Variational Bayesian Learning Theory

Variational Bayesian learning is one of the most popular methods in machine learning. Designed for researchers and graduate students in machine learning, this book summarizes recent developments in the nonasymptotic and asymptotic theory of variational Bayesian learning and suggests how this theory can be applied in practice.

The authors begin by developing a basic framework with a focus on conjugacy, which enables the reader to derive tractable algorithms. Next, it summarizes nonasymptotic theory, which, although limited in application to bilinear models, precisely describes the behavior of the variational Bayesian solution and reveals its sparsity-inducing mechanism. Finally, the text summarizes asymptotic theory, which reveals phase transition phenomena depending on the prior setting, thus providing suggestions on how to set hyperparameters for particular purposes. Detailed derivations allow readers to follow along without prior knowledge of the mathematical techniques specific to Bayesian learning.

Shinichi Nakajima is a senior researcher at Technische Universität Berlin. His research interests include the theory and applications of machine learning, and he has published papers at numerous conferences and in journals such as *The Journal of Machine Learning Research, The Machine Learning Journal, Neural Computation*, and *IEEE Transactions on Signal Processing*. He currently serves as an area chair for Neural Information Processing Systems (NIPS) and an action editor for Digital Signal Processing.

Kazuho Watanabe is an associate professor at Toyohashi University of Technology. His research interests include statistical machine learning and information theory, and he has published papers at numerous conferences and in journals such as *The Journal of Machine Learning Research, The Machine Learning Journal, IEEE Transactions on Information Theory*, and *IEEE Transactions on Neural Networks and Learning Systems*.

Masashi Sugiyama is the director of the RIKEN Center for Advanced Intelligence Project and professor of Complexity Science and Engineering at the University of Tokyo. His research interests include the theory, algorithms, and applications of machine learning. He has written several books on machine learning, including *Density Ratio Estimation in Machine Learning*. He served as program cochair and general cochair of the NIPS conference in 2015 and 2016, respectively, and received the Japan Academy Medal in 2017.

Variational Bayesian Learning Theory

SHINICHI NAKAJIMA
Technische Universität Berlin

KAZUHO WATANABE
Toyohashi University of Technology

MASASHI SUGIYAMA
University of Tokyo

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

79 Anson Road, #06–04/06, Singapore 079906

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/9781107076150
DOI: 10.1017/9781139879354

First published 2019

Printed in the United Kingdom by TJ International Ltd, Padstow Cornwall

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

Library of Congress Cataloging-in-Publication Data
Names: Nakajima, Shinichi, author. | Watanabe, Kazuho, author. | Sugiyama, Masashi, 1974- author.
Title: Variational Bayesian learning theory / Shinichi Nakajima (Technische Universität Berlin), Kazuho Watanabe (Toyohashi University of Technology), Masashi Sugiyama (University of Tokyo).
Description: Cambridge ; New York, NY : Cambridge University Press, 2019. | Includes bibliographical references and index.
Identifiers: LCCN 2019005983| ISBN 9781107076150 (hardback : alk. paper) | ISBN 9781107430761 (pbk. : alk. paper)
Subjects: LCSH: Bayesian field theory. | Probabilities.
Classification: LCC QC174.85.B38 N35 2019 | DDC 519.2/33–dc23
LC record available at https://lccn.loc.gov/2019005983

ISBN 978-1-107-07615-0 Hardback

CAMBRIDGE

Cambridge University Press
978-1-107-07615-0 — Variational Bayesian Learning Theory
Shinichi Nakajima , Kazuho Watanabe , Masashi Sugiyama
Frontmatter
More Information

Contents

Preface

Bayesian learning is a statistical inference method that provides estimators and other quantities computed from the *posterior distribution*—the conditional distribution of unknown variables given observed variables. Compared with *point estimation* methods such as maximum likelihood (ML) estimation and maximum a posteriori (MAP) learning, Bayesian learning has the following advantages:

- Theoretically optimal.
  The posterior distribution is what we can obtain best about the unknown variables from observation. Therefore, Bayesian learning provides most accurate predictions, provided that the assumed model is appropriate.
- Uncertainty information is available.
  Sharpness of the posterior distribution indicates the reliability of estimators. The credible interval, which can be computed from the posterior distribution, provides probabilistic bounds of unknown variables.
- Model selection and hyperparameter estimation can be performed in a single framework.
  The marginal likelihood can be used as a criterion to evaluate how well a statistical model (which is typically a combination of model and prior distributions) fits the observed data, taking account of the flexibility of the model as a penalty.
- Less prone to overfitting.
  It was theoretically proven that Bayesian learning overfits the observation noise less than MAP learning.

On the other hand, Bayesian learning has a critical drawback—computing the posterior distribution is computationally hard in many practical models. This is because Bayesian learning requires *expectation* operations or integral computations, which cannot be analytically performed except for simple cases.

Accordingly, various approximation methods, including deterministic and sampling methods, have been proposed.

Variational Bayesian (VB) learning is one of the most popular deterministic approximation methods to Bayesian learning. VB learning aims to find the closest distribution to the Bayes posterior under some constraints, which are designed so that the expectation operation is tractable. The simplest and most popular approach is the *mean field approximation* where the approximate posterior is sought in the space of *decomposable* distributions, i.e., groups of unknown variables are forced to be independent of each other. In many practical models, Bayesian learning is intractable *jointly* for all unknown parameters, while it is tractable if the dependence between groups of parameters is ignored. Such a case often happens because many practical models have been constructed by combining simple models in which Bayesian learning is analytically tractable. This property is called *conditional conjugacy*, and makes VB learning computationally tractable.

Since its development, VB learning has shown good performance in many applications. Its good aspects and downsides have been empirically observed and qualitatively discussed. Some of those aspects seem inherited from full Bayesian learning, while some others seem to be artifacts by forced independence constraints. We have dedicated ourselves to theoretically clarifying the behavior of VB learning quantitatively, which is the main topic of this book.

This book starts from the formulation of Bayesian learning methods. In Part I, we introduce Bayesian learning and VB learning, emphasizing how conjugacy and conditional conjugacy make the computation tractable. We also briefly introduce other approximation methods and relate them to VB learning. In Part II, we derive algorithms of VB learning for popular statistical models, on which theoretical analysis will be conducted in the subsequent parts.

We categorize the theory of VB learning into two parts, and exhibit them separately. Part III focuses on *nonasymptotic* theory, where we do not assume the availability of a large number of samples. This analysis so far has been applied only to a class of *bilinear* models, but we can make detailed discussions including analytic forms of global solutions and theoretical performance guarantees. On the other hand, Part IV focuses on asymptotic theory, where the number of observed samples is assumed to be large. This approach has been applied to a broad range of statistical models, and successfully elucidated the *phase transition* phenomenon of VB learning. As a practical outcome, this analysis provides a guideline on how to set hyperparameters for different purposes.

Recently, a lot of variations of VB learning have been proposed, e.g., more accurate inference methods beyond the mean field approximation, stochastic gradient optimization for big data analysis, and sampling based update rules for automatic (black-box) inference to cope with general nonconjugate likelihoods including deep neural networks. Although we briefly introduce some of those recent works in Part I, they are not in the central scope of this book. We rather focus on the simplest mean field approximation, of which the behavior has been clarified quantitatively by theory.

This book was completed under the support by many people. Shinichi Nakajima deeply thanks Professor Klaus-Robert Müller and the members in Machine Learning Group in Technische Universität Berlin for their direct and indirect support during the period of book writing. Special thanks go to Sergej Dogadov, Hannah Marienwald, Ludwig Winkler, Dr. Nico Gönitz, and Dr. Pan Kessel, who reviewed chapters of earlier versions, found errors and typos, provided suggestions to improve the presentation, and kept encouraging him in proceeding book writing. The authors also thank Lauren Cowles and her team in Cambridge University Press, as well as all other staff members who contributed to the book production process, for their help, as well as their patience on the delays in our manuscript preparation. Lauren Cowles, Clare Dennison, Adam Kratoska, and Amy He have coordinated the project since its proposal, and Harsha Vardhanan in SPi Global has managed the copy-editing process with Andy Saff.

The book writing project was partially supported by the following organizations: the German Research Foundation (GRK 1589/1) by the Federal Ministry of Education and Research (BMBF) under the Berlin Big Data Center project (Phase 1: FKZ 01IS14013A and Phase 2: FKz 01IS18025A), the Japan Society for the Promotion of Science (15K16050), and the International Research Center for Neurointelligence (WPI-IRCN) at The University of Tokyo Institutes for Advanced Study.

CAMBRIDGE

Cambridge University Press
978-1-107-07615-0 — Variational Bayesian Learning Theory
Shinichi Nakajima , Kazuho Watanabe , Masashi Sugiyama
Frontmatter
More Information

Nomenclature

$a, b, c, \ldots, A, B, C, \ldots$	: Scalars.
$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \ldots$ (bold-faced small letters)	: Vectors.
$\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \ldots$ (bold-faced capital letters)	: Matrices.
$\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots$ (calligraphic capital letters)	: Tensors or sets.

$(\cdot)_{l,m}$	: (l, m)th element of a matrix.
$\top$	: Transpose of a matrix or vector.
$\mathrm{tr}(\cdot)$	: Trace of a matrix.
$\det(\cdot)$	: Determinant of a matrix.
$\odot$	: Hadamard (elementwise) product.
$\otimes$	: Kronecker product.
$\times_n$	: n-mode tensor product.
$\lvert\cdot\rvert$	: Absolute value of a scalar. It applies element-wise for a vector or matrix.
$\mathrm{sign}(\cdot)$	: Sign operator such that $\mathrm{sign}(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ -1 & \text{otherwise.} \end{cases}$ It applies elementwise for a vector or matrix.

$\{\cdots\}$	: Set consisting of specified entities.
$\{\cdots\}^D$	: Dfold Cartesian product, i.e., $\mathbb{X}^D \equiv \{(x_1, \ldots, x_D)^\top ; x_d \in \mathbb{X} \text{ for } d = 1, \ldots, D\}$.
$\#(\cdot)$	: Cardinality (the number of entities) of a set.

$\mathbb{R}$	: The set of all real numbers.
$\mathbb{R}_+$	: The set of all nonnegative real numbers.
$\mathbb{R}_{++}$	: The set of all positive real numbers.
$\mathbb{R}^D$	: The set of all D-dimensional real (column) vectors.

$[\cdot,\cdot]$: The set of real numbers in a range, i.e.,

$$[l,u] = \{x \in \mathbb{R}; l \le x \le u\}.$$

$[\cdot,\cdot]^D$: The set of D-dimensional real vectors whose entries are in a range, i.e., $[l,u]^D \equiv \{\boldsymbol{x} \in \mathbb{R}^D; l \le x_d \le u \text{ for } d = 1,\ldots,D\}$.

$\mathbb{R}^{L\times M}$: The set of all $L \times M$ real matrices.

$\mathbb{R}^{M_1\times M_2\times\cdots\times M_N}$: The set of all $M_1 \times M_2 \times \cdots \times M_N$ real tensors.

$\mathbb{I}$: The set of all integers.

$\mathbb{I}_{++}$: The set of all positive integers.

$\mathbb{C}$: The set of all complex numbers.

$\mathbb{S}^D$: The set of all $D \times D$ symmetric matrices.

$\mathbb{S}^D_+$: The set of all $D \times D$ positive semidefinite matrices.

$\mathbb{S}^D_{++}$: The set of all $D \times D$ positive definite matrices.

$\mathbb{D}^D$: The set of all $D \times D$ diagonal matrices.

$\mathbb{D}^D_+$: The set of all $D \times D$ positive semidefinite diagonal matrices.

$\mathbb{D}^D_{++}$: The set of all $D \times D$ positive definite diagonal matrices.

$\mathbb{H}^{K-1}_N$: The set of all possible histograms for N samples and K categories, i.e., $\mathbb{H}^{K-1}_N \equiv \{\boldsymbol{x} \in \{0,\ldots,N\}^K; \sum_{k=1}^K x_k = N\}$.

Δ^{K-1} : The standard $(K-1)$-simplex, i.e.,

$$\Delta^{K-1} \equiv \{\boldsymbol{\theta} \in [0,1]^K; \textstyle\sum_{k=1}^K \theta_k = 1\}).$$

$(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_M)$: Column vectors of $\boldsymbol{A}$, i.e., $\boldsymbol{A} = (\boldsymbol{a}_1,\ldots,\boldsymbol{a}_M) \in \mathbb{R}^{L\times M}$.

$(\widetilde{\boldsymbol{a}}_1,\ldots,\widetilde{\boldsymbol{a}}_L)$: Row vectors of $\boldsymbol{A}$, i.e., $\boldsymbol{A} = (\widetilde{\boldsymbol{a}}_1,\ldots,\widetilde{\boldsymbol{a}}_L)^\top \in \mathbb{R}^{L\times M}$.

$\mathbf{Diag}(\cdot)$: Diagonal matrix with specified diagonal elements, i.e.,

$$(\mathbf{Diag}(\boldsymbol{x}))_{l,m} = \begin{cases} x_l & \text{if } l = m, \\ 0 & \text{otherwise.} \end{cases}$$

$\mathbf{diag}(\cdot)$: Column vector consisting of the diagonal entries of a matrix, i.e.,

$$(\mathbf{diag}(\boldsymbol{X}))_l = X_{l,l}.$$

$\mathbf{vec}(\cdot)$: Vectorization operator concatenating all column vectors of a matrix into a long column vector, i.e., $\mathbf{vec}(\boldsymbol{A}) = (\boldsymbol{a}_1^\top,\ldots,\boldsymbol{a}_M^\top)^\top \in \mathbb{R}^{LM}$ for a matrix $\boldsymbol{A} = (\boldsymbol{a}_1,\ldots,\boldsymbol{a}_M) \in \mathbb{R}^{L\times M}$.

$\boldsymbol{I}_D$: D-dimensional $(D \times D)$ identity matrix.

$\boldsymbol{\Gamma}$: A diagonal matrix.

$\boldsymbol{\Omega}$: An orthogonal matrix.

$\boldsymbol{e}_k$: One of K expression, i.e., $\boldsymbol{e}_k = (\underbrace{0,\ldots,0,\overbrace{1}^{k\text{th}},0,\ldots,0}_{K})^\top \in \{0,1\}^K$.

$\boldsymbol{1}_K$: K-dimensional vector with all elements equal to one, i.e.,

$$\boldsymbol{e}_k = (\underbrace{1,\ldots,1}_{K})^\top.$$

$\mathrm{Gauss}_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: D-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

$\mathrm{MGauss}_{D_1,D_2}(\boldsymbol{M}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})$: $D_1 \times D_2$ dimensional matrix variate Gaussian distribution with mean $\boldsymbol{M}$ and covariance $\boldsymbol{\Sigma} \otimes \boldsymbol{\Psi}$.

$\mathrm{Gamma}(\alpha, \beta)$: Gamma distribution with shape parameter α and scale parameter β.

$\mathrm{InvGamma}(\alpha, \beta)$: Inverse-Gamma distribution with shape parameter α and scale parameter β.

$\mathrm{Wishart}_D(\boldsymbol{V}, \nu)$: D-dimensional Wishart distribution with scale matrix $\boldsymbol{V}$ and degree of freedom ν.

$\mathrm{InvWishart}_D(\boldsymbol{V}, \nu)$: D-dimensional inverse-Wishart distribution with scale matrix $\boldsymbol{V}$ and degree of freedom ν.

$\mathrm{Multinomial}(\boldsymbol{\theta}, N)$: Multinomial distribution with event probabilities $\boldsymbol{\theta}$ and number of trials N.

$\mathrm{Dirichlet}(\boldsymbol{\phi})$: Dirichlet distribution with concentration parameters $\boldsymbol{\phi}$.

$\mathrm{Prob}(\cdot)$: Probability of an event.

$p(\cdot), q(\cdot)$: Probability distribution (probability mass function for discrete random variables, and probability density function for continuous random variables). Typically p is used for a model distribution and q is used for the true distribution.

$r(\cdot)$: A trial distribution (a variable of a functional) for approximation.

$\langle f(\boldsymbol{x}) \rangle_{p(\boldsymbol{x})}$: Expectation value of $f(\boldsymbol{x})$ over distribution $p(\boldsymbol{x})$, i.e.,

$$\langle f(\boldsymbol{x}) \rangle_{p(\boldsymbol{x})} \equiv \int f(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}.$$

$\widehat{\cdot}$: Estimator for an unknown variable, e.g., $\widehat{\boldsymbol{x}}$ and $\widehat{\boldsymbol{A}}$ are estimators for a vector $\boldsymbol{x}$ and a matrix $\boldsymbol{A}$, respectively.

$\mathbf{Mean}(\cdot)$: Mean of a random variable.

$\mathrm{Var}(\cdot)$: Variance of a random variable.

$\mathbf{Cov}(\cdot)$: Covariance of a random variable.

$\mathrm{KL}(\cdot \| \cdot)$: Kullbuck–Leibler divergence between distributions, i.e.,

$$\mathrm{KL}(p \| q) \equiv \left\langle \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \right\rangle_{p(\boldsymbol{x})}.$$

$\delta(\boldsymbol{\mu}; \widehat{\boldsymbol{\mu}})$: Dirac delta function located at $\widehat{\boldsymbol{\mu}}$. It also denotes its approximation (called Pseudo-delta function) with its entropy finite.

GE : Generalization error.

TE : Training error.

F : Free energy.

$O(f(N))$	: A function such that $\limsup_{N\to\infty} \lvert O(f(N))/f(N)\rvert < \infty$.
$o(f(N))$	: A function such that $\lim_{N\to\infty} o(f(N))/f(N) = 0$.
$\Omega(f(N))$	: A function such that $\liminf_{N\to\infty} \lvert\Omega(f(N))/f(N)\rvert > 0$
$\omega(f(N))$	: A function such that $\lim_{N\to\infty} \lvert\omega(f(N))/f(N)\rvert = \infty$.
$\Theta(f(N))$	: A function such that $\limsup_{N\to\infty} \lvert\Theta(f(N))/f(N)\rvert < \infty$ and $\liminf_{N\to\infty} \lvert\Theta(f(N))/f(N)\rvert > 0$.
$O_{\mathrm{p}}(f(N))$	: A function such that $\limsup_{N\to\infty} \left\lvert O_{\mathrm{p}}(f(N))/f(N)\right\rvert < \infty$ in probability.
$o_{\mathrm{p}}(f(N))$	: A function such that $\lim_{N\to\infty} o_{\mathrm{p}}(f(N))/f(N) = 0$ in probability.
$\Omega_{\mathrm{p}}(f(N))$	: A function such that $\liminf_{N\to\infty} \left\lvert\Omega_{\mathrm{p}}(f(N))/f(N)\right\rvert > 0$ in probability
$\omega_{\mathrm{p}}(f(N))$	: A function such that $\lim_{N\to\infty} \left\lvert\omega_{\mathrm{p}}(f(N))/f(N)\right\rvert = \infty$ in probability.
$\Theta_{\mathrm{p}}(f(N))$	: A function such that $\limsup_{N\to\infty} \left\lvert\Theta_{\mathrm{p}}(f(N))/f(N)\right\rvert < \infty$ and $\liminf_{N\to\infty} \left\lvert\Theta_{\mathrm{p}}(f(N))/f(N)\right\rvert > 0$ in probability.