

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.

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

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
978-1-107-07615-0 — Variational Bayesian Learning Theory
Shinichi Nakajima, Kazuho Watanabe, Masashi Sugiyama
Frontmatter
[More Information](#)

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

© Shinichi Nakajima, Kazuho Watanabe, and Masashi Sugiyama 2019

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

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://lcn.loc.gov/2019005983>

ISBN 978-1-107-07615-0 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Contents

<i>Preface</i>	<i>page</i> ix
<i>Nomenclature</i>	xii
Part I Formulation	1
1 Bayesian Learning	3
1.1 Framework	3
1.2 Computation	10
2 Variational Bayesian Learning	39
2.1 Framework	39
2.2 Other Approximation Methods	51
Part II Algorithm	61
3 VB Algorithm for Multilinear Models	63
3.1 Matrix Factorization	63
3.2 Matrix Factorization with Missing Entries	74
3.3 Tensor Factorization	80
3.4 Low-Rank Subspace Clustering	87
3.5 Sparse Additive Matrix Factorization	93
4 VB Algorithm for Latent Variable Models	103
4.1 Finite Mixture Models	103
4.2 Other Latent Variable Models	115
5 VB Algorithm under No Conjugacy	132
5.1 Logistic Regression	132
5.2 Sparsity-Inducing Prior	135
5.3 Unified Approach by Local VB Bounds	137

Part III	Nonasymptotic Theory	147
6	Global VB Solution of Fully Observed Matrix Factorization	149
6.1	Problem Description	150
6.2	Conditions for VB Solutions	152
6.3	Irrelevant Degrees of Freedom	153
6.4	Proof of Theorem 6.4	157
6.5	Problem Decomposition	160
6.6	Analytic Form of Global VB Solution	162
6.7	Proofs of Theorem 6.7 and Corollary 6.8	163
6.8	Analytic Form of Global Empirical VB Solution	171
6.9	Proof of Theorem 6.13	173
6.10	Summary of Intermediate Results	180
7	Model-Induced Regularization and Sparsity Inducing Mechanism	184
7.1	VB Solutions for Special Cases	184
7.2	Posteriors and Estimators in a One-Dimensional Case	187
7.3	Model-Induced Regularization	195
7.4	Phase Transition in VB Learning	202
7.5	Factorization as ARD Model	204
8	Performance Analysis of VB Matrix Factorization	205
8.1	Objective Function for Noise Variance Estimation	205
8.2	Bounds of Noise Variance Estimator	207
8.3	Proofs of Theorem 8.2 and Corollary 8.3	209
8.4	Performance Analysis	214
8.5	Numerical Verification	228
8.6	Comparison with Laplace Approximation	230
8.7	Optimality in Large-Scale Limit	232
9	Global Solver for Matrix Factorization	236
9.1	Global VB Solver for Fully Observed MF	236
9.2	Global EVB Solver for Fully Observed MF	238
9.3	Empirical Comparison with the Standard VB Algorithm	242
9.4	Extension to Nonconjugate MF with Missing Entries	247
10	Global Solver for Low-Rank Subspace Clustering	255
10.1	Problem Description	255
10.2	Conditions for VB Solutions	258
10.3	Irrelevant Degrees of Freedom	259
10.4	Proof of Theorem 10.2	259

10.5	Exact Global VB Solver (EGVBS)	264
10.6	Approximate Global VB Solver (AGVBS)	267
10.7	Proof of Theorem 10.7	270
10.8	Empirical Evaluation	274
11	Efficient Solver for Sparse Additive Matrix Factorization	279
11.1	Problem Description	279
11.2	Efficient Algorithm for SAMF	282
11.3	Experimental Results	284
12	MAP and Partially Bayesian Learning	294
12.1	Theoretical Analysis in Fully Observed MF	295
12.2	More General Cases	329
12.3	Experimental Results	332
	Part IV Asymptotic Theory	339
13	Asymptotic Learning Theory	341
13.1	Statistical Learning Machines	341
13.2	Basic Tools for Asymptotic Analysis	344
13.3	Target Quantities	346
13.4	Asymptotic Learning Theory for Regular Models	351
13.5	Asymptotic Learning Theory for Singular Models	366
13.6	Asymptotic Learning Theory for VB Learning	382
14	Asymptotic VB Theory of Reduced Rank Regression	385
14.1	Reduced Rank Regression	385
14.2	Generalization Properties	396
14.3	Insights into VB Learning	426
15	Asymptotic VB Theory of Mixture Models	429
15.1	Basic Lemmas	429
15.2	Mixture of Gaussians	434
15.3	Mixture of Exponential Family Distributions	443
15.4	Mixture of Bernoulli with Deterministic Components	451
16	Asymptotic VB Theory of Other Latent Variable Models	455
16.1	Bayesian Networks	455
16.2	Hidden Markov Models	461
16.3	Probabilistic Context-Free Grammar	466
16.4	Latent Dirichlet Allocation	470

17 Unified Theory for Latent Variable Models	500
17.1 Local Latent Variable Model	500
17.2 Asymptotic Upper-Bound for VB Free Energy	504
17.3 Example: Average VB Free Energy of Gaussian Mixture Model	507
17.4 Free Energy and Generalization Error	511
17.5 Relation to Other Analyses	513
<i>Appendix A James–Stein Estimator</i>	516
<i>Appendix B Metric in Parameter Space</i>	520
<i>Appendix C Detailed Description of Overlap Method</i>	525
<i>Appendix D Optimality of Bayesian Learning</i>	527
<i>Bibliography</i>	529
<i>Subject Index</i>	540

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.

Nomenclature

- $a, b, c, \dots, A, B, C, \dots$: Scalars.
- $\mathbf{a}, \mathbf{b}, \mathbf{c}, \dots$ (bold-faced small letters) : Vectors.
- $\mathbf{A}, \mathbf{B}, \mathbf{C}, \dots$ (bold-faced capital letters) : Matrices.
- $\mathcal{A}, \mathcal{B}, \mathcal{C}, \dots$ (calligraphic capital letters) : Tensors or sets.
- $(\cdot)_{l,m}$: (l, m) th element of a matrix.
- \top : Transpose of a matrix or vector.
- $\text{tr}(\cdot)$: Trace of a matrix.
- $\det(\cdot)$: Determinant of a matrix.
- \odot : Hadamard (elementwise) product.
- \otimes : Kronecker product.
- \times_n : n -mode tensor product.
- $|\cdot|$: Absolute value of a scalar. It applies element-wise for a vector or matrix.
- $\text{sign}(\cdot)$: Sign operator such that $\text{sign}(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ -1 & \text{otherwise.} \end{cases}$ It applies elementwise for a vector or matrix.
- $\{\dots\}$: Set consisting of specified entities.
- $\{\dots\}^D$: D fold Cartesian product, i.e.,
 $\mathbb{X}^D \equiv \{(x_1, \dots, x_D)^\top; x_d \in \mathbb{X} \text{ for } d = 1, \dots, D\}$.
- $\#(\cdot)$: 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.,
 $[l, u] = \{x \in \mathbb{R}; l \leq x \leq u\}$.
- $[\cdot, \cdot]^D$: The set of D -dimensional real vectors whose entries are in a range, i.e., $[l, u]^D \equiv \{x \in \mathbb{R}^D; l \leq x_d \leq u \text{ for } d = 1, \dots, D\}$.
- $\mathbb{R}^{L \times M}$: The set of all $L \times M$ real matrices.
- $\mathbb{R}^{M_1 \times M_2 \times \dots \times M_N}$: The set of all $M_1 \times M_2 \times \dots \times M_N$ 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}_N^{K-1} : The set of all possible histograms for N samples and K categories, i.e., $\mathbb{H}_N^{K-1} \equiv \{x \in \{0, \dots, N\}^K; \sum_{k=1}^K x_k = N\}$.
- Δ^{K-1} : The standard $(K-1)$ -simplex, i.e.,
 $\Delta^{K-1} \equiv \{\theta \in [0, 1]^K; \sum_{k=1}^K \theta_k = 1\}$.

- $(\mathbf{a}_1, \dots, \mathbf{a}_M)$: Column vectors of \mathbf{A} , i.e., $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_M) \in \mathbb{R}^{L \times M}$.
- $(\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_L)$: Row vectors of \mathbf{A} , i.e., $\mathbf{A} = (\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_L)^\top \in \mathbb{R}^{L \times M}$.

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

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

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

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

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

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

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

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

\mathbf{e}_k : One of K expression, i.e., $\mathbf{e}_k = \underbrace{(0, \dots, 0, \overset{k\text{th}}{1}, 0, \dots, 0)^\top}_K \in \{0, 1\}^K$.

$\mathbf{1}_K$: K -dimensional vector with all elements equal to one, i.e.,
 $\mathbf{e}_k = \underbrace{(1, \dots, 1)^\top}_K$.

- $\text{Gauss}_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: D -dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
 $\text{MGauss}_{D_1, D_2}(\boldsymbol{M}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})$: $D_1 \times D_2$ dimensional matrix variate Gaussian distribution with mean \boldsymbol{M} and covariance $\boldsymbol{\Sigma} \otimes \boldsymbol{\Psi}$.
 $\text{Gamma}(\alpha, \beta)$: Gamma distribution with shape parameter α and scale parameter β .
 $\text{InvGamma}(\alpha, \beta)$: Inverse-Gamma distribution with shape parameter α and scale parameter β .
 $\text{Wishart}_D(\boldsymbol{V}, \nu)$: D -dimensional Wishart distribution with scale matrix \boldsymbol{V} and degree of freedom ν .
 $\text{InvWishart}_D(\boldsymbol{V}, \nu)$: D -dimensional inverse-Wishart distribution with scale matrix \boldsymbol{V} and degree of freedom ν .
 $\text{Multinomial}(\boldsymbol{\theta}, N)$: Multinomial distribution with event probabilities $\boldsymbol{\theta}$ and number of trials N .
 $\text{Dirichlet}(\boldsymbol{\phi})$: Dirichlet distribution with concentration parameters $\boldsymbol{\phi}$.
- $\text{Prob}(\cdot)$: 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.,

$$\langle f(\boldsymbol{x}) \rangle_{p(\boldsymbol{x})} \equiv \int f(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}.$$
 $\hat{\cdot}$: Estimator for an unknown variable, e.g., $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{A}}$ are estimators for a vector \boldsymbol{x} and a matrix \boldsymbol{A} , respectively.
Mean(\cdot) : Mean of a random variable.
Var(\cdot) : Variance of a random variable.
Cov(\cdot) : Covariance of a random variable.
 $\text{KL}(\cdot \| \cdot)$: Kullback–Leibler divergence between distributions, i.e.,

$$\text{KL}(p \| q) \equiv \left\langle \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \right\rangle_{p(\boldsymbol{x})}.$$
 $\delta(\boldsymbol{\mu}; \widehat{\boldsymbol{\mu}})$: 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 $\limsup_{N \rightarrow \infty} |O(f(N))/f(N)| < \infty$.
 $o(f(N))$: A function such that $\lim_{N \rightarrow \infty} o(f(N))/f(N) = 0$.
 $\Omega(f(N))$: A function such that $\liminf_{N \rightarrow \infty} |\Omega(f(N))/f(N)| > 0$
 $\omega(f(N))$: A function such that $\lim_{N \rightarrow \infty} |\omega(f(N))/f(N)| = \infty$.
 $\Theta(f(N))$: A function such that $\limsup_{N \rightarrow \infty} |\Theta(f(N))/f(N)| < \infty$
 and $\liminf_{N \rightarrow \infty} |\Theta(f(N))/f(N)| > 0$.
 $O_p(f(N))$: A function such that $\limsup_{N \rightarrow \infty} |O_p(f(N))/f(N)| < \infty$
 in probability.
 $o_p(f(N))$: A function such that $\lim_{N \rightarrow \infty} o_p(f(N))/f(N) = 0$ in probability.
 $\Omega_p(f(N))$: A function such that $\liminf_{N \rightarrow \infty} |\Omega_p(f(N))/f(N)| > 0$
 in probability
 $\omega_p(f(N))$: A function such that $\lim_{N \rightarrow \infty} |\omega_p(f(N))/f(N)| = \infty$
 in probability.
 $\Theta_p(f(N))$: A function such that $\limsup_{N \rightarrow \infty} |\Theta_p(f(N))/f(N)| < \infty$
 and $\liminf_{N \rightarrow \infty} |\Theta_p(f(N))/f(N)| > 0$ in probability.

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