論文 / 著書情報
Article / Book Information

題目(和文)	パラメータ制約付き特異モデルの統計的学習理論
Title(English)	Statistical Learning Theory of Parameter-Restricted Singular Models
著者(和文)	林 直輝
Author(English)	Naoki Hayashi
出典(和文)	学位:博士(理学), 学位授与機関:東京工業大学, 報告番号:甲第12028号, 授与年月日:2021年6月30日, 学位の種別:課程博士, 審査員:渡邊 澄夫,高安 美佐子,金森 敬文,山下 真,澄田 範奈
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第12028号, Conferred date:2021/6/30, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

2021FY Ph.D. Dissertation
（令和 3 年度 博士論文）

Statistical Learning Theory of Parameter-Restricted Singular Models

（パラメータ制約付き特異モデルの統計的学習理論）

First Submit	2021/02/17
Revise	2021/03/21, 2021/04/16
Final Submit	2021/05/20
Supervisor	Prof. Sumio Watanabe
Ch. Examiner	Prof. Sumio Watanabe
Examiners	Prof. Takafumi Kanamori
	Prof. Misako Takayasu
	Prof. Makoto Yamashita
	Prof. Hanna Sumita

Affiliation	Tokyo Institute of Technology
	School of Computing
	Department of Mathematical and Computing Science

Naoki Hayashi

Abstract

Statistical models used in machine learning are called learning machines. It is well-known that learning machines are widely applied to predict unknown events and discover knowledge by computers in many fields. Indeed, machine learning has grown over the last several decades. They are used for statistical learning/inference and usually have hierarchical structures. These structures are effective for generalizing to the real world. Statistical learning theory is a theory to clarify the generalization performances of learning machines.

Singular learning theory is a mathematical foundation for statistical inference using singular models. Typical hierarchical models, such as neural networks, tree and forest model, mixture model, matrix factorization, and topic model, are statistically singular since a map from a parameter to a probability density function is not one-to-one. Clarifying generalization behaviors in singular models is an important problem to estimate sufficient sample sizes, design models, and tune hyperparameters. However, conventional statistics theory cannot be applied to these models because their likelihoods cannot be approximated by any normal distribution. Singular learning theory provides a general view for this problem; birational invariants of an analytic set (a.k.a. algebraic variety) determine the generalization error. That is defined by zero of a Kullback-Leibler (KL) divergence between the data-generating distribution and the model. Algebraic structures of statistical models are essential in singular learning theory; thus, it can be interpreted as an intersection between algebraic statistics and statistical learning theory.

One of such invariants is a real log canonical threshold (RLCT). An RLCT is a negative-maximum pole of a zeta function defined by an integral of a KL divergence. Determining an RLCT of a concrete model is performed by resolution of singularities. In fact, algebraic statisticians and machine learning researchers have derived the exact values or upper bounds of the RLCTs for several singular models. The theoretical value of the RLCT is effective in statistical model selection such as sBIC proposed by Drton and Plummer. Besides, Nagata proposed a tuning method using RLCTs for exchange Monte Carlo.

On the other hand, from the practical point of view, the parameter region of the model is often restricted to improve interpretability. Non-negative matrix factorization (NMF) and latent Dirichlet allocation (LDA) are well-known examples of parameter-restricted singular models. In general, such constraints make the generalization error changed. However, for each singular model and condition, the quantitative effect of those constraints has not yet been clarified because the singularities in the above analytic set are also changed by the restriction to the parameter region.

In this dissertation, as a foundation to establish a singular learning theory of parameter-restricted statistical models, we theoretically study the asymptotic behavior of the Bayesian generalization error in NMF and LDA. NMF and LDA are two typical singular models whose parameter regions are constrained. In NMF, we derive an upper bound of the RLCT and a lower bound of the variational approximation error. In LDA, we prove that its RLCT is equal to that of matrix factorization with simplex restriction and clarify the exact asymptotic form of the generalization error, i.e. we determine the exact value of the RLCT of LDA. These results provide quantitative differences of generalization errors from matrix factorization whose parameter space is not restricted.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, we introduce the research investigated by this dissertation. First, in Sec. 1.1, we describe the background of this study. In Sec. 1.2, the aim of this research is stated. In Sec. 1.3, the structure of this thesis is explained. And lastly, in Sec. 1.4, we show symbols commonly used in this dissertation.

1.1 Background

1.1.1 Statistical Learning Theory

Machine learning is a ubiquitous technology that tackles real-world problems such as future prediction and knowledge discovery from data, and it has been developing pattern recognition [14]. It is significant to make scientific or technical judgments based on the data; hence, machine learning has been widely applied as same as statistics. For example, non-negative matrix factorization (NMF) [61, 18] has been used for signal processing [51], text mining [88], bioinformatics [46], and purchase analysis [47]. Another example is latent Dirichlet allocation (LDA). LDA has been applied to text analysis [15, 30], image recognition [52], market research [72], and geology [98]. Statistical learning theory (learning theory) is a theory that aims to build a foundation for solving these problems faced in the real world and for evaluating the results obtained by machine learning. When n-data are obtained from the true distribution (data-generating distribution) $q(x)$ [*1] , we cannot know it in reality. The objectives of learning theory are the following three things. First, to build a theoretical foundation for estimating the true distribution $q(x)$ with a predictive distribution $p^*(x)$ based on the data $X^n = (X_1, \ldots, X_n)$, by learning it with a model $p(x|\theta)$, where x is a variable in the data space $\mathcal{X} \subset \mathbb{R}^N$ and θ is a parameter ($\theta \in \mathcal{W} \subset \mathbb{R}^d$). Second, to clarify the behavior of the error of the learning results. Third, to devise an algorithm (learning algorithm) for estimating the true distribution with high accuracy in practice. Such situations specifically emerge when we create artificial intelligence that recognizes images and sounds by learning from examples, or we conduct knowledge discovery such as reducing dimension and clustering, or future prediction such as anomaly detection or time series prediction.

[*1] Actually, $q(x)$ is a probability density/mass function; however, we call it a probability distribution according to the convention.

Typical learning algorithms include maximum likelihood estimation and Bayesian inference [85]. In maximum likelihood estimation, the predictive distribution becomes $p(x|\theta^*)$, where θ^* is a parameter that maximizes the likelihood (i.e. the probability density when the model observes the data). On the other hand, in Bayesian inference, the prior distribution $\varphi(\theta)$ is defined with the probability model $p(x|\theta)$, i.e. the statistical model is denoted by the simultaneous distribution $p(x|\theta)\varphi(\theta)$. The posterior distribution $\psi(\theta|X^n) \propto p(x|\theta)\varphi(\theta)$ is calculated and the predictive distribution becomes the expectation of $p(x|\theta)$ by $\psi(\theta|X^n)$: $p^*(x) = \int p(x|\theta)\psi(\theta|X^n)\mathrm{d}\theta$. This prediction distribution is called the Bayesian prediction distribution. The predictive distribution is often also written as $p(x|X^n)$, in the sense that it is the probability distribution of x given the data X^n.

Recently-used statistical models and learning machines often have hierarchical structures or latent variables. For example, neural networks and matrix factorization have hierarchical structures, and hidden Markov model and topic model have latent variables. In such models, a map from the parameter set to the probability distribution set is not one-to-one. Its log-likelihood function cannot be approximated by any quadratic form; its likelihood and posterior cannot be approximated by any normal distribution. If the above map is injective, then the statistical model is called regular (or the regular model). The statistical model which is not regular is called singular (or the singular model). Hence, the learning machines mentioned above are singular models. Learning theory for singular models is called singular learning theory. From singular learning theory, neither maximum likelihood estimator nor maximum a posteriori estimator has asymptotic normality and consistency. Besides, Bayesian inference is superior to the maximum likelihood method and maximum a posteriori method in the sense of generalization, i.e. Bayesian inference can make the error between the true distribution and the predictive one (the generalization error) smaller than those point-estimation methods [79, 82, 85]. This fact has been proved not only numerically but also mathematically. Almost all practical learning machines are singular; thus, the determination of the generalization error in singular models is one of the most important issues in machine learning and statistics community.

In Bayesian inference, the asymptotic behavior of the generalization error had been clarified. When the statistical model can realize the true distribution and is regular, the following theorem is proved [4, 85].

Theorem 1.1 *Let n be the number of the data and d be the dimension of the parameter space. The expected Bayesian generalization error $\mathbb{E}[G_n]$ in realizable and regular statistical model has the following asymptotic behavior:*

$$\mathbb{E}[G_n] = \frac{d}{2n} + o\left(\frac{1}{n}\right), \tag{1.1}$$

where the operator $\mathbb{E}[\cdot]$ means expectation on overall datasets: $\mathbb{E}[\cdot] = \int [\cdot] \prod_{i=1}^n q(x_i)\mathrm{d}x_i$.

In the general case, with supposing some technical assumption, the following theorem holds [79, 82, 85]:

Theorem 1.2 *Let n be the number of the data. For the expected Bayesian generalization error $\mathbb{E}[G_n]$ in a realizable statistical model, there exists the constant λ and the following*

asymptotic behavior holds:

$$\mathbb{E}[G_n] = \frac{\lambda}{n} + o\left(\frac{1}{n}\right). \tag{1.2}$$

The constant λ is in the leading term of the asymptotic behavior of the generalization error; thus, it is called the learning coefficient [*2] in learning theory. This coefficient is the same as the real log canonical threshold (RLCT) in algebraic geometry. Conversely, algebraic geometry is needed to prove Theorem 1.2. In the case that the model is regular, the RLCT is equal to $d/2$; thus, Theorem 1.2 includes Theorem 1.1 as a special case. Therefore, singular learning theory can provide general knowledge for statistical modeling and machine learning, in particular when we use Bayesian inference. RLCTs are birational invariants in algebraic geometry and they depend on the algebraic varieties characterized by statistical models. A detailed review of these theories will be given in later chapters. It is important for model selection and hyperparameter tuning to clarify the theoretical generalization error. Besides, if a theoretical relation between the sample size n and the generalization error G_n is known, then we can estimate the sufficient sample size to realize the needed generalization performance. We can also evaluate the correctness of numerical experiments. On the contrary, the absence of such theoretical facts means that there is no method to verify the accuracy of the numerical experiments, therefore the correctness of the empirical results cannot be guaranteed. As a more direct application, a precise model-selection method that uses RLCTs had been proposed [24]. RLCTs are also useful to tune the inverse temperature in exchange Monte Carlo [56]. RLCTs of several models are studied. For example, three-layered neural network [80, 9, 8], reduced rank regression [77, 10], normal mixture model [94], Poisson mixture model [64], hidden Markov model [96], Markov model [100], naive Bayes method [62], Bayesian network [95], Boltzmann machine [97, 6, 7], latent Gaussian tree [23], and etc. RLCTs of them are analyzed by using Atiyah's form [11] of resolution of singularity theorem [39].

The above theorems assume that the statistical model can realize the true distribution. However, since the asymptotic behavior of the generalization error has also been clarified and they are characterized by RLCTs in the case the model is not realizable [82, 85], the realizability assumption cannot hurt the value of our research; determine of the RLCTs of statistical models. Moreover, the realizable case is essential for practical uses because we face a situation that the model is more redundant than the true distribution at the level known from a given sample in real-data analysis situations. The given sample is finite in real-world problems, and the complexity of the appropriate model is finite to the extent that it can be known from that sample. In this case, some models can be redundant, i.e. they seem to include the true distribution. If the model is too simple and suffers from underfitting, we use more complex models. It is the goal of model selection and hypothesis testing to compare appropriate and redundant models in such cases.

1.1.2 Parameter Restriction

From the practical point of view, the parameter region of the statistical model is often restricted in order to have consistency on domain knowledge and improve interpretability. As a toy

[*2] The learning coefficient characterizes the learning curve when the sample size n increases, not the number of epochs. The learning coefficient is not the learning rate in gradient descent methods. In this dissertation, we call it the real log canonical threshold to avoid confusion between learning coefficient and rate.

example, we consider a problem to predict purchase existence for a product by using a logistic regression model. An usual logistic regression treats linear classification. Suppose that there are an objective variable y (purchase existence label) and some covariates like advertisement placements $a_{\mathrm{TVCM}}, a_{\mathrm{DM}}, \ldots$ and ratings r in EC-sites, and the classifier learns a relation between the labels (purchased or not) and these covariates from the sample. The logistic regression model is formalized as the following. That the product is purchased is denoted by $y = 1$. If it is not purchased, then let y be 0. The label $y \in \{0, 1\}$ is considered to be generated by a Bernoulli distribution

$$\mathrm{Ber}(y|u) = u^y(1-u)^{1-y}, \tag{1.3}$$

where

$$u = \sigma(s), \tag{1.4}$$
$$s = \beta_0 + \beta_1 a_{\mathrm{TVCM}} + \beta_2 a_{\mathrm{DM}} + \ldots + \beta_d r, \tag{1.5}$$
$$\sigma(s) = 1/(1+\exp(-s)), \tag{1.6}$$

β_0 is an intercept term, and (β_i) are regression coefficients. Intuitively, the contributions of these covariates are expected to be non-negative [*3]. However, the coefficients of the covariates sometimes become negative. If there is a negative coefficient, the contributions of each covariate cancel each other out, making it difficult to interpret. Non-negative restriction to the coefficients ($\beta_k \geqq 0$ for $k = 1, \ldots, d$) is a solution to resolve such an issue (like Fig. 1.1).

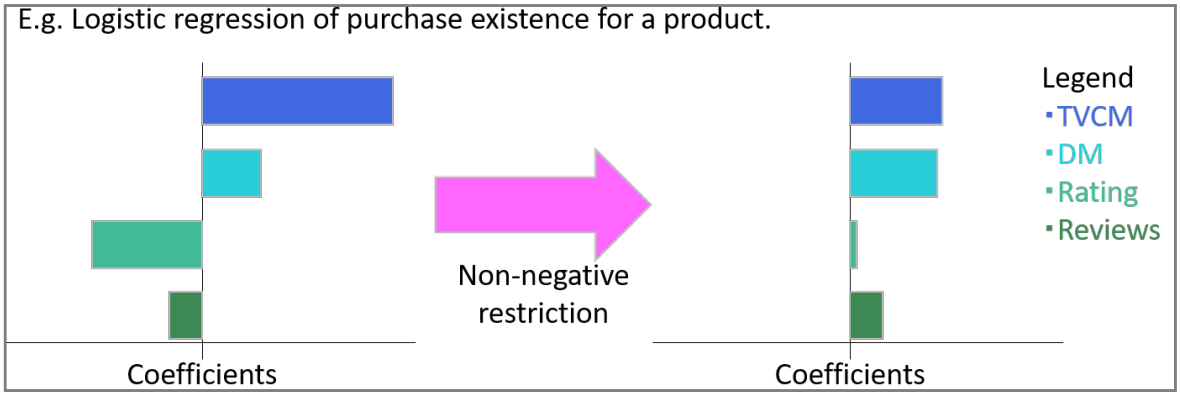


Fig. 1.1: Horizontal bar plots of regression coefficients. Left graph is the non-restricted result and right one is the result with non-negative restriction.

There are several typical restrictions: non-negative and stochastic. Non-negative restriction, as the name implies, constrains the parameter space to non-negative real numbers. Stochastic one constrains the parameter vectors into simplexes, i.e. they are non-negative and the sum of those is equal to one. Practical examples of these restrictions are non-negative matrix factorization (NMF) and latent Dirichlet allocation (LDA). NMF decomposes an observed

[*3] Excessive advertising may discourage purchases; however, this effect is ignored in this toy problem.

matrix into a product of two matrices whose entries are non-negative. Non-negative restriction in NMF makes discovered-latent patterns clearer to interpret than that of non-restricted cases. Kohjima has reviewed that NMF is useful and interpretable for analyzing relations between users and items in purchase logs [48]. LDA is one of the topic models. It represents a word-generating process and its parameters state the probability of words appearance and the topic proportion; thus, the stochastic restriction is imposed on them [15]. LDA is intuitively explained as stochastic matrix factorization (SMF): NMF with stochastic restriction [2, 98]. In fact, a mathematical equivalence between LDA and SMF can be proved in the sense of parameter regions; these RLCTs are the same [38] (see Chap. 5). Moreover, in practice, it seems to be a large demand for parameter restrictions. A probabilistic programming language called Stan [28, 17, 71] provides types of variables in order to restrict the parameter region for statistical models which users make. Thus, parameter-restricted models have been widely applied. However, although theoretical elucidation of the behavior of the generalization error is an important problem in learning theory as mentioned in the previous subsection, it has not yet been clarified how the generalization performance changes with the addition of parameter restrictions.

1.2 Research Goal

In this dissertation, we theoretically clarify the asymptotic behavior of the Bayesian generalization errors in NMF and LDA, as two of the most popular parameter-restricted models. We use a resolution of singularity and mathematically analyze the RLCTs of these models.

Specifically, we theoretically consider the Bayesian inference of NMF and LDA based on the framework described in later chapters and clarify the RLCTs. For NMF, we derive the exact value of the RLCT in some cases, and for the general case, we give a theoretical upper bound by using these exact values. In addition, we clarify the effect of hyperparameters in the case that the prior is gamma distribution, which is often used in NMF [18, 49]. For LDA, we prove that the RLCT of LDA is equal to the one of SMF and determine the exact value of the RLCT in all cases. We also clarify a relationship between LDA and non-restricted matrix factorization (whose RLCT is equal to one of reduced rank regression).

In this study, we provide theorem proofs based on pure theory. Although it is unknown as a direct problem that the assumptions of the theorems are satisfied in real problems, the real problem is solved as an inverse problem; thus, the theoretical results under certain assumptions are essential for solving real issues. Once the theoretical results of certain assumptions are clarified, the structure of the real world can be elucidated by comparing them with the situations we face in reality. NMF and LDA are widely used; however, their mathematical characteristics have not been clarified. Therefore, this dissertation is fundamental research to make statistical inferences with NMF and LDA.

1.3 Dissertation Structure

The structure of the rest of this dissertation is as follows. In Chap. 2, we describe a theoretical framework of Bayesian inference. In Chap. 3, we briefly introduce statistical learning theory for singular models (singular learning theory). In Chap. 4, the results and discussions of the theoretical analysis of NMF are presented. Similarly, in Chap. 5, those of LDA are described.

Lastly, in Chap. 6, we conclude this research. In Appendix A, we summarize the questions and the answers discussed in the defenses of this dissertation.

1.4 Symbols

In this section, we define the symbols often used in this thesis.

Symbols of sets are denoted as follows. Let $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{C}$ be the set of positive integers, real numbers, and complex numbers, respectively. Put $D \subset \mathbb{R}$. Let $\mathbb{R}_{\geqq 0} := \{x \in \mathbb{R} | x \geqq 0\}$ and $\mathbb{R}_{>0} := \{x \in \mathbb{R} | x > 0\}$. $\mathrm{M}(M, N, D)$ is denoted by the set of $M \times N$ matrices whose entries are in D. Put $M, N \in \mathbb{N}$ and $E \subset [0, 1]$. Let $\mathrm{Onehot}(N) := \{w = (w_j) \in \{0, 1\}^N \mid \sum_{j=1}^N w_j = 1\} = \{(1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)\}$ be an N-dimensional one-hot vector set and $\mathrm{Sim}(N, E) := \{c = (c_j) \in E^N \mid \sum_{j=1}^N c_j = 1\}$ be an N-dimensional simplex. Let $\mathrm{S}(M, N, E) = \mathrm{Sim}(M, E)^N$ be a set of $M \times N$ stochastic matrices whose elements are in E.

Besides, symbols of logical operation are defined as follows. Let $\wedge$ and $\vee$ be logical conjunction and disjunction, respectively. Also let $\Rightarrow$ and $\Leftrightarrow$ be implication and logical equivalence, respectively. Iverson bracket is defined as

$$[(\text{proposition})] = \begin{cases} 1 & \text{proposition is true.} \\ 0 & \text{proposition is false.} \end{cases}$$

Let $\sim$ be a binomial relation such that the functions $K_1(w)$ and $K_2(w)$ have same RLCT if $K_1(w) \sim K_2(w)$.

Chapter 2

Bayesian Inference

In this chapter, we describe the framework of Bayesian inference. First, in Sec. 2.1, we explain basic concepts to define what Bayesian inference is. Second, in Sec. 2.2, we state the framework of Bayesian inference and its definition.

The selected references in this chapter are [85] for Sec. 2.1 and 2.2.

2.1 Basic Concepts

In this section, we introduce a part of probability theory and Kullback-Leibler (KL) divergence which we need to define Bayesian inference.

2.1.1 Probability Theory

Let X and Y be random variables in measurable spaces $(\mathbb{R}^M, \mathfrak{B}_M)$ and $(\mathbb{R}^N, \mathfrak{B}_N)$, respectively[*1]. A pair (X, Y) is also a measurable function (i.e. a random variable) for a σ algebra $\mathfrak{B}_M \otimes \mathfrak{B}_N$, where $\mathfrak{B}_k$ is denoted by k-dimensional Borel algebra. First, we define the simultaneous probability distribution $p(x, y)$ of X and Y.

Definition 2.1 (Simultaneous Probability Distribution) *A simultaneous probability density function (or distribution) is defined by the probability density function $p(x, y)$ of (X, Y), where*

$$\int p(x, y) \mathrm{d}x \mathrm{d}y = 1,$$

and a probability of an event $C \in \mathfrak{B}_M \otimes \mathfrak{B}_N$ is

$$\Pr[(X, Y) \in C] = \int_C p(x, y) \mathrm{d}x \mathrm{d}y.$$

Moreover, a simultaneous probability is defined by the above probability $\Pr[(X, Y) \in C]$.

[*1] Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a (complete) probability space. Strictly speaking, a random variable $X : \Omega \to \mathbb{R}^M$ is defined as a $\mathfrak{F}/\mathfrak{B}_M$-measurable map. However, in this dissertation, we do not need to deal with the elementary event $F \in \mathfrak{F}$; thus, we define random variables without explicitly specifying the probability space $(\Omega, \mathfrak{F}, \mathbb{P})$. In fact, statistics and machine learning mainly aim at the data-generating distribution $\mathbb{Q} : \mathfrak{B}_M \to [0, 1]$ ($\mathbb{Q}(A) := \mathbb{P}(X^{-1}(A))$), i.e. a probability measure induced by X and its density $q(x)$ rather than $\mathbb{P}$.

The marginal probability distributions of X and Y are denoted by $p_X(x)$ and $p_Y(y)$, respectively.

Definition 2.2 (Marginal Probability Distribution) *About the above simultaneous probability distribution, put*

$$p_X(x) := \int p(x,y)\mathrm{d}y, \quad p_Y(y) := \int p(x,y)\mathrm{d}x.$$

Then, $p_X(x)$ and $p_Y(y)$ are the marginal probability distribution of X and Y, respectively.

From the above probability distributions, the conditional probability distribution is defined.

Definition 2.3 (Conditional Probability Distribution) *The conditional probability distribution (or density function) of Y given X is defined by*

$$p(y|x) := \frac{p(x,y)}{p_X(x)}.$$

This can be interpreted as a probability density function of y with a parameter x. The conditional probability distribution (or density function) of X given Y is defined in the same way:

$$p(x|y) := \frac{p(x,y)}{p_Y(y)}.$$

Note that the aboves are not defined when the denominator is zero. If the denominator is zero, the conditional density is defined as zero.

From the definition of the conditional probability distribution, we immediately have the following "theorem".

Theorem 2.1 (Bayes's Theorem)

$$p(x,y) = p(y|x)p_X(x) = p(x|y)p_Y(y)$$

and

$$p(y|x) = \frac{p(x|y)p_Y(y)}{p_X(x)}$$

hold.

Here, we define independent-identically-distributed (i.i.d.) random variables.

Definition 2.4 (independent Random Variables) *Let $p_1(x)$ and $p_2(y)$ be the probability density functions of X and Y, respectively. Two random variables X and Y are called independent if the simultaneous probability density function $p(x,y)$ of X and Y is equal to $p_1(x)p_2(y)$.*

Definition 2.5 (Identically-distributed Random Variables) *Assume that random variables X and Y have the same images: $\mathrm{Im}(X) = \mathrm{Im}(Y)$. Let $p_1(x)$ and $p_2(y)$ be the probability density functions of X and Y, respectively. Two random variables X and Y are called identically-distributed if these probability density functions are equal: $p_1 = p_2$.*

Definition 2.6 (I.i.d. Random Variables) *Two random variables X and Y are called i.i.d. if they are independent and identically-distributed.*

In this thesis, we mainly assume that the random variables $X_1, \ldots, X_n$ which mean the sample ($X^n = (X_1, \ldots, X_n)$) are i.i.d. Hence, the simultaneous probability of the sample is $q(x^n) = \prod_{i=1}^{n} q(x_i)$, where $q(x)$ is the data-generating distribution and $x_i \sim q(x)$ for $i = 1, \ldots, n$*[2].

2.1.2 Kullback-Leibler Divergence

Definition 2.7 (Kullback-Leibler Divergence) *Let $p(x)$ and $q(x)$ be probability density functions on a Euclidean space. Suppose that the supports of $p(x)$ and $q(x)$ are equal to each other. The Kullback-Leibler (KL) divergence* KL *is defined by*

$$\mathrm{KL}(q\|p) := \int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x.$$

If the support of $p(x)$ and $q(x)$ is a discrete set $\mathcal{X} = \{x_1, \ldots, x_n, \ldots\}$, i.e. $p(x)$ and $q(x)$ are probability mass functions, then the KL divergence is defined by

$$\mathrm{KL}(q\|p) := \sum_{i=1}^{\infty} q(x_i) \log \frac{q(x_i)}{p(x_i)}.$$

The KL divergence is the same as the relative entropy in statistical mechanics. This is based on a simulation error from an i.i.d. sample generated by the model to the empirical distribution obtained by the true distribution (Sanov's theorem) [63]. Moreover, some "distance-like" properties hold.

Proposition 2.1 *KL divergence satisfies the followings:*

- *Non-negativity:* $\mathrm{KL}(q\|p) \geqq 0$*,*
- *Non-degenerateness:* $\mathrm{KL}(q\|p) = 0$ *if and only if* $q(x) = p(x)$ *almost everywhere (a.e.).*

Thus, the KL divergence has been widely used in machine learning and statistics for a metric of discrepancy from a probability distribution to another one. Note that it is not exchangeable: $\mathrm{KL}(q\|p) \neq \mathrm{KL}(p\|q)$ by the definition.

Example 2.1 (Average Code Length) *In information theory, an information loss from an unknown information source $q(x)$ to a receiver $p^*(x)$ is called an average code length. This is* $\mathrm{KL}(q\|p^*)$ *and means redundancy of the receiver.*

Example 2.2 (Generalization Error) *In statistics and machine learning, an error in the statistical inference of the true distribution $q(x)$ by the predictive distribution $p^*(x)$ is called the generalization error. This is* $\mathrm{KL}(q\|p^*)$.

[2] Let X_i be a random variable in $(\mathbb{R}^M, \mathfrak{B}_M)$ for $i = 1, \ldots, n$. $X_1, \ldots, X_n$ induce the same probability distribution $\mathbb{Q}$ and its density function is $q(x)$. For arbitrary $A_1, \ldots, A_n \in \mathfrak{B}_M$, $\mathbb{Q}(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} \mathbb{Q}(A_i)$ holds.

Here, we prove Proposition 2.1.

Proof of Proposition 2.1. Let f be a function

$$f : \mathbb{R} \to \mathbb{R},$$
$$f(t) = t + e^{-t} - 1.$$

Because of $f'(t) = 1 - e^{-t}$ and $f''(t) = e^{-t} > 0$, f is a convex function and it becomes the minimum $f(0) = 0$ if and only $t = 0$. Thus, $f(t) \geqq 0$ and $f(t) = 0 \Leftrightarrow t = 0$. Here, we have

$$\begin{aligned}
\mathrm{KL}(q\|p) &= \int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x \\
&= \int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x + 1 - 1 \\
&= \int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x + \int p(x) \mathrm{d}x - \int q(x) \mathrm{d}x \\
&= \int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x + \int q(x) \frac{p(x)}{q(x)} \mathrm{d}x - \int q(x) \mathrm{d}x \\
&= \int \left(q(x) \log \frac{q(x)}{p(x)} + q(x) \frac{p(x)}{q(x)} - q(x) \right) \mathrm{d}x \\
&= \int q(x) \left(\log \frac{q(x)}{p(x)} + \frac{p(x)}{q(x)} - 1 \right) \mathrm{d}x \\
&= \int q(x) f \left(\log \frac{q(x)}{p(x)} \right) \mathrm{d}x.
\end{aligned}$$

Owing to the above and $q(x) \geqq 0$, we obtain $\mathrm{KL}(q\|p) \geqq 0$. Besides, since $q(x) = p(x)$, a.e. x if and only if $\log \frac{q(x)}{p(x)} = 0$, a.e. x, we immediately have the second property.

□

2.2 Framework of Bayesian Inference

With the above preparations, we define Bayesian inference. Let a sample $X^n = (X_1, \ldots, X_n)$ be a collection of i.i.d. random variables subject to a fixed probability distribution $q(x)$. The sample X^n is also a random variable subject to

$$q(x_1, \ldots, x_n) = \prod_{i=1}^{n} q(x_i),$$

and we call $q(x)$ the true probability distribution or the true probability density function. Simply, it is called the true distribution. We estimate it from the model and the prior distribution (prior), mentioned below. Let $\mathcal{W} \subset \mathbb{R}^d$ be a parameter set. The conditional probability distribution of $x \in \mathbb{R}^N$ given $\theta \in \mathcal{W}$ is written as $p(x|\theta)$ and it is called the probability model (or the model, simply). Besides, we define the prior distribution (prior) as a probability distribution of $\theta \in \mathcal{W}$.

The structure of the framework of Bayesian inference is as follows. First, we define the posterior distribution (posterior) by using the model, the prior, and the sample. Second, the Bayesian predictive distribution is defined by an expectation of the model by the posterior. Bayesian inference is to infer that "the true distribution $q(x)$ may be the Bayesian predictive distribution $p^*(x)$"; hence, Bayesian inference is a distributional estimation.

We define the posterior as follows.

Definition 2.8 (Posterior Distribution) *The posterior distribution (posterior) of the parameter θ given the sample X^n is defined by*

$$\psi(\theta|X^n) := \frac{1}{Z_n}\varphi(\theta)\prod_{i=1}^{n} p(X_i|\theta),$$

where Z_n is the normalizing constant which makes $\psi(\theta|X^n)$ satisfy $\int \psi(\theta|X^n)\mathrm{d}\theta = 1$:

$$Z_n := \int_{\mathcal{W}} \varphi(\theta)\prod_{i=1}^{n} p(X_i|\theta)\mathrm{d}\theta.$$

The normalizing constant Z_n is also called the marginal likelihood. In statistical mechanics, it is called the partition function.

The Bayesian predictive distribution is defined by the following.

Definition 2.9 (Bayesian Predictive Distribution) *The expectation of the model on the parameter with regard to the posterior*

$$p^*(x) = \int_{\mathcal{W}} p(x|\theta)\psi(\theta|X^n)\mathrm{d}\theta$$

is called the Bayesian predictive distribution (or the predictive, simply). To emphasize that the predictive is depend on the sample X^n, it is also represented by $p(x|X^n)$.

Bayesian inference is statistical inference, not propositional logic-based reasoning. Therefore, it is not correct [*3]. However, we can mathematically evaluate the error of the inferred result: how different the predictive is from the true distribution. The generalization loss and the free energy are typical criteria for that.

Definition 2.10 (Generalization Loss and Free Energy) *The generalization loss $\overline{G}_n$ and the free energy $\overline{F}_n$ are respectively defined by*

$$\overline{G}_n := -\int q(x)\log\left(\int_{\mathcal{W}} p(x|\theta)\psi(\theta|X^n)\mathrm{d}\theta\right)\mathrm{d}x = -\int q(x)\log p(x|X^n)\mathrm{d}x,$$

$$\overline{F}_n := -\log Z_n = -\log\int_{\mathcal{W}} \varphi(\theta)\prod_{i=1}^{n} p(X_i|\theta)\mathrm{d}\theta.$$

Obviously, minimizing $\overline{F}_n$ is equivalent to maximizing Z_n. On the other hand, minimizing $\overline{G}_n$ is neither a necessary nor sufficient condition for minimizing $\overline{F}_n$. Two methods have been

[*3] There is no "correct" statistical inference. The philosophical interpretation of probability (subjective v.s. objective or Bayesian v.s. frequentist) never makes either one "correct".

devised to select an appropriate model: minimizing generalization losses and minimizing free energy and they are nowadays used as the accuracy of prediction and the certainty of knowledge discovery, respectively.

Both of the generalization loss and free energy can be normalized [*4] in the sense that they remove the term that depends only on the true distribution. The normalized generalization loss is called the generalization error.

Definition 2.11 (Normalized Generalization Loss and Free Energy) *The normalized generalization loss, i.e. the generalization error G_n is defined by*

$$G_n := -\int q(x) \log \left\{ \int_{\mathcal{W}} \psi(\theta|X^n) \exp\left(-\log \frac{q(x)}{p(x|\theta)} \right) \mathrm{d}\theta \right\} \mathrm{d}x.$$

The normalized free energy F_n is defined as

$$F_n := -\log \int_{\mathcal{W}} \varphi(\theta) \exp\left(-\log \prod_{i=1}^{n} \frac{q(X_i)}{p(X_i|\theta)} \right) \mathrm{d}\theta.$$

The meaning of the above normalizing is formalized by the following proposition. All we have to do is analyzing the generalization error G_n and the normalized free energy F_n since this normalizing decomposes them the model-depend term and the term which only depends on the true distribution $q(x)$. The behavior of G_n is described in Theorem 1.2.

Proposition 2.2 *Let S and S_n be the entropy and the empirical entropy, respectively:*

$$S := -\int q(x) \log q(x) \mathrm{d}x,$$
$$S_n := -\frac{1}{n} \sum_{i=1}^{n} \log q(X_i).$$

The following equalities hold:

$$\overline{G}_n = S + G_n,$$
$$\overline{F}_n = nS_n + F_n.$$

Proof. Since

$$\log p(x|\theta) = \log q(x) - \log \frac{q(x)}{p(x|\theta)},$$

i.e.

$$p(x|\theta) = q(x) \exp\left(-\log \frac{q(x)}{p(x|\theta)} \right)$$

[*4] Generally, even when dealing with models that are not realizable, the generalization losses and free energies can be normalized similarly. See also the third chapter in [85].

holds, we obtain

$$\begin{aligned}
\overline{G}_n &= -\int q(x) \log \left(\int_{\mathcal{W}} p(x|\theta)\psi(\theta|X^n)\mathrm{d}\theta \right) \mathrm{d}x \\
&= -\int q(x) \log \left(\int_{\mathcal{W}} q(x) \exp\left(-\log \frac{q(x)}{p(x|\theta)} \right) \psi(\theta|X^n)\mathrm{d}\theta \right) \mathrm{d}x \\
&= -\int q(x) \log \left(q(x) \int_{\mathcal{W}} \exp\left(-\log \frac{q(x)}{p(x|\theta)} \right) \psi(\theta|X^n)\mathrm{d}\theta \right) \mathrm{d}x \\
&= -\int q(x) \left\{ \log q(x) + \log \left(\int_{\mathcal{W}} \exp\left(-\log \frac{q(x)}{p(x|\theta)} \right) \psi(\theta|X^n)\mathrm{d}\theta \right) \right\} \mathrm{d}x \\
&= -\int q(x) \log q(x) \mathrm{d}x - \int q(x) \log \left(\int_{\mathcal{W}} \exp\left(-\log \frac{q(x)}{p(x|\theta)} \right) \psi(\theta|X^n)\mathrm{d}\theta \right) \mathrm{d}x \\
&= S + G_n.
\end{aligned}$$

Besides, we have

$$\begin{aligned}
Z_n &= \int_{\mathcal{W}} \varphi(\theta) \prod_{i=1}^n p(X_i|\theta)\mathrm{d}\theta \\
&= \int_{\mathcal{W}} \varphi(\theta) \prod_{i=1}^n q(X_i) \exp\left(-\log \frac{q(X_i)}{p(X_i|\theta)} \right) \mathrm{d}\theta \\
&= \int_{\mathcal{W}} \varphi(\theta) \left(\prod_{i=1}^n q(X_i) \right) \left\{ \exp\left(-\sum_{i=1}^n \log \frac{q(X_i)}{p(X_i|\theta)} \right) \right\} \mathrm{d}\theta \\
&= \left(\prod_{i=1}^n q(X_i) \right) \int_{\mathcal{W}} \varphi(\theta) \left\{ \exp\left(-\sum_{i=1}^n \log \frac{q(X_i)}{p(X_i|\theta)} \right) \right\} \mathrm{d}\theta.
\end{aligned}$$

Thus, the following hold:

$$\begin{aligned}
\overline{F}_n &= -\log Z_n \\
&= -\log \left[\left(\prod_{i=1}^n q(X_i) \right) \int_{\mathcal{W}} \varphi(\theta) \left\{ \exp\left(-\sum_{i=1}^n \log \frac{q(X_i)}{p(X_i|\theta)} \right) \right\} \mathrm{d}\theta \right] \\
&= -\sum_{i=1}^n \log q(X_i) - \log \int_{\mathcal{W}} \varphi(\theta) \left\{ \exp\left(-\sum_{i=1}^n \log \frac{q(X_i)}{p(X_i|\theta)} \right) \right\} \mathrm{d}\theta \\
&= nS_n + F_n.
\end{aligned}$$

□

The generalization error in Definition 2.2 seems to be different from that in Definition 2.11; however, we can prove that they are equal if the model is realizable.

Example 2.2. From Definition 2.11, we have

$$\begin{aligned}
G_n &= -\int q(x) \log \left\{ \int_{\mathcal{W}} \psi(\theta|X^n) \exp\left(-\log \frac{q(x)}{p(x|\theta)} \right) \mathrm{d}\theta \right\} \mathrm{d}x \\
&= -\int q(x) \log \left(\int_{\mathcal{W}} \frac{p(x|\theta)}{q(x)} \psi(\theta|X^n) \mathrm{d}\theta \right) \mathrm{d}x \\
&= -\int q(x) \log \left(\frac{\int_{\mathcal{W}} p(x|\theta)\psi(\theta|X^n)\mathrm{d}\theta}{q(x)} \right) \mathrm{d}x \\
&= -\int q(x) \log \frac{p^*(x)}{q(x)} \mathrm{d}x \\
&= \int q(x) \log \frac{q(x)}{p^*(x)} \mathrm{d}x.
\end{aligned}$$

Besides, from Definition 2.7, the most-right-hand-side term is $\mathrm{KL}(q\|p^*)$. Therefore, we have Example 2.2.

□

Chapter 3

Singular Learning Theory

In this chapter, we outline statistical learning theory for singular models: singular learning theory. This chapter is organized as follows. In Sec. 3.1, we mention the motivations behind applying algebro-geometric methods to the theory of statistical learning. In Sec. 3.2, as a mathematical preparation, we depict the framework of analyzing F_n and G_n through the use of algebraic geometry. Thirdly, in Sec. 3.3, we introduce the key consequences of singular learning theory, which shows the relationship between RLCTs and F_n and G_n. Lastly, in Sec. 3.4, as applications of singular learning theory, we describe information criteria: WAIC, WBIC, sBIC, and WsBIC.

The selected textbooks as the references of this chapter are [81] and [85]. For introducing sBIC and WsBIC, the author refers the original papers [24, 43]. In addition, note that much of this chapter is taken from [38], one of the author's papers.

3.1 Motivation

First, we explain motivation why we apply algebraic geometry to statistical learning theory. As described in the above chapter, statistical learning encounters a situation that the true distribution $q(x)$ is not known although a plurality of data (a.k.a. sample) X^n can be obtained, where the number of data or the sample size is n. Researchers and practitioners design learning machines or statistical models $p(x|\theta)$ to estimate $q(x)$ by making the predictive distribution $p(x|X^n)$. There is a problem, "How different are our model and the true distribution?" This issue can be characterized as the model selection problem, "Which model is suitable?" This "suitableness" criteria are the normalized free energy F_n and the generalization error G_n, as mentioned above [*1]. However, calculating $\overline{F}_n$ is very high cost for computers and $\overline{G}_n$ cannot be computed since $q(x)$ is unknown. These normalized values F_n and G_n also

[*1] In actual data analysis situations, the model should be evaluated according to the purpose of analysis and the domain knowledge of the data. In singular learning theory, we consider F_n and G_n as typical evaluation criteria used in generic situations [73, 20, 40, 50, 54]. Watanabe (2021) [86], "their proposal is widely accepted in statistics, data science, and machine learning, on which many statistical systems and learning machines are being applied to scientific and practical problems" (p. 2).

depend on $q(x)$. We should estimate them from the data. Assume that the likelihood function $\mathcal{L}(\theta) = \prod_{l=1}^{n} p(X_l|\theta)$ and the posterior distribution $\psi(\theta|X^n)$ can be approximated by a Gaussian function of θ. This case is called regular and the model in regular case is called a regular model. For a regular model, we can estimate F_n and G_n, by using Bayesian information criterion (BIC) [66] and Akaike information criterion (AIC) [3], respectively. AIC and BIC are respectively defined by

$$AIC = -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i|\hat{\theta}) + \frac{d}{n}$$

and

$$BIC = -\sum_{i=1}^{n} \log p(X_i|\hat{\theta}) + \frac{d}{2}\log n,$$

where $\hat{\theta}$ is the maximum likelihood estimator or the maximum posterior estimator and d is the parameter dimension. AIC and BIC are derived by not using algebraic geometry; however, they are asymptotically equal to $\overline{G}_n$ and $\overline{F}_n$ if $\mathcal{L}(\theta)$ and $\psi(\theta|X^n)$ can be approximated by a normal distribution. In order to briefly describe the situation, we define the following error functions for learning theory.

Definition 3.1 (Error Functions) *Suppose $p(x|\theta)$ can realize $q(x)$, i.e. there exists θ_0 such that $p(x|\theta_0) = q(x)$. The function*

$$K(\theta) := \int q(x) \log \frac{q(x)}{p(x|\theta)} \mathrm{d}x = \mathrm{KL}(q\|p)$$

is called an average error function. In addition, the function

$$K_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i|\theta)}$$

is called an empirical error function.

In a regular case, the zero point θ_0 of $K(\theta)$ [*2] is unique and $\nabla^2 K(\theta_0)$ is a strictly positive definite matrix. Therefore, we can approximate $K(\theta)$ as

$$K(\theta) = \frac{1}{2}(\theta - \theta_0)^T \nabla^2 K(\theta^*)(\theta - \theta_0) \tag{3.1}$$

by Taylor's theorem, where there exists $t \in \mathbb{R}$ $(0 < t < 1)$ such that $\theta^* = t\theta + (1-t)\theta_0$ [*3]. By using this approximation, Akaike had derived AIC and Schwarz had done BIC, respectively. However, in general, we cannot estimate G_n and F_n by using AIC and BIC. This is because the above approximation does not hold since $\nabla^2 K(\theta_0)$ has eigenvalues which are zeros. θ_0 is not uniquely determined. $K^{-1}(0)$ includes singularities in the parameter space. Thus, we need algebraic geometry to study the singularities.

[*2] Because the model can realize the true distribution, $\mathrm{argmin} K(\theta)$ is equal to $K^{-1}(0)$. If $p(x|\theta)$ is not realizable $q(x)$, we can obtain similar results by considering $\mathrm{argmin} K(\theta)$ instead of $K^{-1}(0)$. For example, Takeuchi information criterion (TIC) [70, 85] and widely applicable information criterion (WAIC) [82] do not need the realizableness assumption.

[*3] From the definition, $K(\theta_0) = 0$ and $\nabla K(\theta_0) = 0$ hold.

3.2 Singularity Resolution Theorem and Zeta Function

Second, the framework of analyzing G_n and F_n, which uses algebraic geometry, is explained. We consider $K(\theta)$ and its zero set $K^{-1}(0)$: this is an algebraic variety which can have singularities. We use the following form by [11] of the singularities resolution theorem [39]. This form was originally derived by Atiyah for the analysis of distributions (hyperfunctions); however, Watanabe proved that it is useful for constructing singular learning theory [79, 82, 81, 85].

Theorem 3.1 (Singularity Resolution Theorem) *Let F be a non-negative analytic function on the open set $\mathcal{W}' \subset \mathbb{R}^d$ and assume that there exists $\theta \in \mathcal{W}'$ such that $F(\theta) = 0$. Then, there are d-dimensional manifold $\mathcal{M}$ and an analytic map $g : \mathcal{M} \to \mathcal{W}'$ such that for each local chart of $\mathcal{M}$,*

$$F(g(u)) = u_1^{2k_1} \dots u_d^{2k_d},$$
$$|g'(u)| = b(u)|u_1^{h_1} \dots u_d^{h_d}|,$$

where $|g'(u)|$ is the Jacobian of g, k_j and h_j are non-negative integers and $b : \mathcal{M} \to \mathbb{R}$ is strictly positive analytic: $b(u) > 0$.

A pair of the above manifold and map $(\mathcal{M}, g)$ is called a resolution of singularity.

This theorem does not remove singularities but makes them easier to handle. In general, a zero set of a non-negative analytic function F has singularities; however, $(F \circ g)^{-1}(0)$ has only singularities which are the following form: the zero set of $u_1^{2k_1} \dots u_d^{2k_d}$. Such singularities are called normal-crossing singularities.

Thanks to Theorem 3.1, the following analytic theorem is proved [11, 12, 65].

Theorem 3.2 *Let $F : \mathbb{R}^d \to \mathbb{R}$ be an analytic function of a variable $\theta \in \mathbb{R}^d$. $a : \mathcal{W} \to \mathbb{R}$ is denoted by a C^∞-function with compact support $\mathcal{W}$. Then*

$$\zeta(z) = \int_{\mathcal{W}} |F(\theta)|^z a(\theta) \mathrm{d}\theta$$

is a holomorphic function on $\mathrm{Re}(z) > 0$. Moreover, $\zeta(z)$ can be analytically continued to a unique meromorphic function on the entire complex plane $\mathbb{C}$. The poles of the extended function are all negative rational numbers.

The KL divergence is non-negative and analytic; thus, we can apply Theorem 3.1 to $K(\theta)$ on $K^{-1}(0) \cap \mathcal{W}'$. Then, we obtain

$$K(g(u)) = u_1^{2k_1} \dots u_d^{2k_d},$$
$$|g'(u)| = b(u)|u_1^{h_1} \dots u_d^{h_d}|.$$

Assuming the domain of prior $\varphi(\theta)$ is $\mathcal{W}$ and it satisfies $\mathcal{W} \subset \mathcal{W}'$, we can also apply Theorem 3.2 to $(K(\theta), \varphi(\theta))$ and obtain the following zeta function.

Definition 3.2 (Zeta Function in Statistical Learning Theory) *Let $K(\theta)$ be the average error function. Suppose that the support of the prior $\varphi(\theta)$ is compact. The zeta function in statistical learning theory is defined by the following univariate complex function:*

$$\zeta(z) = \int_{\mathcal{W}} K(\theta)^z \varphi(\theta) \mathrm{d}\theta. \tag{3.2}$$

If the prior can be zero or infinity on the zero set $K^{-1}(0)$ of the average error function, then we must consider the zeta function above as defined by Definition 3.2. However, if the prior is positive and bounded on $K^{-1}(0)$, then the prior has no effect on the maximum pole. With this in mind, we can consider the following zeta function:

$$\zeta(z) = \int_{\mathcal{W}} K(\theta)^z \mathrm{d}\theta. \tag{3.3}$$

Although $\int K(\theta)^z \varphi(\theta) \mathrm{d}\theta$ is a holomorphic function on $\mathrm{Re}(z) > 0$, we can prove that the zeta function in statistical learning theory has an analytic continuation on $\mathbb{C}$ as a unique meromorphic function and its poles are negative rational numbers. Applying Theorem 3.1 to $K(\theta)\varphi(\theta)$, there exists a manifold $\mathcal{M}$ and an analytic map g such that

$$K(g(u)) = u_1^{2k_1} \dots u_d^{2k_d},$$
$$\varphi(g(u))|g'(u)| = b(u)|u_1^{h_1} \dots u_d^{h_d}|$$

for each chart of $\mathcal{M}$, where $|g'(u)|$ is a Jacobian of g and b is a positive-analytic function. Let $k = (k_1, \dots, k_d)$ and $h = (h_1, \dots, h_d)$ be non-negative multi-indexes. Put $u_1^{2k_1} \dots u_d^{2k_d} =: u^{2k}$ and $u_1^{h_1} \dots u_d^{h_d} =: u^h$. Considering a partition of unity $\sum_a \phi_a(u)$ for $\mathcal{M}$, the zeta function becomes as follows:

$$\begin{aligned}
\zeta(z) &= \sum_a \int K(g(u))^z |g'(u)| \phi_a(u) \mathrm{d}u \\
&= \sum_a \int u^{2kz} |u^h| b(u) \phi_a(u) \mathrm{d}u \\
&= \sum_a \int_{[0,1]^d} u^{2kz} u^h b(u) \mathrm{d}u.
\end{aligned}$$

Since $b(u)$ is positive, it does not effect the maximum pole of $\zeta(z)$. Hence, we only have to treat

$$\int_{[0,1]^d} u^{2kz} u^h \mathrm{d}u = \prod_{j=1}^d \int_0^1 u_j^{2k_j z + h_j} \mathrm{d}u_j.$$

Calculating the integral, we have

$$\begin{aligned}
\int_{[0,1]^d} u^{2kz} u^h \mathrm{d}u &= \prod_{j=1}^d \int_0^1 u_j^{2k_j z + h_j} \mathrm{d}u_j \\
&= \prod_{j=1}^d \frac{1}{2k_j z + h_j + 1}.
\end{aligned}$$

Thus, allowing for duplication, the poles $(-\lambda_j)_{j=1}^d$ are as follows:

$$\begin{aligned} z = -\lambda_1 &:= -\frac{h_1+1}{2k_1} \\ &\vdots \\ z = -\lambda_j &:= -\frac{h_j+1}{2k_j} \\ &\vdots \\ z = -\lambda_d &:= -\frac{h_d+1}{2k_d}. \end{aligned}$$

The zeta function in statistical learning theory is a meromorphic function which is the result of analytic continuation. Hence, its poles are negative rational numbers. The definition of a real log canonical threshold is the negative maximum pole of it.

Definition 3.3 (Real Log Canonical Threshold and its Multiplicity) *Let $k = (k_1, \ldots, k_d)$ and $h = (h_1, \ldots, h_d)$ be non-negative d-dimensional multi-indexes. Assume that $k = (k_1, \ldots, k_d)$ is at least one of them positive. For a sequence*

$$\frac{h_j+1}{2k_j} \quad (j = 1, \ldots, d),$$

if $k_j = 0$, we define

$$\frac{h_j+1}{2k_j} = \infty.$$

The minimum of the above sequence

$$\lambda := \min_{j=1}^{d} \frac{h_j+1}{2k_j}$$

is called a real log canonical threshold (RLCT) and the number m of elements which are equal to λ is called its multiplicity:

$$m := \sum_{j=1}^{d} \left[\frac{h_j+1}{2k_j} = \lambda \right],$$

where $[(\text{proposition})]$ constitutes Iverson bracket.

From the above discussion of the zeta function and the definition of an RLCT, an RLCT λ is a sign reversal of the maximum pole $(-\lambda)$ of the zeta function $\zeta(z)$, and its multiplicity m is equal to the order of the maximum pole of the zeta function. Namely, the maximum pole is corresponding to the deepest singularity in the analytic set $K^{-1}(0)$. Furthermore, an RLCT is independent of methods of singularity resolution; an RLCT is a birational invariant in algebraic geometry. The RLCT is determined only for the statistical model $p(x|\theta)$, the prior $\varphi(\theta)$ and the true distribution $q(x)$ although there are infinite pairs $(\mathcal{M}, g)$ to resolve the singularities in the zero set of the given analytic function $K(\theta) = \int q(x) \log(q(x)/p(x|\theta)) \mathrm{d}x$.

An RLCT can be interpreted as a volume dimension of the analytic set $K^{-1}(0)$ in the parameter space $\mathcal{W}$ [81]. Let $t > 0$.

Theorem 3.3 *Let $V(t)$ be the volume of the set $\{\theta \in \mathcal{W} \mid K(\theta) < t\}$ in the sense of the measure $\varphi(\theta)\mathrm{d}\theta$, i.e.*

$$V(t) := \int_{K(\theta)<t} \varphi(\theta)\mathrm{d}\theta.$$

Then, the following equality holds:

$$\lambda = \lim_{t \to +0} \frac{\log V(t)}{\log t}.$$

This is a base of a method to numerically compute an RLCT by using Monte Carlo simulation [89]. Besides, the above limit is similar to a Minkowski dimension, one of fractal dimensions.

Definition 3.4 (Minkowski Dimension) *Let S be a subset of $\mathbb{R}^d$. The Minkowski dimension of S is defined as the following d^*:*

$$d^* := d - \lim_{t \to +0} \frac{\log V_{\mathrm{nbhd}}(S, t)}{\log t},$$

where $V_{\mathrm{nbhd}}(S, t)$ is the volume of the t-neighborhood of S:

$$V_{\mathrm{nbhd}}(S, t) := \int_{\mathrm{dist}(S,\theta)<t} \mathrm{d}\theta, \quad \mathrm{dist}(S, \theta) = \inf\{\|s - \theta\| \mid s \in S\}.$$

λ and d^* are similar but different concepts because λ is a positive rational number whereas d^* can be an irrational number. However, both of λ and d^* can be interpreted as intrinsic dimensions of subsets included by a Euclidean space. Indeed, they are useful for learning theory. As described later, the RLCT of $K(\theta) = \mathrm{KL}(q||p)$ dominates the Bayesian generalization error and the free energy (see Theorem 3.4). On the other hand, convergence rates of the approximation and generalization errors by deep neural networks depend on the Minkowski dimension of the data and the rates are optimal in the minimax sense [58]. Thus, we can refer an RLCT to an intrinsic dimension of the model.

In order to compute the exact value of an RLCT for a particular statistical model, one needs to find the manifold $\mathcal{M}$ and the analytic map g in the Singularity Resolution Theorem. By using the case when it is easy to calculate the exact value of the RLCT, some studies have elucidated the theoretical upper bound of the RLCT. As mentioned in Sec. 1.1, one of statistical models whose RLCT is well-known is reduced rank regression [10]. In this case, the upper bound of the RLCT was derived before its exact value was determined [77]. In almost all of the others, upper bounds are only clarified, i.e. it is much more difficult to find the exact value of the RLCT than to derive a non-trivial upper bound of that. Even clarifying a non-conservative upper bound of the RLCT is challenging since there is no standard method to find a resolution of singularities $(\mathcal{M}, g)$ for a collection of analytic functions like the KL divergence of statistical models. Instead, researchers have been studying RLCTs by developing novel methods to analyze RLCTs for each statistical model [80, 9, 8, 94, 64, 96, 100, 62, 95, 97, 6, 7, 23] (see also Sec. 1.1).

3.3 Key Results of Singular Learning Theory

We introduce the theorem which shows the relationship between RLCTs and G_n and F_n [79, 81, 85].

Theorem 3.4 (Watanabe) *Let $q(x)$, $p(x|\theta)$, and $\varphi(\theta)$ be the true distribution, the learning machine, and the prior distribution, where x is a point of $\mathbb{R}^N$ and θ is an element of the compact subset W of $\mathbb{R}^d$. Let $K(\theta)$ be the average error function and λ is denoted by the RLCT of $(K(\theta), \varphi(\theta))$. If there exists at least one θ_0 such that $q(x) = p(x|\theta_0)$ (i.e. the model can realize the true distribution), then the asymptotic behavior of the generalization error G_n and the normalized free energy F_n is as belows:*

$$\mathbb{E}[G_n] = \frac{\lambda}{n} - \frac{m-1}{n \log n} + o\left(\frac{1}{n \log n}\right),$$
$$F_n = \lambda \log n - (m-1) \log \log n + O_p(1).$$

Theorem 3.5 (Watanabe) *If there exists at least one θ_0 such that $q(x) = p(x|\theta_0)$ and maximum likelihood or posterior method is applied (i.e. the predictive distribution is $p^*(x) = p(x|\hat{\theta})$, where $\hat{\theta}$ is the maximum likelihood or posterior estimator), then there is a constant $\mu > d/2$ such that the asymptotic behavior of the generalization error G_n is as belows:*

$$\mathbb{E}[G_n] = \frac{\mu}{n} + o\left(\frac{1}{n}\right).$$

$K(\theta)$ depends on $q(x)$ and $p(x|\theta)$; thus, Theorem 3.4 can be understood as we can clarify G_n and F_n if the RLCT is clarified, which is determined by $(q(x), p(x|\theta), \varphi(\theta))$. As introduced in Chap. 1, there are several researches to find the RLCT of a statistical model by analyzing the maximum pole of the zeta function. Their studies are based on Theorem 3.4 and the zeta function derived by Theorem 3.2. The researchers have found the singularity resolution map g for the exact value or an upper bound of $\Phi(\theta)$, and have obtained the one of the RLCT since the RLCT is order isomorphic: if $\Phi(\theta) \leqq \Psi(\theta)$, then $\lambda_\Phi \leqq \lambda_\Psi$, where $(-\lambda_\Phi)$ and $(-\lambda_\Psi)$ are the maximum pole of $\zeta_1(z) = \int \Phi(\theta)^z d\theta$ and $\zeta_2(z) = \int \Psi(\theta)^z d\theta$, respectively [81].

Recently, RLCTs of statistical models have been used in other ways. Analyzing and tuning exchange probabilities in exchange Monte Carlo methods [56], deriving the asymptotic behavior of the Bayesian estimation accuracy for latent variables [90, 91], devising singular Bayesian information criterion [24] (see Sec. 3.4) and constructing deep learning theory [55] are such examples. Singular learning theory assumes that the optimal distribution is essentially unique [85]; however, in some of cases when that assumption is not satisfied, RLCTs and its multiplicities also determine the behavior of the free energy [57]. Moreover, singular learning theory in non-i.i.d. cases is also studied such for conditional independent samples [84, 85], in the exchangeable case [87], and for structured data [93].

From the practical point of view, Theorem 3.5 shows that Bayesian inference makes the free energy and the generalization error smaller than maximum likelihood or posterior method in singular case since $\mu > d/2 \geqq \lambda$ [85]. Hence, if the RLCT is clarified, then we can draw the $\mathbb{E}[G_n]$-n learning curve and estimate the sample size which satisfies the required inference performance.

There is no standard method to find an RLCT for a statistical model (family of functions). Here, we show a fundamental method to find the RLCT for a non-negative analytic function: this is called blowing-up [39]. We explain blowing-up which is used to study learning machines, based on a concrete example and [81, 85]. If a reader needs the rigorous definition of blowing-up, then see [39]. Let

$$K(\theta) = \theta_1^2 + \ldots + \theta_d^2 \tag{3.4}$$

and θ_i, $i = 1, \ldots, d$ be independent parameters. Especially, we treat the case $d = 2$. Blowing-up of $K(\theta)$ is a transformation of the coordinate that is defined

$$\begin{cases} \theta_1 = \theta_1^{(1)} = \theta_1^{(2)}\theta_2^{(2)}, \\ \theta_2 = \theta_1^{(1)}\theta_2^{(1)} = \theta_2^{(2)} \end{cases} .$$

Using this blowing-up,

$$K(\theta) = (\theta_1^{(1)})^2\{1 + (\theta_2^{(1)})^2\} = (\theta_2^{(2)})^2\{(\theta_1^{(2)})^2 + 1\},$$

and the absolute of Jacobian $|J|$ of this transformation is

$$|J| = |\theta_1^{(1)}| = |\theta_2^{(2)}|.$$

From the applied mathematical point of view, $1 + (\theta_i^{(j)})^2$ is strictly positive thus the RLCT can be calculated. The zeta function $\zeta(z)$ is

$$\zeta(z) = \int (\theta_1^{(1)})^{2z+1}\{1 + (\theta_2^{(1)})^2\}^z \mathrm{d}\theta^{(1)} = \int (\theta_2^{(2)})^{2z+1}\{1 + (\theta_1^{(2)})^2\}^z d\theta^{(2)}$$

and it is immediately proved that $1 + (\theta_i^{(j)})^2$ does not effect the RLCT λ. Then all we have to consider is the function

$$\zeta(z) = \int (\theta_1^{(1)})^{2z+1} \mathrm{d}\theta^{(1)} = \int (\theta_2^{(2)})^{2z+1} \mathrm{d}\theta^{(2)},$$

which are analytically connected to $\mathbb{C}$ as a unique meromorphic function, respectively. Therefore, we get

$$\zeta(z) = \frac{c_1}{2(z+1)} = \frac{c_2}{2(z+1)},$$

and

$$\lambda = \min\{1, 1\} = 1,$$

where c_1 and c_2 are positive constants. By the same way, in general d, the RLCT λ is equal to

$$\lambda = \frac{d}{2}. \tag{3.5}$$

If the statistical model is regular, then its average error function can be approximated by a quadratic form as Eq. (3.1). By diagonalizing the positive-definite matrix $\nabla^2 K(\theta_0)$ and centralizing that form, we arrive at Eq. (3.4). Because of Eq. (3.5), we can reconstruct the asymptotic behavior of the Bayesian generalization error in regular case: Theorem 1.1.

Let $\sim$ be a binomial relation such that the functions $K_1(w)$ and $K_2(w)$ have same RLCT if $K_1(w) \sim K_2(w)$. There are some propositions to evaluate RLCTs [81].

Proposition 3.1 *Let $\Phi : \mathbb{R}^d \to [0, \infty)$ be a non-negative analytic function. Let λ and m be the RLCT and its multiplicity defined by the following zeta function*

$$\zeta(z) = \int \Phi(\theta)^z \varphi(\theta) \mathrm{d}\theta,$$

respectively. Let $a(\theta) > 0$ and $b(\theta) > 0$ be smooth functions of θ which are strictly positive on the neighborhood of $\Phi^{-1}(0)$. Then, the following zeta function has the same maximum pole as that of the above: the RLCT and its multiplicity are same

$$\zeta_{a,b}(z) = \int (a(\theta)\Phi(\theta))^z b(\theta)\varphi(\theta)\mathrm{d}\theta.$$

Proposition 3.2 *RLCTs save the orders: if $\Phi(\theta) \leqq \Psi(\theta)$ and $\varphi_1(\theta) = \varphi_2(\theta)$, then $\lambda_\Phi \leqq \lambda_\Psi$, where $(-\lambda_\Phi)$ and $(-\lambda_\Psi)$ are the maximum pole of $\zeta_1(z) = \int \Phi(\theta)^z \varphi_1(\theta)\mathrm{d}\theta$ and $\zeta_2(z) = \int \Psi(\theta)^z \varphi_2(\theta)\mathrm{d}\theta$, respectively.*

From the aboves, the following is immediately derived.

Proposition 3.3 *Let λ_Φ and λ_Ψ be RLCTs in the above proposition. If $\varphi_1(\theta) = \varphi_2(\theta)$ and there are two positive constants $c_1 > 0$ and $c_2 > 0$ such that*

$$c_1\Phi(\theta) \leqq \Psi(\theta) \leqq c_2\Phi(\theta),$$

then $\lambda_\Phi = \lambda_\Psi$ holds and their multiplicities are also same.

By using the relationship between ideals and analytic sets, the following property can be proved.

Proposition 3.4 *Suppose $s, t \in \mathbb{N}$, and let $f_1(w), \ldots, f_s(w), g_1(w), \ldots, g_t(w)$ be real polynomials. Furthermore, let*

$$I := \langle f_1, \ldots, f_s \rangle, \quad J := \langle g_1, \ldots, g_t \rangle$$

be the generated ideals of $(f_1, \ldots, f_s)$ and $(g_1, \ldots, g_t)$, respectively. We put

$$F(w) := \sum_{i=1}^{s} f_i(w)^2, \quad G(w) := \sum_{j=1}^{t} g_j(w)^2.$$

Then, $F \sim G$ if $I = J$.

Corollary 3.1 *Assume that $F(w) = \sum_{i=1}^{s} f_i(w)^2$. Then*

$$F(w) + \left(\sum_{i=1}^{s} f_i(w) \right)^2 \sim F(w).$$

The prior can affect to the RLCT as the following.

Proposition 3.5 *Let $\lambda_1 := \lambda_\Phi$ and $\lambda_2 := \lambda_\Psi$ be RLCTs in Proposition 3.2. If $\Phi(\theta) = \Psi(\theta)$ and $\varphi_1(\theta) \leqq \varphi_2(\theta)$, then $\lambda_1 \geqq \lambda_2$ holds.*

3.4 Information Criteria from Singular Learning Theory

From the practical point of view, singular learning theory provides novel information criteria: WAIC [82] and WBIC [83]. They are useful even if the posterior cannot be approximated by any normal distribution, i.e. neither AIC nor BIC can be applied.

Definition 3.5 (Widely Applicable Information Criterion (WAIC)) *Widely applicable information criterion (WAIC) is defined by the following random variable W_n:*

$$W_n := T_n + V_n/n,$$

where T_n is the empirical loss and V_n is the functional variance:

$$T_n = -\frac{1}{n}\sum_{i=1}^{n} \log p^*(X_i) = -\frac{1}{n}\sum_{i=1}^{n} \log \mathbb{E}_\theta[p(X_i|\theta)],$$
$$V_n = \sum_{i=1}^{n} \left[\mathbb{E}_\theta[(\log p(X_i|\theta))^2] - \{\mathbb{E}_\theta[\log p(X_i|\theta)]\}^2\right] = \sum_{i=1}^{n} \mathbb{V}_\theta[\log p(X_i|\theta)].$$

$W_n - S_n$ is called the WAIC error.

WAIC asymptotically approximates the Bayesian generalization loss in the general case [*4] in the sense of the expectation [82].

Theorem 3.6 *The expected WAIC is asymptotically equal to the expected generalization loss and these difference is $O(1/n^2)$:*

$$\mathbb{E}[W_n] = \mathbb{E}[\overline{G}_n] + O(1/n^2).$$

This means that WAIC is asymptotically equal to leave-one-out cross validation (LOOCV) loss. LOOCV loss is collapsed in the case when the data has leverage points if LOOCV loss is calculated by the importance sampling (ISCV) [25]; however, WAIC seems consistent in that case [54]. Some outliers can become leverage points in practical uses. Thus, WAIC is considered to be numerically more robust than the ISCV.

The mathematical property of the variance of WAIC is as follows.

Theorem 3.7 *The WAIC error has same variance as the Bayesian generalization error:*

$$W_n - S_n + G_n = 2\lambda/n + o_p(1/n).$$

Note that $W_n - S_n$ and G_n are random variables but the summation of them is asymptotically deterministic $2\lambda/n$. This property means that the WAIC error has an inverse correlation to the Bayesian generalization error.

Widely applicable Bayesian information criterion (WBIC) is defined as follows. It uses the expectation by the tempered posterior distribution.

[*4] Some technical assumptions (such the relatively finite variance of the model) are needed but they are practically consistent [85]. As far as we know, widely used statistical models are considered to satisfy the assumptions.

Definition 3.6 (Tempered Posterior) *Let $\beta > 0$ be a positive constant. The tempered posterior $\psi^\beta(\theta|X^n)$ is defined by*

$$\psi^\beta(\theta|X^n) := \frac{1}{Z_n(\beta)}\varphi(\theta)\prod_{i=1}^n p(X_i|\theta)^\beta,$$

where the partition function $Z_n(\beta)$ is equal to

$$Z_n(\beta) = \int_{\mathcal{W}} \varphi(\theta)\prod_{i=1}^n p(X_i|\theta)^\beta \mathrm{d}\theta$$

and the scale constant β is called the inverse temperature in statistical mechanics.

Definition 3.7 (Widely Applicable Bayesian Information Criterion (WBIC)) *Let $\mathbb{E}_\theta^\beta$ be the expectation operator by the tempered posterior:*

$$\mathbb{E}_\theta^\beta[\cdot] := \int_{\mathcal{W}} [\cdot]\psi^\beta(\theta|X^n)\mathrm{d}\theta.$$

Widely applicable Bayesian information criterion (WBIC) is defined by the following random variable W_n':

$$W_n' := \mathbb{E}_\theta^\beta[nL_n(\theta)],$$

where $L_n(\theta)$ is the negative log likelihood

$$L_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log p(X_i|\theta)$$

and $\beta = 1/\log n$.

WBIC asymptotically approximates the free energy.

Theorem 3.8 *WBIC W_n' satisfies the following asymptotic equality:*

$$W_n' = \overline{F}_n + O_p(\sqrt{\log n}).$$

The probability that the true model is selected by minimizing WBIC converges to one when $n \to \infty$ Moreover, if the model is regular, then the difference between WBIC and BIC converges to zero in probability.

This is because $W_n' = nS_n + \lambda \log n + O_p(\sqrt{\log n})$ and $\overline{F}_n = nS_n + \lambda \log n - (m-1)\log\log n + O_p(1)$ have been proved. In this way, WBIC matches the free energy only up to the leading term. WBIC tends to underestimate the free energy and the accuracy of the model selection may be low. The underestimating term has been theoretically analyzed and an adjustment term has been proposed in order to make WBIC asymptotically unbiased [44]. Also, more precise approximation methods have been proposed: sBIC [24] and WsBIC [43].

Singular Bayesian information criterion (sBIC) have been proposed by Drton and Plummer [24]. sBIC is derived as a novel expansion of BIC based on Theorem 3.4. Let $\hat{\theta}$ be the (local)

maximum likelihood estimator. Namely, sBIC is defined as the first and second leading term of the free energy:

$$\text{sBIC} \triangleq nL_n(\hat{\theta}) + \lambda \log n - (m-1)\log\log n$$

and the sBIC seems to match the free energy higher-order terms than WBIC, where $A \triangleq B$ means that A is namely (intuitively) equal to B but is not mathematically equal to it*5. However, in fact, sBIC is calculated by solving fixed point simultaneous equation system and has consistency, i.e. the probability that the true model is selected converges to one when $n \to \infty$ [24]. sBIC uses the theoretical value of λ and m. In general, they are depend on the true distribution; however, in computing sBIC, the true distribution is referred to one of submodels. For example, we theoretically consider a model selection problem: selecting the number of hidden units i in three-layered neural network model (more general, the considered model is dominated by a control variable $i \in \mathbb{N} \cup \{0\}$ and all submodels are monotonically included: model $j \subset$ model i if $j < i$). For the sake of simplicity, the model whose number of hidden units is i is called the model i (written as $p_i(x|\theta)$) or that the model is i. In the same way, let $\varphi_i(\theta)$ be the prior when the model is i. When the candidate models are set to $i = 0, \ldots, H$, for each i, the submodels are considered to $j = 0, \ldots, i$. Let λ_{ij} and m_{ij} be the RLCT and its multiplicity when the model is i and the true distribution is the model j, respectively. If the RLCT and its multiplicity are clarified, we can immediately have the following two matrices:

$$\Lambda = \begin{bmatrix} \lambda_{00} & 0 & \ldots & 0 \\ \lambda_{10} & \lambda_{11} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{H0} & \lambda_{H1} & \ldots & \lambda_{HH} \end{bmatrix}, \tag{3.6}$$

$$\mathscr{M} = \begin{bmatrix} m_{00} & 0 & \ldots & 0 \\ m_{10} & m_{11} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ m_{H0} & m_{H1} & \ldots & m_{HH} \end{bmatrix}. \tag{3.7}$$

If an upper bound of the RLCT is only clarified, then the matrix Λ is constructed by the bounds and the positive entries of $\mathscr{M}$ are set to one. Let $\hat{\theta}_i$ be the (local) maximum likelihood estimator (MLE) of the model i and L'_{ij} satisfy

$$-\log L'_{ij} = nL_n^{(i)}(\hat{\theta}_i) + \lambda_{ij}\log n - (m_{ij}-1)\log\log n, \tag{3.8}$$

where the likelihood of the model i is denoted by $P_i(X^n|\theta) = \prod_{l=1}^n p_i(X_l|\theta)$ and $L_n^{(i)}(\theta) = -\frac{1}{n}\log P_i(X^n|\theta)$. Obviously, we have

$$L'_{ij} := P_i(X^n|\hat{\theta}_i)\frac{(\log n)^{m_{ij}-1}}{n^{\lambda_{ij}}}. \tag{3.9}$$

In this case, sBIC is calculated as Algorithm 1 *6.

*5 In scare quotes sense, A is "equal" to B (but $A \neq B$).

*6 The original paper defines sBIC as approximation of the log marginal likelihood $-\overline{F}_n$. Thus, in this thesis, the sign is reversed in some parts compared to the original version.

Algorithm 1 Calculation Algorithm of sBIC

Require: H:candidate model size, Λ:matrix of RLCTs, $\mathscr{M}$:matrix of multiplicities
P_i:likelihood of the model i, X^n:data, $\hat{\theta}_i$: MLE for the model i
Ensure: Calculation of sBIC
Allocate $L'_{ij} := P_i(X^n|\hat{\theta}_i)(\log n)^{m_{ij}-1}/n^{\lambda_{ij}}$
Allocate arrays $L, \mathrm{sBIC} \leftarrow$ initialize
for $i = 0$ to H **do**
  if $i == 0$ **then**
    $L[i] \leftarrow L'_{ii}$
    $sBIC[i] \leftarrow -\log L[i]$
  else
    $b \leftarrow -L'_{ii} + \sum_{j<i} L[j] \frac{\varphi_j(\theta_j)}{\varphi_i(\theta_i)}$
    $c \leftarrow \sum_{j<i} L[j] L'_{ij} \frac{\varphi_j(\theta_j)}{\varphi_i(\theta_i)}$
    $L[i] \leftarrow (-b + \sqrt{b^2 + 4c})/2$
    $\mathrm{sBIC}[i] \leftarrow -\log L[i]$
  end if
end for
return sBIC

In this algorithm, if $i \neq 0$, we solve the quadratic equation system

$$L[i]^2 + bL[i] - 4c = 0,\ i = 1, \ldots, H. \tag{3.10}$$

It is proved that this simultaneous equation has a unique positive solution [24]. Therefore, we compute and plug-in $L[i] \leftarrow (-b + \sqrt{b^2 + 4c})/2$. The value of sBIC for the model i is $\mathrm{sBIC}[i]$ in the above pseudo-code. Thus, the model $\mathrm{argmin}\{\mathrm{sBIC}\}$ is selected. Actually sBIC requires theoretical RLCTs; however, if they are clarified, it is empirically known that sBIC is more precise than WBIC in the sense of the accuracy of the model selection [24, 43].

For real computation, the value of the likelihood becomes small not to correctly evaluate the result of floating-point arithmetics in a computer. However, if the purpose of the computation is to solve the model selection problem by sBIC, we need the order of the maximal likelihoods $\{P_i(X^n|\hat{\theta}_i)\}$ of the model i as well as the exact maximal value of the maximal likelihood $\mathrm{ML}_n := \max_i\{P_i(X^n|\hat{\theta}_i)\}$. Hence, we can normalize sBIC by using $P_i(X^n|\hat{\theta}_i)/\mathrm{ML}_n$ instead of $P_i(X^n|\hat{\theta}_i)$ for $i = 0, \ldots, H$ in order to make the computation stable (private communication with Mathias Drton and Fumito Nakamura).

It is well-known that MLEs of singular models are neither stable nor unique even if the sample size is large enough to be treated as an asymptotic scale. Thus, there are some heuristics to make sBIC more precise. One of them is using variational Bayes estimator (VBE) instead of MLE [35]. For some of singular models like Gaussian mixture model (GMM), we can derive expectation-maximization (EM) algorithm to calculate its MLE; however, it is proved that there is no MLE in such mixture models [32]. As far as we know, when EM algorithm can be derived, deterministic variational Bayes method can also be constructed, for example GMM; hence, this heuristics does little to reduce the applicability of sBIC. According to the technical report [35], the model selection with the original (using MLE calculated by EM)

sBIC failed and that with the proposed (using VBE) method succeeded when the model is GMM with non-diagonal variance-covariance matrices. Another heuristics is considering the difference of the RLCTs and sBIC for $i = 0, \ldots, H$ [69]. This heuristic method returns the first point when the slope of the sequence $(i, \mathrm{sBIC}[i])$ becomes negative as the model selection result.

Lastly, we introduce widely applicable singular Bayesian information criterion (WsBIC). WsBIC is defined by a variant of sBIC which uses numerically-calculated RLCTs instead of theoretical ones [43]. In order to obtain numerical RLCTs, Imai proposed a consistent estimator $\hat{\lambda}^s_{\mathbb{V}}$ of an RLCT, where $s \in \mathbb{N}$ is the number of independent simulations.

Definition 3.8 (Imai's estimator of an RLCT) *Let $\mathbb{V}^\beta_\theta$ be the variance operator along with the tempered posterior:*

$$\mathbb{V}^\beta_\theta[\cdot] = \left[\mathbb{E}^\beta_\theta[(\cdot)^2] - \left\{\mathbb{E}^\beta_\theta[\cdot]\right\}^2\right].$$

$\hat{\lambda}^1_{\mathbb{V}}(X^n)$ is defined as

$$\hat{\lambda}^1_{\mathbb{V}}(X^n) = \beta^2 \mathbb{V}^\beta_\theta[\log P(X^n|\theta)],$$

where $P(X^n|\theta)$ is the likelihood: $\log P(X^n|\theta) = -nL_n(\theta)$.

Assume that there are (simulated) independent datasets $(\mathcal{D}_1, \ldots, \mathcal{D}_s)$ and these size is same as $n_{\mathrm{simulated}}$, i.e. $|\mathcal{D}_k| = n_{\mathrm{simulated}}$ and $\mathcal{D}_k$ is generated by the submodel for $k = 1, \ldots, s$. Let $\hat{\lambda}^s_{\mathbb{V}}$ be defined as

$$\hat{\lambda}^s_{\mathbb{V}} = \frac{1}{s}\sum_{k=1}^{s} \hat{\lambda}^1_{\mathbb{V}}(\mathcal{D}_k).$$

Theorem 3.9 (Imai) *Let λ be the theoretical RLCT of the considered model. $\hat{\lambda}^1_{\mathbb{V}}(X^n)$ is a consistent estimator, i.e. we have*

$$\hat{\lambda}^1_{\mathbb{V}}(X^n) = \lambda + O_p(1/\sqrt{\log n}).$$

Besides, $\hat{\lambda}^s_{\mathbb{V}}$ is asymptotically normal. In other words, the asymptotic distribution of $\hat{\lambda}^s_{\mathbb{V}}$ is determined as

$$\hat{\lambda}^s_{\mathbb{V}} \to_d \mathcal{N}(\lambda, \sigma_\lambda),$$

where

$$\sigma^2_\lambda = \frac{v_M}{4s \log n_{\mathrm{simulated}}},$$

v_M is a positive constant and $\mathcal{N}(\mu, \sigma)$ be a normal distribution whose mean and standard deviation are μ and σ, respectively.

Imai's estimator contains useful properties as the above. Moreover, it is also proved that there is a unique inverse temperature $\beta_0 = 1/\log n + o_p(1/\log n)$ which is depend on $(q(x), p(x|\theta), \varphi(\theta))$ and makes $\hat{\lambda}^1_{\mathbb{V}}(X^n)$ an unbiased estimator of λ. On the other hand, it requires Markov chain Monte Carlo (MCMC) sampling from the tempered posterior; thus, it cannot directly apply to Gibbs sampling which is difficult to sample from the tempered posterior in general.

WsBIC is obtained by replacing the theoretical RLCT λ to Imai's estimator $\hat{\lambda}^s_{\mathbb{V}}$. $\hat{\lambda}^s_{\mathbb{V}}$ can be calculated if the tempered posterior is realized; thus, the applicability of WsBIC is nearly equal to that of WBIC. Moreover, WsBIC is based on sBIC; thus, it is also more precise than WBIC nevertheless sBIC is less widely applicable than WBIC. Except for the computational cost, WsBIC outperforms sBIC and WBIC in terms of accuracy and applicability.

Chapter 4

Bayesian Generalization Error in Non-negative Matrix Factorization

In this chapter, we report the result of theoretical analysis for Bayesian generalization error in NMF. This chapter consists of five parts. First, in Sec. 4.1, we describe motivation of this theoretical study. Second, in Sec. 4.2, we state the main theorem with regard to Bayesian generalization error in NMF. Third, in Sec. 4.3, we prepare lemmas for the proof of the theorem. Fourth, in Sec 4.4, we prove the main theorem. Lastly, in Sec. 4.5, we discuss the theoretical results.

In the followings, $\theta = (U, V)$ is a parameter and $x = X$ is an observed random variable. Note that this chapter is based on the author's papers [37, 36, 33].

4.1 Motivation

Non-negative matrix factorization (NMF) [61, 18] has been applied to text mining [88], signal processing [51], bioinformatics [46], consumer analysis [47], and recommender systems [16]. NMF experiments discover the knowledge and predict the future unknown structures in the real world, however, the method suffers from many local minima and seldom reaches the global minimum. In addition, the results of numerical experiments strongly depend on the initial values; a rigorous method has not yet been established.

In order to resolve this difficulty, Bayesian inference for NMF has been established [18]. It uses, for numerical calculation of the Bayesian posterior distribution, Gibbs sampling method which is a kind of Markov chain Monte Carlo method (MCMC). Bayesian NMF is known as a more robust method than usual recursive methods of NMF since it numerically realizes the posterior distribution; the parameters are subject to a probability distribution and that makes it possible to grasp the degree of fluctuation of the learning/inference result. As is described later, in general, Bayesian method has higher estimation accuracy than maximum likelihood estimation and maximum posterior estimation if the model has hierarchical structures or hidden variables, like NMF.

On the other hand, the variational Bayesian algorithm (VB) for NMF has also been established [18], with being inspired the mean field approximation. The variational Bayesian NMF algorithm (VBNMF) also results more numerically stable than usual recursive algorithms as

VB approximates the Bayesian posterior distribution. Moreover, VBNMF computes faster than usual Bayesian inference such an MCMC. However, its free energy (called the variational free energy) is larger than the Bayesian free energy, since VB ascends the evidence lower bound but it is not the true model evidence. Note that the marginal likelihood is also called the model evidence and the negative logarithm value of the evidence lower bound is equal to the variational free energy. From the above, it is important to clarify the approximation error of variational inference for not only theoretical reasons but also practical points of view.

As mentioned Chap. 3, researchers have studied RLCTs of many statistical models to clarify theoretical behaviors of the generalization error and the free energy. Moreover, for several statistical models, the variational free energy was proved that it asymptotically equals $nS_n + \lambda_{\rm vb} \log n + O_p(1)$,where $\lambda_{\rm vb}$ is a learning coefficient and it depends on the model. Normal mixture models [78], hidden Markov models [42], and NMF [49] are such examples. Kohjima proved the following theorem about the learning coefficient of VBNMF [49].

Theorem 4.1 (Kohjima) *Let the elements of the data matrices x_{ij} $(i = 1, \ldots, M;\ j = 1, \ldots, N)$ be independently generated from the Poisson distribution whose mean is equal to the (i, j) element of U_0V_0, where the number of columns in U_0 $(=$ the number of rows in $V_0)$ is H_0; called the non-negative rank of U_0V_0 [19].*

Let the likelihood model and the prior be the following Poisson and gamma distributions, respectively:

$$p(X|U,V) = \prod_{i=1}^{M}\prod_{j=1}^{N} \frac{((UV)_{ij})^{x_{ij}}}{x_{ij}!} e^{-(UV)_{ij}},$$

$$\varphi(U,V) = \prod_{i=1}^{M}\prod_{k=1}^{H} \left(\frac{\theta_U^{\phi_U}}{\Gamma(\theta_U)} u_{ik}^{\phi_U - 1} e^{-\theta_U u_{ik}} \right) \prod_{k=1}^{H}\prod_{j=1}^{N} \left(\frac{\theta_V^{\phi_V}}{\Gamma(\theta_V)} v_{kj}^{\phi_V - 1} e^{-\theta_V v_{kj}} \right),$$

where $\phi_U, \theta_U, \phi_V, \theta_V > 0$ are hyperparameters, the size of U and V are $M \times H$ and $H \times N$, and $(UV)_{ij}$ is the (i, j) entry of UV, respectively.

Then, the variational free energy $\bar{F}_n^{\rm vb}$ satisfies the following asymptotic equality:

$$\bar{F}_n^{\rm vb} = nS_n + \lambda_{\rm vb} \log n + O_p(1)\ (n \to \infty),$$

where

$$\lambda_{\rm vb} = \begin{cases} (H - H_0)(M\phi_U + N\phi_V) + \frac{1}{2}H_0(M + N), & \text{if } M\phi_U + N\phi_V < \frac{M+N}{2} \\ \frac{1}{2}H(M + N), & \text{otherwise}. \end{cases}$$

In general, the learning coefficient of VB may not be equal to but becomes an upper bound of the RLCT: $\lambda_{\rm vb} \geqq \lambda$, since the variational free energy is larger than the free energy even if the sample size diverges infinity. Unfortunately, the variational generalization error is *not* equal to $\lambda_{\rm vb}/n$, asymptotically. Besides, as described below, the variational inference seeks the mode of the true posterior and the variational posterior distributions tend to concentrate in a fraction of true posterior distributions. One might say that a variational posterior distribution is approximately equivalent to a posterior distribution; however, for these reasons, variational inference is just an approximation of Bayesian inference and they are not equivalent. As far as we know, there has been no direct theoretical comparison between them, except for [92]. The

difference between VB and the local variational approximation, which is an approximation of VB, has been studied theoretically [76]; however, for Bayesian inference and VB, there exists few theoretical comparison.

VBNMF has been devised [18], and the exact learning coefficient of VBNMF has been derived [49]. Nevertheless, the variational approximation error has not been clarified since the RLCT of NMF has been unknown. If the prior distribution is strictly and entirely positive and bounded analytic function on the domain, then an upper bound of the RLCT of NMF has been proved [37, 36]. If the non-negative restriction is not assumed for matrix factorization, then the exact value of the RLCT has been clarified as those of reduced rank regression models [10]. However, the RLCT has been unknown in the case of that the prior is a gamma distribution, which may be zero.

In this chapter, we give an upper bound of the RLCT of NMF λ when the prior is a gamma distribution, which determines an upper bound of the Bayesian generalization error and the free energy in that case. Moreover, by comparing λ with $\lambda_{\rm vb}$, we also derive a lower bound of the variational approximation error in VBNMF.

4.2 Main Theorem

Let $\mathrm{M}(M,N,C)$ be a set of $M \times N$ matrices whose elements are in C, where C is a subset of $\mathbb{R}$. Let K be a compact subset of $\mathbb{R}_{\geqq 0} = \{x \in \mathbb{R} | x \geqq 0\}$ and let K_0 be a compact of subset $\mathbb{R}_{>0} = \{x \in \mathbb{R} | x > 0\}$. We denote that $U \in \mathrm{M}(M,H,K)$, $V \in \mathrm{M}(H,N,K)$ and $U_0 \in \mathrm{M}(M,H_0,K_0)$, $V_0 \in \mathrm{M}(H_0,N,K_0)$ are the NMF result of U_0V_0 such that H_0 is the non-negative rank [19] of U_0V_0, i.e. they give the minimal H_0, where $H \geqq H_0$ and $\{(x,y,a,b) \in K \times K_0 | xy = ab\} \neq \emptyset$.

Definition 4.1 (RLCT of NMF) *Assume that the largest pole of the function of one complex variable z,*

$$\zeta(z) = \int_{\mathrm{M}(M,H,K)} \mathrm{d}U \int_{\mathrm{M}(H,N,K)} \mathrm{d}V \Big(\|UV - U_0V_0\|^2 \Big)^z \varphi(U,V)$$

is equal to $(-\lambda)$. Then λ is said to be the RLCT of NMF.

According to one of the first Bayesian NMF paper [18], we set the prior as gamma distributions:

$$\varphi(U,V) = \mathrm{Gam}(U|\phi_U,\theta_U)\mathrm{Gam}(V|\phi_V,\theta_U),$$

where

$$\mathrm{Gam}(U|\phi_U,\theta_U) = \prod_{i=1}^{M}\prod_{k=1}^{H} \frac{\theta_U^{\phi_U}}{\Gamma(\theta_U)} u_{ik}^{\phi_U - 1} e^{-\theta_U u_{ik}},$$

$$\mathrm{Gam}(V|\phi_V,\theta_V) = \prod_{k=1}^{H}\prod_{j=1}^{N} \frac{\theta_V^{\phi_V}}{\Gamma(\theta_V)} v_{ik}^{\phi_V - 1} e^{-\theta_V v_{ik}},$$

and $\phi_U, \theta_U, \phi_V, \theta_V > 0$.

In this chapter, we prove the following theorems.

Theorem 4.2 *If the prior is the above gamma distributions, then the RLCT of NMF λ satisfies the following inequality:*

$$\lambda \leqq \frac{1}{2}\left[(H-H_0)\min\{M\phi_U, N\phi_V\} + H_0(M+N-2) + \delta_{H_0}\right],$$

where $\delta_{H_0} = [H_0 \equiv 1 (\mathrm{mod}\ 2)]$. The equality holds when $H = H_0 = 1$ or 2.

We prove this theorem in the next section. As two applications of this theorem, we obtain an upper bounds of the free energy and Bayesian generalization error of NMF in this case. The following theorem shows a statistical bound of Bayesian estimation of NMF.

Theorem 4.3 *Let the probability density functions of $X \in \mathrm{M}(M,N,K)$ be $q(X)$ and $p(X|U,V)$, which represent a true distribution and a learning machine respectively defined by*

$$q(X) = \mathrm{Poi}(X|U_0V_0),$$
$$p(X|U,V) = \mathrm{Poi}(X|UV),$$

where

$$\mathrm{Poi}(X|A) = \prod_{i=1}^{M}\prod_{j=1}^{N}\frac{(a_{ij})^{x_{ij}}}{x_{ij}!}e^{-a_{ij}},\ X = (x_{ij})_{i=1,j=1}^{M,N},\ A = (a_{ij})_{i=1,j=1}^{M,N}.$$

Also let $\varphi(U,V) = \mathrm{Gam}(U|\phi_U,\theta_U)\mathrm{Gam}(V|\phi_V,\theta_U)$. Then, the normalized free energy F_n and the expected generalization error $\mathbb{E}[G_n]$ satisfies the following inequality:

$$F_n \leqq \frac{1}{2}\left[(H-H_0)\min\{M\phi_U, N\phi_V\} + H_0(M+N-2) + \delta_{H_0}\right]\log n + O_p(1),$$
$$\mathbb{E}[G_n] \leqq \frac{1}{2n}\left[(H-H_0)\min\{M\phi_U, N\phi_V\} + H_0(M+N-2) + \delta_{H_0}\right] + o\left(\frac{1}{n}\right).$$

In Theorem 4.3, we study a case when a set of random matrices $X^n = X_1, X_2, ..., X_n$ are observed and the true decomposition U_0 and V_0 are statistically estimated. Actually sometimes NMF has studied in the case when only one target matrix is decomposed, however, in general, decomposition of a set of independent matrices should be studied because target matrices are often obtained daily, monthly, or different places for purchase study [47]. Other situation for observing multiple data matrices is traffic data analysis [21]. In such cases, decomposition of a set of matrices results in statistical inference. We consider this situation and it is common to [49]. A statistical model $p(X|U,V)$ which has parameters (U,V) are employed for estimation. Then the free energy and generalization error of Bayesian estimation is given by this theorem.

Moreover, we also derive a lower bound of the variational approximation error in VBNMF as Theorem 4.4 in the following. In VB, the posterior $\psi(\theta|X^n)$ is approximated by the variational posterior $\psi^{\mathrm{vb}}(\theta|X^n)$ which is a distribution of independent parameters like $\psi^{\mathrm{vb}}(\theta|X^n) := \psi_1(\theta_1|X^n)\psi_2(\theta_2|X^n)$. The variational approximation error is defined by the KL divergence

from the variational posterior to the true one: $\mathrm{KL}(\psi^{\mathrm{vb}}\|\psi)$. Developing it, we have

$$\mathrm{KL}(\psi^{\mathrm{vb}}\|\psi) = \int \mathrm{d}\theta\psi^{\mathrm{vb}}(\theta|X^n)\log\frac{\psi^{\mathrm{vb}}(\theta|X^n)}{\psi(\theta|X^n)} \tag{4.1}$$

$$= \int \mathrm{d}\theta\psi^{\mathrm{vb}}(\theta|X^n)\left(\log\psi^{\mathrm{vb}}(\theta|X^n) - \log\frac{\prod_{i=1}^n p(X_i|\theta)\varphi(\theta)}{Z_n}\right) \tag{4.2}$$

$$= \int \mathrm{d}\theta\psi^{\mathrm{vb}}(\theta|X^n)\left(\log\psi^{\mathrm{vb}}(\theta|X^n) - \log\prod_{i=1}^n p(X_i|\theta)\varphi(\theta) + \log Z_n\right) \tag{4.3}$$

$$= \int \mathrm{d}\theta\psi^{\mathrm{vb}}(\theta|X^n)\log\frac{\psi^{\mathrm{vb}}(\theta|X^n)}{\prod_{i=1}^n p(X_i|\theta)\varphi(\theta)} - \overline{F_n}. \tag{4.4}$$

The first term in Eq. (4.4) is called the variational free energy $\bar{F}_n^{\mathrm{vb}}$. Because of $\mathrm{KL}(\psi^{\mathrm{vb}}\|\psi) \geqq 0$, $\bar{F}_n^{\mathrm{vb}} \geqq \overline{F}_n$. It is the objective function of VB. If θ_1 and θ_2 such that $\theta = (\theta_1, \theta_2)$ were independent, then the variational posterior would be same as the true posterior and the variational approximation error would become zero. However, in general, θ_1 and θ_2 such that $\theta = (\theta_1, \theta_2)$ are not independent. Thus, VB minimizes the variational free energy.

Theorem 4.4 *Let the variational free energy of VBNMF be $\bar{F}_n^{\mathrm{vb}}$. Then, the following inequality is attained:*

$$\bar{F}_n^{\mathrm{vb}} - \overline{F}_n \geqq \underline{\lambda}\log n + O_p(1),$$

where

$$\underline{\lambda} = \begin{cases} \frac{1}{2}[(H-H_0)(M\phi_U+N\phi_V+\max\{M\phi_U, N\phi_V\})-\delta_{H_0}]+H_0, & \text{if } M\phi_U+N\phi_V<\frac{M+N}{2} \\ \frac{1}{2}[(H-H_0)(M+N-\min\{M\phi_U, N\phi_V\})-\delta_{H_0}]+H_0, & \text{otherwise.} \end{cases}$$

The framework of VB and the definition of the variational free energy are described in Appendix. Theorem 4.4 gives a lower bound of the difference of the free energy between the variational approximation and the true.

4.3 Preparation

In order to prove Theorem 4.2, we use the following five lemmas.

Lemma 4.1 *Let $q(X)$ and $p(X|U,V)$ be the probability density functions of a non-negative matrix X, which represent a true distribution and a learning machine respectively defined by*

$$q(X) = \mathrm{Poi}(X|U_0V_0),$$
$$p(X|U,V) = \mathrm{Poi}(X|UV),$$

where $\mathrm{Poi}(X|W)$ is a probability density function of the Poisson distribution with average W. Then, the RLCT of the KL divergence $\sum_X q(X)\log\frac{q(X)}{p(X|U,V)}$ is equal to the RLCT of NMF in Definition 4.1:

$$\|UV - U_0V_0\|^2 \sim \sum_X q(X)\log\frac{q(X)}{p(X|U,V)}.$$

Lemma 4.2 *Let λ be the absolute value of the maximum pole of*

$$\zeta(z) = \iint dUdV \left(\|UV\|^2\right)^z \mathrm{Gam}(U|\phi_U, \theta_U)\mathrm{Gam}(V|\phi_V, \theta_V).$$

Then,

$$\lambda = \frac{H \min\{M\phi_U, N\phi_V\}}{2}$$

holds; this is the equality of Theorem 4.2 in the case $H_0 = 0$.

Lemma 4.3 *If $H_0 = H = 1$, the equal sign of the Theorem4.2 holds:*

$$\lambda = \frac{M + N - 1}{2}.$$

Lemma 4.4 *If $H_0 = H = 2$, the equal sign of the Theorem4.2 holds:*

$$\lambda = M + N - 2.$$

Lemma 4.5 *If $H = H_0$, the Theorem4.2 is attained:*

$$\lambda \leqq \frac{H_0(M + N - 2) + [H_0 \equiv 1(\mathrm{mod}\ 2)]}{2}.$$

Let the entries of the matrices (U, V) be

$$U = (u_1, \ldots, u_H),\ u_k = (u_{ik})_{i=1}^M,$$
$$V = (v_1, \ldots, v_H)^T,\ v_k = (v_{kj})_{j=1}^N,$$

and the ones of (U_0, V_0) be

$$U_0 = (u_1^0, \ldots, u_{H_0}^0),\ u_k^0 = (u_{ik}^0)_{i=1}^M,$$
$$V_0 = (v_1^0, \ldots, v_{H_0}^0)^T,\ v_k^0 = (v_{kj}^0)_{j=1}^N,$$

respectively.

In the following, these lemmas are proved.

Proof of Lemma 4.1. Let $(w_{ij}) = W = UV$ and $(w_{ij}^0) = W_0 = U_0V_0$. Calculating the KL divergence $\sum_X q(X) \log \frac{q(X)}{p(X|U,V)}$, we have

$$\sum_X q(X) \log \frac{q(X)}{p(X|U,V)} \tag{4.5}$$

$$= \sum_X q(X) \log \frac{\prod_{i=1}^M \prod_{j=1}^N \frac{(w_{ij}^0)^{x_{ij}}}{x_{ij}!} e^{-w_{ij}^0}}{\prod_{i=1}^M \prod_{j=1}^N \frac{(w_{ij})^{x_{ij}}}{x_{ij}!} e^{-w_{ij}}} \tag{4.6}$$

$$= \sum_X q(X) \sum_{i=1}^M \sum_{j=1}^N \{x_{ij} \log w_{ij}^0 - \log x_{ij}! - w_{ij}^0 - (x_{ij} \log w_{ij} - \log x_{ij}! - w_{ij})\} \tag{4.7}$$

$$= \sum_X q(X) \sum_{i=1}^M \sum_{j=1}^N \left\{ x_{ij} \log \frac{w_{ij}^0}{w_{ij}} - (w_{ij}^0 - w_{ij}) \right\}. \tag{4.8}$$

Because of

$$\sum_{x_{ij}=0}^{\infty} \frac{(w_{ij}^0)^{x_{ij}}}{x_{ij}!} e^{-w_{ij}^0} x_{i'j'} = \begin{cases} w_{ij}^0 & (i,j) = (i',j'), \\ x_{i'j'} & (i,j) \neq (i',j'), \end{cases} \tag{4.9}$$

we have

$$\sum_X q(X) \log \frac{q(X)}{p(X|U,V)} \tag{4.10}$$

$$= \sum_X q(X) \sum_{i=1}^{M} \sum_{j=1}^{N} \left\{ x_{ij} \log \frac{w_{ij}^0}{w_{ij}} - (w_{ij}^0 - w_{ij}) \right\} \tag{4.11}$$

$$= \sum_X \sum_{i=1}^{M} \sum_{j=1}^{N} \left(\prod_{i=1}^{M} \prod_{j=1}^{N} \frac{(w_{ij}^0)^{x_{ij}}}{x_{ij}!} e^{-w_{ij}^0} \right) \left\{ x_{ij} \log \frac{w_{ij}^0}{w_{ij}} - (w_{ij}^0 - w_{ij}) \right\} \tag{4.12}$$

$$= \sum_X \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{(w_{ij}^0)^{x_{ij}}}{x_{ij}!} e^{-w_{ij}^0} \left\{ x_{ij} \log \frac{w_{ij}^0}{w_{ij}} - (w_{ij}^0 - w_{ij}) \right\} \tag{4.13}$$

$$= \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{x_{ij}=0}^{\infty} \frac{(w_{ij}^0)^{x_{ij}}}{x_{ij}!} e^{-w_{ij}^0} \left\{ x_{ij} \log \frac{w_{ij}^0}{w_{ij}} - (w_{ij}^0 - w_{ij}) \right\} \tag{4.14}$$

$$= \sum_{i=1}^{M} \sum_{j=1}^{N} \left\{ w_{ij}^0 \log \frac{w_{ij}^0}{w_{ij}} - (w_{ij}^0 - w_{ij}) \right\}. \tag{4.15}$$

All we have to prove is

$$(w_{ij}^0 - w_{ij})^2 \sim w_{ij}^0 \log \frac{w_{ij}^0}{w_{ij}} - (w_{ij}^0 - w_{ij}) \tag{4.16}$$

for $i = 1, \ldots, M$ and $j = 1, \ldots, N$, since $\|W - W_0\|^2 = \sum_{i=1}^{M} \sum_{j=1}^{N} (w_{ij}^0 - w_{ij})^2$.

Let $f(a,b) = (a-b)^2$ and

$$g(a,b) = a \log \frac{a}{b} - (a-b) \tag{4.17}$$

on $(a,b) \in (\mathbb{R}_{\geqq 0})^2$. Owing to

$$\partial_a g(a,b) = \log a - \log b,$$
$$\partial_b g(a,b) = 1 - a/b,$$

and that a log function is monotone increasing, we obtain

$$\partial_a g(a,b) = \partial_b g(a,b) = 0 \Leftrightarrow a = b. \tag{4.18}$$

Signs of the above partial derivations are

$$\partial_a g(a,b) > 0 \wedge \partial_b g(a,b) < 0 \,\text{in case of a} > \text{b}, \tag{4.19}$$

$$\partial_a g(a,b) < 0 \wedge \partial_b g(a,b) > 0 \,\text{in case of a} < \text{b}. \tag{4.20}$$

On account of the above and smoothness, the increase or decrease and convexity of $g(a,b)$ is the same as those of $f(a,b)$. Hence there exists $c_1, c_2 > 0$ such that

$$c_1 f(a,b) \leqq g(a,b) \leqq c_2 f(a,b). \tag{4.21}$$

i.e. $K(a,b) \sim (b-a)^2$. Therefore, we have

$$\|UV - U_0V_0\|^2 \sim \sum_X q(X) \log \frac{q(X)}{p(X|U,V)}. \tag{4.22}$$

□

Put $\Phi(U,V) = \|UV\|^2$. The prior $\varphi(U,V)$ is gamma distributions, hence,

$$\varphi^{-1}(0) = \left(\bigcup_{i=1}^{M} \bigcup_{k=1}^{H} \{(U,V) \mid u_{ik} = 0\} \right) \cup \left(\bigcup_{k=1}^{H} \bigcup_{j=1}^{N} \{(U,V) \mid v_{kj} = 0\} \right).$$

If $\Phi^{-1}(0) \neq \emptyset$, $\varphi^{-1}(0) \neq \emptyset$, and $\Phi^{-1}(0) \cap \varphi^{-1}(0) = \emptyset$, then $\varphi(U.V)$ does not correspond to the maximum pole of the zeta function. However, in the case Lemma 4.2, the set $\Phi^{-1}(0)$ has intersections with $\varphi^{-1}(0)$. Here, we prove Lemma 4.2.

Proof of Lemma 4.2. We consider simultaneous resolution $\|UV\|^2 = 0$ and $\varphi(U,V) = 0$ since $\{(U,V) \mid \|UV\|^2 = 0\}$ has intersections with $\varphi^{-1}(0)$.

The zeta function is equal to

$$\zeta(z) = \iint \mathrm{d}U\mathrm{d}V \left(\|UV\|^2\right)^z \mathrm{Gam}(U|\phi_U, \theta_U)\mathrm{Gam}(V|\phi_V, \theta_V). \tag{4.23}$$

Since elements of matrices is nonnegative, the inequality

$$\sum_{k=1}^{H} u_{ik}^2 v_{kj}^2 \leqq \left(\sum_{k=1}^{H} u_{ik} v_{kj} \right)^2 \leqq H \sum_{k=1}^{H} u_{ik}^2 v_{kj}^2 \tag{4.24}$$

holds. Thus, we have

$$\sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{H} u_{ik}^2 v_{kj}^2 \leqq \|UV\|^2 \leqq H \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{H} u_{ik}^2 v_{kj}^2. \tag{4.25}$$

As a log canonical threshold is not changed by any constant coefficient and it is order isomorphic; $\exists (c_1, c_2) \in \mathbb{R}^2$ s.t. $c_1 F \leqq G \leqq c_2 F \Rightarrow F \sim G$, all we have to do is calculating the RLCT of

$$\sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{H} u_{ik}^2 v_{kj}^2 = \sum_{k=1}^{H} \left(\sum_{i=1}^{M} u_{ik}^2 \right) \left(\sum_{j=1}^{N} v_{kj}^2 \right). \tag{4.26}$$

The RLCT λ becomes a sum of each ones about k. For each k, we blow-up of variables (u_{ik}), (v_{kj}) each other. Simultaneously, we also consider blowing-ups for $\prod_{k=1}^{H}(\prod_{i=1}^{M} u_{ik}^{\phi_U-1}\prod_{j=1}^{N} v_{kj}^{\phi_V-1})$ and the determinant $|J|$ of their Jacobi matrix J. $|J|$ is called Jacobian.

Let A be either U or V and $\mathfrak{b}_{rc}[A] : (U,V) \mapsto (U'.V')$ be a blowing-up such that the (i,k)-entry of U' and the (k,j)-entry of V' are equal to U'_{ik} and V'_{kj}:

$$\begin{cases} U'_{ik} = \begin{cases} u_{ik} & (i=r \text{ and } k=c) \text{ or } k \neq c \\ u_{rc}u_{ik} & i \neq r \text{ and } k=c \end{cases}, \quad V'_{kj} = v_{kj} & A = U, \\ U'_{ik} = u_{ik}, \quad V'_{kj} = \begin{cases} v_{kj} & k=r \text{ and } j=c \\ v_{rc}v_{kj} & \text{otherwise} \end{cases} & A = V \end{cases}. \tag{4.27}$$

For example, let $M = N = H = 2$ and apply $\mathfrak{b}_{11}[U]$ to

$$\|UV\|^2 \sim \sum_{k=1}^{H}\left(\sum_{i=1}^{M} u_{ik}^2\right)\left(\sum_{j=1}^{N} v_{kj}^2\right). \tag{4.28}$$

Then, we have

$$U' = \begin{pmatrix} u_{11} & u_{12} \\ u_{11}u_{21} & u_{22} \end{pmatrix}, \quad V' = \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix} \tag{4.29}$$

and

$$\sum_{k=1}^{2}\left(\sum_{i=1}^{2} u_{ik}^2\right)\left(\sum_{j=1}^{2} v_{kj}^2\right) \mapsto u_{11}^2(1+u_{21}^2)(v_{11}^2+v_{12}^2) + (u_{12}^2+u_{22}^2)(v_{21}^2+v_{22}^2), \tag{4.30}$$

$$\prod_{k=1}^{2}(\prod_{i=1}^{2} u_{ik}^{\phi_U-1}\prod_{j=1}^{2} v_{kj}^{\phi_V-1}) \mapsto u_{11}^{2(\phi_U-1)}u_{21}^{\phi_U-1}v_{11}^{\phi_V-1}v_{12}^{\phi_V-1}u_{12}^{\phi_U-1}u_{22}^{\phi_U-1}v_{21}^{\phi_V-1}v_{22}^{\phi_V-1}. \tag{4.31}$$

The Jacobi matrix is an 8×8 matrix as the below:

$$J = \frac{\partial(U',V')}{\partial(U,V)} = \begin{pmatrix} 1 & u_{12} & u_{21} & u_{22} & v_{11} & v_{12} & v_{21} & v_{22} \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & u_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{4.32}$$

and its determinant is $|J| = (u_{11})^1$.

We arbitrarily take $k \in \{1, \ldots, H\}$ and fix it. In the general case, because of the symmetry of the variables, only the blowing-ups $\mathfrak{b}_{1k}[U]$ and $\mathfrak{b}_{k1}[V]$ should be treated. Hence, we have

$$\left(\sum_{i=1}^{M} u_{ik}^2\right)\left(\sum_{j=1}^{N} v_{kj}^2\right) \mapsto u_{1k}^2 v_{k1}^2 \left(1 + \sum_{i=2}^{M} u_{ik}^2\right)\left(1 + \sum_{j=2}^{N} v_{kj}^2\right), \tag{4.33}$$

$$\prod_{i=1}^{M} u_{ik}^{\phi_U - 1} \prod_{j=1}^{N} v_{kj}^{\phi_V - 1} \mapsto u_{1k}^{M\phi_U - M} v_{k1}^{N\phi_V - N} \prod_{i=2}^{M} u_{ik}^{\phi_U - 1} \prod_{j=2}^{N} v_{kj}^{\phi_V - 1}. \tag{4.34}$$

The Jacobian $|J|_k$ is equal to $|J|_k = u_{1k}^{M-1} v_{k1}^{N-1}$. The term $\left(1 + \sum_{i=2}^{M} u_{ik}^2\right)\left(1 + \sum_{j=2}^{N} v_{kj}^2\right)$ is strictly positive; thus, we should consider the maximum pole of the following meromorphic function:

$$\tilde{\zeta}(z) = \iint \mathrm{d}U\mathrm{d}V (u_{1k}^2 v_{k1}^2)^z \left(u_{1k}^{M\phi_U - M} v_{k1}^{N\phi_V - N} \prod_{i=2}^{M} u_{ik}^{\phi_U - 1} \prod_{j=2}^{N} v_{kj}^{\phi_V - 1} \right) |J|_k \tag{4.35}$$

$$= C \iint \mathrm{d}u_{1k}\mathrm{d}v_{k1} u_{1k}^{2z} v_{k1}^{2z} u_{1k}^{M\phi_U - M} v_{k1}^{N\phi_V - N} u_{1k}^{M-1} v_{k1}^{N-1} \tag{4.36}$$

$$= C \iint \mathrm{d}u_{1k}\mathrm{d}v_{k1} u_{1k}^{2z + M\phi_U - 1} v_{k1}^{2z + N\phi_V - 1} \tag{4.37}$$

$$= D \frac{1}{2z + M\phi_U} \times \frac{1}{2z + N\phi_V}, \tag{4.38}$$

where C and D are positive constants. The poles are $z = -M\phi_U/2, -N\phi_V/2$; therefore, the RLCT λ_k is

$$\lambda_k = \frac{\min\{M\phi_U, N\phi_V\}}{2}.$$

Let U_k be $(u_{ik})_{i=1}^{M}$, V_k be $(v_{kj})_{j=1}^{N}$, and $\Phi_k(U_k, V_k)$ be $\left(\sum_{i=1}^{M} u_{ik}^2\right)\left(\sum_{j=1}^{N} v_{kj}^2\right)$ for each k. The RLCT λ becomes a sum of each ones about k;

$$\lambda = \sum_{k=1}^{H} \lambda_k, \tag{4.39}$$

since $\Phi(U, V) \sim \sum_{k=1}^{H} \Phi_k(U_k, V_k)$ and the RLCT of Φ_k is λ_k. Thus, we have

$$\lambda = \frac{H \min\{M\phi_U, N\phi_V\}}{2}. \tag{4.40}$$

□

Remark 4.1 *Lemma 4.2 means that the equality in Main Theorem holds if $U_0 V_0 = O$, i.e. $H_0 = 0$.*

Next, the exact values of the RLCTs in the case $H = H_0 = 1$ and $H = H_0 = 2$ (Lemma 4.3 and 4.4) are derived. To prove Lemma 4.3 and 4.4, we show the following lemma.

Lemma 4.6 *The RLCT of reduced rank regression $\lambda_{\rm MF}$,which is equal to the RLCT of matrix factorization, is a lower bound of the RLCT of NMF λ, i.e.*

$$\lambda_{\rm MF} \leqq \lambda. \tag{4.41}$$

Proof of Lemma 4.6. Let A, B, A_0 and B_0 be $M \times H$, $H \times N$, $M \times r$ and $r \times N$ matrices, respectively. The RLCT of matrix factorization is defined by the absolute maximum pole of the zeta function

$$\zeta(z) = \iint \|AB - A_0B_0\|^{2z} \varphi_{\rm MF}(A,B) {\rm d}A{\rm d}B, \tag{4.42}$$

where $\varphi_{\rm MF}$ is a prior distribution of (A,B) whose support includes negative real numbers (see also Definition 4.2 in the below). On the other hand, φ is a prior distribution of non-negative matrices (U,V); thus, there exists $c>0$ such that

$$\varphi(A,B) \leqq c\varphi_{\rm MF}(A,B). \tag{4.43}$$

Because of the property of the maximum pole of the zeta function [81], we obtain $\lambda_{\rm MF} \leqq \lambda$.

□

Remark 4.2 *As compared to the RLCT of reduced rank regression, the domain of the prior of NMF is considered that includes negative real numbers as 0 density:*

$$\varphi(A,B) = \begin{cases} \varphi(U,V) & A=U \geqq 0 \wedge B=V \geqq 0, \\ 0 & A \text{ or } B \text{ has negative entries}. \end{cases} \tag{4.44}$$

The RLCT of reduced rank regression was clarified in the all case [10] as the follwoing theorem.

Theorem 4.5 (Aoyagi) *The RLCT of reduced rank regression $\lambda_{\rm MF}$ is as follows:*
(1) *If $N+r \leqq M+H \wedge M+r \leqq N+H \wedge H+r \leqq M+N$,*
(1−1) *in the case $M+H+N+r$ is even,*

$$\lambda_{\rm MF} = \frac{-(H+r)^2 - M^2 - N^2 + 2\{(H+r)(M+N)+MN\}}{8}.$$

(1−2) *in the case $M+H+N+r$ is odd,*

$$\lambda_{\rm MF} = \frac{1-(H+r)^2 - M^2 - N^2 + 2\{(H+r)(M+N)+MN\}}{8}.$$

(2) *Else if $M+H<N+r$,*

$$\lambda_{\rm MF} = \frac{HM - Hr + Nr}{2}.$$

(3) *Else if $N+H<M+r$,*

$$\lambda_{\rm MF} = \frac{HN - Hr + Mr}{2}.$$

(4) *Or else, i.e. in the case* $M + N < H + r$*,*

$$\lambda_{\rm MF} = \frac{MN}{2}.$$

Especially, if $H = r$*, then*

$$\lambda_{\rm MF} = \frac{H(M + N - H)}{2}.$$

In the case $(1-2)$*, the multiplicity is two:* $m_{\rm MF} = 2$*. Otherwise, it equals one:* $m_{\rm MF} = 1$*.*

By using Theorem 4.5 and Lemma 4.6, the following inequalities hold.

Corollary 4.1 *If* $H_0 = H = 1$*, then*

$$\frac{M + N - 1}{2} \leqq \lambda.$$

Besides, if $H_0 = H = 2$*, then*

$$M + N - 2 \leqq \lambda.$$

In the case $H = H_0$, we can treat the entries of U and V as positive numbers since the entries of U_0 and V_0 are positive for any $(U_0, V_0) \in \Phi^{-1}(0)$ and there exists a neighborhood such that its all elements are positive matrices owing to continuity of real numbers. Thus, in this case, the prior $\varphi(U, V)$ is positive and bounded; it does not affect the RLCT. Then, because of Corollary 4.1, we only have to prove inversed inequalities of the aboves in order to derive Lemma 4.3 and 4.4.

Put $\Phi(U, V) = \|UV - U_0V_0\|^2$. In general, if $\Phi^{-1}(0)$ has an R-dimensional sub-manifold, then the rank of its Hesse matrix $\nabla^2\Phi$ is at most $H(M + N) - R$. According to [81] (by using implicit function theorem), we have

$$\lambda \leqq \frac{H(M + N) - R}{2}. \tag{4.45}$$

Proof of Lemma 4.3. Let $H = H_0 = 1$ and $p > 0$. Because we consider the situation $\Phi(U, V) = \|UV - U_0V_0\|^2 = 0$ and U and V are vectors, we can treat the elements in U and V as strictly positive. Hence, the entries of pU and $p^{-1}V$ are also strictly positive. We put $\mathcal{P} := \{p \in \mathbb{R} | p > 0\}$ and it is a 1-dimensional sub-manifold of $\Phi^{-1}(0)$; thus, by using the inequality (4.45), we have

$$\lambda \leqq \frac{M + N - 1}{2}. \tag{4.46}$$

Owing to Corollary 4.1, we obtain

$$\lambda = \frac{M + N - 1}{2}. \tag{4.47}$$

□

Remark 4.3 *The author initially derived Lemma 4.3 in a different way: by using degenerating of ideal and mathematical induction about* M *and* N*. The initial proof of Lemma 4.3 is described in Appendix.*

In a way similar to Proof of Lemma 4.3, we prove Lemma 4.4 in the following.

Proof of Lemma 4.4. Let $H = H_0 = 2$ and arbitrarily take a parameter matrix pair (U, V) in a sufficient small neighborhood $\mathcal{N}$ of $\Phi^{-1}(0)$. A set of 2×2 real regular matrices is denoted by $\mathrm{GL}(2, \mathbb{R})$. In $\mathcal{N}$, we can treat the entries of U and V as positive numbers. Let

$$\mathcal{P} := \{P \in \mathrm{GL}(2, \mathbb{R}) | UP > 0, P^{-1}V > 0\}. \tag{4.48}$$

$\mathcal{P}$ is a sub-manifold of $\Phi^{-1}(0)$; thus, all we have to do is prove that $\mathcal{P}$ is a 4-dimensional variety. $P \in \mathcal{P}$ is denoted by

$$P := \begin{pmatrix} p & q \\ r & s \end{pmatrix} \tag{4.49}$$

and we put $\Delta := \det P$. Then, we immediately get

$$P^{-1} = \frac{1}{\Delta} \begin{pmatrix} s & -q \\ -r & p \end{pmatrix}. \tag{4.50}$$

Hence, we have

$$UP = \begin{pmatrix} u_{11}p + u_{12}r & u_{11}q + u_{12}s \\ \vdots & \vdots \\ u_{M1}p + u_{M2}r & u_{M1}q + u_{M2}s \end{pmatrix}, \tag{4.51}$$

$$P^{-1}V = \frac{1}{\Delta} \begin{pmatrix} -v_{21}q + v_{11}s & \dots & -v_{2N}q + v_{1N}s \\ v_{21}p - v_{11}r & \dots & v_{2N}p - v_{1N}r \end{pmatrix}. \tag{4.52}$$

Without loss of generality, we can consider the case $\Delta > 0$ because we can derive the lemma in the case $\Delta < 0$ by using inversed inequalities. We can represent the condition $UP > 0$ and $P^{-1}V > 0$ as the following simultaneous inequalities:

$$\begin{cases} u_{11}p + u_{12}r > 0 \\ \vdots \\ u_{M1}p + u_{M2}r > 0, \end{cases} \tag{4.53}$$

$$\begin{cases} u_{11}q + u_{12}s > 0 \\ \vdots \\ u_{M1}q + u_{M2}s > 0, \end{cases} \tag{4.54}$$

$$\begin{cases} -v_{21}q + v_{11}s > 0 \\ \vdots \\ -v_{2N}q + v_{1N}s > 0, \end{cases} \tag{4.55}$$

$$\begin{cases} v_{21}p - v_{11}r > 0 \\ \vdots \\ v_{2N}p - v_{1N}r > 0, \end{cases} \tag{4.56}$$

for $u_{ik} > 0$ and $v_{kj} > 0$ $(i = 1, \ldots, M; k = 1, 2; j = 1, \ldots, N)$. Therefore, (p, q, r, s) must be in upper domains of lines in rp-plain and qs-plain. The domains in rp-plain and qs-plain are 2-dimensional manifold, respectively. The Cartesian product of these is also a manifold and its dimension is $2 + 2 = 4$. Hence, we have

$$\lambda \leqq \frac{2(M+N)-4}{2} = M + N - 2. \tag{4.57}$$

By applying Corollary 4.1, Lemma 4.4 is proved:

$$\lambda = \frac{2(M+N)-4}{2} = M + N - 2. \tag{4.58}$$

□

As the last part of this section, we drive an upper bound of the RLCT of NMF when $H = H_0$. Put

$$U_k := \begin{pmatrix} u_{1(2k-1)} & u_{1(2k)} \\ \vdots & \vdots \\ u_{M(2k-1)} & u_{M(2k)} \end{pmatrix},$$

$$V_k := \begin{pmatrix} v_{(2k-1)1} & \cdots & v_{(2k-1)N} \\ v_{H(2k-1)} & \cdots & v_{H(2k)} \end{pmatrix},$$

$$U_k^0 := \begin{pmatrix} u^0_{1(2k-1)} & u^0_{1(2k)} \\ \vdots & \vdots \\ u^0_{M(2k-1)} & u^0_{M(2k)} \end{pmatrix},$$

$$V_k^0 := \begin{pmatrix} v^0_{(2k-1)1} & \cdots & v^0_{(2k-1)N} \\ v^0_{H(2k-1)} & \cdots & v^0_{H(2k)} \end{pmatrix}.$$

Proof of Lemma 4.5. We divide this proof into two cases: whether $H = H_0$ is even.

Case (1): $H{=}H_0{\equiv}0 \pmod 2$, i.e. there exists $H'{\in}\mathbb{N}$ such that $H{=}H_0{=}2H'$.

There is a positive constant $C > 0$ such that

$$\|UV - U_0V_0\|^2 \tag{4.59}$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N}(u_{i1}v_{1j} + ... + u_{iH}v_{Hj} - u^0_{i1}v^0_{1j} - u^0_{iH}v^0_{Hj})^2 \tag{4.60}$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N}\left(\sum_{k=1}^{H}(u_{ik}v_{kj} - u^0_{ik}v^0_{kj})\right)^2 \tag{4.61}$$

$$\leqq C\sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{H'}(u_{i(2k-1)}v_{(2k-1)j}+u_{i(2k)}v_{(2k)j}-u^0_{i(2k-1)}v^0_{(2k-1)j}-u^0_{i(2k)}v^0_{(2k)j})^2 =: \overline{\Phi}_1. \tag{4.62}$$

Thus, we have

$$\overline{\Phi}_1 \sim \sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{H'}(u_{i(2k-1)}v_{(2k-1)j} + u_{i(2k)}v_{(2k)j} - u^0_{i(2k-1)}v^0_{(2k-1)j} - u^0_{i(2k)}v^0_{(2k)j})^2 \tag{4.63}$$

$$= \sum_{k=1}^{H'}\left(\sum_{i=1}^{M}\sum_{j=1}^{N}(u_{i(2k-1)}v_{(2k-1)j}+u_{i(2k)}v_{(2k)j}-u^0_{i(2k-1)}v^0_{(2k-1)j}-u^0_{i(2k)}v^0_{(2k)j})^2\right) \tag{4.64}$$

$$= \sum_{k=1}^{H'}\left\|U_kU_k - U^0_kV^0_k\right\|^2 =: \overline{\Phi}_2. \tag{4.65}$$

Let $\overline{\lambda}_k$ be the RLCT of $\left\|U_kU_k - U^0_kV^0_k\right\|^2$ for $k = 1, \ldots, H'$. The RLCT of $\overline{\Phi}_2$ is denoted by $\overline{\lambda}$. As all intersections between $\{(U_k, V_k)\}$ and $\{(U_{k'}, V_{k'})\}$ are empty sets if $k \neq k'$,

$$\overline{\lambda} = \sum_{k=1}^{H'}\overline{\lambda}_k \tag{4.66}$$

holds. For each k, because of Lemma 4.4, we have $\overline{\lambda}_k = M + N - 2$.

Thus, on account of $H' = H_0/2$, we obtain

$$\overline{\lambda} = \sum_{k=1}^{H'}(M + N - 2) \tag{4.67}$$

$$= H'(M + N - 2) \tag{4.68}$$

$$= H_0\frac{M + N - 2}{2}. \tag{4.69}$$

Case (2): $H{=}H_0{\equiv}1 \pmod 2$, i.e. there exists $H'{\in}\mathbb{N}$ such that $H{=}H_0{=}2H'{-}1$.
In the same way as Case (1), there exists a positive constant $C' > 0$ such that

$$\|UV - U_0V_0\|^2 \tag{4.70}$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N}\left(\sum_{k=1}^{H}(u_{ik}v_{kj} - u^0_{ik}v^0_{kj})\right)^2 \tag{4.71}$$

$$\leqq C'\sum_{i=1}^{M}\sum_{j=1}^{N}\left(\sum_{k=1}^{H'-1}(u_{i(2k-1)}v_{(2k-1)j} + u_{i(2k)}v_{(2k)j} - u^0_{i(2k-1)}v^0_{(2k-1)j} - u^0_{i(2k)}v^0_{(2k)j})^2\right. \tag{4.72}$$

$$\left. + (u_{i(2H'-1)}v_{(2H'-1)j} - u^0_{i(2H'-1)}v^0_{(2H'-1)j})^2\right) \tag{4.73}$$

$$\sim \sum_{i=1}^{M}\sum_{j=1}^{N}\Bigg(\sum_{k=1}^{H'-1}(u_{i(2k-1)}v_{(2k-1)j}+u_{i(2k)}v_{(2k)j}-u^0_{i(2k-1)}v^0_{(2k-1)j}-u^0_{i(2k)}v^0_{(2k)j})^2 \tag{4.74}$$

$$+(u_{i(2H'-1)}v_{(2H'-1)j}-u^0_{i(2H'-1)}v^0_{(2H'-1)j})^2\Bigg) \tag{4.75}$$

$$=\sum_{k=1}^{H'-1}\Bigg(\sum_{i=1}^{M}\sum_{j=1}^{N}(u_{i(2k-1)}v_{(2k-1)j}+u_{i(2k)}v_{(2k)j}-u^0_{i(2k-1)}v^0_{(2k-1)j}-u^0_{i(2k)}v^0_{(2k)j})^2\Bigg) \tag{4.76}$$

$$+\sum_{i=1}^{M}\sum_{j=1}^{N}(u_{i(2H'-1)}v_{(2H'-1)j}-u^0_{i(2H'-1)}v^0_{(2H'-1)j})^2 \tag{4.77}$$

$$=\sum_{k=1}^{H'-1}\left\|U_kU_k-U^0_kV^0_k\right\|^2+\|u_{2H'-1}(v_{2H'-1})^T-u^0_{2H'-1}(v^0_{2H'-1})^T\|^2=:\overline{\Phi}'. \tag{4.78}$$

Let $\overline{\lambda}'_1$ and $\overline{\lambda}'_2$ be the RLCT of $\sum_{k=1}^{H'-1}\left\|U_kU_k-U^0_kV^0_k\right\|^2$ and $\|u_{2H'-1}(v_{2H'-1})^T-u^0_{2H'-1}(v^0_{2H'-1})^T\|^2$, respectively. The former is calculated by Lemma 4.4 as same as in Case (1) and the latter is derived by Lemma 4.3. Hence, we have

$$\overline{\lambda}'_1=\sum_{k=1}^{H'-1}(M+N-2)=(H'-1)(M+N-2), \tag{4.79}$$

$$\overline{\lambda}'_2=\frac{M+N-1}{2}. \tag{4.80}$$

The RLCT of $\overline{\Phi}'$ is denoted by $\overline{\lambda}'$. By the definition, $H'=(H_0+1)/2$ holds. Since all combinations of intersections in $(\{(U_k,V_k)\},\{(U_{k'},V_{k'})\},\{(u_{2H'-1},v_{2H'-1})\})$ are empty sets if $k\neq k'$, we have

$$\overline{\lambda}'=(H'-1)(M+N-2)+\frac{M+N-1}{2} \tag{4.81}$$

$$=\frac{(H_0-1)(M+N-2)+(M+N-1)}{2} \tag{4.82}$$

$$=\frac{(H_0-1)(M+N-2)+(M+N-2)+1}{2} \tag{4.83}$$

$$=\frac{H_0(M+N-2)+1}{2}. \tag{4.84}$$

The RLCT is order isomorphic: $\lambda\leqq\overline{\lambda}[H_0\equiv 0\,(\mathrm{mod}\,2)]+\overline{\lambda}'[H_0\equiv 1\,(\mathrm{mod}\,2)]$. Therefore, summarizing Case (1) and (2), we obtain

$$\lambda\leqq\frac{1}{2}\left\{H_0(M+N-2)+[H_0\equiv 1\,(\mathrm{mod}\,2)]\right\}. \tag{4.85}$$

□

4.4 Proof of Main Theorem

Based on the above preparation, Theorem 4.2, 4.3 and 4.4 are proved.

Proof of Theorem 4.2. Put $\delta_{H_0} := [H_0 \equiv 1 (\mathrm{mod}\ 2)]$. There exists a positive constant $D > 0$ such that

$$\|UV - U_0V_0\|^2 \tag{4.86}$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N}(u_{i1}v_{1j} + \ldots + u_{iH}v_{Hj} - u^0_{i1}v^0_{1j} - \ldots - u^0_{iH_0}v^0_{H_0j})^2 \tag{4.87}$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N}\left(\sum_{k=1}^{H}u_{ik}v_{kj} - \sum_{k=1}^{H_0}u^0_{ik}v^0_{kj}\right)^2 \tag{4.88}$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N}\left(\sum_{k=1}^{H_0}(u_{ik}v_{kj} - u^0_{ik}v^0_{kj}) + \sum_{k=H_0+1}^{H}u_{ik}v_{kj}\right)^2 \tag{4.89}$$

$$\leqq D\sum_{k=1}^{(H_0-\delta_{H_0})/2}\left\|U_kV_k - U^0_kV^0_k\right\|^2 + \delta_{H_0}D\|u_{H_0}(v_{H_0})^T - u^0_{H_0}(v^0_{H_0})^T\|^2 \tag{4.90}$$

$$+ D\left\|\begin{pmatrix} u_{1(H_0+1)} & \cdots & u_{1H} \\ \vdots & \ddots & \vdots \\ u_{M(H_0+1)} & \cdots & u_{MH}\end{pmatrix}\begin{pmatrix} v_{(H_0+1)1} & \cdots & v_{(H_0+1)N} \\ \vdots & \ddots & \vdots \\ v_{H1} & \cdots & v_{HN}\end{pmatrix}\right\|^2. \tag{4.91}$$

Let $\overline{K}_1$, $\overline{K}_2$ and $\overline{K}_3$ be

$$\overline{K}_1 = \sum_{k=1}^{(H_0-\delta_{H_0})/2}\left\|U_kV_k - U^0_kV^0_k\right\|^2, \tag{4.92}$$

$$\overline{K}_2 = \|u_{H_0}(v_{H_0})^T - u^0_{H_0}(v^0_{H_0})^T\|^2, \tag{4.93}$$

$$\overline{K}_3 = \left\|\begin{pmatrix} u_{1(H_0+1)} & \cdots & u_{1H} \\ \vdots & \ddots & \vdots \\ u_{M(H_0+1)} & \cdots & u_{MH}\end{pmatrix}\begin{pmatrix} v_{(H_0+1)1} & \cdots & v_{(H_0+1)N} \\ \vdots & \ddots & \vdots \\ v_{H1} & \cdots & v_{HN}\end{pmatrix}\right\|^2, \tag{4.94}$$

respectively. The RLCT of $\overline{K}_1$, $\overline{K}_2$ and $\overline{K}_3$ are respectively denoted by $\overline{\lambda}_1$, $\overline{\lambda}_2$ and $\overline{\lambda}_3$. The constant D does not affect the RLCTs; thus, we only have to consider $\overline{\lambda}_1$, $\overline{\lambda}_2$ and $\overline{\lambda}_3$.

According to Lemma 4.5, we have

$$\overline{\lambda}_1 + \overline{\lambda}_2 = \frac{1}{2}\{H_0(M + N - 2) + \delta_{H_0}\}. \tag{4.95}$$

Moreover, as $\overline{K}_3$ is equivalent to $\|UV\|^2$ replaced from H to $H - H_0$ in Lemma 4.2, we obtain

$$\overline{\lambda}_3 = \frac{1}{2}(H - H_0)\min\{M\phi_U, N\phi_V\}. \tag{4.96}$$

Summarizing the aboves, the RLCT of NMF λ satisfies

$$\lambda \leqq \overline{\lambda}_1 + \overline{\lambda}_2 + \overline{\lambda}_3 \tag{4.97}$$
$$= \frac{1}{2}\{H_0(M+N-2) + \delta_{H_0}\} + \frac{1}{2}(H-H_0)\min\{M\phi_U, N\phi_V\} \tag{4.98}$$
$$= \frac{1}{2}\left\{(H-H_0)\min\{M\phi_U, N\phi_V\} + H_0(M+N-2) + \delta_{H_0}\right\}. \tag{4.99}$$

Therefore, Theorem 4.2 is proved:

$$\lambda \leqq \frac{1}{2}\left\{(H-H_0)\min\{M\phi_U, N\phi_V\} + H_0(M+N-2) + \delta_{H_0}\right\}. \tag{4.100}$$

□

Proof of Theorem 4.3. Let $\overline{\lambda}$ be the upper bound in Theorem 4.2:

$$\overline{\lambda} = \frac{1}{2}\left\{(H-H_0)\min\{M\phi_U, N\phi_V\} + H_0(M+N-2) + \delta_{H_0}\right\}. \tag{4.101}$$

By using Theorem 3.4 and Theorem 4.2 $\lambda \leqq \overline{\lambda}$, we obtain

$$\mathbb{E}[G_n] = \frac{\lambda}{n} - \frac{m-1}{n\log n} + o\left(\frac{1}{n\log n}\right), \tag{4.102}$$
$$\leqq \frac{\overline{\lambda}}{n} + o\left(\frac{1}{n}\right) \tag{4.103}$$

and

$$F_n = \lambda \log n - (m-1)\log\log n + O_p(1) \tag{4.104}$$
$$\leqq \overline{\lambda}\log n + O_p(1). \tag{4.105}$$

Thus, Theorem 4.3 is derived.

□

Proof of Theorem 4.4. Let S_n be the empirical entropy. By using Proposition 2.2 and Theorem 4.3, we have

$$\overline{F_n} = nS_n + F_n \tag{4.106}$$
$$\leqq nS_n + \overline{\lambda}\log n + O_p(1). \tag{4.107}$$

Also, because of Theorem 4.1,

$$\bar{F}_n^{\mathrm{vb}} = nS_n + \lambda_{\mathrm{vb}}\log n + O_p(1) \tag{4.108}$$

holds, where

$$\lambda_{\mathrm{vb}} = \begin{cases} (H-H_0)(M\phi_U + N\phi_V) + \frac{1}{2}H_0(M+N), & M\phi_U + N\phi_V < \frac{M+N}{2} \\ \frac{1}{2}H(M+N), & M\phi_U + N\phi_V \geqq \frac{M+N}{2}. \end{cases} \tag{4.109}$$

Thus, we compute their difference

$$\bar{F}_n^{\rm vb} - \overline{F}_n = (\lambda_{\rm vb} - \lambda)\log n + O_p(1) \tag{4.110}$$
$$\geqq (\lambda_{\rm vb} - \overline{\lambda})\log n + O_p(1). \tag{4.111}$$

Case(1): $M\phi_U + N\phi_V < \frac{M+N}{2}$ holds. We have

$$\lambda_{\rm vb} - \overline{\lambda} = (H - H_0)(M\phi_U + N\phi_V) + \frac{1}{2}H_0(M+N) \tag{4.112}$$
$$- \frac{1}{2}\left[(H - H_0)\min\{M\phi_U, N\phi_V\} + H_0(M+N-2) + \delta_{H_0}\right] \tag{4.113}$$
$$= (H - H_0)\left[M\phi_U + N\phi_V - \frac{1}{2}\min\{M\phi_U, N\phi_V\}\right] \tag{4.114}$$
$$+ \frac{1}{2}H_0(M+N-M-N+2) - \frac{\delta_{H_0}}{2} \tag{4.115}$$
$$= \frac{1}{2}[(H-H_0)(M\phi_U+N\phi_V+M\phi_U+N\phi_V-\min\{M\phi_U, N\phi_V\})+\delta_{H_0}]+H_0 \tag{4.116}$$
$$= \frac{1}{2}[(H - H_0)(M\phi_U + N\phi_V + \max\{M\phi_U, N\phi_V\}) + \delta_{H_0}] + H_0. \tag{4.117}$$

Case(2): $M\phi_U + N\phi_V \geqq \frac{M+N}{2}$ holds. We have

$$\lambda_{\rm vb}-\overline{\lambda} = \frac{1}{2}H(M+N)-\frac{1}{2}\left[(H-H_0)\min\{M\phi_U, N\phi_V\}+H_0(M+N-2)+\delta_{H_0}\right] \tag{4.118}$$
$$= \frac{1}{2}H(M+N)-\frac{1}{2}(H-H_0)\min\{M\phi_U, N\phi_V\}-\frac{1}{2}H_0(M+N)-\frac{\delta_{H_0}}{2}+H_0 \tag{4.119}$$
$$= \frac{1}{2}(H - H_0)(M+N) - \frac{1}{2}(H - H_0)\min\{M\phi_U, N\phi_V\} - \frac{\delta_{H_0}}{2} + H_0 \tag{4.120}$$
$$= \frac{1}{2}[(H - H_0)(M+N-\min\{M\phi_U, N\phi_V\}) + \delta_{H_0}] + H_0. \tag{4.121}$$

Therefore, we obtain Theorem 4.4.

□

4.5 Discussion

Here, we will discuss the results of this chapter from four points of view. After that, we will describe the numerical behavior of the theoretical result by conducting numerical experiments.

4.5.1 Application to Model Selection

First, we will explain an application of the Main Theorems. In this paper, we theoretically clarify the difference between the variational free energy and the usual free energy in NMF. From a practical point of view, the free energy $\overline{F}_n$ can be calculated from the data; however,

the entailed numerical integration is very hard and the sampling approximation, such an exchange Monte Carlo method, spends a long time for finding $\overline{F}_n$. On the other hand, we can compute the variational one $\bar{F}_n^{\rm vb}$ more easily than $\overline{F}_n$. If the estimator of VBNMF is found, all we have to do is to substitute it for the functional whose minimum value is equal to $\bar{F}_n^{\rm vb}$.

It has not been clarified how much the variational free energy differs from the free energy; however, the Main Theorems give the lower bound. We can use the lower bound to approximate the free energy from the variational one. Namely, when $\bar{F}_n^{\rm vb}$ is known, we have the approximation

$$\overline{F}_n \approx \bar{F}_n^{\rm vb} - \underline{\lambda} \log n.$$

The usual VBNMF gives $\overline{F}_n \approx \bar{F}_n^{\rm vb}$; here though, we can obtain a more accurate value[*1]. In this way we should be able to more accurately select the model in VBNMF by using $\bar{F}_n^{\rm vb} - \underline{\lambda} \log n$.

4.5.2 Generalization Error

Second, we describe the generalization error in NMF. Theorem 4.3 also gives an upper bound of the generalization error G_n as well as the free energy F_n. Generally speaking, the learning coefficients that control the asymptotic behavior of the F_n and G_n are the same RLCTs [85]; hence, we can clarify both behaviors at once. Since the situation in which the probability model $p(X|U,V)$ is a Poisson distribution and the prior $\varphi(U,V)$ is a gamma distribution is a case where the Gibbs sampling [18] of NMF is performed, it can be regarded that not only F_n but also G_n are theoretically clarified when Gibbs sampling is applied.

By contrast, in Theorem 4.1, only the learning coefficient of the variational free energy $\bar{F}_n^{\rm vb}$ is determined. This is because the learning coefficient of the variational generalization error is *not* equal to the one of $\bar{F}_n^{\rm vb}$. Generally, in the case of VB, no zeta function is capable of uniformly handling F_n and G_n, and the RLCT cannot obtain the learning coefficient[*2]. For example, in VB of three-layered linear neural networks, the asymptotic behaviors are clarified not only with the variational free energy but also the variational generalization error [60], and their learning coefficients are different. A linear neural network is also known as a reduced rank regression, a dimension reduction model, and the parameters are equivalent to a matrix factorization model without a non-negative value constraint. In contrast in Bayesian inference in matrix factorization and NMF, the RLCT of matrix factorization is a lower bound for the RLCT of NMF, and it is known that the non-negative rank is dominant rather than the rank of the matrix in NMF, as described in [37, 36]. Therefore, we cannot directly apply the results of linear neural networks to VBNMF.

In this way, theoretical generalization error in VB is rarely clarified, although that in Bayesian inference has been clarified with the free energy. The Main Theorems show that Gibbs sampling is more reliable than VB, in the sense that it gives a theoretical guarantee not only about the free energy but also the generalization error. We can estimate the sample size to achieve the needed inference performance and tune the hyperparameters. Although various factors determine whether Gibbs sampling or VB is appropriate, our research can answer the question of whether or not the theoretical generalization error is clarified.

[*1] Actually $\underline{\lambda}$ has the true non-negative rank H_0; however, in the same way as sBIC [24], we can avoid using the true knowledge by considering $H_0 = 0, \ldots, H$.

[*2] The learning coefficient of VB is not equal to the RLCT.

4.5.3 Effect of Non-negative Restriction

Third, the effect of non-negative restriction to the parameter region is considered. In NMF, the entries of the parameter matrices (U, V) are non-negative. On the other hand, we can consider non-restricted matrix factorization: the entries of the parameter matrices can be negative. Usually, this is just called matrix factorization (MF) and its RLCT is defined as follows.

Definition 4.2 (RLCT of MF) *Let C be a compact set of $\mathbb{R}$ and let $U^{\mathrm{MF}} \in \mathrm{M}(M, H, C)$, $V^{\mathrm{MF}} \in \mathrm{M}(H, N, C)$, $U_0^{\mathrm{MF}} \in \mathrm{M}(M, r, C)$ and $V_0^{\mathrm{MF}} \in \mathrm{M}(r, N, C)$. The largest pole of the following univariate complex function*

$$\zeta(z)=\int_{\mathrm{M}(M,H,C)} \mathrm{d}U^{\mathrm{MF}} \int_{\mathrm{M}(H,N,C)} \mathrm{d}V^{\mathrm{MF}} \left(\|U^{\mathrm{MF}}V^{\mathrm{MF}}-U_0^{\mathrm{MF}}V_0^{\mathrm{MF}}\|^2\right)^z \varphi_{\mathrm{MF}}(U^{\mathrm{MF}}, V^{\mathrm{MF}})$$

is denoted by $(-\lambda_{\mathrm{MF}})$ and its order is denoted by m_{MF}, where φ_{MF} is a prior distribution of $(U^{\mathrm{MF}}, V^{\mathrm{MF}})$. Then λ_{MF} is called the RLCT of MF and m_{MF} is called its multiplicity.

The RLCT of MF was clarified in [10] as that of reduced rank regression (see Theorem 4.5). As proved above, it is a lower bound of the RLCT of NMF (see Theorem 4.6). We discuss what the case $\lambda = \lambda_{\mathrm{MF}}$ holds or what the case $\lambda > \lambda_{\mathrm{MF}}$ holds. The control variable of MF is $H = \mathrm{rank}UV$ and it is compared with $r = \mathrm{rank}U_0V_0$.

To discuss relationship between NMF and MF, there is an important concept; the smallest inner matrix dimension of NMF is called a non-negative rank [19]. A useful property of a non-negative rank is as follows.

Theorem 4.6 (Cohen) *Suppose $W \in \mathrm{M}(M, N, K)$. Let rank_+W be the nonnegative rank of W. The following inequality holds:*

$$\mathrm{rank}W \leqq \mathrm{rank}_+W \leqq \min\{M, N\}.$$

A sufficient condition of $\mathrm{rank}W = \mathrm{rank}_+W$ is $\mathrm{rank}W \leqq 2 \vee M \leqq 3 \vee N \leqq 3$.

From the above theorem, if $r \leqq H_0 \leqq 2$, then $r = H_0$. An example of non-negative matrices whose non-negative rank is strictly larger than its rank is as follows:

$$W = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

The above 4×4 non-negative matrix W satisfies $\mathrm{rank}W = 3$ and $\mathrm{rank}_+W = 4$ [19].

In the case $H = H_0 = 1$ or $H = H_0 = 2$, $H_0 = r = 1$ or $H_0 = r = 2$ hold, respectively; thus, the RLCT of NMF is equal to that of MF in these case. On the other hand, the case that $\lambda > \lambda_{\mathrm{MF}}$ holds is also included in Theorem 4.2. If $H_0 = 0$, then λ can be strictly greater than λ_{MF}. We prove this property below.

Proof. According to Theorem 4.6, $r = 0$ holds from $H_0 = 0$. Hence, we consider Theorem 4.5 in the case $r = 0$ and Theorem 4.2 in the case $H_0 = 0$. Then, we have

$$\lambda = \frac{H \min\{M\phi_U, N\phi_V\}}{2}. \tag{4.122}$$

To keep the condition consistent with Theorem 4.5, let $\phi_U = \phi_V = 1$. Here, we calculate $\lambda_{\rm MF}$ subject to Theorem 4.5 in the followings.

Case (1): $N \leqq M + H \wedge M \leqq N + H \wedge H \leqq M + N \wedge M + H + N \equiv 0 \pmod 2$.

We assume $N \leqq M$, i.e. $\lambda = HM/2$. Owing to Theorem 4.5,

$$\lambda_{\rm MF} = \left\{2H(M+N) - (M-N)^2 - H^2\right\}/8 \tag{4.123}$$
$$= \left\{2(MH + HN + NM) - M^2 - N^2 - H^2\right\}/8 \tag{4.124}$$

holds; thus we have

$$\lambda_{\rm MF} - \lambda = -\left\{M^2 + H^2 + N^2 - 2(MH + HN + NM) - 4HM\right\}/8 \tag{4.125}$$
$$= -\left\{M^2 + H^2 + N^2 - 2(-MH + HN + NM)\right\}/8 \tag{4.126}$$
$$= -\left[(-M)^2 + (-H)^2 + N^2 + 2\{(-M)(-H) + (-H)N + N(-M)\}\right]/8 \tag{4.127}$$
$$= -(N - M - H)^2/8. \tag{4.128}$$

Because of $N \leqq M$ and $H \geqq 1$, $N < M + H$ holds. Therefore, $\lambda_{\rm MF} < \lambda$. If $N > M$, that can be derived in the same way as above.

Case (2): $N \leqq M + H \wedge M \leqq N + H \wedge H \leqq M + N \wedge M + H + N \equiv 1 \pmod 2$.

We assume $N \leqq M$, i.e. $\lambda = HM/2$. In the same way as Case (1), we have

$$\lambda_{\rm MF} - \lambda = 1/8 - (N - M - H)^2/8 \tag{4.129}$$
$$= 1^2/8 - (M + H - N)^2/8 \tag{4.130}$$
$$= -(M + H - N + 1)(M + H - N - 1)/8. \tag{4.131}$$

We derive $N + 1 \leqq M + H$ by reductio ad absurdum. We suppose $M + H < N + 1$. Using the assumption of Case (2), $N < M + H < N + 1$ holds. $M + H$ and N are positive integers; thus the above inequality is inconsistent. That is why $N + 1 \leqq M + H$ and $-(M + H - N + 1)(M + H - N - 1) \leqq 0$. Therefore, $\lambda_{\rm MF} \leqq \lambda$. If $N > M$, that can be derived in the same way as above.

Case (3): $N + H < M$ i.e. $N < N + H < M$.

On account of $N < M$,$\lambda = HM/2 = \lambda_{\rm MF}$.

Case (4): $M + H < N$ i.e. $M < M + H < N$.

In the same way as Case (4), $\lambda = HN/2 = \lambda_{\rm MF}$.

Case (5): $M + N < H$ i.e. $N < M + N < H \wedge M < M + N < H$.

On account of $M < H \wedge N < H$, $MN < HN \wedge MN < HM$ i.e. $MN < H \min\{M, N\}$ holds. Thus, we have $\lambda = H \min\{M, N\}/2 > MN/2 = \lambda_{\rm MF}$.

□

4.5.4 Robustness on Probability Distributions

Third, let us discuss the true distribution and the model of the data. In this study, we consider the case in which the probability model $p(X|U,V)$ is a Poisson distribution and the prior $\varphi(U,V)$ is a gamma distribution in the same way as in the derivation of the Gibbs sampling algorithm of NMF by Cemgil [18]. These assumptions are necessary for Gibbs sampling and the derivation of VB; however, other models can be considered when using other MCMC methods. For example, we may want to set that the entries of the data matrix X are non-negative real numbers. Is the main result applicable to these cases?

To tell the truth, several distributions satisfy the condition that the RLCT of NMF is equal to the absolute value of the maximum pole of the following zeta function

$$\zeta(z) = \int_{\mathrm{M}(M,H,K)} \mathrm{d}U \int_{\mathrm{M}(H,N,K)} \mathrm{d}V \Big(\|UV - U_0V_0\|^2 \Big)^z \varphi(U,V).$$

Specifically, when the elements of the data matrix follow a normal distribution, a Poisson distribution, an exponential distribution, or a Bernoulli distribution, the asymptotic behavior of the free energy and the generalization error can be described using the same RLCT defined by the above zeta function: Theorem 4.2. Therefore, if the prior distribution is a gamma distribution, Theorem 4.2 and Theorem 4.3 hold not only when the probability model and the true distribution are Poisson distributions but also when they are normal distributions, exponential distributions, or Bernoulli distributions.

We also consider the case in which X is generated by an exponential distribution. Then, the KL divergence has the same RLCT as the square error if elements of UV are positive.

Proposition 4.1 *Let the probability density functions of $X \in \mathrm{M}(M,N,K)$ be $q(X)$ and $p(X|U,V)$, which represent the true distribution and the model respectively defined by*

$$q(X) \propto \mathrm{Exp}(X|U_0V_0),$$
$$p(X|U,V) \propto \mathrm{Exp}(X|UV),$$

where $\mathrm{Exp}(X|W)$ is a probability density function of an exponential distribution with average W:

$$\mathrm{Exp}(X|W) := \prod_{i=1}^{M}\prod_{j=1}^{N} \frac{e^{-x_{ij}/w_{ij}}}{w_{ij}},\ X = (x_{ij}),\ W = (w_{ij}).$$

Also let $\varphi(X,Y)$ be a probability density function such that it is positive on a compact subset of $\mathrm{M}(M,H,K_0) \times \mathrm{M}(H,N,K_0)$. Then, the Kullback-Leibler divergence has same RLCT as the square error.

Proof. Let $x > 0$, $a > 0$ and $b > 0$. We put

$$p(x|a) := \frac{e^{-x/a}}{a}, \tag{4.132}$$

$$g(a,b) := \int p(x|a) \log \frac{p(x|a)}{p(x|b)} \mathrm{d}x. \tag{4.133}$$

By the similar way to the proof of Lemma 4.1, all we have to do is prove $g(a,b) \sim (a-b)^2$. By using

$$\log \frac{p(x|a)}{p(x|b)} = \log \frac{b}{a} e^{-x/a+x/b} \tag{4.134}$$

$$= \log b - \log a - \frac{x}{a} + \frac{x}{b}, \tag{4.135}$$

we have

$$\int p(x|a) \mathrm{d}x = 1, \tag{4.136}$$

$$\int x p(x|a) \mathrm{d}x = \mathbb{E}[x] = a. \tag{4.137}$$

Developing the terms, we obtain

$$g(a,b) = \int p(x|a) \left(\log b - \log a - \frac{x}{a} + \frac{x}{b} \right) \mathrm{d}x$$

$$= \log b - \log a - 1 + \frac{a}{b}. \tag{4.138}$$

Hence, we immediately get

$$\partial_a g(a,b) = \frac{1}{b} - \frac{1}{a},$$

$$\partial_b g(a,b) = \frac{1}{b} - \frac{a}{b^2}$$

$$= \frac{b-a}{b^2}$$

and increase or decrease of $g(a,b)$. Thus, this proposition can be proved in the same way as Lemma 4.1.

□

Remark 4.4 *The right sides of Eqs.* (4.17) *and* (4.138) *are respectively equal to* "I-divergence"[74] *and* "Itakura-Saito-divergence"[45] *which are used as criteria of difference between observed matrix and reproduced matrix in NMF*[51, 26, 27]. *Moreover, in the same way as the proof of Theorem 4.5* [10]*, we can prove that the square error* $\|UV - U_0V_0\|^2$ *has the same RLCT as that of the KL divergence when the entries of the data matrix are subject to a normal distribution.*

As a result, we can apply Theorem 4.2 to Bayesian inference if we use a Kullback-Leibler divergence or a square error as a criterion of difference between UV and U_0V_0 in cases where elements of matrices are generated by normal, Poisson or exponential distributions. Thus, the upper bound of the expected generalization error $\mathbb{E}[G_n]$ and the free energy $\overline{F_n}$ can be clarified if the size and inner dimension of the observed matrix and reproduced matrix are given.

4.5.5 Experiment

Here, we run numerical experiments to check the numeric behavior of the theoretical results. Theorem 4.1 gives the exact value of the learning coefficient $\lambda_{\rm vb}$ of VBNMF and its validity was confirmed in Kohjima's previous research [49]. The core result of this chapter is Theorem 4.2. Therefore, we only have to run experiments for it; i.e., the RLCT λ of Bayesian NMF is calculated by using Gibbs sampling.

Let $\mathbb{E}_\theta$ be an expectation operator of the posterior: $\mathbb{E}_\theta[\cdot] = \int \mathrm{d}\theta \psi(\theta|\mathcal{D})[\cdot]$. Let $\hat{\lambda}$ be the numerically calculated RLCT. The widely applicable information criterion (WAIC) [82] is defined by the following random variable W_n:

$$W_n = T_n + V_n/n,$$

where T_n is the empirical loss and V_n is the functional variance:

$$T_n = -\frac{1}{n}\sum_{i=1}^{n} \log p^*(X_i) = -\frac{1}{n}\sum_{i=1}^{n} \log \mathbb{E}_\theta[p(X_i|\theta)], \tag{4.139}$$

$$V_n = \sum_{i=1}^{n} \left[\mathbb{E}_\theta[(\log p(X_i|\theta))^2] - \{\mathbb{E}_\theta[\log p(X_i|\theta)]\}^2 \right] = \sum_{i=1}^{n} \mathbb{V}_\theta[\log p(X_i|\theta)]. \tag{4.140}$$

Even if the posterior distribution cannot be approximated by any normal distribution (i.e., the model is singular), the expected WAIC $\mathbb{E}[W_n]$ is asymptotically equal to the expected generalization loss $\mathbb{E}[G_n + S]$ [82];

$$\mathbb{E}[G_n + S] = \mathbb{E}[W_n] + o(1/n^2).$$

Moreover, the generalization error and the WAIC error $W_n - S_n$ have the same variance [85]:

$$G_n + W_n - S_n = 2\lambda/n + o_p(1/n). \tag{4.141}$$

Eq. (4.141) is useful for computing $\hat{\lambda}$ because the leading term $2\lambda/n$ is deterministic. Nevertheless, the left hand side is probabilistic. This means that the needed number of simulation D is smaller than that in the case using this approximation: $G_n \approx \frac{1}{n}\sum_{t=1}^{n_T} \log \frac{q(X_t)}{p^*(X_t)}$, where n_T is the number of the test data and $(X^*)^{n_T} = (X_1^*, \ldots, X_{n_T}^*)$ is the test data generated by $q(X)$.

The method was as follows. First, the training data X^n was generated from the true distribution $q(X)$. Second, the posterior distribution was calculated by using Gibbs sampling [18] (see also Algorithm 3). Third, G_n and $W_n - S_n$ were computed by using the training data X^n and the artificial test data $(X^*)^{n_T}$ generated from $q(X)$. These three steps were repeated and each value of $n(G_n + W_n - S_n)/2$ was saved. After all repetitions have been completed, $n(G_n + W_n - S_n)/2$ was averaged over the simulations. This average was $\hat{\lambda}$.

The pseudo-code is listed in Algorithm 2, where n_K is the sample size of the parameter subject to the posterior. We used the programing language named Julia 1.3.0 [13] for this experiment. The implementation is available at the following github page: `https://github.com/chijan-nh/LearningCoefficient-RLCT-ofNMF-usingGS`.

Algorithm 2 How to Compute $\hat{\lambda}$

Require: $\phi = (\phi_U, \theta_U, \phi_V, \theta_V) > 0$: the hyperparameters,
U_0: the true parameter matrix whose size is (M, H_0),
V_0: the true parameter matrix whose size is (H_0, N),
GS: the Gibbs sampling function whose return value consists of the samples from the posterior. See also Algorithm 3.
Ensure: The numerical computed RLCT $\hat{\lambda}$.
Allocate an array $\Lambda[D]$.
for $d = 1$ to D **do**
Generate $X^n \sim q(X) = p(X|w_0)$, where $w_0 = (U_0, V_0)$.
Allocate arrays $\mathcal{U}[M, H, n_K]$ and $\mathcal{V}[H, N, n_K]$.
Get $\mathcal{U}, \mathcal{V} \leftarrow \mathrm{GS}(X^n, \phi)$.
Generate $(X^*)^{n_T} \sim q(X)$.
Calculate $G_n \approx \frac{1}{n_T} \sum_{t=1}^{n_T} \log \frac{q(X_t^*)}{\mathbb{E}_\theta[p(X_t^*|\theta)]}$, $S_n = -\frac{1}{n} \sum_{i=1}^{n} \log q(X_i)$,
and $W_n \approx -\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta[p(X_i|\theta)] + \frac{1}{n} \sum_{i=1}^{n} \mathbb{V}_\theta[\log p(X_i|\theta)]$,
where $\mathbb{E}_\theta[f(\theta)] \approx \frac{1}{n_K} \sum_{k=1}^{n_K} f(\theta_k)$ and $\theta_k = (\mathcal{U}[:,:,k], \mathcal{V}[:,:,k])$.
Save $\Lambda[d] \leftarrow n(G_n + W_n - S_n)/2$.
end for
Calculate $\hat{\lambda} = \frac{1}{D} \sum_{d=1}^{D} \Lambda[d]$.

We set $M = N = 4$, $H = 2$, $H_0 = 1$, and $n_T = 100n$. To examine the behavior given different sample sizes, we set $n = 500, 1000$ and $n_K = 1000, 2000$, respectively. To decrease the probabilistic effect of Eq. (4.141), we conducted the simulations twenty times: $D = 20$.

The hyperparameters were set to $\theta_U = \theta_V = 1$ and

$$(\phi_U, \phi_V) = (0.25, 0.25), (0.5, 0.5), (1, 1), (2, 2).$$

We chose four pairs of (ϕ_U, ϕ_V) in view of a theoretical point: the critical lines of the RLCT of NMF and the learning coefficient in VBNMF. The RLCT or the learning coefficient drastically changes on these lines. Figure 4.1 shows this phenomenon called phase transition.

Under the condition $M = N$, $\phi_U + \phi_V = 1$ is the phase transition line (see Theorem 4.1). Each point on the straight line $\phi_U = \phi_V$ is characterized as follows. $(0.25, 0.25)$ is before the phase transition line, and $(0.5, 0.5)$ is a phase transition point. $(1, 1)$ is a case where the prior distribution $\varphi(U, V)$ is strictly positive and bounded. $(2, 2)$ is one of points beyond the critical line.

In the Gibbs sampling, we had to conduct a burn-in to decrease the effect of the initial values and thin the samples in order to break the correlations. The sample size for the burn-in was 20000, while the sample size for the thinning was 20; thus, the sample sizes of the parameter was $20000 + 20K = 40000$ and 60000 ($K = 1000$ and 2000) and we used the $(20000 + 20k)$-th sample as the entry of $\mathcal{U}[:,:,k]$ and $\mathcal{V}[:,:,k]$ for $k = 1$ to K.

The experimental results are shown in Table 4.1. The symbol $\overline{\lambda}$ denotes the theoretical upper bound of the RLCT λ in Theorem 4.2. There are columns for each sample size, and each row contains the hyperparameter (Hyperparam.) and the learning coefficient (Coeff.). The experimental values have three significant digits.

Algorithm 3 Gibbs Sampling for NMF
$\mathrm{GS}(X^n, \phi, K = K, \mathrm{burnin} = 20000, \mathrm{thin} = 20)$

Require: $X^n = (X_l)_{l=1}^n$: the data where $X_l$is an $M \times N$ non-negative integer matrix,
$\phi = (\phi_U, \theta_U, \phi_V, \theta_V) \in \mathbb{R}^4_{>0}$: the hyperparameter of the Gamma priors for the non-negative matrices U and V,
$\mathrm{Gam}(W|\Phi, \Theta)$: a Gamma distribution of a matrix whose (i, j)-entry is generated by $\mathrm{Gam}(w|\Phi[i, j], \Theta[i, j])$,
$\mathrm{Multi}(s|x, \pi)$: a Multinomial distribution whose trials and event probabilities are $x \in \mathbb{N}$ and $\pi \in \mathrm{Sim}(H, [0, 1])$, respectively.
Ensure: Sampling non-negative matrices from the numerical posterior.
Let $\mathrm{iter} = \mathrm{butnin} + \mathrm{thin} * K$.
Allocate arrays $\mathcal{U}[M, H, K]$, $\overline{\mathcal{U}}[M, H, \mathrm{iter}]$, $\mathcal{V}[H, N, K]$ and $\overline{\mathcal{V}}[H, N, \mathrm{iter}]$.
Inital sampling for U and V from the prior:
Generate $U, V \sim \mathrm{Gam}(U|\phi_U, \theta_U)\mathrm{Gam}(V|\phi_V, \theta_V)$.
# Sampling from the posterior:
for $k = 1$ to iter **do**
Sampling the hidden variable s.
Allocate an array $s[M, H, N, n]$.
for $i = 1$, $j = 1$, and $l = 1$ to M, N, and n **do**
for $h = 1$ to H **do**
Let $\pi[i, h, j] = U[i, h]V[h, j] / \sum_{h=1}^{H} U[i, h]V[h, j]$.
end for
Generate $s[i, :, j, l] \sim \mathrm{Multi}(s|X_l[i, j], \pi[i, :, j])$.
end for
Sampling the non-negative matrix U.
for $i = 1$ and $h = 1$ to M and H **do**
Let $\hat{\phi}_U[i, h] = \sum_{l=1}^{n} \sum_{j=1}^{N} s[i, h, j, l] + \phi_U$.
Let $\hat{\theta}_U[i, h] = n \sum_{j=1}^{N} V[h, j] + \theta_U$.
end for
Generate $U \sim \mathrm{Gam}(U|\hat{\phi}_U, \hat{\theta}_U)$.
Put $\overline{\mathcal{U}}[M, H, k] \leftarrow U$.
Sampling the non-negative matrix V.
for $h = 1$ and $j = 1$ to H and N **do**
Let $\hat{\phi}_V[h, j] = \sum_{l=1}^{n} \sum_{i=1}^{M} s[i, h, j, l] + \phi_V$.
Let $\hat{\theta}_V[h, j] = n \sum_{i=1}^{M} U[i, h] + \theta_V$.
end for
Generate $V \sim \mathrm{Gam}(V|\hat{\phi}_V, \hat{\theta}_V)$.
Put $\overline{\mathcal{V}}[H, N, k] \leftarrow V$.
end for
# Burn-in and thinning.
for $k = 1$ to K **do**
$\mathcal{U}[M, H, k] \leftarrow \overline{\mathcal{U}}[M, H, \mathrm{burnin} + \mathrm{thin} * k]$.
$\mathcal{V}[H, N, k] \leftarrow \overline{\mathcal{V}}[H, N, \mathrm{burnin} + \mathrm{thin} * k]$.
end for
Return $\mathcal{U}, \mathcal{V}$.

Table 4.1: Numerically Calculated and Theoretical Values of the Learning Coefficients

Hyperparam. / Coeff.		$n = 500$	$n = 1000$
ϕ_U=ϕ_V=0.25 θ_U=θ_V=1	λ_{vb}	6	6
	$\overline{\lambda}$	4	4
	$\hat{\lambda}$	3.74 ± 0.0412	3.73 ± 0.0508
	$\overline{\lambda} - \hat{\lambda}$	0.260 ± 0.0412	0.268 ± 0.0508
ϕ_U=ϕ_V=0.5 θ_U=θ_V=1	λ_{vb}	8	8
	$\overline{\lambda}$	9/2	9/2
	$\hat{\lambda}$	4.05 ± 0.0706	4.15 ± 0.0842
	$\overline{\lambda} - \hat{\lambda}$	0.450 ± 0.0706	0.346 ± 0.0842
ϕ_U=ϕ_V=1 θ_U=θ_V=1	λ_{vb}	8	8
	$\overline{\lambda}$	11/2	11/2
	$\hat{\lambda}$	4.52 ± 0.0492	4.54 ± 0.0628
	$\overline{\lambda} - \hat{\lambda}$	0.976 ± 0.0492	0.965 ± 0.0628
ϕ_U=ϕ_V=2 θ_U=θ_V=1	λ_{vb}	8	8
	$\overline{\lambda}$	15/2	15/2
	$\hat{\lambda}$	4.76 ± 0.0454	4.79 ± 0.0477
	$\overline{\lambda} - \hat{\lambda}$	2.74 ± 0.0454	2.71 ± 0.0477

As shown in Table 4.1, all numerically calculated values are smaller than the theoretical upper bound, thus, the Theorem 4.2 is consistent with the experimental result. Since $\hat{\lambda}$ is larger in the case $\phi_U = \phi_V = 2$ than that in the case $\phi_U = \phi_V = 1$, it is conjectured that the larger ϕ_U and ϕ_V are, the larger λ will be, while λ_{vb} saturates to $H(M + N)/2 = 8$ (owing to Theorem 4.1) and the upper bound of λ looks less tight than that in the other case. In the case $\phi_U = \phi_V = 0.5$, when the hyperparameter is on the phase transition line, $\hat{\lambda}$ fluctuates more than in the other case. We can see that learning is unstable on the critical line. Hence, the hyperparameters should be set to avoid the neighborhood of the phase transition line.

4.6 Conclusion

We give an upper bound of the RLCT for NMF whose priors are gamma distributions (Theorem 4.2) and describe theoretical applications to Bayesian and variational inference. According to Theorem 4.3, a theoretical upper bounds of the generalization error and the free energy are derived and hyperparameters make them change: there are a phase transition structure. Moreover, owing to Theorem 4.4, the variational approximation error, i.e., the difference between the variational free energy and the free energy, are quantitatively evaluated. These difference depends on the true non-negative rank and the hyperparameters from the gamma prior distributions. The numerical results are consistent with the theoretical results and they suggest the exact values of the RLCT and the stability of learning. Future tasks include conducting large-scale experiments and clarifying the exact value of the RLCT.

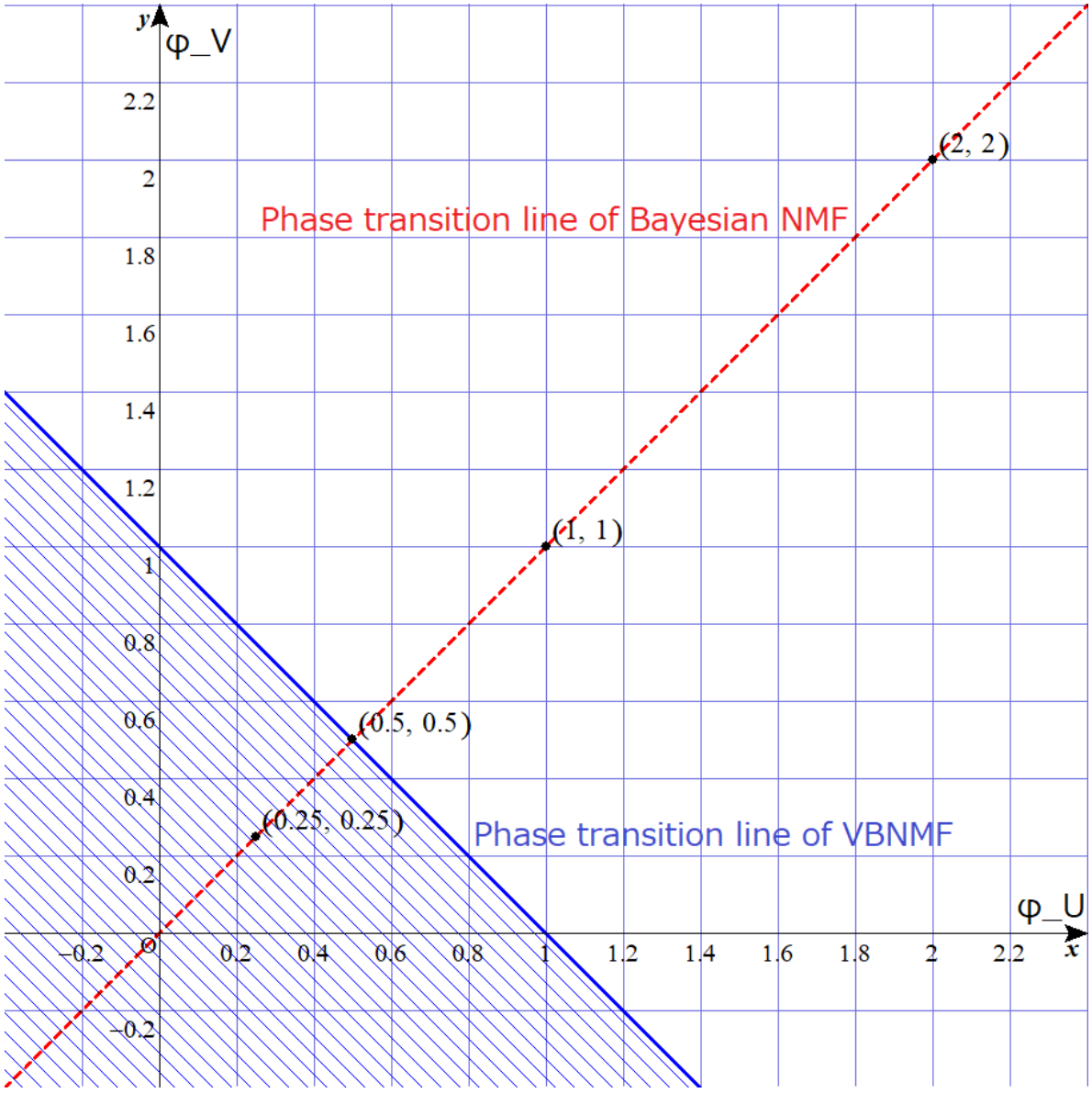


Fig. 4.1: This image represents the critical lines of the RLCT of NMF and the learning coefficient in VBNMF when $M = N$. The horizontal and the vertical axises are corresponding to the hyperparameter ϕ_U and ϕ_V, respectively. The used four points $\{(0.25, 0.25), (0.5, 0.5), (1, 1), (2, 2)\}$ are plotted as black disks. The red-dashed-line is the critical line of the RLCT of NMF by Theorem 4.2. On the other hand, the blue-line is that of the learning coefficient in VBNMF by Theorem 4.1. Obviously, they are thoroughly different: VB is not equivalent to Bayesian inference.

Chapter 5

Bayesian Generalization Error in Latent Dirichlet Allocation

In this chapter, we report the result of theoretical analysis for Bayesian generalization error in LDA. This chapter consists of five parts. First, in Sec. 5.1, we describe motivation of this theoretical research. Second, in Sec. 5.2, we state the main theorem with regard to Bayesian generalization error in LDA. Third, in Sec. 5.3, we prepare for the proof of the theorem. Fourth, in Sec 5.4, we prove the main theorem. Lastly, in Sec. 5.5, we discuss the theoretical results.

In the followings, $\theta = (A, B)$ is a parameter and x is an observed random variable. Note that this chapter is based on the author's papers [38, 34].

5.1 Motivation

Topic model [29] is a ubiquitous learning machine used in many research areas, including text mining [15, 30], computer vision [52], marketing research [72], and geology [98]. Latent Dirichlet allocation (LDA) [15] is one of the most popular Bayesian topic models. It has been devised for text analysis, and it can extract essential information in documents by defining the topics of the words. The topics are formulated as one-hot vectors subject to categorical distributions which are different for each document (Fig. 5.1). This formulation refers documents to bags of words: a text is simply referred to a multiset of words [31]. For example, a text "I think that that that that that boy wrote is wrong" is referred to a multiset $\{\text{I, think, that, that, that, that, that, boy wrote, is, wrong}\}_{\text{m}}$. To consider the frequency of appearance of each word, a muliset allows duplication. By using one-hot-encoding, this multiset is transformed to the vectors

$$\left(\begin{bmatrix}1\\0\\0\\0\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\\0\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\\0\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\\0\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\\0\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\\0\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\0\\1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\0\\0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\0\\0\\0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\0\\0\\0\\0\\1\end{bmatrix} \right). \tag{5.1}$$

If the text is the given document, LDA models that the above vectors are generated from the *mixture* of categorical mixture distribution in Fig. 5.1 in the case $N = 1$. Practically, a single document has more words and the number of documents is also larger $*^1$. In fact, topic model (including LDA) can treat bags of anythings by formulating the raw data to the bags; thus, it has many application described above. The standard inference algorithms, such as Gibbs sampling [30] and the variational Bayesian method [15], require the appropriate number of the topics to be set in advance. If the chosen number of topics is too small, then LDA suffers from underfitting. On the other hand, if the chosen number of topics is too large, it suffers from overfitting on the training data. In practical applications, neither the optimal number of topics of the ground truth nor the true distribution is known; thus, researchers and practitioners face a situation in which the number of topics they set may be larger than the optimal one. Thus, clarifying the behavior of the generalization error is necessary as a theoretical foundation to resolve model selection problems. However, the mathematical property of LDA has not yet been clarified, because it has a hierarchical structure; the regular statistical theory can not be applied. To rephrase, LDA is a singular model.

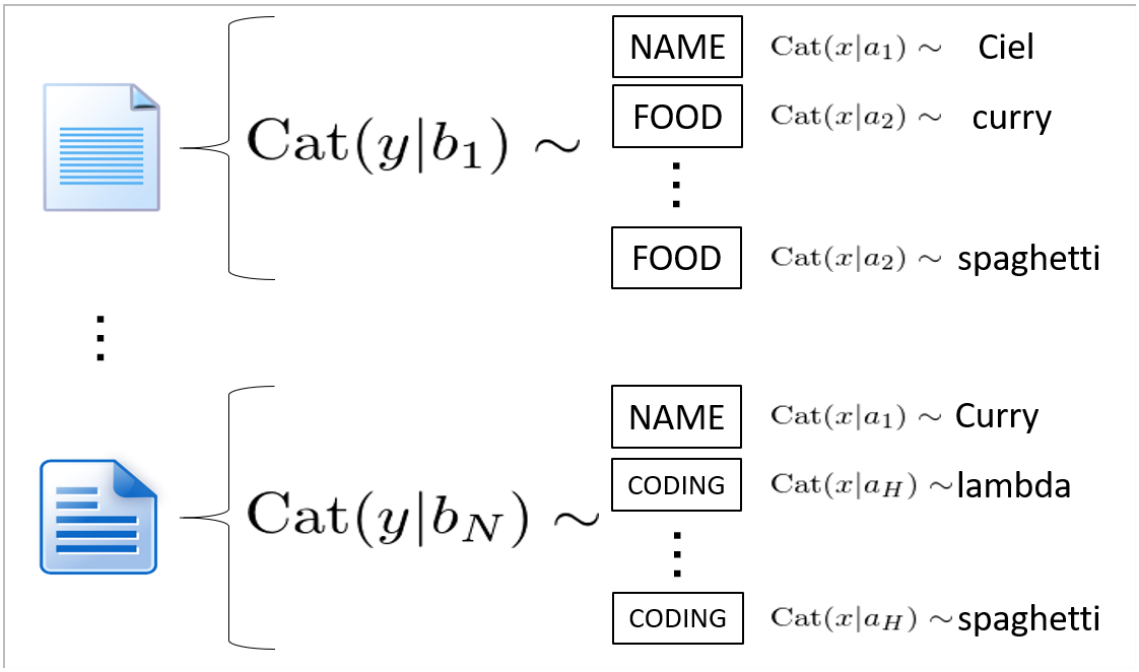


Fig. 5.1: This figure gives an overview of LDA. The categorical distributions Cat that depend on the documents. Words in the uppercase such as NAME, FOOD, and CODING are topics. There are categorical distributions that are different for each topic; the words (Ciel, curry, lambda, ...) are generated from them. Hence, we can explain LDA as a *mixture* of categorical mixture models. This model has a hierarchical structure. This figure is quoted and modified from the author's works [38, 34]

Matrix factorization (MF) is also used in machine learning frequently. MF decomposes the data matrix into a product of two matrices and discovers hidden structures or patterns,

$*^1$ Besides, for practical uses, one may remove articles (e.g. "a" and "the") and demonstratives (e.g. "this" and "that") to make a language model [68]. Also, one may refer a text to a bag of n-grams and vectorize it.

hence it has been experimentally used for knowledge discovery in many fields. However, MF has no guarantee of reaching the unique factorization, and it is sensitive to the initial value of the numerical calculation. This non-uniqueness interferes with data-driven inference and interpretations of the results. Besides, the sensitivity to the initial value causes the factorization result to have low reliability. From the viewpoint of data-based prediction, this instability may lead to incorrect predictions. To improve interpretability, non-negative matrix factorization (NMF) [61, 51] has been devised; it is a restricted MF wherein the elements of the matrix are non-negative. Thanks to the non-negativity constraint, the extracted factors are readily interpretable, therefore NMF is frequently used for extracting latent structures and patterns in many fields (see also Sec. 4.1).

Stochastic matrix factorization (SMF) was devised by Adams [2]; it can be understood as a restriction of NMF in which at least one matrix factor is "stochastic": the elements of the matrix factors are non-negative and the sum of the elements in a column is equal to one. In other words, the columns of the matrix are in a simplex. By making two further assumptions, Adams proved the uniqueness of the results of SMF [1, 2]. For a statement of these two conditions, let us consider a data matrix X whose size is $M \times N$ and factor matrices A and B which are "stochastic" and whose sizes are $M \times H$ and $H \times N$, respectively. H might be the rank of X but the "stochastic" condition makes this determination non-trivial. In other words, SMF can be viewed as a method that finds a factor matrices pair (A, B) such that $X = AB$ for a given X and H. The non-uniqueness property has been paraphrased as the existence of $H \times H$ regular matrix $P \neq I_H$ such that

$$X = APP^{-1}B, \tag{5.2}$$

where I_H is an $H \times H$ identity matrix. Thus, uniqueness means that Eq. (5.2) is attained if and only if $P = I_H$. Adams assumed that

$$AP \geqq 0 \text{ and } P^{-1}B \geqq 0, \tag{5.3}$$

i.e., the elements of AP and $P^{-1}B$ are non-negative, and $P^{-1}B =: (b'_{kj})$ satisfies

$$\sum_{k=1}^{H} b'_{kj} = 1 \text{ or } \sum_{j=1}^{N} b'_{kj} = 1. \tag{5.4}$$

Adams claimed that these assumptions are "natural" and applied SMF to image recognition (the same problem analyzed in [51]) and text mining[2]. It is emphasized that, in this thesis, we consider that it is not clear whether Adam's assumptions (5.3) and (5.4) are mathematically "natural". In the following, we do not assume Eqs. (5.3) and (5.4).

The MF methods described so far, including SMF, are understood as a deterministic procedure. As will be shown later, for hierarchical learning machines such as MF, we study probabilistic procedures, because Bayesian inference has higher predictive accuracy than deterministic methods or maximum likelihood estimation. The same is also true regarding the accuracy of the discovered knowledge. Moreover, the probabilistic or statistical view gives a wider application. Indeed, Bayesian NMF [75, 18] has been applied to image recognition [18], audio signal processing [75], and recommender systems [16]. From a statistical point of view, the data matrices are random variables subject to the true distribution. As mentioned in

Chap. 4, the factorization of a set of independent matrices should be studied because the target matrices are often obtained daily, monthly, or in different places. More importantly, as proved later, the SMF has the same learning curve as LDA; if the Bayesian generalization error in SMF has been clarified, then that of LDA is also determined. That is why the decomposition of a set of matrices with stochastic restriction is considered to be a statistical inference in order to clarify the Bayesian generalization error in LDA.

In this chapter, we derive the exact asymptotic form of the Bayesian generalization error by determination of the exact real log canonical threshold λ (RLCT; see also Chap. 3 for the general definition and property) in LDA. For finding λ, we prove that the RLCT of LDA is equal to that of SMF, i.e. LDA is equivalent to SMF in the sense of generalization.

5.2 Main Theorem

Now let us introduce the main result of this chapter.

Let K be a compact subset of $[0,1] = \{x \in \mathbb{R} | 0 \leqq x \leqq 1\}$ and let K_0 be a compact subset of $(0,1) = \{x \in \mathbb{R} | 0 < x < 1\}$. Let $\mathrm{Onehot}(N) := \{w = (w_j) \in \{0,1\}^N \mid \sum_{j=1}^N w_j = 1\} = \{(1,0,\ldots,0),\ldots,(0,\ldots,0,1)\}$ be an N-dimensional one-hot vector set and $\mathrm{Sim}(N,K) := \{c = (c_j) \in K^N \mid \sum_{j=1}^N c_j = 1\}$ be an N-dimensional simplex. Let $\mathrm{S}(M,N,E) = \mathrm{Sim}(M,E)^N$ be a set of $M \times N$ stochastic matrices whose elements are in E, where E is a subset of $[0,1]$, and $M, N \in \mathbb{N}$. In addition, we set $H, H_0 \in \mathbb{N}$ and $H \geqq H_0$.

In LDA terminology, the number of documents and the vocabulary size are denoted by N and M, respectively. Let H_0 be the true or optimal number of topics and H be the chosen one. In this situation, the sample size n is the number of words in all of the given documents. See also Table 5.1. This table is quoted and modified from the author's works [38, 34].

Table 5.1: Description of Variables in LDA Terminology

Variable	Description	Index
b_j=$(b_{kj}) \in \mathrm{Sim}(H,K)$	probability that topic is k when document is j	k=1, ..., H
a_k=$(a_{ik}) \in \mathrm{Sim}(M,K)$	probability that word is i when topic is k	i=1, ..., M
x=$(x_i) \in \mathrm{Onehot}(M)$	word i is defined by $x_i = 1$	i=1, ..., M
y=$(y_k) \in \mathrm{Onehot}(H)$	topic k is defined by $y_k = 1$	k=1, ..., H
z=$(z_j) \in \mathrm{Onehot}(N)$	document j is defined by $z_j = 1$	j=1, ..., N
$*_0$ and $*^0$	true or optimal variable corresponding to $*$	-

Assume that $M \geqq 2$, $N \geqq 2$, and $H \geqq H_0 \geqq 1$. We define $A = (a_{ik}) \in \mathrm{S}(M,H,K)$ and $B = (b_{kj}) \in \mathrm{S}(H,N,K)$, and assume that $A_0 = (a^0_{ik}) \in \mathrm{S}(M,H_0,K_0)$ and $B_0 = (b^0_{kj}) \in \mathrm{S}(H_0,N,K_0)$ are SMFs such that they give the minimal factorization of A_0B_0.

Let $q(x|z)$ and $p(x|z,A,B)$ be conditional probability density functions of $x \in \mathrm{Onehot}(N)$ given $z \in \mathrm{Onehot}(M)$, which represent the true distribution and the model of LDA, respec-

tively,

$$q(x|z) = \prod_{j=1}^{N} \left(\sum_{k=1}^{H_0} b_{kj}^0 \prod_{i=1}^{M} (a_{ik}^0)^{x_i} \right)^{z_j}, \tag{5.5}$$

$$p(x|z, A, B) = \prod_{j=1}^{N} \left(\sum_{k=1}^{H} b_{kj} \prod_{i=1}^{M} (a_{ik})^{x_i} \right)^{z_j}. \tag{5.6}$$

These distributions are the marginalized ones of the following simultaneous ones with respect to the topics $y^0 \in \mathrm{Onehot}(H_0)$ and $y \in \mathrm{Onehot}(H)$:

$$q(x, y^0|z) = \prod_{j=1}^{N} \left[\prod_{k=1}^{H_0} \left(b_{kj}^0 \prod_{i=1}^{M} (a_{ik}^0)^{x_i} \right)^{y_k^0} \right]^{z_j}, \tag{5.7}$$

$$p(x, y|z, A, B) = \prod_{j=1}^{N} \left[\prod_{k=1}^{H} \left(b_{kj} \prod_{i=1}^{M} (a_{ik})^{x_i} \right)^{y_k} \right]^{z_j}. \tag{5.8}$$

In practical cases, the topics are not observed; thus, we use Eqs. (5.5) and (5.6). Besides, let $\varphi(A, B) > 0$ be a probability density function such that it is positive on a compact subset of $\mathrm{S}(M, H, K) \times \mathrm{S}(H, N, K)$ including $\Phi^{-1}(0)$ i.e. (A_0, B_0).

Definition 5.1 (RLCT of LDA) *Let* $\mathrm{KL}(A, B)$ *be the KL divergence from* $q(x|z)$ *to* $p(x|z, A, B)$ *in the aboves:*

$$\mathrm{KL}(A, B) := \sum_{z \in \mathrm{Onehot}(M)} \sum_{x \in \mathrm{Onehot}(N)} q(x|z) q'(z) \log \frac{q(x|z)}{p(x|z, A, B)},$$

where $q'(z)$ *is the true distribution of the document. In LDA,* $q'(z)$ *is not observed and assumed that it is positive and bounded.*

Then, the holomorphic function of one complex variable z $(\mathrm{Re}(z) > 0)$

$$\zeta(z) = \int_{\mathrm{S}(M,H,K)} \mathrm{d}A \int_{\mathrm{S}(H,N,K)} \mathrm{d}B \, \mathrm{KL}(A, B)^z$$

can be analytically continued to a unique meromorphic function on the entire complex plane $\mathbb{C}$ *and all of its poles are rational and negative (see also Theorem* 3.2*). If the largest pole is* $(-\lambda)$*, then* λ *is said to be the RLCT of LDA.*

Definition 5.2 (RLCT of SMF) *Set* $\Phi(A, B) = \|AB - A_0 B_0\|^2$*. Then the holomorphic function of one complex variable* z $(\mathrm{Re}(z) > 0)$

$$\zeta(z) = \int_{\mathrm{S}(M,H,K)} \mathrm{d}A \int_{\mathrm{S}(H,N,K)} \mathrm{d}B \, \Phi(A, B)^z$$

can also be analytically continued to a unique meromorphic function on $\mathbb{C}$ *and its all poles are rational and negative. If the largest pole is* $z = -\lambda$*, then* λ *is the RLCT of SMF.*

In this paper, we prove the following two theorems.

Theorem 5.1 (Equivalence of the LDA and SMF) *Let $\lambda_{\rm SMF}$ be the RLCT of SMF and $\lambda_{\rm LDA}$ be the RLCT of LDA. Then, the following equality holds:*

$$\lambda_{\rm SMF} = \lambda_{\rm LDA}.$$

In the same way,

$$m_{\rm SMF} = m_{\rm LDA}$$

holds where $m_{\rm SMF}$ and $m_{\rm LDA}$ are the multiplicities of SMF and LDA, respectively.

In order to prepare to state Main Theorem of this chapter, we define an intrinsic value r of the true distribution defined by stochastic matrices A_0 and B_0. Let $\tilde{A}_0^{\backslash (M,H_0)}$ and $\tilde{B}_0^{\backslash (H_0,1)}$ be matrices defined as

$$\tilde{A}_0^{\backslash (M,H_0)} = (a_{ik}^0)_{i=1,k=1}^{M-1,H_0-1} \tag{5.9}$$

and

$$\tilde{B}_0^{\backslash (H_0,1)} = (b_{kj}^0)_{k=1,j=2}^{H_0-1,N}, \tag{5.10}$$

respectively. Also, $\tilde{A}_0^{(\backslash M),H_0}$ and $\tilde{B}_0^{(\backslash H_0),1}$ are denoted by matrices whose column vectors are same such that

$$\tilde{A}_0^{(\backslash M),H_0} = (a_{(1:M-1)H_0}^0, \cdots a_{(1:M-1)H_0}^0) = (a_{iH_0}^0)_{i=1,k=1}^{M-1,H_0-1} \tag{5.11}$$

and

$$\tilde{B}_0^{(\backslash H_0),1} = (b_{(1:H_0-1)N}^0, \cdots b_{(1:H_0-1)1}^0) = (b_{k1}^0)_{k=1,l=1}^{H_0-1,N-1}, \tag{5.12}$$

where $a_{(1:M-1)H_0}^0 = (a_{iH_0}^0)_{i=1}^{M-1}$ and $b_{(1:H_0-1)1}^0 = (b_{k1}^0)_{k=1}^{H_0-1}$ are an $(M-1)$-dimensional vector and an (H_0-1)-dimensional vector, respectively. Then, let

$$U_0 = \tilde{A}_0^{\backslash (M,H_0)} - \tilde{A}_0^{(\backslash M),H_0}, \tag{5.13}$$

$$V_0 = \tilde{B}_0^{\backslash (H_0,1)} - \tilde{B}_0^{(\backslash H_0),1}, \tag{5.14}$$

and $r = \mathrm{rank}(U_0V_0)$. Obviously, r depends on H_0.

The main result of this paper is the following theorem.

Theorem 5.2 (Main Theorem) *Suppose $M \geqq 2$, $N \geqq 2$, and $H \geqq H_0 \geqq 1$. Let r be the rank of U_0V_0 which is a product of two matrices (U_0, V_0) defined in Eqs. (5.13) and (5.14). The RLCT of LDA $\lambda = \lambda_{\rm LDA}$ and its multiplicity $m = m_{\rm LDA}$ are as follows.*

1. *If $N + r + 1 \leqq M + H$ and $M + r + 1 \leqq N + H$ and $H + r + 1 \leqq M + N$,*
   (a) and if $M + N + H + r$ is odd, then

$$\lambda = \frac{1}{8}\{2(H+r+1)(M+N) - (M-N)^2 - (H+r+1)^2\} - \frac{1}{2}N,\ m = 1.$$

   (b) and if $M + N + H + r$ is even, then

$$\lambda = \frac{1}{8}\{2(H+r+1)(M+N) - (M-N)^2 - (H+r+1)^2 + 1\} - \frac{1}{2}N,\ m = 2.$$

2. *Else if* $M + H < N + r + 1$*, then*

$$\lambda = \frac{1}{2}\{MH + N(r+1) - H(r+1) - N\},\ m = 1.$$

3. *Else if* $N + H < M + r + 1$*, then*

$$\lambda = \frac{1}{2}\{NH + M(r+1) - H(r+1) - N\},\ m = 1.$$

4. *Else (i.e.* $M + N < H + r + 1$*), then*

$$\lambda = \frac{1}{2}(MN - N),\ m = 1.$$

We prove Main Theorem in the next section. As two applications of this theorem, we obtain the exact form of the free energy and Bayesian generalization error of LDA by applying Theorem 3.4.

Theorem 5.3 *Let* F_n *be the normalized free energy and* $\mathbb{E}[G_n]$ *be the expected generalization error in LDA, respectively. Then, these asymptotic forms are as the followings:*

$$F_n = \lambda \log n - (m-1)\log\log n + O_p(1),$$
$$\mathbb{E}[G_n] = \frac{\lambda}{n} - \frac{m-1}{n\log n} + o\left(\frac{1}{n\log n}\right),$$

where λ *and* m *are the RLCT of LDA and its multiplicity which are determined in Theorem* 5.2*.*

Theorem 5.3 can be immediately derived if Theorem 5.2 is proved. Thus, we prove Theorem 5.2 in the following Secs. 5.3 and 5.4. To prove Theorem 5.2, we derive Theorem 5.1 in the next section.

5.3 Preparation

In order to prove Theorem 5.1, we show the following facts.

Proposition 5.1 *Let* $w = (w_i)_{i=1}^M \in \mathrm{Sim}(M, K)$ *and* $w_0 = (w_i^0)_{i=1}^M \in \mathrm{Sim}(M, K_0)$*. Then, there exist positive constants* $c_1 > 0$ *and* $c_2 > 0$ *such that*

$$c_1 \sum_{i=1}^M (w_i - w_i^0)^2 \leqq \sum_{i=1}^M w_i \log \frac{w_i^0}{w_i} \leqq c_2 \sum_{i=1}^M (w_i - w_i^0)^2.$$

I.e. $\sum_{i=1}^M w_i \log \frac{w_i^0}{w_i}$ *has the same RLCT as* $\sum_{i=1}^M (w_i - w_i^0)^2$*.*

This proposition was derived in [53] in order to clarify the RLCT of categorical mixture model. By using Proposition 5.1, we prove Theorem 5.1.

Proof of Theorem 5.1. Without loss of generality, we can rewrite the notation of $q(x|d)$ and $p(x|d, A, B)$ as follows:

$$q(x|z_i = 1) = \sum_{k=1}^{H_0} b_{kj}^0 \prod_{i=1}^{M} (a_{ik}^0)^{x_i}, \quad p(x|z_i = 1, A, B) = \sum_{k=1}^{H} b_{kj} \prod_{i=1}^{M} (a_{ik})^{x_i}. \tag{5.15}$$

The word x is a one-hot vector; hence, we obtain

$$q(x_j = 1|z_i = 1) = \sum_{k=1}^{H_0} a_{ik}^0 b_{kj}^0, \quad p(x_j = 1|z_i = 1, A, B) = \sum_{k=1}^{H} a_{ik} b_{kj}. \tag{5.16}$$

Then, the conditional Kullback-Leibler divergence between $q(x|d)$ and $p(x|d, A, B)$ is equal to

$$\begin{aligned}
\mathrm{KL}(A, B) &= \sum_{z \in \mathrm{Onehot}(M)} \sum_{x \in \mathrm{Onehot}(N)} q(x|z) q'(z) \log \frac{q(x|z)}{p(x|z, A, B)} && (5.17)\\
&= \sum_{i=1}^{M} \sum_{j=1}^{N} q(x_j = 1|z_i = 1) q'(z_i = 1) \log \frac{q(x_j = 1|z_i = 1)}{p(x_j = 1|z_i = 1, A, B)} && (5.18)\\
&= \sum_{j=1}^{N} q'(z_i = 1) \sum_{i=1}^{M} \left(\sum_{k=1}^{H_0} a_{ik}^0 b_{kj}^0 \right) \log \frac{\sum_{k=1}^{H_0} a_{ik}^0 b_{kj}^0}{\sum_{k=1}^{H} a_{ik} b_{kj}}. && (5.19)
\end{aligned}$$

Owing to $A = (a_{ik}) \in \mathrm{S}(M, H, K)$, $B = (b_{kj}) \in \mathrm{S}(H, N, K)$, $A_0 = (a_{ik}^0) \in \mathrm{S}(M, H_0, K_0)$, and $B_0 = (b_{kj}^0) \in \mathrm{S}(H_0, N, K_0)$, the (i, j) entries of AB and $A_0 B_0$ are $(AB)_{ij} := \sum_{k=1}^{H} a_{ik} b_{kj}$ and $(A_0 B_0)_{ij} := \sum_{k=1}^{H_0} a_{ik}^0 b_{kj}^0$. We have

$$\mathrm{KL}(A, B) = \sum_{j=1}^{N} q'(z_i = 1) \sum_{i=1}^{M} (A_0 B_0)_{ij} \log \frac{(A_0 B_0)_{ij}}{(AB)_{ij}}. \tag{5.20}$$

According to Proposition 5.1, $\sum_{i=1}^{M} (A_0 B_0)_{ij} \log \frac{(A_0 B_0)_{ij}}{(AB)_{ij}}$ in Eq. (5.20) has the same RLCT of $\sum_{i=1}^{M} ((AB)_{ij} - (A_0 B_0)_{ij})^2$. In addition, $q'(z_i = 1)$ is positive and bounded. Accordingly, we have

$$\begin{aligned}
\mathrm{KL}(A, B) &= \sum_{j=1}^{N} q'(z_i = 1) \sum_{i=1}^{M} (A_0 B_0)_{ij} \log \frac{(A_0 B_0)_{ij}}{(AB)_{ij}} && (5.21)\\
&\sim \sum_{j=1}^{N} q'(z_i = 1) \sum_{i=1}^{M} ((AB)_{ij} - (A_0 B_0)_{ij})^2 && (5.22)\\
&\sim \sum_{j=1}^{N} \sum_{i=1}^{M} ((AB)_{ij} - (A_0 B_0)_{ij})^2 = \|AB - A_0 B_0\|^2. && (5.23)
\end{aligned}$$

Therefore, $\mathrm{KL}(A, B) \sim \|AB - A_0 B_0\|^2$; i.e., the RLCT of LDA equals the RLCT of SMF.

□

The RLCT of MF was clarified in the all case [10] as the follwoing theorem. Note that this theorem is same as Theorem 4.5 in Chap. 4; however, for self-containedness of this chapter, we repeat it.

Theorem 5.4 (Aoyagi) *Let M, N, H be positive integers. Matrices U and V are denoted by an $M \times H$ and an $H \times N$ matrix whose entries are real numbers. Let U_0 and V_0 be an $M \times *$ matrix and an $* \times N$ matrix, respectively. Suppose U_0 and V_0 are constants and put $r = \mathrm{rank}(U_0 V_0)$ and $\Phi(U,V) = \|UV - U_0 V_0\|^2$. Let λ_{MF} and m_{MF} be the RLCT of MF and its multiplicity, respectively. I.e. $(-\lambda_{\mathrm{MF}})$ is the maximum pole of the following zeta function and m_{MF} is its order:*

$$\zeta_{\mathrm{MF}}(z) = \iint \mathrm{d}U \mathrm{d}V \Phi(U,V)^z.$$

Then, λ_{MF} and m_{MF} are as follows:

(1) *If $N + r \leqq M + H \wedge M + r \leqq N + H \wedge H + r \leqq M + N$,*

($1-1$) *in the case $M + H + N + r$ is even,*

$$\lambda_{\mathrm{MF}} = \frac{-(H+r)^2 - M^2 - N^2 + 2\{(H+r)(M+N) + MN\}}{8}.$$

($1-2$) *in the case $M + H + N + r$ is odd,*

$$\lambda_{\mathrm{MF}} = \frac{1-(H+r)^2 - M^2 - N^2 + 2\{(H+r)(M+N) + MN\}}{8}.$$

(2) *Else if $M + H < N + r$,*

$$\lambda_{\mathrm{MF}} = \frac{HM - Hr + Nr}{2}.$$

(3) *Else if $N + H < M + r$,*

$$\lambda_{\mathrm{MF}} = \frac{HN - Hr + Mr}{2}.$$

(4) *Or else, i.e. in the case $M + N < H + r$,*

$$\lambda_{\mathrm{MF}} = \frac{MN}{2}.$$

In the case $(1-2)$, the multiplicity is two: $m_{\mathrm{MF}} = 2$. Otherwise, it equals one: $m_{\mathrm{MF}} = 1$.

The RLCT of MF and its multiplicity depend on (M, N, H, r); thus, we write them $\lambda_{\mathrm{MF}} = \lambda_{\mathrm{MF}}(M, N, H, r)$ and $m_{\mathrm{MF}} = m_{\mathrm{MF}}(M, N, H, r)$, respectively.

In the next section, by using changes of variables, we come down Theorem 5.2 (Main Theorem) to Theorem 5.4. This is why we consider the rank of $U_0 V_0$, where U_0 and V_0 are defined in Eqs. (5.13) and (5.14).

5.4 Proof of Main Theorem

Based on the above preparation, Theorem 5.2 and 5.3 are proved.

The structure of the proof of Main Theorem is as follows. First, we summarize terms in $\|AB - A_0B_0\|^2$ and consider degeneration of a polynomial ideal. Second, we resolve the non-negative restriction by variable transformations which are isomorphic maps. Third, we verify that the problem can come down to finding the RLCT of reduced rank regression. Lastly, we calculate the concrete value of the RLCT in each case.

Proof of Main Theorem. Because of Theorem 5.1 we only have to consider the analytic set defined by

$$\{(A, B) \mid \|AB - A_0B_0\|^2 = 0, A \text{ and } B \text{ are stochastic matrices.}\}$$

to determine the RLCT of LDA λ and its multiplicity m.

The first part is same as the first half of the proof of Appendix A in our previous research [38]. For the sake of self-containedness, we write down the process of developing the terms in the above paper. Let $\sim$ be a binomial relation such that the functions $K_1(w)$ and $K_2(w)$ have same RLCT if $K_1(w) \sim K_2(w)$. Summarizing the terms, we have

$$\|AB - A_0B_0\|^2 = \sum_{j=1}^{N}\sum_{i=1}^{M-1}\sum_{k=1}^{H} a_{ik}b_{kj} - \sum_{k=1}^{H_0} a^0_{ik}b^0_{kj} + \sum_{j=1}^{N}\sum_{k=1}^{H} a_{Mk}b_{kj} - \sum_{k=1}^{H_0} a^0_{Mk}b^0_{kj}. \tag{5.24}$$

Put

$$K_{ij} := \sum_{k=1}^{H} a_{ik}b_{kj} - \sum_{k=1}^{H_0} a^0_{ik}b^0_{kj}, \tag{5.25}$$

$$L_j := \sum_{k=1}^{H} a_{Mk}b_{kj} - \sum_{k=1}^{H_0} a^0_{Mk}b^0_{kj}, \tag{5.26}$$

then we get

$$\|AB - A_0B_0\|^2 = \sum_{j=1}^{N}\sum_{i=1}^{M-1} K_{ij}^2 + \sum_{j=1}^{N} L_j^2.$$

Using $a_{Mk} = 1 - \sum_{i=1}^{M-1} a_{ik}$, $b_{Hj} = 1 - \sum_{k=1}^{H-1} b_{kj}$, $a^0_{Mk} = 1 - \sum_{i=1}^{M-1} a^0_{ik}$, and $b^0_{H_0j} = 1 - \sum_{k=1}^{H_0-1} b^0_{kj}$, we have

$$K_{ij} = \sum_{k=1}^{H-1}(a_{ik} - a_{iH})b_{kj} - \sum_{k=1}^{H_0-1}(a^0_{ik} - a^0_{iH_0})b^0_{kj} + (a_{iH} - a^0_{iH_0}), \tag{5.27}$$

$$L_j = -\sum_{i=1}^{M-1}\sum_{k=1}^{H-1}(a_{ik} - a_{iH})b_{kj} + \sum_{i=1}^{M-1}\sum_{k=1}^{H_0-1}(a^0_{ik} - a^0_{iH_0})b^0_{kj} - \sum_{i=1}^{M-1}(a_{iH} - a^0_{iH_0}), \tag{5.28}$$

thus

$$L_j^2 = \left(\sum_{i=1}^{M-1} K_{ij} \right)^2 .$$

Therefore

$$\|AB - A_0B_0\|^2 = \sum_{j=1}^{N} \sum_{i=1}^{M-1} K_{ij}^2 + \sum_{j=1}^{N} L_j^2 \tag{5.29}$$

$$= \sum_{j=1}^{N} \sum_{i=1}^{M-1} K_{ij}^2 + \sum_{j=1}^{N} \left(\sum_{i=1}^{M-1} K_{ij} \right)^2 . \tag{5.30}$$

Since the polynomial $\sum_{i=1}^{M-1} K_{ij}$ is contained in the ideal generated from $(K_{ij})_{i=1,j=1}^{M-1,N}$, we have

$$\|AB - A_0B_0\|^2 \sim \sum_{j=1}^{N} \sum_{i=1}^{M-1} K_{ij}^2,$$

i.e.

$$\|AB - A_0B_0\|^2 \tag{5.31}$$

$$\sim \sum_{j=1}^{N} \sum_{i=1}^{M-1} \left\{ \sum_{k=1}^{H-1} (a_{ik} - a_{iH}) b_{kj} - \sum_{k=1}^{H_0-1} (a_{ik}^0 - a_{iH_0}^0) b_{kj}^0 + (a_{iH} - a_{iH_0}^0) \right\}^2 \tag{5.32}$$

$$= \sum_{j=1}^{N} \sum_{i=1}^{M-1} \left[\sum_{k=1}^{H_0-1} \{(a_{ik} - a_{iH}) b_{kj} - (a_{ik}^0 - a_{iH_0}^0) b_{kj}^0\} + \sum_{k=H_0}^{H-1} (a_{ik} - a_{iH}) b_{kj} + (a_{iH} - a_{iH_0}^0) \right]^2 . \tag{5.33}$$

$$\text{Let} \begin{cases} a_{ik} = a_{ik} - a_{iH}, & k < H \\ c_i = a_{iH} - a_{iH_0}^0, \\ b_{kj} = b_{kj} \end{cases} \tag{5.34}$$

and put $a_{ik}^0 = a_{ik}^0 - a_{iH_0}^0$. Then we have

$$\|AB - A_0B_0\|^2 \sim \sum_{j=1}^{N} \sum_{i=1}^{M-1} \left\{ \sum_{k=1}^{H_0-1} (a_{ik} b_{kj} - a_{ik}^0 b_{kj}^0) + \sum_{k=H_0}^{H-1} a_{ik} b_{kj} + c_i \right\}^2 . \tag{5.35}$$

We had derived an upper bound of λ by using some inequalities of Frobenius norm and the exact value of λ in special cases [38]. However, in this paper, we use changes of variables which resolve non-negative restrictions and find the RLCT in the all cases.

The transformation (5.34) resolves the non-negative restrictions of $a_{ik}(k < H)$ and c_i for $i = 1, \ldots, M-1$. The changed variables $a_{ik}(k < H)$ and c_i can be negative. We call the

determinant of the Jacobian matrix Jacobian for the sake of simplicity. The Jacobian of the transformation (5.34) equals one.

$$\text{Let} \begin{cases} a_{ik} = a_{ik}, & k < H \\ x_i = c_i + \sum_{k=1}^{H_0-1}(a_{ik}b_{k1} - a_{ik}^0 b_{k1}^0) + \sum_{k=H_0}^{H-1} a_{ik}b_{k1}, & \\ b_{kj} = b_{kj}. & \end{cases} \tag{5.36}$$

It is immediately derived that the Jacobian of this map is equal to one. About the transform (5.36), for $j = 2, \ldots, N$, we have

$$\begin{aligned} &\sum_{k=1}^{H_0-1}(a_{ik}b_{kj} - a_{ik}^0 b_{kj}^0) + \sum_{k=H_0}^{H-1} a_{ik}b_{kj} + c_i \\ &= x_i - \sum_{k=1}^{H_0-1}(a_{ik}b_{k1} - a_{ik}^0 b_{k1}^0) - \sum_{k=H_0}^{H-1} a_{ik}b_{k1} + \sum_{k=1}^{H_0-1}(a_{ik}b_{kj} - a_{ik}^0 b_{kj}^0) + \sum_{k=H_0}^{H-1} a_{ik}b_{kj}. \end{aligned} \tag{5.37}$$

Substituting this for $\sum_{k=1}^{H_0-1}(a_{ik}b_{kj} - a_{ik}^0 b_{kj}^0) + \sum_{k=H_0}^{H-1} a_{ik}b_{kj} + c_i$ in Eq. (5.35), we have

$$\begin{aligned} &\|AB - A_0B_0\|^2 \\ &\sim \sum_{i=1}^{M-1}\left\{\sum_{k=1}^{H_0-1}(a_{ik}b_{k1} - a_{ik}^0 b_{k1}^0) + \sum_{k=H_0}^{H-1} a_{ik}b_{k1} + c_i\right\}^2 \\ &\quad + \sum_{j=2}^{N}\sum_{i=1}^{M-1}\left\{\sum_{k=1}^{H_0-1}(a_{ik}b_{kj} - a_{ik}^0 b_{kj}^0) + \sum_{k=H_0}^{H-1} a_{ik}b_{kj} + c_i\right\}^2 \\ &= \sum_{i=1}^{M-1} x_i^2 + \sum_{j=2}^{N}\sum_{i=1}^{M-1}\left\{x_i - \sum_{k=1}^{H_0-1}(a_{ik}b_{k1} - a_{ik}^0 b_{k1}^0) - \sum_{k=H_0}^{H-1} a_{ik}b_{k1}\right. \\ &\quad \left. + \sum_{k=1}^{H_0-1}(a_{ik}b_{kj} - a_{ik}^0 b_{kj}^0) + \sum_{k=H_0}^{H-1} a_{ik}b_{kj}\right\}^2 \\ &= \sum_{i=1}^{M-1} x_i^2 + \sum_{j=2}^{N}\sum_{i=1}^{M-1}\left[x_i + \sum_{k=1}^{H_0-1}\{a_{ik}(b_{kj}-b_{k1}) - a_{ik}^0(b_{kj}^0 - b_{k1}^0)\} + \sum_{k=H_0}^{H-1} a_{ik}(b_{kj}-b_{k1})\right]^2. \end{aligned} \tag{5.38}$$

Put

$$g_{ij} = \sum_{k=1}^{H_0-1}\{a_{ik}(b_{kj} - b_{k1}) - a_{ik}^0(b_{kj}^0 - b_{k1}^0)\} + \sum_{k=H_0}^{H-1} a_{ik}(b_{kj} - b_{k1}).$$

From Eq. (5.38), we have

$$\|AB - A_0B_0\|^2 \sim \sum_{i=1}^{M-1} x_i^2 + \sum_{j=2}^{N}\sum_{i=1}^{M-1}(x_i + g_{ij})^2. \tag{5.39}$$

Let J be a polynomial ideal $\langle (x_i)_{i=1}^{M-1}, (g_{ij})_{i=1,j=2}^{M-1,N} \rangle$. On account of $x_i + g_{ij} \in J$, we have

$$\sum_{i=1}^{M-1} x_i^2 + \sum_{j=2}^{N} \sum_{i=1}^{M-1} (x_i + g_{ij})^2 \sim \sum_{i=1}^{M-1} x_i^2 + \sum_{j=2}^{N} \sum_{i=1}^{M-1} (x_i^2 + g_{ij}^2),$$

i.e.

$$\begin{aligned}
&\|AB - A_0B_0\|^2 \\
&\sim \sum_{j=2}^{N} \sum_{i=1}^{M-1} (x_i^2 + g_{ij}^2) \\
&\sim \sum_{i=1}^{M-1} x_i^2 + \sum_{j=2}^{N} \sum_{i=1}^{M-1} g_{ij}^2 \\
&= \sum_{i=1}^{M-1} x_i^2 + \sum_{j=2}^{N} \sum_{i=1}^{M-1} \left[\sum_{k=1}^{H_0-1} \{a_{ik}(b_{kj} - b_{k1}) - a_{ik}^0(b_{kj}^0 - b_{k1}^0)\} + \sum_{k=H_0}^{H-1} a_{ik}(b_{kj} - b_{k1}) \right]^2 .
\end{aligned} \tag{5.40}$$

$$\text{Let} \begin{cases} a_{ik} = a_{ik}, & k < H \\ x_i = x_i, & \\ b_{k1} = b_{k1}, & \\ b_{kj} = b_{kj} - b_{k1} & j > 1. \end{cases} \tag{5.41}$$

For $k = 1, \ldots, H-1$ and $j = 2, \ldots, N$, non-negative restrictions of b_{kj} can be resolved. The Jacobian of the transformation (5.41) is one. Apply this map to Eq. (5.40) and put $b_{kj}^0 = b_{kj}^0 - b_{k1}^0$. Then, we have

$$\begin{aligned}
\|AB - A_0B_0\|^2 &\sim \sum_{i=1}^{M-1} x_i^2 + \sum_{j=2}^{N} \sum_{i=1}^{M-1} \left\{ \sum_{k=1}^{H_0-1} (a_{ik}b_{kj} - a_{ik}^0 b_{kj}^0) + \sum_{k=H_0}^{H-1} a_{ik}b_{kj} \right\}^2 \\
&= \sum_{i=1}^{M-1} x_i^2 + \sum_{j=2}^{N} \sum_{i=1}^{M-1} \left(\sum_{k=1}^{H-1} a_{ik}b_{kj} - \sum_{k=1}^{H_0-1} a_{ik}^0 b_{kj}^0 \right)^2 .
\end{aligned} \tag{5.42}$$

There are not b_{k1} $(k = 1, \ldots, H-1)$ in the right hand side; thus, we can regard the non-negative restrictions of the all variables are resolved after applying the transformation (5.41).

Real matrices U, V, U_0, and V_0 are denoted by $U := (u_{ik})_{i=1,k=1}^{M-1,H-1}$, $V := (v_{kl})_{k=1,l=1}^{H-1,N-1}$, $U_0 := (u_{ik}^0)_{i=1,k=1}^{M-1,H_0-1}$, and $V_0 := (v_{kl}^0)_{k=1,l=1}^{H_0-1,N-1}$, respectively. Here, we have $u_{ik} = a_{ik}$, $v_{kl} = v_{k(j-1)} = b_{kj}$, $u_{ik}^0 = a_{ik}^0$, and $v_{kl}^0 = v_{k(j-1)}^0 = b_{kj}^0$ for $i = 1, \ldots, M-1$, $k = 1, \ldots, H-1$ and $j = 2, \ldots, N$. Note that these U_0 and V_0 are same as in Eqs. (5.13) and (5.14) because of the above transformations for the entries of A_0 and B_0. Therefore, $r := \mathrm{rank}(U_0V_0)$ is also equal to that of Main Theorem.

Now, let us start coming down the problem from LDA to reduced rank regression.

$$\|UV - U_0V_0\|^2 = \sum_{l=1}^{N-1}\sum_{i=1}^{M-1}\left(\sum_{k=1}^{H-1} u_{ik}v_{kl} - \sum_{k=1}^{H_0-1} u^0_{ik}v^0_{kl}\right)^2$$
$$= \sum_{j=2}^{N}\sum_{i=1}^{M-1}\left(\sum_{k=1}^{H-1} a_{ik}b_{kj} - \sum_{k=1}^{H_0-1} a^0_{ik}b^0_{kj}\right)^2 \tag{5.43}$$

holds; thus, from Eq. (5.42) and (5.43), we have

$$\|AB - A_0B_0\|^2 \sim \sum_{i=1}^{M-1} x_i^2 + \|UV - U_0V_0\|^2. \tag{5.44}$$

Let (λ_1, m_1) and (λ_2, m_2) be pairs of the RLCT and its multiplicity of the first and the second term, respectively. There is no common variable between $\{(x_i)_{i=1}^M\}$ and $\{(U, V)\}$; hence, we have

$$\lambda = \lambda_1 + \lambda_2, \tag{5.45}$$
$$m = m_1 + m_2 - 1. \tag{5.46}$$

By simple calculation, $\lambda_1 = (M-1)/2$ and $m_1 = 1$ hold. Besides, the entries of the matrices U and V can be real as well as non-negative. Thus, λ_2 is the RLCT of non-restricted MF, i.e. that of reduced rank regression [10]. The same is true for the multiplicity m_2. Therefore, we obtain

$$\lambda = \frac{M-1}{2} + \lambda_{\mathrm{MF}}(M-1, N-1, H-1, r), \tag{5.47}$$
$$m = m_{\mathrm{MF}}(M-1, N-1, H-1, r), \tag{5.48}$$

where $r = \mathrm{rank}(U_0V_0)$.

Finally, we concretely calculate λ and m. According to [10], the RLCT and its multiplicity of MF are as follows.

(1) If $N + r + 1 \leqq M + H$ and $M + r + 1 \leqq N + H$ and $H + r + 1 \leqq M + N$ and $M + N + H + r + 1$ is even ($M + N + H + r$ is odd), then

$$\lambda_{\mathrm{MF}}(M-1, N-1, H-1, r) = \frac{1}{8}\{2(H+r-1)(M+N-2) - (M-N)^2 - (H+r-1)^2\}, \tag{5.49}$$
$$m_{\mathrm{MF}}(M-1, N-1, H-1, r) = 1. \tag{5.50}$$

(2) Else if $N + r + 1 \leqq M + H$ and $M + r + 1 \leqq N + H$ and $H + r + 1 \leqq M + N$ and $M + N + H + r + 1$ is odd ($M + N + H + r$ is even), then

$$\lambda_{\mathrm{MF}}(M-1, N-1, H-1, r) = \frac{1}{8}\{2(H+r-1)(M+N-2) - (M-N)^2 - (H+r-1)^2 + 1\}, \tag{5.51}$$
$$m_{\mathrm{MF}}(M-1, N-1, H-1, r) = 2. \tag{5.52}$$

(3) Else if $M+H<N+r+1$, then

$$\lambda_{\mathrm{MF}}(M-1,N-1,H-1,r)=\frac{1}{2}\{(M-1)(H-1)+(N-1)r-(H-1)r\}, \quad (5.53)$$
$$m_{\mathrm{MF}}(M-1,N-1,H-1,r)=1. \quad (5.54)$$

(4) Else if $N+H<M+r+1$, then

$$\lambda_{\mathrm{MF}}(M-1,N-1,H-1,r)=\frac{1}{2}\{(N-1)(H-1)+(M-1)r-(H-1)r\}, \quad (5.55)$$
$$m_{\mathrm{MF}}(M-1,N-1,H-1,r)=1. \quad (5.56)$$

(5) Else (i.e. $M+N<H+r+1$), then

$$\lambda_{\mathrm{MF}}(M-1,N-1,H-1,r)=\frac{1}{2}(M-1)(N-1), \quad (5.57)$$
$$m_{\mathrm{MF}}(M-1,N-1,H-1,r)=1. \quad (5.58)$$

Since the multiplicity is clear, we find the RLCT. We develop the terms in each case by using Eq. (5.47).
In the case (1), we have

$$\lambda=\frac{M-1}{2}+\frac{1}{8}\{2(H+r-1)(M+N-2)-(M-N)^2-(H+r-1)^2\} \quad (5.59)$$
$$=\frac{M-1}{2}+\frac{1}{8}\{2(H+r)(M+N)-2(M+N)-4(H+r)+4$$
$$-(M-N)^2-(H+r)^2+2(H+r)-1\} \quad (5.60)$$
$$=\frac{1}{8}\{4M-4+2(H+r)(M+N)-(M-N)^2-(H+r)^2-2(M+N)-2(H+r)+3\} \quad (5.61)$$
$$=\frac{1}{8}\{2(H+r)(M+N)-(M-N)^2-(H+r)^2+2(M+N)-2(H+r)-1-4N\} \quad (5.62)$$
$$=\frac{1}{8}\{2(H+r+1)(M+N)-(M-N)^2-(H+r+1)^2\}-\frac{N}{2}. \quad (5.63)$$

In the case (2), by the same way as the case (1), we have

$$\lambda=\frac{1}{8}\{2(H+r+1)(M+N)-(M-N)^2-(H+r+1)^2+1\}-\frac{N}{2}. \quad (5.64)$$

In the case (3), we have

$$\lambda = \frac{M-1}{2} + \frac{1}{2}\{(M-1)(H-1) + (N-1)r - (H-1)r\} \tag{5.65}$$
$$= \frac{M-1}{2} + \frac{1}{2}\{MH - (M+H) + 1 + Nr - r - Hr + r\} \tag{5.66}$$
$$= \frac{1}{2}(MH - M - H + 1 + Nr - Hr + M - 1 + N - N) \tag{5.67}$$
$$= \frac{1}{2}\{MH + 1 + N(r+1) - H(r+1) - N\} \tag{5.68}$$
$$= \frac{1}{2}\{MH + 1 + N(r+1) - H(r+1)\} - \frac{N}{2}. \tag{5.69}$$

In the case (4), by the same way as the case (3), we have

$$\lambda = \frac{1}{2}\{NH + 1 + M(r+1) - H(r+1)\} - \frac{N}{2}. \tag{5.70}$$

In the case (5), we have

$$\lambda = \frac{M-1}{2} + \frac{1}{2}(M-1)(N-1) \tag{5.71}$$
$$= \frac{1}{2}(M-1)N \tag{5.72}$$
$$= \frac{1}{2}MN - \frac{N}{2}. \tag{5.73}$$

From the above, Main Theorem is proved. Comparing the RLCT of MF [10], we also obtain

$$\lambda = \lambda_{\mathrm{MF}}(M, N, H, r+1) - \frac{N}{2}.$$

Therefore, we obtain Theorem 5.2.

□

5.5 Discussion

Here, we will discuss the results of this chapter from three points of view. After that, we will describe the numerical behavior of the theoretical result by conducting numerical experiments.

5.5.1 Parameter Restriction

The RLCT of LDA can be represented by using that of MF. Namely, Main Theorem can be interpreted as that the learning coefficient of LDA is that of the unconstrained MF minus the penalty due to the simplex constraint. In fact, it can be proved that

$$\lambda = \lambda_{\mathrm{MF}}(M, N, H, r+1) - \frac{N}{2} \tag{5.74}$$

holds (see also the rigorous proof of Main Theorem in Appendix). The dimension of the stochastic matrix AB with the degrees of freedom is $(M-1)N = MN - N$. The subtracted N is the dimension of the parameter that can be uniquely determined from the parameters of the other $(M-1)N$ dimensions in the matrix AB. This part can be regarded as an N-dimensional regular statistical model, whose RLCT is $N/2$. This is the reason of the above statement. Note that Main Theorem and its proof are not trivial. A hermeneutic explanation cannot be a mathematical proof. In addition, the actual parameter dimension is $(M-1)H + (H-1)N = (M+N-1)H - N$ because we have to consider the matrices A and B rather than AB. We cannot reach the result of this paper simply by maintaining consistency of the degrees of freedom. Algebraic geometrical methods are used to solve this problem in learning theory: what the learning coefficient of LDA is.

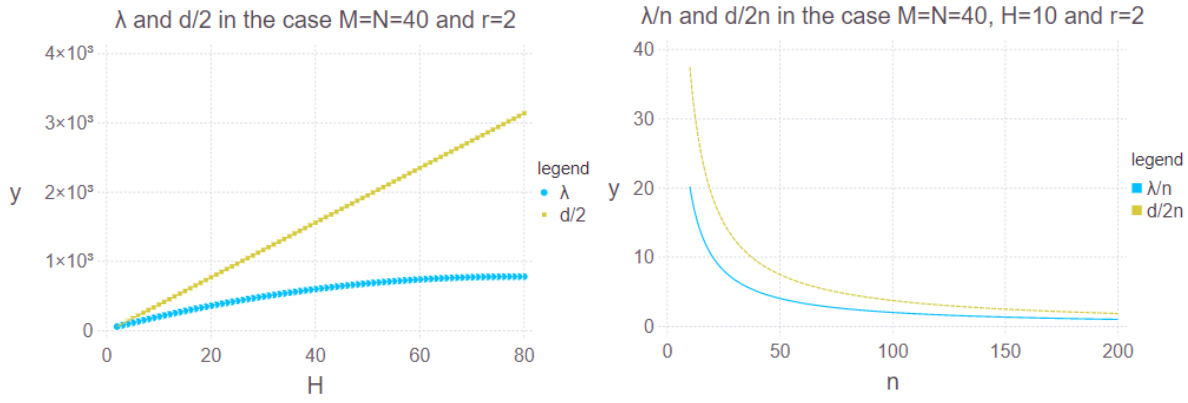
5.5.2 Theoretical Application

Since LDA is a knowledge discovery method, marginal-likelihood-based model selection often tends to be preferred. However, BIC [66] cannot be used for LDA because it is a singular statistical model. Although Gibbs sampling is usually used for full Bayesian inference of LDA, it is difficult to achieve a tempered posterior distribution; thus, we need other Markov chain Monte Carlo method (MCMC) to calculate WBIC [83] and WsBIC [43]. The result of this study allows us to perform a rigorous model selection of LDA with sBIC [24], which is MCMC-free. Even when the marginal likelihoods are computed directly by the exchange Monte Carlo method, our result is useful for the design of the exchange probability [56]. Furthermore, it may be possible to evaluate how precise MCMC approximates the posterior with the exact values of that [85, 43].

One may use BIC for model selection of LDA; however, using it causes that too small models are chosen. This is because there exists a large difference in values and behaviors between $d/2$ and λ. In a regular statistical model, the learning coefficient is half of the parameter dimension $d/2$. In LDA, $d/2 = (M+N-1)H/2 - N/2$ holds; hence, it linearly increases as the number of topics H does. On the other hand, the RLCT of LDA λ does not. In addition, λ is much smaller than $d/2$. Fig. 5.2a shows how the RLCT of LDA λ behaves when the number of topics H increases, with λ-value in the vertical axis and H-value in the horizontal axis. If λ was equal to $d/2$, then it would linearly increase (the square markers dotted plot in Fig. 5.2a). However, in fact, λ is given by Main Theorem and its curve is obviously non-linear (the circles dotted plot in Fig. 5.2a). Hence, their values and behaviors are very different. BIC is based on $d/2$ from the asymptotics of regular statistical models. In contrast, the foundation of sBIC is singular learning theory; thus, it uses λ instead of $d/2$. That is why sBIC is theoretically recommended for LDA.

5.5.3 Behavior of Learning Curve

We can draw the theoretical learning curve like the solid line in Fig. 5.2b, with $\mathbb{E}[G_n]$-value in the vertical axis and n-value in the horizontal axis. We also namely draw a curve like the dashed line in Fig. 5.2b. This dashed curve is not only an upper bound of the learning curve of LDA in Bayesian inference but also a lower bound of that in maximum likelihood or posterior estimation methods. Let G_n^{MAP} and μ be the generalization error and the learning coefficient of LDA in maximum posterior (MAP) methods, respectively. This is well-defined,



(a) (b)

Fig. 5.2: (a) In this chapter, we give the exact value of the learning coefficient of LDA λ. The learning coefficient is smaller than half of the parameter dimension $d/2$, since LDA is a singular statistical model. The dotted blue line drawn by the circles in this figure represents the learning coefficients of LDA when the number of topics H is increased. If LDA was a regular statistical model, its learning coefficient would be the dotted yellow line drawn by the squares. The behavior of them are so different.
(b) This figure shows the theoretical learning curve of LDA and that of a regular statistical model whose parameter dimension d is same as LDA. The former is the solid blue line and the latter is the dashed yellow line. The vertical axis means the expected generalization error $\mathbb{E}[G_n]$ and the horizontal one is the sample size n. This is based on Theorem 3.4 and the exact value of λ which is clarified by our result.

i.e. $\mathbb{E}[G_n^{\mathrm{MAP}}] = \mu/n + o(1/n)$ holds (see also Theorem 3.5). On the basis of the same prior distribution, Watanabe proved the following inequality [85]:

$$\lambda < d/2 < \mu. \tag{5.75}$$

This means $\mathbb{E}[G_n^{\mathrm{MAP}}] > \mathbb{E}[G_n] + o(1/n)$ and the leading term of these difference is $(\mu - \lambda)/n > (d - 2\lambda)/2n$. Owing to Main Theorem, we immediately have the exact value of $d - 2\lambda$. Therefore, our result shows at least how much Bayesian inference improves the generalization performance of LDA compared to MAP method. If the prior distribution is a uniform one, then μ equals the learning coefficient of LDA in maximum likelihood estimation. Hence, the above consideration can be applied to maximum likelihood estimation.

5.5.4 Experiment

Now, we run numerical experiments to check the behavior of Main Theorem when the sample size is finite. Theorem 5.2 gives the exact asymptotic form of Bayesian generalization error in LDA by using Theorem 3.4. We calculate the Bayesian generalization error in LDA by using Gibbs sampling and compare the numerically-calculated RLCT with the theoretical one. Our experimental approach and its description in this section is based on [33] since it also treats the numerical experiment to compute the RLCT by using Gibbs sampling to verify the numerical

behavior of its theoretical result.

Let $\mathbb{E}_\theta$ be an expectation operator of the posterior: $\mathbb{E}_\theta[\cdot] = \int \mathrm{d}\theta \psi(\theta|\mathcal{D})[\cdot]$. Let $\hat{\lambda}$ be the numerically calculated RLCT. The widely applicable information criterion (WAIC) [82] is defined by the following random variable W_n:

$$W_n = T_n + V_n/n, \tag{5.76}$$

where T_n is the empirical loss and V_n is the functional variance:

$$T_n = -\frac{1}{n}\sum_{i=1}^{n} \log p^*(X_i) = -\frac{1}{n}\sum_{i=1}^{n} \log \mathbb{E}_\theta[p(X_i|\theta)], \tag{5.77}$$

$$V_n = \sum_{i=1}^{n} \left[\mathbb{E}_\theta[(\log p(X_i|\theta))^2] - \{\mathbb{E}_\theta[\log p(X_i|\theta)]\}^2 \right] = \sum_{i=1}^{n} \mathbb{V}_\theta[\log p(X_i|\theta)]. \tag{5.78}$$

Even if the posterior distribution cannot be approximated by any normal distribution (i.e., the model is singular), the expected WAIC $\mathbb{E}[W_n]$ is asymptotically equal to the expected generalization loss $\mathbb{E}[G_n + S]$ [82];

$$\mathbb{E}[G_n + S] = \mathbb{E}[W_n] + O(1/n^2). \tag{5.79}$$

Moreover, the generalization error and the WAIC error $W_n - S_n$ have the same variance [82, 85]:

$$G_n + W_n - S_n = 2\lambda/n + O_p(1/n). \tag{5.80}$$

We need to repeat the simulations to compute $\hat{\lambda}$ to decrease the random effect caused by G_n, W_n and S_n. Thus, Eq. (5.80) is useful for computing $\hat{\lambda}$ because the leading term $2\lambda/n$ is deterministic, nevertheless the left hand side is probabilistic. This means that the needed number of simulations D can be less than that in the case using $\lambda \approx n\mathbb{E}[G_n]$ from Theorem (3.4).

The method was as follows. First, the training data $\mathcal{D}$ was generated from the true distribution $q(X|Z)$. Second, the posterior distribution was calculated by using Gibbs sampling [30] (see also Algorithm 5). Third, G_n and $W_n - S_n$ were computed by using the training data $\mathcal{D}$ and the artificial test data $\mathcal{D}^* = (X_t^*)$ generated from $q(X|Z)$. These three steps were repeated and each value of $n(G_n + W_n - S_n)/2$ was saved. After all repetitions have been completed, $n(G_n + W_n - S_n)/2$ was averaged over the simulations. This average was $\hat{\lambda}$.

The pseudo-code is listed in Algorithm 4, where K is the sample size of the parameter subject to the posterior and n_T is the sample size of the synthesized test data. We used the programing language named Julia [13] for this experiment.

We set $M = 10$, $N = 5$, $H_0 = 2$, $r = 1$, $n = 1000$, and $n_T = 200n = 200000$. To examine the numerical behavior when the number of topics $H \geqq H_0$ in the model increases, we set $H = 2, 3, 4, 5$ and carried out experiments in each case. To decrease the probabilistic effect of Eq. (5.80), we conducted the simulations one hundred times, i.e. $D = 100$.

In the Gibbs sampling, we had to conduct a burn-in to decrease the effect of the initial values and thin the samples in order to break the correlations. The length of the burn-in was 10000, while the length of the thinning was 20; thus, the sample sizes of the parameter was $10000 +$

Algorithm 4 How to Compute $\hat{\lambda}$

Require: D: the number of simulations,
A_0: the true parameter matrix whose size is (M, H_0),
B_0: the true parameter matrix whose size is (H_0, N),
GS: the Gibbs sampling function whose return value consists of the samples from the posterior. See also Algorithm 5.
Ensure: The numerical computed RLCT $\hat{\lambda}$.
Allocate an array $\Lambda[D]$.
for $d = 1$ to D **do**
Generate $\mathcal{D}$ from the true distribution.
Allocate arrays $\mathcal{A}[M, H, K]$ and $\mathcal{B}[H, N, K]$.
Get $\mathcal{A}, \mathcal{B} \leftarrow \mathrm{GS}(\mathcal{D})$.
Generate $\mathcal{D}^*$ from the true distribution.
Calculate $G_n \approx \frac{1}{n_T}\sum_{t=1}^{n_T} \log \frac{q(X_t^*)}{\mathbb{E}_\theta[p(X_t^*|\theta)]}$, $S_n = -\frac{1}{n}\sum_{i=1}^{n} \log q(X_i)$,
and $W_n \approx -\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\theta[p(X_i|\theta)] + \frac{1}{n}\sum_{i=1}^{n} \mathbb{V}_\theta[\log p(X_i|\theta)]$,
where $\mathbb{E}_\theta[f(\theta)] \approx \frac{1}{K}\sum_{k=1}^{K} f(\theta_k)$ and $\theta_k = (\mathcal{A}[:,:,k], \mathcal{B}[:,:,k])$.
Save $\Lambda[d] \leftarrow n(G_n + W_n - S_n)/2$.
end for
Calculate $\hat{\lambda} = \frac{1}{D}\sum_{d=1}^{D} \Lambda[d]$.

$20K = 50000$ ($K = 2000$) and we used the $(10000 + 20k)$-th sample as the entry of $\mathcal{A}[:,:,k]$ and $\mathcal{B}[:,:,k]$ for $k = 1$ to K. The implementation is available at the following github page: `https://github.com/chijan-nh/LearningCoefficient-RLCT-ofLDA-usingGS`.

The experimental results are shown in Table 5.2 and visualized in Fig. 5.3. The symbol λ denotes the exact value of the RLCT λ in Theorem 5.2. There are three columns for each H, and each row contains the model settings (Settings), symbols of calculated values (RLCTs), and the theoretical or numerical values (Values). The experimental values have four significant digits. The numerically-calculated RLCT $\hat{\lambda}$ is an average of $n(G_n + W_n - S_n)/2$ obtained from each simulations; hence, we also show the standard deviation of it as the right next to the plus-minus sign $\pm$ in each setting.

As shown in Table 5.2 and Fig. 5.3, all numerically calculated values are nearly equal to the theoretical RLCTs, i.e. these differences are sufficiently smaller than the standard deviation overall simulations. Note that the parameter dimensions $(M-1)H + (H-1)N$ are 23, 37, 51 and 65 for $H = 2$, 3, 4 and 5; thus, we consider that the sample size $n = 1000$ is not an asymptotic scale. Moreover, we also consider that it is natural to fix the sample size while the number of topics increases because we compare some models for a dataset (the sample size is fixed) in practical situations. Although the sample size is finite (not an asymptotic scale) and fixed, the theoretical values are included in the 1-standard deviation ranges for each case. Besides, because of Fig. 5.3, the standard deviations are sufficiently small for the scale of the RLCTs. Therefore, Theorem 5.2 is consistent with the experimental result.

Algorithm 5 Gibbs Sampling for LDA
$\mathrm{GS}(\mathcal{D}, \alpha = 1_H, \beta = 1_N, K = K, \mathrm{burnin} = 10000, \mathrm{thin} = 20)$

Require: $\mathcal{D} = \{(z_l, x_l)\}_{l=1}^{L(=n/N)}$: the data where $z_l \in \mathrm{Onehot}(N)$ and $x_l \in \mathrm{Onehot}(M)$ are a document and a word, respectively. When $z_{lj}=1$, x_l is the l-th word in the document j.
$\alpha \in \mathbb{R}_{>0}^H$: the hyperparameter of the Dirichlet prior for the stochastic matrix A,
$\beta \in \mathbb{R}_{>0}^N$: the hyperparameter of the Dirichlet prior for the stochastic matrix B,
$\mathrm{Dir}(C|\Gamma)$: a Dirichlet distribution of a stochastic matrix whose h-th column is generated by $\mathrm{Dir}(c|\Gamma[:, h])$.
Ensure: Sampling stochastic matrices from the numerical posterior.
Let $\mathrm{iter} = \mathrm{butnin} + \mathrm{thin} * K$.
Allocate arrays $\mathcal{A}[M, H, K]$, $\overline{\mathcal{A}}[M, H, \mathrm{iter}]$, $\mathcal{B}[H, N, K]$ and $\overline{\mathcal{B}}[H, N, \mathrm{iter}]$.
Inital sampling for A and B from the prior:
Generate $A, B \sim \mathrm{Dir}(A|\alpha)\mathrm{Dir}(B|\beta)$.
# Sampling from the posterior:
for $k = 1$ to iter **do**
  ## Sampling the topic y.
  Allocate an array $y[L, H]$.
  for $l = 1$ to L **do**
    for $h = 1$ to H **do**
      Let $\eta[l, h] = \exp(\sum_{j=1}^N z_{lj}(\log b_{hj} + \sum_{i=1}^M x_{li} \log a_{ih}))$.
      Put $\eta[l, h] \leftarrow \eta[l, h] / \sum_{h=1}^H \eta[l, h]$.
    end for
    Generate $y[l, :] \sim \mathrm{Cat}(y|\eta[l, :])$.
  end for
  ## Sampling the stochastic matrix A.
  for $h = 1$ and $i = 1$ to H and M **do**
    Let $\hat{\alpha}[i, h] = \sum_{l=1}^L y[l, h]x[l, i] + \alpha_h$.
  end for
  Generate $A \sim \mathrm{Dir}(A|\hat{\alpha})$.
  Put $\overline{\mathcal{A}}[M, H, k] \leftarrow A$.
  ## Sampling the stochastic matrix B.
  for $j = 1$ and $h = 1$ to N and H **do**
    Let $\hat{\beta}[h, j] = \sum_{l=1}^L z[l, j]y[l, h] + \beta_j$.
  end for
  Generate $B \sim \mathrm{Dir}(B|\hat{\beta})$.
  Put $\overline{\mathcal{B}}[H, N, k] \leftarrow B$.
end for
# Burn-in and thinning.
for $k = 1$ to K **do**
  $\mathcal{A}[M, H, k] \leftarrow \overline{\mathcal{A}}[M, H, \mathrm{burnin} + \mathrm{thin} * k]$.
  $\mathcal{B}[H, N, k] \leftarrow \overline{\mathcal{B}}[H, N, \mathrm{burnin} + \mathrm{thin} * k]$.
end for
Return $\mathcal{A}, \mathcal{B}$.

Table 5.2: Numerically-Calculated and Theoretical Values of RLCTs

Settings	RLCTs	Values
$H = 2$	Theoretical: λ	21/2
$(M = 10, N = 5$	Numerical: $\hat{\lambda}$	10.79 ± 0.8591
$H_0 = 2, r = 1)$	Difference: $\lvert\lambda - \hat{\lambda}\rvert$	0.2901 ± 0.8591
$H = 3$	Theoretical: λ	12
$(M = 10, N = 5$	Numerical: $\hat{\lambda}$	12.25 ± 0.9510
$H_0 = 2, r = 1)$	Difference: $\lvert\lambda - \hat{\lambda}\rvert$	0.2534 ± 0.9510
$H = 4$	Theoretical: λ	27/2
$(M = 10, N = 5$	Numerical: $\hat{\lambda}$	13.57 ± 1.036
$H_0 = 2, r = 1)$	Difference: $\lvert\lambda - \hat{\lambda}\rvert$	0.07114 ± 1.036
$H = 5$	Theoretical: λ	15
$(M = 10, N = 5$	Numerical: $\hat{\lambda}$	14.80 ± 1.143
$H_0 = 2, r = 1)$	Difference: $\lvert\lambda - \hat{\lambda}\rvert$	0.2049 ± 1.143

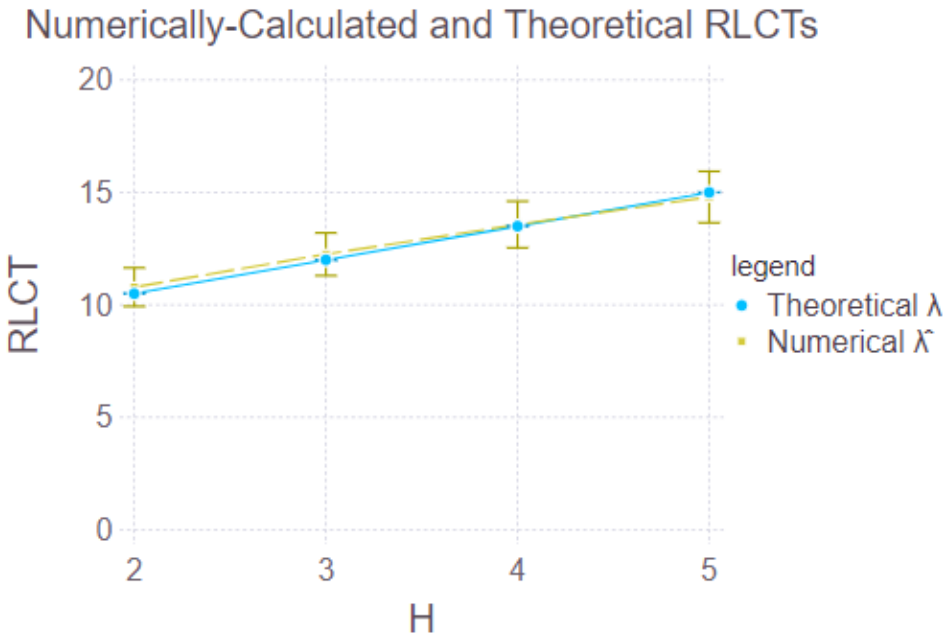


Fig. 5.3: This figure is drawn based on Table 5.2 and Theorem 5.2. It compares numerically-calculated RLCTs $\hat{\lambda}$ (Numerical $\hat{\lambda}$, as the dashed yellow line with the error bars) and theoretical ones λ (Theoretical λ, as the solid blue line) for $H = 2, 3.4.5$. The horizontal line means the number of topics H and the vertical one is the numerically-calculated or theoretical value of the RLCT. Each error bar of experimental results is the 1-standard deviation range. The line of Numerical $\hat{\lambda}$ and that of Theoretical λ are very close and the standard deviations are sufficiently smaller than the scale of the RLCTs.

5.6 Conclusion

We determine the RLCT of LDA in general cases (Theorem 5.2) and describe theoretical applications to Bayesian inference. According to Theorem 5.3, Theorem 5.2 means that the exact asymptotic form of the generalization error and the free energy are theoretically derived for LDA.

The RLCT of LDA monotonically increases when the number of topic increases (Fig 5.2a), although its representation has complex branches as Theorem 5.2. Besides, the RLCT saturates to a constant which is not depend on the number of the model's topics and the rank determined by the true distribution. Hence, its behavior is non-linear and bounded. On the other hand, the parameter dimension d monotonically and linearly increases and it is not bounded. If LDA was regular, its RLCT would be $d/2$. However, LDA is singular and its RLCT is less than $d/2$. The theoretical result shows how they are different. Moreover, these difference $d/2 - \lambda$ gives a lower bound of the difference between MAP and Bayesian generalization error. We numerically compute the RLCT when the sample size is finite (not asymptotic scale) and the experimental result is consistent to the main result.

One of future works is clarifying the effect of the hyperparameter to the generalization error and the free energy. This is formulated to finding simultaneous resolution of singularities when the prior distribution is a Dirichlet distribution. A density function of a non-uniform Dirichlet distribution has zero or diverged points; thus, the hyperparameter (i.e. the parameter of the Dirichlet distribution) may affect the RLCT of LDA.

Chapter 6

Conclusion

6.1 Conclusion

In this dissertation, the author aims to establish statistical learning theory when the parameter of the model region is restricted. As a foundation of that, we studied the asymptotic behaviors of the Bayesian generalization errors in non-negative matrix factorization (NMF) and latent Dirichlet allocation (LDA). NMF and LDA are two typical singular models whose parameter regions are restricted.

Singular learning theory is a mathematical foundation for statistical inference using singular models. It describes the asymptotic behaviors of the Bayesian generalization error and the free energy by using algebraic geometry; a real log canonical threshold (RLCT) rules them. Thus, we analyzed those of NMF and LDA in the framework of singular learning theory. The contributions of the author's works are as follows:

- An upper bound of the RLCT of NMF is theoretically derived. The numerical experiment was consistent with the theoretical result and provided knowledge about the stability of learning.
  - The upper bound provides theoretical upper bounds of the Bayesian generalization error and the free energy in NMF.
  - The bound includes the exact value in some cases and depends on the hyperparameter.
  - The Bayesian generalization error and the free energy in NMF become strictly larger than that of non-restricted matrix factorization when the entries of the true parameter matrices are zero.
  - The difference between the learning coefficient of VBNMF and the RLCT of NMF gives the variational approximation error. Thus, using the upper bound, we also derived a lower bound of the variational approximation error in NMF.
  - This theoretical analysis for NMF shows a phase transition structure; there is a critical line of hyperparameters. The Bayesian generalization error and the variational approximation error drastically changes beyond that line.
  - The phase transition line we found is different from that of variational Bayesian NMF. That is because the variational posterior of NMF is different from the Bayesian posterior distribution of NMF.
- The exact value of the RLCT is mathematically determined in general cases for LDA. The numerical experiment was consistent with the theoretical result.

- The RLCT provides the exact asymptotic form of the Bayesian generalization error and the free energy in LDA.
- We clarify that the RLCT of LDA is equal to that of stochastic matrix factorization (SMF). SMF is one of the matrix factorization models whose parameter region is restricted. By using this fact, we obtain the exact value of the RLCT.
- SMF is NMF whose matrices have columns in simplexes. Thus, the main result shows the effect of the simplex restriction in matrix factorization.
- When the number of topics increases, the RLCT monotonically and non-linearly grows but is bounded, whereas the parameter dimension linearly does and is not bounded.
- This mathematical study for LDA shows the RLCT of LDA is much smaller than that of a regular model whose parameter dimension is the same as LDA. This fact provides at least how much the generalization performance improves when we use Bayesian inference.

6.2 Future Work

The future research topics for establishing singular learning theory of parameter-restricted models include the followings. There are two layers of topics: general theory and problems with concrete models such as NMF and LDA.

First, we briefly mention the research topics for creating a general theory when there exists a constraint for the parameter region. Even if the sample is i.i.d. from the data-generating distribution, the effect to the generalization error caused by parameter restriction has not yet been clarified. In general, like Lemma 4.6, when the model is the same, the support of the prior distribution is larger, the RLCT is smaller. This fact causes that parameter restriction which does not decrease the dimension, such as non-negative restriction in NMF, increases the Bayesian generalization error. However, this is only a qualitative evaluation. Thus, there has not been a quantitative evaluation method. Besides, the restriction that decreases the dimension, such as simplex restriction in LDA and SMF, does not generally increase the RLCT. In fact, in SMF, there are simplex restrictions to the parameter region and the RLCT of SMF is smaller than that of non-restricted matrix factorization.

Because of the difficulty in the general case, a possible policy is to construct a theory for each class of algebraic varieties that characterize RLCTs. Vandermonde matrix type singularities, which give RLCTs of mixture distribution models and neural networks, are known and these RLCTs have been studied [5, 9, 8]. However, the quantitative effect of the constraints to these parameter regions is still unknown except for the RLCT of the Poisson mixture model [64]; the average parameter of each component is non-negative. In this thesis, the strategy was to deal with singularities formed by the squared error of matrix factorization. For these singularities, we considered two typical constraints (NMF and LDA). Future topics include other restrictions such as semi non-negative [22] and convolutive non-negative [67].

Second, we describe future works for singular learning theory of NMF and LDA.

For NMF, clarifying the exact value of the RLCT is considered one of the most significant but demanding open problems. The non-negative rank of a non-negative matrix is generally different from the usual rank of that; thus, we cannot apply the proof of Aoyagi's result [10] to the RLCT of NMF. Moreover, analysis and comparison of non-parametric Bayesian NMF

[41] is also included in future tasks.

For LDA, determining the RLCT when the prior is a Dirichlet distribution is considered as one of the most important future research directions. This situation is formulated to finding simultaneous resolution of singularities. A density function of a non-uniform Dirichlet distribution has zero or diverged points; thus, the hyperparameter (the parameter of the Dirichlet distribution) may affect the RLCT of LDA. If we conducted the simultaneous resolution, the effect of the hyperparameter on the RLCT would be clarified. Moreover, variational approximation error could be quantitatively evaluated in the same way as Theorem 4.4 since the variational free energy and its learning coefficient in LDA has been determined [59]. Analysis and comparison of non-parametric LDA [99] is also considered as one of the future research topics.

Note that the extracted clusters strongly depend on the hyperparameter in both non-parametric NMF and LDA. Hence, we have to reveal the free energy and the generalization error as functions of the hyperparameter. For singular models, the effect of the hyperparameter is complex since they have phase transition structures. Besides, in the framework of non-parametric Bayesian inference, the model size such as the non-negative rank and the number of topics is referred to as infinity; thus, we may not be able to directly apply singular learning theory to the analysis of non-parametric NMF and LDA. Hence, we consider that the analysis of non-parametric cases is challenging.

Appendix A

Questions and Answers in Defense

In this chapter, the author writes the questions and the answers at the defense of this dissertation. The defenses were held on 2021/02/22 (pre-examination) and 2021/03/29 (public presentation).

A.1 Singular Learning Theory

Q. 1 **What is the benefit of determining RLCTs of statistical models in the situation of real data analysis?**

(a) One of the most important benefits is that we can use sBIC [24] (see also Sec. 3.4) to select the appropriate model in knowledge discovery sense. The criterion sBIC uses the theoretical value of the RLCT and its multiplicity. If they are unknown, WsBIC [43] is one of the choices because it estimates the RLCT. However, the calculation cost of estimating the RLCT is very high and this is just estimation: we cannot obtain the exact value. Thus, clarifying the RLCT and its multiplicity is useful for model selection.

Furthermore, a tuning method for the inversed temperature parameter in exchange Monte Carlo has been proposed, which uses the theoretical RLCT [56]. It has been also considered that the correctness of MCMC is verified by comparing the estimated RLCT and the theoretical one [85, 43]. We can estimate the sufficient sample size n^* such that $\mathbb{E}[G_n] < \varepsilon$ by using $\lambda/n^* < \varepsilon \Leftrightarrow n^* > \lambda/\varepsilon$ if the theoretical value of the RLCT λ is clarified. This fact means that there is a potential application to propose how much data should be collected.

Q. 2 **Let n be the sample size. Can singular learning theory treat the limit of the model size D such that $D, n \to \infty$ and $D/n = \gamma > 0$? For example, D is the size of the data matrix in NMF or the number of topics in LDA.**

(a) At present, no it cannot. In singular learning theory, the model and the prior are fixed to the sample. However, we can consider the limit of $D \to \infty$ in the RLCT. For example, in LDA, when the number of topics H tends to the infinity, the RLCT of LDA λ converges to a constant:

$$\lim_{H \to \infty} \lambda = (MN - N)/2,$$

where M and N are the number of vocabulary and that of documents, respectively (see Chap. 5).

Q. 3 **Does singular learning theory treat only the case that the parameter region is not restricted?**

(a) Singular learning theory can treat not only the case that the parameter region is not restricted but also that is constrained, even if the model is singular, a map from a parameter to a probability distribution is not injective and its likelihood and posterior cannot be approximated by any normal distribution. However, the effect of the parameter restriction on the generalization error has yet been unknown. Thus, this dissertation aims to construct statistical learning theory for parameter-restricted singular models. As a foundation of it, the author studies two typical models: NMF and LDA.

Q. 4 **Is it really that the parameter restriction increases the generalization error? The questioner guesses that the constraint seems to make the estimation result fit a good area.**

(a) In Chap. 4, it is mathematically proved that the non-negative restriction strictly increases the RLCT in some cases. Intuitively, this is because the constraint makes the parameter space narrow without decreasing in dimension, which eliminates regions that are more amenable to generalization.

Q. 5 **Some models are often trained with a penalty term to its loss function to relax their non-identifiability. Can we consider a singular model as a regular model by applying norm constraints with a penalty?**

(a) No, we cannot. When the loss function corresponds to the likelihood of the model, the penalty is referred to as the prior distribution. For example, the L2 penalty to the negative log-likelihood function is equivalent to the standard normal prior: $\log \mathcal{N}(\theta|0,1) \propto \|\theta\|^2$. Although the model $p(x\|\theta)$ and the prior $\varphi(\theta)$ characterize the singularity of the model, their contributions are significantly different. The scale of the effect from the model is larger than that from the prior. Typically, the average error function is determined by the model and its RLCT rules the behavior of the free energy and the generalization error (see Chap. 3).

Q. 6 **How do we derive a non-trivial upper bound of the RLCT?**

(a) We find an upper bound $U(\theta)$ of the average error function $K(\theta)$. It is desired that the RLCT of the bounding function $U(\theta)$ can be easily calculated. If we can determine the RLCT $\overline{\lambda}$ of $U(\theta)$, we obtain an upper bound of the RLCT λ of $K(\theta)$ by using Proposition 3.2. This proposition means that the RLCT is order isomorphic; thus, $K(\theta) \leqq U(\theta) \Rightarrow \lambda \leqq \overline{\lambda}$ is derived. The tighter $U(\theta)$ is, the tighter $\overline{\lambda}$ is. Let d be the parameter dimension. If the prior is strictly positive and bounded, a trivial bounding function is $U(\theta) \propto \|\theta\|^2 = \sum_{i=1}^{d} \theta_i^2$: $\overline{\lambda} = d/2$. If the model is regular, $\lambda = d/2$. Thus, if the found bound satisfies $\overline{\lambda} > d/2$, the upper bound is vacuous.

A.2 Main Results

Q. 1 **In the work of NMF, where does the non-negative restriction essentially affect the main result?**

(a) The non-negative restriction is used to prove Lemmas 4.2, 4.3 and 4.4. In partic-

ular, when the true non-negative rank is zero ($H_0 = 0$, Lemma 4.2), the RLCT of NMF is strictly larger than that of non-restricted matrix factorization. By using these lemmas, the upper bound is derived for the general case (Theorem 4.2), and the bound immediately provides the upper bound of the Bayesian generalization error and the free energy in NMF (Theorem 4.3).

Q. 2 **In the work of NMF, what is the novel point of this research comparing with the prior study by Kohjima [49]?**

(a) This work is a theoretical analysis for inference methods different from [49]. Kohjima's previous work [49] has determined the exact learning coefficient of the variational free energy in NMF (see also Theorem 4.1). In this dissertation, the author treats Bayesian NMF and evaluates the Bayesian generalization error and the free energy in NMF by analyzing the RLCT. The RLCT rules the asymptotic behavior of the Bayesian generalization error and the free energy. On the other hand, the RLCT is not equal to (is smaller than) the learning coefficient of the variational free energy and the coefficient does not dominate the variational generalization error. Besides, with merging this study and Kohjima's result, the variational approximation error is also theoretically evaluated: a lower bound of it is derived (Theorem 4.4).

Q. 3 **In the work of NMF, how do we understand the fact that the variational posterior is essentially different from the true one, in particular when we carry out real data analysis (like making a predictive model)?**

(a) Let $\theta = (\theta_1, \theta_2)$ be a parameter of the model. One might think that the variational posterior $\psi_1(\theta_1)\psi_2(\theta_2)$ is equivalent to the true one $\psi(\theta_1, \theta_2)$, i.e. variational inference is faster and equally accurate with comparing to MCMC (full-Bayesian inference). However, this is false if the parameters θ_1 and θ_2 are not independent. Thus, if we want to realize the posterior, then the variational inference can be inappropriate (MCMC is recommended). If we want a fast algorithm rather than an accurate one, then the variational inference can be helpful but it cannot realize the true posterior.

Q. 4 **In the work of NMF, when $M = N$, can it be immediately proved that the phase transition lines of Bayesian inference and variational Bayes method are orthogonal?**

(a) Yes. Let l_1 and l_2 be the phase transition lines of Bayesian NMF and VBNMF in Theorem 4.4 (see also Fig. 4.1), respectively. In the (ϕ_U, ϕ_V)-plane, l_1 and l_2 are parameterized as $M\phi_U = N\phi_V$ and $M\phi_U + N\phi_V = (M+N)/2$, respectively. Because of $M = N$, we have $l_1 : \ \phi_V = \phi_U$ and $l_2 : \ \phi_V = -\phi_U + 1$. The slope of l_1 is 1 and that of l_2 is -1; thus, they are orthogonal. Note that they are neither orthogonal nor parallel if M and N ($M \neq N$) are finite but they are formally parallel when either $M \to \infty \ \wedge \ N < \infty$ or $N \to \infty \ \wedge \ M < \infty$.

Q. 5 **In the work of NMF, what order is the term "phase transition"?**

(a) The asymptotic (variational) free energy is dominated by the learning coefficient who has a phase transition line of the hyperparameters (ϕ_U, ϕ_V). On the phase transition lines l_1 and l_2 described above, the learning coefficient is continuous but not differentiable. Hence, this is a second-order phase transition.

Q. 6 **In the work of LDA, why does not the author analyze the variational approximation error by the same way of NMF?**

(a) Actually, the variational free energy in LDA is theoretically analyzed and its learning coefficient is determined [59]. However, the effect of the hyperparameters of the Dirichlet prior distribution has not been clarified. Therefore, at this point, it is difficult to compare the variational free energy and the true one in LDA and to evaluate the variational approximation error. The author considers that this is an important future task (see Sec. 6.2).

Q. 7 **In the work of LDA, the main result immediately derives at least how much Bayesian inference improves the generalization performance compared to maximum a posteriori (MAP) method and maximum likelihood (ML) method. On the other hand, what about the computational cost?**

(a) In general, MAP and ML methods suffer from non-identifiability and hardly reach the global minima for singular models. They are numerically unstable and hard to converge. In this sense, their computational cost is not low although they carry out point estimations. Moreover, like Fig. 5.2a, the gap between half of the parameter dimension $d/2$ and the RLCT λ is large, and $d/2 - \lambda$ gives a lower bound of the difference of the generalization errors of MAP (or ML) and Bayesian inference. Thus, even if the computational cost is reduced by avoiding MCMC and using MAP method, the generalization error will be so large that it is not worth the advantage of reducing the cost.

Q. 8 **Both the work of NMF and LDA, to prove the main results, do the prior distributions have to be set to specific distributions (gamma and Dirichlet)?**

(a) For the main theorem of NMF, the distribution form is important; however, it is not limited to the gamma distribution. The gamma distribution $\mathrm{Gam}(w|\phi,\theta) = \frac{\theta^{\phi}}{\Gamma(\theta)} w^{\phi-1} e^{-\theta w}$ can become zero or unbounded because of the power term $w^{\phi-1}$ and this term affects the upper bound (see Theorem 4.2). Therefore, if the prior has the same power term and the other terms are positive and bounded, the main theorem holds.

(b) For the main result of LDA, the prior must not affect the RLCT; thus, it must be positive and bounded. In other words, if the prior is positive and bounded, the distribution form of it is not limited to the uniform Dirichlet distribution (like the Gaussian distribution on the simplex).

Q. 9 **What is the correspondence between the author's publications and the theorems in this dissertation?**

(a) The author's peer-reviewed publications which construct this dissertation are as follows: [37], [36], [33], [38] and [34] (see also Bibliography and List of Publications).

(b) The two journal papers [37, 33] and one international conference paper [36] correspond to the study of NMF (Chap. 4). The paper [37] is the first study for the RLCT of NMF. In [37], Lemmas 4.2 and 4.3 was proved in the case $\phi_U = \phi_V = 1$. Besides, by using these lemmas, an upper bound of the RLCT was also derived. In the paper [36], Lemma 4.4 and a tighter upper bound was obtained and the bound is equal to the upper bound in Theorem 4.2 in the case $\phi_U = \phi_V = 1$. In [33], the effect of the hyperparameter was clarified, i.e. the case $\phi_U > 0$ and $\phi_V > 0$ was treated. Moreover, a lower bound of the variational approximation error in NMF was also derived (Theorem 4.4. Rigorously speaking, [33] shows a

looser upper bound of the RLCT than that in this dissertation because it could not be proved that Lemma 4.4 for general hyperparameter $\phi_U > 0$ and $\phi_V > 0$ at that time. In fact, Lemma 4.4 is true in that case; thus, Theorem 4.2 is proved in this dissertation.

(c) The two journal papers [38, 34] are corresponding to the work of LDA (Chap. 5). The paper [38] is the first paper for the RLCT of LDA. In [38], it was proved that the RLCT of stochastic matrix factorization is equal to that of LDA (Theorem 5.1. By using this relation, an upper bound of the RLCT of LDA was also derived. In the paper [34], the exact value of the RLCT of LDA was determined (Theorem 5.2).

Acknowledgement

I would like to express my great appreciation for the following help, supports, and encouragement. I had carried out this research under the laboratory of Prof. Sumio Watanabe, Department of Mathematical and Computing Science, Tokyo Institute of Technology. Prof. Watanabe taught me the basics of statistical learning theory and algebraic geometry and has widely and profoundly supervised me studying since I was a master's student in 2016. I learned the illustriousness of investigating the nature and establishing mathematical theory from Prof. Watanabe. From research to real life and career counseling, Prof. Watanabe encouraged me with gentle words. Some of the travels and papers for this research were partially supported by the Ministry of Education, Science, Sports and Culture in Japan, Grant-in-Aid for Scientific Research 15K00331 from Prof. Watanabe. Besides, this study was also partially supported by the expenses of NTT DATA Mathematical Systems Inc. where I have worked as a data scientist. I thank them for their patience. Former and present members of Watanabe Group, Ph.D. Fumito Nakamura, Ph.D. Masahiro Kohjima, Ph.D. Natsuki Kariya, M.S. Natsuki Kohda, M.Eng. Toshihiko Kondo, M.S. Shoji Sugai, M.Eng. Rito Takeuchi, M.S. Hikaru Fujimura, M.S. Kohya Ohashi, M.S. Takashi Okano, M.S. Satoshi Kataoka, M.S. Kenichiro Sato, M.S. Raiki Tanaka, M.S. Atsuyoshi Muta, M.S. Shuya Nagayasu, M.S. Toshiki Mikami, M.S. Kento Yoshimura, B.S. Takumi Watanabe, B.S. Tatsuyuki Baba, B.S. Rui Hirose, B.S. Satoshi Nakagawa, B.S. Joe Hirose, B.S. Nozomi Maki, and B.S. Ko Takemura (degree, grade, and alphabetical order), lively discussed this research and other science topics with me. Prof. Mathias Drton, Department of Mathematics, Technical University of Munich, kindly answered my questions about sBIC and encouraged this research. Prof. Miki Aoyagi, Department of Mathematics, Nihon University, taught me in detail about the Vandermonde matrix type singularity theory, which is one of the most advanced topics in singular learning theory. Prof. Kota Matsui, Department of Biostatistics, Nagoya University, invited me to StatsML Symposium'20. I was blessed with the valuable opportunity having discussed and introduced the singular learning theory and the contents of my research in that invited talk. Prof. Takafumi Kanamori, Prof. Misako Takayasu, Prof. Makoto Yamashita, and Prof. Hanna Sumita carefully reviewed this dissertation. Lastly, I greatly appreciate my parents for warmly raising and supporting me, my friends for kindly encouraging each other, and all the other people who took care of me though I cannot write in this chapter.

Acknowledgement (in Japanese)

本研究は様々な支援を受けました．指導教官である渡邊澄夫教授からは統計的学習理論及び代数幾何学・特異点論の基礎から具体的かつ詳細な研究指導に至るまで広く深くご教授いただき，実世界を探索し数理科学の理論を構築することの尊さを学ぶことができました．修士課程から博士課程まで，途中就職により研究室に在籍していないときでも，渡邊澄夫教授は筆者の研究相談に乗ってくださいました．研究から実生活・進路相談に至るまで，渡邊澄夫教授は優しい言葉をかけて私を励ましてくださいました．一部の論文投稿や学会のための渡航などについても，渡邊澄夫教授から科学研究費補助金 15K00331 の援助を受けました．また，筆者が所属する株式会社 NTT データ数理システムの経費も，一部の論文投稿や学会参加の費用として用いました．研究室で時を同じくして過ごした中村文士博士，幸島匡宏博士，仮屋夏樹博士，香田夏輝修士，近藤稔彦修士，須貝将士修士，竹内理人修士，大橋耕也修士，岡野高志修士，藤村光修士，片岡論史修士，佐藤件一郎修士，田中来輝修士，永安修也修士，牟田篤兄修士，三上敬生修士，吉村健斗修士，渡邉匠学士，中川哲学士，馬場達之学士，廣瀬瑠伊学士，竹村航学士，広瀬青学士，槇望学士（学位・学年・あいうえお順）は様々なディスカッションに付き合ってくださいました．ミュンヘン工科大学の Mathias Drton 教授は sBIC に関する筆者の拙い質問に丁寧に答えてくださり，本研究を応援してくださいました．日本大学の青柳美樹教授からは特異学習理論における最先端理論である Vandermonde 行列型特異点論について詳細にご教示いただきました．名古屋大学の松井孝太講師からは「第 5 回統計・機械学習若手シンポジウム」に招待いただき，特異学習理論と本研究内容を紹介してディスカッションを行う貴重な機会が得られました．金森敬文教授，高安美佐子教授，山下真教授，澄田範奈講師には本博士論文を審査いただきました．最後に，筆者を育ててくださった両親，博士課程期間中に励まし合うことができた友人たち，その他本章に書き記しきれないお世話になりましたすべての方々に感謝します．

Bibliography

[1] Christopher P. Adams. Finite mixture models with one exclusion restriction. *The Econometrics Journal*, Vol. 19, No. 2, pp. 150–165, 2016.

[2] Christopher P. Adams. Stochastic matrix factorization. *SSRN Electronic Journal*, pp. 1–24, 2016. Available at SSRN: https://ssrn.com/abstract=2840852.

[3] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pp. 199–213. Springer, 1998.

[4] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, Vol. 19, No. 6, pp. 716–723, 1974.

[5] Miki Aoyagi. A bayesian learning coefficient of generalization error and vandermonde matrix-type singularities. Communications in Statistics—Theory and Methods, Vol. 39, No. 15, pp. 2667–2687, 2010.

[6] Miki Aoyagi. Stochastic complexity and generalization error of a restricted boltzmann machine in bayesian estimation. *Journal of Machine Learning Research*, Vol. 11, No. Apr, pp. 1243–1272, 2010.

[7] Miki Aoyagi. Learning coefficient in bayesian estimation of restricted boltzmann machine. *Journal of Algebraic Statistics*, Vol. 4, No. 1, pp. 30–57, 2013.

[8] Miki Aoyagi. Learning coefficient of vandermonde matrix-type singularities in model selection. *Entropy*, Vol. 21, No. 6, p. 561, 2019.

[9] Miki Aoyagi and Kenji Nagata. Learning coefficient of generalization error in bayesian estimation and vandermonde matrix-type singularity. *Neural Computation*, Vol. 24, No. 6, pp. 1569–1610, 2012.

[10] Miki Aoyagi and Sumio Watanabe. Stochastic complexities of reduced rank regression in bayesian estimation. *Neural Networks*, Vol. 18, No. 7, pp. 924–933, 2005.

[11] Michael Francis Atiyah. Resolution of singularities and division of distributions. *Communications on pure and applied mathematics*, Vol. 23, No. 2, pp. 145–150, 1970.

[12] Joseph Bernstein. The analytic continuation of generalized functions with respect to a parameter. *Funktsional'nyi Analiz i ego Prilozheniya*, Vol. 6, No. 4, pp. 26–40, 1972.

[13] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, Vol. 59, No. 1, pp. 65–98, 2017.

[14] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[15] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022, 2003.

[16] Jesús Bobadilla, Rodolfo Bojorque, Antonio Hernando Esteban, and Remigio Hurtado. Recommender systems clustering using bayesian non negative matrix factorization. *IEEE Access*, Vol. 6, pp. 3549–3564, 2018.

[17] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell.

Stan: a probabilistic programming language. *Grantee Submission*, Vol. 76, No. 1, pp. 1–32, 2017.

[18] Ali T. Cemgil. Bayesian inference in non-negative matrix factorisation models. *Computational Intelligence and Neuroscience*, Vol. 2009, No. 4, p. 17, 2009. Article ID 785152.

[19] Joel E. Cohen and Uriel G. Rothblum. Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. *Linear Algebra and Its Applications*, Vol. 190, pp. 149–168, 1993.

[20] Peter D Congdon. *Bayesian hierarchical models: with applications using R*. CRC Press, 2019.

[21] Dingxiong Deng, Cyrus Shahabi, Ugur Demiryurek, Linhong Zhu, Rose Yu, and Yan Liu. Latent space model for road networks to predict time-varying traffic. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1525–1534, 2016.

[22] Chris HQ Ding, Tao Li, and Michael I Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 32, No. 1, pp. 45–55, 2008.

[23] Mathias Drton, Shaowei Lin, Luca Weihs, Piotr Zwiernik, et al. Marginal likelihood and model selection for gaussian latent tree and forest models. *Bernoulli*, Vol. 23, No. 2, pp. 1202–1232, 2017.

[24] Mathias Drton and Martyn Plummer. A bayesian information criterion for singular models. *Journal of the Royal Statistical Society Series B*, Vol. 79, pp. 323–380, 2017. with discussion.

[25] Ilenia Epifani, Steven N MacEachern, Mario Peruggia, et al. Case-deletion importance sampling estimators: Central limit theorems and related results. *Electronic Journal of Statistics*, Vol. 2, pp. 774–806, 2008.

[26] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural computation*, Vol. 21, No. 3, pp. 793–830, 2009.

[27] Lorenzo Finesso and Peter Spreij. Nonnegative matrix factorization and i-divergence alternating minimization. *Linear Algebra and its Applications*, Vol. 416, No. 2-3, pp. 270–287, 2006.

[28] Andrew Gelman, Daniel Lee, and Jiqiang Guo. Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, Vol. 40, No. 5, pp. 530–543, 2015.

[29] Daniel Gildea and Thomas Hofmann. Topic-based language models using em. In *Sixth European Conference on Speech Communication and Technology*, 1999.

[30] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, Vol. 101, No. suppl 1, pp. 5228–5235, 2004.

[31] Zellig S. Harris. Distributional structure. *WORD*, Vol. 10, No. 2-3, pp. 146–162, 1954.

[32] JA Hartigan. A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley Conference in Honor of J. Neyman and J. Kiefer, 1985*, pp. 807–810, 1985.

[33] Naoki Hayashi. Variational approximation error in non-negative matrix factorization. *Neural Networks*, Vol. 126, pp. 65–75, 2020.

[34] Naoki Hayashi. The exact asymptotic form of bayesian generalization error in latent

dirichlet allocation. *Neural Networks*, Vol. 137, pp. 127–137, 2021.

[35] Naoki Hayashi and Fumito Nakamura. Experimental analysis of variational bayesian method in model selection of gaussian mixture model by singular bayesian information criterion. In *Information-Based Induction Sciences and Machine Learning (IBISML), IEICE Technical Report*, Vol. 117, pp. 19–26, 9 2017. in Japanese.

[36] Naoki Hayashi and Sumio Watanabe. Tighter upper bound of real log canonical threshold of non-negative matrix factorization and its application to bayesian inference. In *IEEE Symposium Series on Computational Intelligence (IEEE SSCI)*, pp. 718–725, 11 2017.

[37] Naoki Hayashi and Sumio Watanabe. Upper bound of bayesian generalization error in non-negative matrix factorization. *Neurocomputing*, Vol. 266C, No. 29 November, pp. 21–28, 2017.

[38] Naoki Hayashi and Sumio Watanabe. Asymptotic bayesian generalization error in latent dirichlet allocation and stochastic matrix factorization. *SN Computer Science*, Vol. 1, No. 2, pp. 1–22, 2020.

[39] Heisuke Hironaka. Resolution of singularities of an algbraic variety over a field of characteristic zero. *Annals of Mathematics*, Vol. 79, pp. 109–326, 1964.

[40] N Thompson Hobbs and Mevin B Hooten. *Bayesian models: a statistical primer for ecologists*. Princeton University Press, 2015.

[41] Matthew D Hoffman, David M Blei, and Perry R Cook. Bayesian nonparametric matrix factorization for recorded music. In *ICML*, 2010.

[42] Tikara Hoshino, Kazuho Watanabe, and Sumio Watanabe. Stochastic complexity of variational bayesian hidden markov models. In *International Joint Conference on Neural Networks*, Vol. 2, pp. 1114–1119 vol. 2, July 2005.

[43] Toru Imai. Estimating real log canonical thresholds. *arXiv preprint arXiv:1906.01341*, 2019.

[44] Toru Imai. On the overestimation of widely applicable bayesian information criterion. *arXiv preprint arXiv:1908.10572*, 2019.

[45] F. Itakura and S. Saito. Analysis synthesis telephony based on the maximum likelihood method. In *In Proc. 6th of the International Congress on Acoustics*, 1968.

[46] Hyunsoo Kim and Haesun Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, Vol. 23, No. 12, pp. 1495–1502, 2007. doi:10.1093/bioinformatics/btm134. PMID 17483501.

[47] Masahiro Kohjima, Tatsushi Matsubayashi, and Hiroshi Sawada. Probabilistic non-negative inconsistent-resolution matrices factorization. In *Proceeding of CIKM '15 Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, Vol. 1, pp. 1855–1858, 2015.

[48] Masahiro Kohjima, Tatsushi Matsubayashi, and Hiroshi Sawada. Multiple data analysis and non-negative matrix/tensor factorization [i]: multiple data analysis and its advances. *The journal of the Institute of Electronics, Information and Communication Engineers (IEICE)*, Vol. 99, No. 6, pp. 543–550, 2016. in Japanese.

[49] Masahiro Kohjima and Sumio Watanabe. Phase transition structure of variational bayesian nonnegative matrix factorization. In *International Conference on Artificial Neural Networks*, pp. 146–154. Springer, 2017.

[50] Ben Lambert. A student' s guide to Bayesian statistics. Sage, 2018.

[51] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature*, Vol. 401, pp. 788–791, 1999.

[52] Fei-Fei Li and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, CVPR '05, pp. 524–531, Washington, DC, USA, 2005. IEEE Computer Society.

[53] Ken Matsuda and Sumio Watanabe. Weighted blowup and its application to a mixture of multinomial distributions. *IEICE Transactions*, Vol. J86-A, No. 3, pp. 278–287, 2003. in Japanese.

[54] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press, 2016.

[55] Daniel Murfet, Susan Wei, Mingming Gong, Hui Li, Jesse Gell-Redman, and Thomas Quella. Deep learning is singular, and that's good. *arXiv preprint arXiv:2010.11560*, 2020.

[56] Kenji Nagata and Sumio Watanabe. Asymptotic behavior of exchange ratio in exchange monte carlo method. *Neural Networks*, Vol. 21, No. 7, pp. 980–988, 2008.

[57] Shuya Nagayasu and Sumio Watanabe. Asymptotic behavior of free energy when optimal probability distribution is not unique. *arXiv preprint arXiv:2012.08338*, 2020.

[58] Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, Vol. 21, No. 174, pp. 1–38, 2020.

[59] Shinichi Nakajima, Issei Sato, Masashi Sugiyama, Kazuho Watanabe, and Hiroko Kobayashi. Analysis of variational bayesian latent dirichlet allocation: Weaker sparsity than map. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, Vol. 27, pp. 1224–1232. Curran Associates, Inc., 2014.

[60] Shinichi Nakajima and Sumio Watanabe. Variational bayes solution of linear neural networks and its generalization performance. *Neural Computation*, Vol. 19, No. 4, pp. 1112–1153, 2007.

[61] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, Vol. 5, No. 2, pp. 111–126, 1994. doi:10.1002/env.3170050203.

[62] Dmitry Rusakov and Dan Geiger. Asymptotic model selection for naive bayesian networks. *Journal of Machine Learning Research*, Vol. 6, No. Jan, pp. 1–35, 2005.

[63] Ivan N Sanov. *On the probability of large deviations of random variables*. United States Air Force, Office of Scientific Research, 1958.

[64] Kenichiro Sato and Sumio Watanabe. Bayesian generalization error of poisson mixture and simplex vandermonde matrix type singularity. *arXiv preprint arXiv:1912.13289*, 2019.

[65] Mikio Sato and Takuro Shintani. On zeta functions associated with prehomogeneous vector spaces. *Annals of Mathematics*, pp. 131–170, 1974.

[66] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, Vol. 6, No. 2, pp. 461–464, 1978.

[67] Paris Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *International Conference on Independent Component Analysis and Signal Separation*, pp. 494–499. Springer, 2004.

[68] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, Vol. 27, No. 4, p. 521âĂŞ544, December 2001.

[69] Sayaka Suzuki, Souta Shina, and Miki Aoyagi. The improving method of singular bayesian information criterion by analyzing learning coefficients. In *Information-Based Induction Sciences and Machine Learning (IBISML), IEICE Technical Report*, Vol. 117, pp. 71–76, 3 2018. in Japanese.

[70] Kei Takeuchi. Jouhou-toukeiryou no bunpu to model no tekisetsu-sa no kijun. *Suurikagaku*, Vol. 14, No. 3, pp. 12–18, 1976. in Japanese.

[71] Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual*, 2.26 edition, 2019. https://mc-stan.org.

[72] Seshadri Tirunillai and Gerard J Tellis. Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, Vol. 51, No. 4, pp. 463–479, 2014.

[73] M Antónia Amaral Turkman, Carlos Daniel Paulino, and Peter Müller. *Computational Bayesian Statistics: An Introduction*, Vol. 11. Cambridge University Press, 2019.

[74] G. Tusnady, I. Csiszar, and G. Tusnády. Information geometry and alternating minimization procedures. *Statistics and Decisions: Supplement Issues*, Vol. 1, pp. 205–237, 1984.

[75] Tuomas Virtanen, A Taylan Cemgil, and Simon Godsill. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 1825–1828. IEEE, 2008.

[76] Kazuho Watanabe. An alternative view of variational bayes and asymptotic approximations of free energy. *Machine learning*, Vol. 86, No. 2, pp. 273–293, 2012.

[77] Kazuho Watanabe and Sumio Watanabe. Upper bounds of bayesian generalization errors in reduced rank regression. *IEICE Transactions*, Vol. J86-A, No. 3, pp. 278–287, 2003. in Japanese.

[78] Kazuho Watanabe and Sumio Watanabe. Stochastic complexities of gaussian mixtures in variational bayesian approximation. *Journal of Machine Learning Research*, Vol. 7, No. Apr, pp. 625–644, 2006.

[79] Sumio Watanabe. Algebraic analysis for non-regular learning machines. *Advances in Neural Information Processing Systems*, Vol. 12, pp. 356–362, 2000. Denver, USA.

[80] Sumio Watanabe. Algebraic geometrical methods for hierarchical learning machines. *Neural Networks*, Vol. 13, No. 4, pp. 1049–1060, 2001.

[81] Sumio Watanabe. *Algebraix Geometry and Statistical Learning Theory*. Cambridge University Press, 2009.

[82] Sumio Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, Vol. 11, No. Dec, pp. 3571–3594, 2010.

[83] Sumio Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, Vol. 14, No. Mar, pp. 867–897, 2013.

[84] Sumio Watanabe. Difference between bayes cross validation and waic for conditional independent samples. In *Tenth Workshop on Information Theoretic Methods in Science and Engineering (WITMSE2017)*, pp. 38–41, 2017.

[85] Sumio Watanabe. *Mathematical theory of Bayesian statistics*. CRC Press, 2018.

[86] Sumio Watanabe. Information criteria and cross validation for bayesian inference in regular and singular cases. *Japanese Journal of Statistics and Data Science*, pp. 1–19, 2021.

[87] Sumio Watanabe. Waic and wbic for mixture models. *Behaviormetrika*, pp. 1–17, 2021.

[88] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. Association for Computing Machinery*, pp. 267–273, 2003.

[89] Keisuke Yamazaki. *Asymptotic Expansion of Stochastic Complexities in Singular Learning Machines*. PhD thesis, Tokyo Institute of Technology, 2003.

[90] Keisuke Yamazaki. Asymptotic accuracy of bayes estimation for latent variables with redundancy. *Machine Learning*, Vol. 102, No. 1, pp. 1–28, 2016.

[91] Keisuke Yamazaki. Bayesian estimation of multidimensional latent variables and its asymptotic accuracy. *Neural Networks*, Vol. 105, pp. 14–25, 2018.

[92] Keisuke Yamazaki and Daisuke Kaji. Comparing two bayes methods based on the free energy functions in bernoulli mixtures. *Neural Networks*, Vol. 44, pp. 36–43, 2013.

[93] Keisuke Yamazaki and Yoichi Motomura. Hidden node detection between observable nodes based on bayesian clustering. *Entropy*, Vol. 21, No. 1, p. 32, 2019.

[94] Keisuke Yamazaki and Sumio Watanabe. Singularities in mixture models and upper bounds of stochastic complexity. *Neural Networks*, Vol. 16, No. 7, pp. 1029–1038, 2003.

[95] Keisuke Yamazaki and Sumio Watanabe. Stochastic complexity of bayesian networks. In *Uncertainty in Artificial Intelligence (UAI'03)*, 2003.

[96] Keisuke Yamazaki and Sumio Watanabe. Algebraic geometry and stochastic complexity of hidden markov models. *Neurocomputing*, Vol. 69, pp. 62–84, 2005. issue 1-3.

[97] Keisuke Yamazaki and Sumio Watanabe. Singularities in complete bipartite graph-type boltzmann machines and upper bounds of stochastic complexities. *IEEE Transactions on Neural Networks*, Vol. 16, pp. 312–324, 2005. issue 2.

[98] Kenta Yoshida, Tatsu Kuwatani, Takao Hirajima, Hikaru Iwamori, and Shotaro Akaho. Progressive evolution of whole-rock composition during metamorphism revealed by multivariate statistical analyses. *Journal of Metamorphic Geology*, Vol. 36, No. 1, pp. 41–54, 2018.

[99] Elias Zavitsanos, Georgios Paliouras, and George A Vouros. Non-parametric estimation of topic hierarchies from texts with hierarchical dirichlet processes. *Journal of Machine Learning Research*, Vol. 12, No. 10, 2011.

[100] Piotr Zwiernik. An asymptotic behaviour of the marginal likelihood for general markov models. *Journal of Machine Learning Research*, Vol. 12, No. Nov, pp. 3283–3310, 2011.

List of Publications

Peer-reviewed Journal Papers

1. Naoki Hayashi, Sumio Watanabe. "Upper Bound of Bayesian Generalization Error in Non-Negative Matrix Factorization", Neurocomputing, Volume 266C, 29 November 2017, pp.21-28. doi: 10.1016/j.neucom.2017.04.068.
2. Naoki Hayashi, Sumio Watanabe. "Asymptotic Bayesian Generalization Error in Latent Dirichlet Allocation and Stochastic Matrix Factorization", SN Computer Science, Volume 1, 69 (2020), pp.1-22. doi: 10.1007/s42979-020-0071-3.
3. Naoki Hayashi. "Variational Approximation Error in Non-negative Matrix Factorization", Neural Networks, Volume 126, June 2020, pp.65-75. doi: 10.1016/j.neunet.2020.03.009.
4. Keita Harada, Naoki Hayashi, Katsushi Kagaya. "Individual behavioral type captured by a Bayesian model comparison of cap making by sponge crabs", PeerJ 8:e9036, pp.1-26. doi: 10.7717/peerj.9036.
5. Naoki Hayashi. "The Exact Asymptotic Form of Bayesian Generalization Error in Latent Dirichlet Allocation". Neural Networks, Volume 137, May 2021, pp.127-137. doi: 10.1016/j.neunet.2021.01.024.

Peer-reviewed International Conferences

1. Naoki Hayashi, Sumio Watanabe."Tighter upper bound of real log canonical threshold of non-negative matrix factorization and its application to Bayesian inference," 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, 2017, pp.1-8, doi: 10.1109/SSCI.2017.8280811.
2. Akira Ito, Masaru Okutsu, Masatoshi Yukishima, Ryosuke Matsushita, Naoki Hayashi, Aiko Furukawa, Gaku Shoji, Takanobu Suzuki. "Trial of damage prediction for telecommunication conduits using machine learning". 17 World Conference on Earthquake Engineering (17 WCEE), Sendai, Miyagi, 2020, pp.1-10, 6a-0011.

Not-peer-reviewed International Conferences

1. Naoki Hayashi. "Bayesian Generalization Error and Real Log Canonical Threshold in Non-negative Matrix Factorization and Latent Dirichlet Allocation". Algebraic Statistics 2020, Online (virtual mini conference). June. 22 - 26, 2020.

Not-peer-reviewed Domestic Conferences (in Japanese)

1. 林直輝, 渡邊澄夫. "非負値行列分解の実対数閾値と Bayes 学習への応用", 第 19 回情報論的学習理論ワークショップ (IBIS2016), 京都, 信学技報, Vol.116, No.300, pp.215-220. (2016/11/17 発表).
   - (ENG) Naoki Hayashi, Sumio Watanabe."A Real Log Canonical Threshold of Nonnegative Matrix Factorization and Its Application to Bayesian Learning", IBIS2016, IEICE Technical Report Vol.116, No.300, pp.215-220. (2016/11/17).
2. 林直輝, 渡邊澄夫. "非負値行列分解における実対数閾値の実験的考察", ニューロコンピューティング研究会 (NC), 東京, 信学技報, Vol.116, No.521, pp.85-90. (2017/3/13 発表).
   - (ENG) Naoki Hayashi, Sumio Watanabe. "Experimental Analysis of Real Log Canonical Threshold in Non-negative Matrix Factorization", Neurocomputing (NC), IEICE Technical Report, Vol.116, No.521, pp.85-90. (2017/3/13).
3. 林直輝, 中村文士. "特異 Bayes 情報量規準による混合正規分布のモデル選択における変分 Bayes 法の実験的考察", 情報論的学習理論と機械学習（IBISML）, 東京, 信学技報, Vol.117, No.211, pp.19-26. (2017/9/15 発表).
   - (ENG) Naoki Hayashi, Fumito Nakamura. "Experimental Analysis of Variational Bayesian Method in Model Selection of Gaussian Mixture Model by Singular Bayesian Information Criterion", Information-Based Induction Sciences and Machine Learning (IBISML), IEICE Technical Report, Vol.117, No.211, pp.19-26. (2017/9/15).
4. 林直輝, 渡邊澄夫. "確率行列分解の実対数閾値と Bayes 学習への応用", 第 20 回情報論的学習理論ワークショップ (IBIS2017), 東京, 信学技報, Vol.117, No.293, pp.23-30. (2017/11/9 発表).
   - (ENG) Naoki Hayashi, Sumio Watanabe. "Real Log Canonical Threshold of Stochastic Matrix Factorization and its Application to Bayesian Learning", IBIS2017, IEICE Technical Report, Vol.117, No.293, pp.23-30. (2017/11/9).
5. 林直輝, 渡邊澄夫. "ハミルトニアンモンテカルロ法を用いた確率行列分解における実対数閾値の実験的考察", ニューロコンピューティング研究会 (NC), 東京. 信学技報, vol. 117, no. 508, NC2017-89, pp. 127-131. (2018/3/14 発表).
   - (ENG) Naoki Hayashi, Sumio Watanabe. "Experimental Analysis of Real Log Canonical Threshold in Stochastic Matrix Factorization using Hamiltonian Monte Carlo Method", Neurocomputing (NC), Tokyo. IEICE Technical Report, vol. 117, no. 508, NC2017-89, pp. 127-131. (2018/3/14).
6. 林直輝. "非負値行列分解における変分近似精度の理論解析", 第 21 回情報論的学習理論ワークショップ (IBIS2018), 札幌. 信学技報, vol. 118, no. 284, IBISML2018-51, pp. 53-60. (2018/11/5 発表).
   - (ENG) Naoki Hayashi. "Variational Approximation Accuracy in Non-negative Matrix Factorization", IBIS2018, Sapporo. IEICE Technical Report, vol. 118, no. 284, IBISML2018-51, pp. 53-60. (2018/11/5).
7. 西郷彰, 林直輝, 伊藤孝太朗. "YOLOv3 とドメイン知識を用いた CT 画像の病変部位検出", 第 33 回日本人工知能学会全国大会 (JSAI2019), 新潟. (2019/6/5 発表).
   - (ENG) Akira Saigo, Naoki Hayashi, Kotaro Ito. "Lesion Detection in Computed

Tomography Images using YOLOv3 and Domain Knowledge", The 33rd Annual Conference of the Japanese Society for Artificial Intelligence 2019(JSAI2019), Niigata. (2019/6/5).

8. 林直輝．"LDA における汎化誤差の厳密な漸近形"，第 23 回情報論的学習理論ワークショップ (IBIS2020), オンライン. (2020/11/26 発表). 動画リンク：https://www.youtube.com/watch?v=e8t3ZGluc6U. 優秀発表賞ファイナリスト.
    - (ENG) Naoki Hayashi. "The Exact Asymptotic Form of Generalization Error in LDA", IBIS2020, Online. (2020/11/26). Outstanding Presentation Award Finalist.

Domestic Invited Talks (in Japanese)

1. 林直輝．"パラメータ制約付き行列分解のベイズ汎化誤差解析"，第 5 回統計・機械学習若手シンポジウム (StatsML Symposium'20), オンライン. (2020/12/5 発表). 動画リンク：https://www.youtube.com/watch?v=ZbVfah9pnb4.
    - (ENG) Naoki Hayashi. "Theoretical Analysis of Bayesian Generalization Error in Parameter-restricted Matrix Factorization", StatsML Symposium'20, Online. (2020/12/5).

Awards

1. Outstanding Presentation Award Finalist at IBIS2020 (2020/11/26 published, 2021/1/8 gotten).