Probabilistic Numerics

Probabilistic numerical computation formalises the connection between machine learning and applied mathematics. Numerical algorithms approximate intractable quantities from computable ones. They estimate integrals from evaluations of the integrand, or the path of a dynamical system described by differential equations from evaluations of the vector field. In other words, they infer a latent quantity from data. This book shows that it is thus formally possible to think of computational routines as learning machines, and to use the notion of Bayesian inference to build more flexible, efficient, or customised algorithms for computation.

The text caters for Masters' and PhD students, as well as postgraduate researchers in artificial intelligence, computer science, statistics, and applied mathematics. Extensive background material is provided along with a wealth of figures, worked examples, and exercises (with solutions) to develop intuition.

Philipp Hennig holds the Chair for the Methods of Machine Learning at the University of Tübingen, and an adjunct position at the Max Planck Institute for Intelligent Systems. He has dedicated most of his career to the development of Probabilistic Numerical Methods. Hennig's research has been supported by Emmy Noether, Max Planck and ERC fellowships. He is a co-Director of the Research Program for the Theory, Algorithms and Computations of Learning Machines at the European Laboratory for Learning and Intelligent Systems (ELLIS).

Michael A. Osborne is Professor of Machine Learning at the University of Oxford, and a co-Founder of Mind Foundry Ltd. His research has attracted £10.6M of research funding and has been cited over 15,000 times. He is very, very Bayesian.

Hans P. Kersting is a postdoctoral researcher at INRIA and École Normale Supérieure in Paris, working in machine learning with expertise in Bayesian inference, dynamical systems, and optimisation.

'This impressive text rethinks numerical problems through the lens of probabilistic inference and decision making. This fresh perspective opens up a new chapter in this field, and suggests new and highly efficient methods. A landmark achievement!'

– Zoubin Ghahramani, University of Cambridge

'In this stunning and comprehensive new book, early developments from Kac and Larkin have been comprehensively built upon, formalised, and extended by including modern-day machine learning, numerical analysis, and the formal Bayesian statistical methodology. Probabilistic numerical methodology is of enormous importance for this age of data-centric science and Hennig, Osborne, and Kersting are to be congratulated in providing us with this definitive volume.'

– Mark Girolami, University of Cambridge and The Alan Turing Institute

'This book presents an in-depth overview of both the past and present of the newly emerging area of probabilistic numerics, where recent advances in probabilistic machine learning are used to develop principled improvements which are both faster and more accurate than classical numerical analysis algorithms. A must-read for every algorithm developer and practitioner in optimization!'

– Ralf Herbrich, Hasso Plattner Institute

'Probabilistic numerics spans from the intellectual fireworks of the dawn of a new field to its practical algorithmic consequences. It is precise but accessible and rich in wide-ranging, principled examples. This convergence of ideas from diverse fields in lucid style is the very fabric of good science.'

– Carl Edward Rasmussen, University of Cambridge

'An important read for anyone who has thought about uncertainty in numerical methods; an essential read for anyone who hasn't'

– John Cunningham, Columbia University

'This is a rare example of a textbook that essentially founds a new field, re-casting numerics on stronger, more general foundations. A tour de force.'

– David Duvenaud, University of Toronto

'The authors succeed in demonstrating the potential of probabilistic numerics to transform the way we think about computation itself.'

– Thore Graepel, Senior Vice President, Altos Labs

PHILIPP HENNIG
Eberhard–Karls–Universität Tübingen, Germany

MICHAEL A. OSBORNE
University of Oxford

HANS P. KERSTING
École Normale Supérieure, Paris

PROBABILISTIC NUMERICS

COMPUTATION AS MACHINE LEARNING

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title:www.cambridge.org/9781107163447
DOI: 10.1017/9781316681411

First published 2022

Printed in the United Kingdom by TJ Books Limited, Padstow Cornwall

A catalogue record for this publication is available from the British Library.

ISBN 978-1-107-16344-7 Hardback

To our families.

Measurement owes its existence to Earth
Estimation of quantity to Measurement
Calculation to Estimation of quantity
Balancing of chances to Calculation
and Victory to Balancing of chances.

Sun Tzu – *The Art of War*
§4.18: Tactical Dispositions
Translation by Lionel Giles, 1910

Contents

Acknowledgements

Many people helped in the preparation of this book. We, the authors, extend our gratitude to the following people, without whom this book would have been impossible.

We are particularly grateful to Mark Girolami for his involvement during the early stages of this book as a project. Though he could not join as an author in the end, he provided a lot of support and motivation to make this book a reality.

We would like to deeply thank the many people who offered detailed and thoughtful comments on drafts of the book: Ondrej Bajgar, Nathanael Bosch, Jon Cockayne, Michael Cohen, Paul Duckworth, Nina Effenberger, Carl Henrik Ek, Giacomo Garegnani, Roman Garnett, Alexandra Gessner, Saad Hamid, Marius Hobbhahn, Toni Karvonen, Nicholas Krämer, Emilia Magnani, Chris Oates, Jonathan Schmidt, Sebastian Schulze, Thomas Schön, Arno Solin, Tim J. Sullivan, Simo Särkkä, Filip Tronarp, Ed Wagstaff, Xingchen Wan, and Richard Wilkinson.

Philipp Hennig

I would like to thank my research group, not just for thorough proof-reading, but for an intense research effort that contributed substantially to the results presented in this book. And, above all, for wagering some of their prime years, and their career, on me, and on the idea of probabilistic numerics:

Edgar Klenske, Maren Mahsereci, Michael Schober, Simon Bartels, Lukas Balles, Alexandra Gessner, Filip de Roos, Frank Schneider, Emilia Magnani, Niklas Wahl and Hans-Peter Wieser, Felix Dangel, Frederik Künstner, Jonathan Wenger, Agustinus Kristiadi, Nicholas Krämer, Nathanael Bosch, Lukas Tatzel, Thomas Gläßle, Julia Grosse, Katharina Ott, Marius Hobbhahn, Motonobu Kanagawa, Filip Tronarp, Robin Schmidt, Jonathan Schmidt, Marvin Pförtner, Nina Effenberger, and Franziska Weiler.

I am particularly grateful to those among this group who have contributed significantly to the development of the `probnum` library, which would not exist at all, not even in a minimal state, without their commitment.

Last but not least, I am grateful to my wife Maike Kaufman. And to my daughters Friederike und Iris. They both arrived while we worked on this book and drastically slowed down progress on it in the most wonderful way possible.

Michael A. Osborne

I would like to thank Isis Hjorth, for being the most valuable source of support I have in life, and our amazing children Osmund and Halfdan – I wonder what you will think of this book in a few years?

Hans P. Kersting

I would like to thank my postdoc adviser, Francis Bach, for giving me the freedom to allocate sufficient time to this book.

I am grateful to Dana Babin, my family, and my friends for their continuous love and support.

Symbols and Notation

Bold symbols ($\boldsymbol{x}$) are used for vectors, but only where the fact that a variable is a vector is relevant. Square brackets indicate elements of a matrix or vector: if $\boldsymbol{x} = [x_1, \dots, x_N]$ is a row vector, then $[\boldsymbol{x}]_i = x_i$ denotes its entries; if $A \in \mathbb{R}^{n \times m}$ is a matrix, then $[A]_{ij} = A_{ij}$ denotes its entries. Round brackets $(\cdot)$ are used in most other cases (as in the notations listed below).

Notation	Meaning
$a \propto c$	a is **proportional to** c: there is a constant k such that $a = k \cdot c$.
$A \wedge B$, $A \vee B$	The **logical conjunctions** "and" and "or"; i.e. $A \wedge B$ is true iff both A and B are true, $A \vee B$ is true iff $\neg A \wedge \neg B$ is false.
$A \otimes B$	The **Kronecker product** of matrices A, B. See Eq. (15.2).
$A \circledS B$	The **symmetric Kronecker product**. See Eq. (19.16).
$A \odot B$	The **element-wise product** (aka Hadamard product) of two matrices A and B of the same shape, i.e. $[A \odot B]_{ij} = [A]_{ij} \cdot [B]_{ij}$.
$\overrightarrow{A}$, $\natural\overrightarrow{A}$	$\overrightarrow{A}$ is the vector arising from stacking the elements of a matrix A row after row, and its inverse ($A = \natural\overrightarrow{A}$). See Eq. (15.1).
$\operatorname{cov}_p(x, y)$	The **covariance** of x and y under p. That is, $\operatorname{cov}_p(x, y) := \mathbb{E}_p(x \cdot y) - \mathbb{E}_p(x)\mathbb{E}_p(y)$.
$C^q(V, \mathbb{R}^d)$	The set of q-times **continuously differentiable functions** from V to $\mathbb{R}^d$, for some $q, d \in \mathbb{N}$.
$\delta(x - y)$	The **Dirac delta**, heuristically characterised by the property $\int f(x)\delta(x - y)\,dx = f(y)$ for functions $f : \mathbb{R} \to \mathbb{R}$.
δ_{ij}	The **Kronecker symbol**: $\delta_{ij} = 1$ if $i = j$, otherwise $\delta_{ij} = 0$.
$\det(A)$	The **determinant** of a square matrix A.
$\operatorname{diag}(\boldsymbol{x})$	The **diagonal matrix** with entries $[\operatorname{diag}(\boldsymbol{x})]_{ij} = \delta_{ij}[x]_i$.
$d\omega_t$	The notation for an **an Itô integral** in a **stochastic differential equation**. See Definition 5.4.
$\operatorname{erf}(x)$	The **error function** $\operatorname{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\,\mathrm{d}t$.
$\mathbb{E}_p(f)$	The **expectation** of f under p. That is, $\mathbb{E}_p(f) := \int f(x)\,dp(x)$.
$\mathbb{E}_{\mid Y}(f)$	The expectation of f under $p(f \mid Y)$.
$\Gamma(z)$	The **Gamma function** $\Gamma(z) := \int_0^\infty x^{z-1}\exp(-x)\,\mathrm{d}x$. See Eq. (6.1).
$\mathcal{G}(\cdot; a, b)$	The **Gamma distribution** with shape $a > 0$ and rate $b > 0$, with probability density function $\mathcal{G}(z; a, b) := \frac{b^a z^{a-1}}{\Gamma(a)} e^{-bz}$.
$\mathcal{GP}(f; \mu, k)$	The **Gaussian process** measure on f with mean function μ and covariance function (kernel) k. See §4.2
$\mathbb{H}_p(x)$	The (differential) **entropy** of the distribution $p(x)$. That is, $\mathbb{H}_p(x) := -\int p(x)\log p(x)\,\mathrm{d}x$. See Eq. (3.2).
$\mathbb{H}(x \mid y)$	The (differential) entropy of the cond. distribution $p(x \mid y)$. That is, $\mathbb{H}(x \mid y) := \mathbb{H}_{p(\cdot \mid y)}(x)$.
$I(x\,; y)$	The **mutual information** between random variables X and Y. That is, $I(x\,; y) := \mathbb{H}(x) - \mathbb{H}(x \mid y) = \mathbb{H}(y) - \mathbb{H}(y \mid x)$.

Notation	Meaning
I, I_N	The **identity** matrix (of dimensionality N): $[I]_{ij} = \delta_{ij}$.
$\mathbb{I}(\cdot \in A)$	The **indicator function** of a set A.
K_ν	The **modified Bessel function** for some parameter $\nu \in \mathbb{C}$. That is, $K_\nu(x) := \int_0^\infty \exp(-x \cdot \cosh(t)) \cosh(\nu t)\, \mathrm{d}t$.
$\mathcal{L}$	The **loss** function of an optimization problem (§26.1), or the log-likelihood of an inverse problem (§41.2).
$\mathcal{M}$	The **model** $\mathcal{M}$ capturing the probabilistic relationship between the latent object and computable quantities. See §9.3.
$\mathbb{N}, \mathbb{C}, \mathbb{R}, \mathbb{R}_+$	The **natural** numbers (excluding zero), the **complex** numbers, the **real** numbers, and the **positive real** numbers, respectively.
$\mathcal{N}(x; \mu, \Sigma) = p(x)$	The vector x has the **Gaussian probability density function** with mean vector μ and covariance matrix Σ. See Eq. (3.1).
$\mathcal{N}(\mu, \Sigma) \sim X$	The random variable X is distributed according to a Gaussian distribution with mean μ and covariance Σ.
$\mathcal{O}(\cdot)$	Landau **big-Oh**: for functions f, g defined on $\mathbb{N}$, the notation $f(n) = \mathcal{O}(g(n))$ means that $f(n)/g(n)$ is bounded for $n \to \infty$.
$p(y \mid x)$	The **conditional** the probability density function for variable Y having value y conditioned on variable X having value x.
$\mathrm{rk}(A)$	The **rank** of a matrix A.
$\mathrm{span}\{x_1, \dots, x_n\}$	The **linear span** of $\{x_1, \dots, x_n\}$.
$\mathrm{St}(\cdot; \mu, \lambda_1, \lambda_1)$	The **Student's-*t*** probability density function with parameters $\mu \in \mathbb{R}$ and $\lambda_1, \lambda_2 > 0$, see Eq. (6.9).
$\mathrm{tr}(A)$	The **trace** of matrix A, That is, $\mathrm{tr}(A) = \sum_i [A]_{ii}$.
$A^\intercal$	The **transpose** of matrix A: $[A^\intercal]_{ij} = [A]_{ji}$.
$\mathbb{U}_{a,b}$	The **uniform distribution** with probability density function $p(u) := \mathbb{I}(u \in (a, b))$, for $a < b$.
$\mathbb{V}_p(x)$	The **variance** of x under p. That is, $\mathbb{V}_p(x) := \mathrm{cov}_p(x, x)$.
$\mathbb{V}_{\mid Y}(f)$	The variance of f under $p(f \mid Y)$. That is, $\mathbb{H}(x \mid y) := -\int \log p(x \mid y)\, dp(x \mid y)$.
$\mathcal{W}(V, \nu)$	The **Wishart distribution** with probability density function $\mathcal{W}(x; V, \nu) \propto \lvert x \rvert^{(\nu-N-1)/2} e^{-1/2\, \mathrm{tr}(V^{-1}x)}$. See Eq. (19.1).
$x \perp y$	x is **orthogonal** to y, i.e. $\langle x, y \rangle = 0$.
$x := a$	The object x is **defined to be** equal to a.
$x \overset{\Delta}{=} a$	The object x is equal to a by virtue of its definition.
$x \leftarrow a$	The object x is **assigned the value** of a (used in pseudo-code).
$X \sim p$	The random variable X is **distributed according** to p.
$\mathbf{1}, \mathbf{1}_d$	A column **vector of** d **ones**, $\mathbf{1}_d := [1, \dots, 1]^\intercal \in \mathbb{R}^d$.
$\nabla_x f(x, t)$	The **gradient** of f w.r.t. x. (We omit subscript x if redundant.)