$ of u.

Let $\mathcal{U}$, $\mathcal{Y}$, and $\mathcal{Q}$ be measurable spaces, let Y and Q be measurable maps, and let $\mathcal{P}_{\mathcal{U}}$ etc. denote the corresponding sets of probability distributions on these spaces. Let $\mathsf{Q}_\sharp \colon \mathcal{P}_{\mathcal{U}} \to \mathcal{P}_{\mathcal{Q}}$ denote the push-forward[2] of the map Q, and define $\mathsf{Y}_\sharp$ etc. similarly.

Definition 1 (Cockayne et al., 2019a, Section 2)

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

I.e., given a belief μ about u, $\boldsymbol{\beta}(\mu, \cdot)$ converts data $y \in \mathcal{Y}$ about u into a belief $\boldsymbol{\beta}(\mu, y) \in \mathcal{P}_{\mathcal{Q}}$ about $\mathsf{Q}(u)$, as illustrated by the dashed arrow:

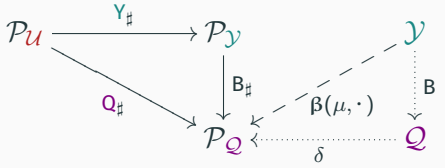


(5)

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

- *Some* PNMs $\boldsymbol{\beta}$ have point estimators (e.g. mean, median, or mode) that are closely related to standard deterministic numerical methods $\mathbf{B}$. This aspect is present in works of Schober et al. (2014), who consider probabilistic ODE solvers with Runge–Kutta schemes as their posterior means, and Cockayne et al. (2016, 2017), who consider PDE solvers with the symmetric collocation method as the posterior mean.
- Another desideratum for a PNM $\boldsymbol{\beta}$ is that the spread (e.g. the variance) of the distributional output should fairly reflect the accuracy of the approximation of the quantity of interest. In the statistics literature this amounts to a deside for credible intervals to be well calibrated (Robins and van der Vaart, 2006; Cockayne et al., 2020).

Diagram (4), when it commutes, characterises the "ideal" classical numerical method B; there is, as yet, no closed loop in the PNM diagram



which we would need in order to describe an "ideal" PNM $\boldsymbol{\beta}$. This missing map here is intimately related to the notion of a *Bayesian* PNM (Cockayne et al., 2019a).

Given a prior belief expressed as a probability distribution $\mu \in \mathcal{P}_{\mathcal{U}}$ and the information operator $\mathsf{Y} \colon \mathcal{U} \to \mathcal{Y}$, a Bayesian practitioner has a privileged map from $\mathcal{Y}$ into $\mathcal{P}_{\mathcal{U}}$ to add to diagram (5), i.e. conditioning.

Bayesian conditioning 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. A Bayesian has no choice in her/his belief $\beta(\mu, y)$ about $\mathsf{Q}(u)$: it must be nothing other than the image under Q of μ^y.

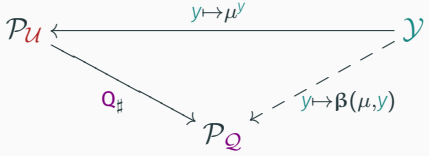
Definition 2

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

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

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

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



(6)

- A Bayesian PNM need not actually calculates μ^y and then compute the push-forward; we demand only that the output of the PNM is equal to $\mathsf{Q}_\sharp\mu^y$.
  Being Bayesian is specific to the quantity of interest Q — a PNM $\boldsymbol{\beta}(\mu, \cdot)$ can be Bayesian for some priors μ yet be non-Bayesian for other choices of μ.
  Interestingly, about half of the papers published on PN can be viewed as being (at least approximately) Bayesian.
- A key advantage of Bayesian PNMs is that they are closed under composition. 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): strictly Bayesian models can be brittle under model misspecification, whereas non-Bayesianity confers additional robustness.

The conditioning operation $y \mapsto \mu^y$ is interpreted in the sense of a disintegration (Chang and Pollard, 1997); this is needed in order to make rigorous sense of the operation of conditioning on the μ-negligible event that $\mathsf{Y}(u) = y$. 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 $\mathsf{Y}(u) = y$, i.e. $\mu^y(\{u \in \mathcal{U} \mid \mathsf{Y}(u) \neq y\}) = 0$;
- 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,

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

Under mild conditionssuch a disintegration always exists, and is unique up to modification on $\mathsf{Y}_\sharp\mu$-null sets.

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

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

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

$$m^y(t) = m(t) + k_T(t)^\top k_{TT}^{-1}(z_T - m_T), \tag{7}$$
$$k^y(t,t') = k(t,t') - k_T(t)^\top k_{TT}^{-1} k_T(t'), \tag{8}$$

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

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

Bayesian PNM output $\boldsymbol{\beta}(\mu, y) = \mathsf{Q}_\sharp \mu^y = \mathcal{N}(\overline{m}^y, (\overline{\sigma}^y)^2)$ with

$$\overline{m}^y = \int_a^b m(t)\,\mathrm{d}t + \left[\int_a^b k_T(t)\,\mathrm{d}t\right]^\top k_{TT}^{-1}(z_T - m_T),$$

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

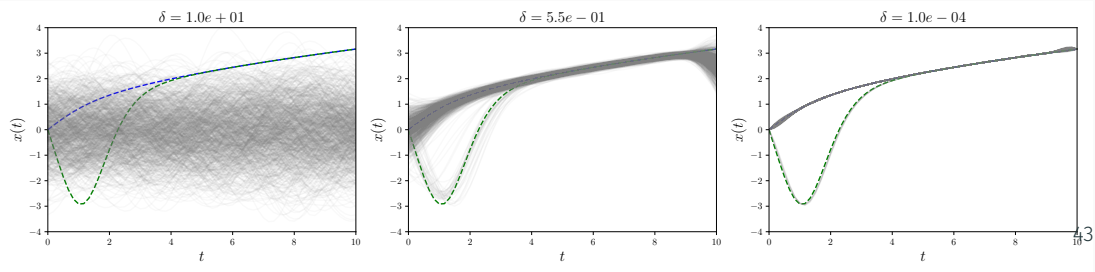
From a practical perspective, k is typically taken to have a parametric form k_θ and the parameters θ are adjusted in a data-dependent manner, for example to maximise the marginal likelihood of the information y under the Gaussian model.

For the Brownian covariance kernel $k(t,t') = \min(t,t')$, the posterior $\mathsf{Q}_\sharp \mu$ for $\int_a^b u(t)\,\mathrm{d}t$ is given by Larkin's trapezoidal rule, the variance of which is clearly minimised by an equally-spaced point set $\{t_j\}_{j=1}^J$. See O'Hagan (1991) for variance minimisation for more general kernels k.

- For PDEs that lack unique solutions, the Bayesian approach offers an attractive selection mechanism (Cockayne et al., 2019a).
- In the absence of any helpful structure, the computational implementation boils down to a statistical sampling problem. In this numerical disintegration approach, one enforces the PDE more strongly while sending a tempering parameter $\delta \to 0$.
- Example: Bayesian solution of Painlevé's first transcendental

$$u''(x) - u^2(x) = -x \quad + \text{boundary conditions}$$

with a centred Gaussian prior.



Discussion and outlook

- Greatest area of success to date is Bayesian global optimisation (Snoek et al., 2012; The MathWorks Inc.; Acerbi, 2018; Paul et al., 2018), a high-profile example being Bayesian optimisation in AlphaGo (Chen et al., 2018).
- Other applications of probabilistic methods for cubature in computer graphics (Marques et al., 2013; Xi et al., 2018) 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 in an industrial context (Oates et al., 2019).

- It has been suggested that PN 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 nonlinear PDEs that 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; see e.g. 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, 2019).

- The discussion above did not cover adaptive PNMs, e.g. sequential selection of integration nodes. This is a major open area.
- In the deterministic world, 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).
- How this interacts with Bayesianity and the composition of PNMs into pipelines is still open, as are connections to *empirical Bayes methods* (Carlin and Louis, 2000; Casella, 1985). Some early work in this direction includes Schober et al. (2018) and Jagadeeswaran and Hickernell (2019).

- The IBC literature intensively studies (i) optimal information operators Y for a given task, and (ii) 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).
- Open challenge: can these principles can be used to design optimal numerical methods automatically (rather than by inspired guesswork on the mathematician's part, à la Larkin)? Cf. the automation of statistical reasoning envisioned by Wald and subsequent commentators on his work (Owhadi and Scovel, 2017b).

- A major challenge is the interdisciplinary gap between numerical analysts and statisticians.
- Caricature: 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.
- 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.

- 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.
- A way forward: a more careful statement of the approach being taken to address the numerical task, e.g. variable precision, noisiness of output, ...
- 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, cf. the tradeoff in multilevel methods (Giles, 2015).

Closing remarks

Closing remarks

- Probabilistic approaches to numerical tasks have a long history, and keep coming around, especially as computer power advances.
- What appears to be new this time is more engagement between numerical analysts and statisticians, and computing paradigms that demand the crossover.
- Formal structures to describe PNMs and their relationship to (Bayesian) inference are now established.
- Standard libraries/implementations are starting to be developed.
- Are PNMs here to stay this time?

  "Det er vanskeligt at spaa, især naar det gælder Fremtiden."

  — *Danish proverb*

- Probabilistic approaches to numerical tasks have a long history, and keep coming around, especially as computer power advances.
- What appears to be new this time is more engagement between numerical analysts and statisticians, and computing paradigms that demand the crossover.
- Formal structures to describe PNMs and their relationship to (Bayesian) inference are now established.
- Standard libraries/implementations are starting to be developed.
- Are PNMs here to stay this time?

  "Det er vanskeligt at spaa, især naar det gælder Fremtiden."

  — *Danish proverb*

Thank You

References I

A. Abdulle and G. Garegnani. Random time step probabilistic methods for uncertainty quantification in chaotic and geometric numerical integration. *Stat. Comput.*, 30(4):907–932, 2020. doi:10.1007/s11222-020-09926-w.

L. Acerbi. 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.

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

S. Bartels and P. Hennig. Probabilistic approximate least-squares. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 676–684, 2016. http://proceedings.mlr.press/v51/bartels16.pdf.

S. Bartels, J. Cockayne, I. C. F. Ipsen, and P. Hennig. Probabilistic linear solvers: A unifying view. *Stat. Comput.*, 29(6): 1249–1263, 2019. doi:10.1007/s11222-019-09897-7.

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

F.-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic. Probabilistic integration: A role in statistical computation? (with discussion and rejoinder). *Stat. Sci.*, 34(1):1–22, 2019. doi:10.1214/18-STS660.

B. P. Carlin and T. A. Louis. Empirical Bayes: past, present and future. *J. Amer. Stat. Assoc.*, 95(452):1286–1289, 2000. doi:10.2307/2669771.

G. Casella. An introduction to empirical Bayes data analysis. *Amer. Stat.*, 39(2):83–87, 1985. doi:10.2307/2682801.

References II

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

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

Y. Chen, A. Huang, Z. Wang, I. Antonoglou, J. Schrittwieser, D. Silver, and N. de Freitas. Bayesian optimization in AlphaGo, 2018. arXiv:1812.06855.

O. A. Chkrebtii, D. A. Campbell, B. Calderhead, and M. A. Girolami. Bayesian solution uncertainty quantification for differential equations. *Bayesian Anal.*, 11(4):1239–1267, 2016. doi:10.1214/16-BA1017.

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

J. Cockayne, C. Oates, T. J. Sullivan, and M. Girolami. Probabilistic numerical methods for PDE-constrained Bayesian inverse problems. In G. Verdoolaege, editor, *Proceedings of the 36th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 1853 of *AIP Conference Proceedings*, pages 060001–1–060001–8, 2017. doi:10.1063/1.4985359.

J. Cockayne, C. Oates, T. J. Sullivan, and M. Girolami. Bayesian probabilistic numerical methods. *SIAM Rev.*, 61(4):756–789, 2019a. doi:10.1137/17M1139357.

J. Cockayne, C. J. Oates, I. C. F. Ipsen, and M. Girolami. A Bayesian conjugate gradient method. *Bayesian Anal.*, 2019b. doi:10.1214/19-BA1145.

References III

J. Cockayne, M. M. Graham, C. J. Oates, and T. J. Sullivan. Testing whether a learning procedure is calibrated, 2020. arXiv:2012.12670.

P. R. Conrad, M. Girolami, S. Särkkä, A. Stuart, and K. Zygalakis. Statistical analysis of differential equations: introducing probability measures on numerical solutions. *Stat. Comp.*, 27(4):1065–1082, 2017. doi:10.1007/s11222-016-9671-0.

C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *J. Amer. Stat. Assoc.*, 86(416):953–963, 1991. doi:10.1080/01621459.1991.10475138.

P. Diaconis. Bayesian numerical analysis. In *Statistical decision theory and related topics, IV, Vol. 1 (West Lafayette, Ind., 1986)*, pages 163–175, New York, 1988. Springer. doi:10.1007/978-1-4613-8768-8_20.

C. Duffin, E. Cripps, T. Stemler, and M. Girolami. Statistical finite elements for misspecified models. *Proc. Nat. Acad. Sci.*, 118 (2), 2021. doi:10.1073/pnas.2015006118.

M. Ehler, M. Gräf, and C. J. Oates. Optimal Monte Carlo integration on closed manifolds. *Stat. Comput.*, 29(6):1203–1214, 2019. doi:10.1007/s11222-019-09894-w.

M. B. Giles. Multilevel Monte Carlo methods. *Acta Numer.*, 24:259–328, 2015. doi:10.1017/S096249291500001X.

J. M. Ginoux and C. Gerini. *Henri Poincaré: A Biography Through the Daily Papers*. World Scientific, 2013. doi:10.1142/8956.

M. Girolami, E. Febrianto, G. Yin, and F. Cirak. The statistical finite element method (statFEM) for coherent synthesis of observation data and model predictions. *Comput. Methods Appl. Mech. Engrg.*, 375:113533, 2021. doi:10.1016/j.cma.2020.113533.

References iv

T. Gunter, M. A. Osborne, R. Garnett, P. Hennig, and S. J. Roberts. Sampling for inference in probabilistic models with fast Bayesian quadrature. In *Advances in Neural Information Processing Systems 27*, pages 2789–2797, 2014. https://papers.nips.cc/paper/5483-sampling-for-inference-in-probabilistic-models-with-fast-bayesian-quadrature.

S. Hauberg, M. Schober, M. Liptrot, P. Hennig, and A. Feragen. A random Riemannian metric for probabilistic shortest-path tractography. volume 9349 of *Lecture Notes in Computer Science*, pages 597–604. 2015. doi:10.1007/978-3-319-24553-9_73.

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

P. Hennig and M. Kiefel. Quasi-Newton methods: A new direction. *J. Mach. Learn. Research*, 14(Mar):843–865, 2013. http://www.jmlr.org/papers/volume14/hennig13a/hennig13a.pdf.

P. Hennig, M. A. Osborne, and M. Girolami. Probabilistic numerics and uncertainty in computations. *Proc. R. Soc. A*, 471 (2179):20150142, 2015. doi:10.1098/rspa.2015.0142.

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

P. Henrici. *Error Propagation for Difference Method*. John Wiley & Sons, Inc., New York-London, 1963.

N. J. Higham and T. Mary. A new approach to probabilistic rounding error analysis. *SIAM J. Sci. Comput.*, 41(5):A2815–A2835, 2019. doi:10.1137/18M1226312.

C. D. Hollings. *Scientific Communication Across the Iron Curtain*. SpringerBriefs in History of Science and Technology. Springer, Cham, 2016. doi:10.1007/978-3-319-25346-6.

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

References v

P. E. Jacob, L. M. Murray, C. C. Holmes, and C. P. Robert. Better together? Statistical learning in models made of modules, 2017. arXiv:1708:08719.

R. Jagadeeswaran and F. J. Hickernell. Fast automatic Bayesian cubature using lattice sampling. *Stat. Comput.*, 29(6): 1215–1229, 2019. doi:10.1007/s11222-019-09895-9.

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

S. Karlin. 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)*, pages 447–466. Academic Press, New York, 1969.

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

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

S. Karlin. *Studies in Spline Functions and Approximation Theory*, chapter On a class of best nonlinear approximation problems and extended monosplines, pages 19–66. Academic Press, New York, 1976.

T. Karvonen, C. J. Oates, and S. Särkkä. 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.

T. Karvonen, M. Kanagawa, and S. Särkkä. On the positivity and magnitudes of Bayesian quadrature weights. *Stat. Comput.*, 29(6):1317–1333, 2019a. doi:10.1007/s11222-019-09901-0.

References vi

T. Karvonen, S. Särkkä, and C. J. Oates. Symmetry exploits for Bayesian cubature methods. *Stat. Comput.*, 29(6):1231–1248, 2019b. doi:10.1007/s11222-019-09896-8.

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

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

H. Kersting, T. J. Sullivan, and P. Hennig. Convergence rates of Gaussian ODE filters. *Stat. Comput.*, 30(6):1791–1816, 2020. doi:10.1007/s11222-020-09972-4.

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

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

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

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

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

References vii

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

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

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

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

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

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

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

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

H. C. Lie, A. M. Stuart, and T. J. Sullivan. Strong convergence rates of probabilistic integrators for ordinary differential equations. *Stat. Comput.*, 29(6):1265–1283, 2019. doi:10.1007/s11222-019-09898-6.

References viii

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

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

U. Manber and P. Norvig. 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.

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

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

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

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

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

References ix

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

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

C. Oates, S. Niederer, A. Lee, F.-X. Briol, and M. Girolami. Probabilistic models for integration error in the assessment of functional cardiac models. In *Advances in Neural Information Processing Systems 30*, pages 110–118, 2017. http://papers.nips.cc/paper/6616-probabilistic-models-for-integration-error-in-the-assessment-of-functional-cardiac-models.

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

C. J. Oates, J. Cockayne, R. G. Aykroyd, and M. Girolami. Bayesian probabilistic numerical methods in time-dependent state estimation for industrial hydrocyclone equipment. *J. Amer. Stat. Assoc.*, 114(528):1518–1531, 2019. doi:10.1080/01621459.2019.1574583.

C. J. Oates, J. Cockayne, D. Prangle, T. J. Sullivan, and M. Girolami. Optimality criteria for probabilistic numerical methods. In F. J. Hickernell and P. Kritzer, editors, *Multivariate Algorithms and Information-Based Complexity*, volume 27 of *Radon Series on Computational and Applied Mathematics*, pages 65–88. Berlin/Boston: De Gruyter, 2020. doi:10.1515/9783110635461-005.

References x

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

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

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

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

A. Owen. Unreasonable effectiveness of Monte Carlo. *Stat. Sci.*, 2019.

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

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

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

H. Owhadi and C. Scovel. Toward Machine Wald. In *Handbook of Uncertainty Quantification*, pages 157–191. Springer International Publishing, 2017b. doi:10.1007/978-3-319-12385-1_3.

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

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

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

P. Piiroinen. *Statistical Measurements, Experiments and Applications*. PhD thesis, University of Helsinki, 2005.

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

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

Queen's University at Kingston. Frederick Michael Larkin (1936–1982), 11 Feb. 1982. https://grahamlarkin.files.wordpress.com/2018/12/fmlarkin_obit.pdf.

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

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

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

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

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

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

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

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

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

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

F. Schäfer, T. J. Sullivan, and H. Owhadi. Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity. *Multiscale Model. Simul.*, 2021. To appear. arXiv:1706.02205.

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

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

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

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

J. Skilling. Bayesian solution of ordinary differential equations. In *Maximum Entropy and Bayesian Methods*, pages 23–37. Springer, 1992. doi:10.1007/978-94-017-2219-3.

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

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

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

G. Strang. Variational crimes in the finite element method. In *The mathematical foundations of the finite element method with applications to partial differential equations (Proc. Sympos., Univ. Maryland, Baltimore, MD, 1972)*, pages 689–710, 1972.

References xiv

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

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

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

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

A. V. Sul'din. 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)*, pages 80–82. Izdat. Kazan. Univ., Kazan, 1963b.

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

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

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

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

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

O. Teymur, H. C. Lie, T. J. Sullivan, and B. Calderhead. 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. Accessed December 2018.

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

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

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

F. Tronarp, H. Kersting, S. Särkkä, and P. Hennig. Probabilistic solutions to ordinary differential equations as non-linear Bayesian filtering: A new perspective. *Stat. Comput.*, 29(6):1297–1315, 2019. doi: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.

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

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

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

G. R. Yoo and H. Owhadi. De-noising by thresholding operator adapted wavelets. *Stat. Comput.*, 29(6):1185–1201, 2019. doi:10.1007/s11222-019-09893-x.

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