論文 / 著書情報
Article / Book Information

題目(和文)	統計的学習における汎化損失の不偏推定量の研究
Title(English)	Unbiased estimator of generalization loss in statistical learning : theory and practice
著者(和文)	山田耕史
Author(English)	Koshi Yamada
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9047号, 授与年月日:2013年3月26日, 学位の種別:課程博士, 審査員:渡邊 澄夫
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第9047号, Conferred date:2013/3/26, Degree Type:Course doctor, Examiner:
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Unbiased Estimator of Generalization Loss in Statistical Learning : Theory and Practice

Koshi Yamada[1]

February 14, 2013

[1]Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Mail box:G5-19, 4259 Nagatsuta, Midori-ku, Yokohama, 226-8502, Japan
E-mail:yamada.k.am@m.titech.ac.jp

Abstract

Statistical learning is now widely used in various fields of science and engineering. It is important in practice to design a statistical model or a learning machine so that it attains good generalization performance, hence it plays a central role in model selection problem to estimate generalization loss from obtained samples. The Akaike information criterion, AIC, is the seminal work for this problem, which provides an asymptotically unbiased estimator of generalization loss under a regularity condition. However, it has been pointed out that many learning machines with hierarchical structure, such as layered neural networks, normal mixture, hidden Markov model, Bayes network, do not satisfy this regularity condition, resulting that maximum likelihood estimator is not subject to normal distribution even asymptotically and Bayes posterior distribution does not converge to normal distibution in law. These learning machines are called non-regular or singular and conventional information criteria, such as AIC, BIC or DIC, do not have theoretical foundation for them.

Recently, a learning theory which also holds for these singular learning machines has been constructed based on algebraic geometrical method. That theory is called singular learning theory. In singular learning theory, it has been clarified that the asymptotic behavior of expected Bayes generalization loss and training loss can be described by two birational invariants, real log canonical threshold and singular fluctuation. Besides, an information criterion, WAIC, has been proposed as an asymptotically unbiased estimator of Bayes generalization loss which can be used even in singular cases.

The goal of this thesis is to study this unbiased estimator of generalization loss in statistical learning from both theoretical and practical viewpoints. From theoretical perspective, we focus on singular fluctuation, which is a birational invariant firstly discovered in singular learning theory and its mathematical properties are left totally unknown. In this thesis, we propose a new concept in statistical learning theory, *quasi-regular case*, in which the exact value of singular fluctuation is firstly clarified for non-regular cases. Quasi-regular case is characterized as a class of singular cases which has similar properties as regular cases. From practical perspective, we propose a method to apply WAIC for model selection problem in variational Bayes learning. Variational Bayes learning gives the accurate statistical estimation as Bayes learning with smaller computational cost. However, it has been difficult to estimate its generalization loss, because learning machines used in variational Bayes are not regular but singular. Then

we propose a new information criterion for variational Bayes learning, WAIC-VB, which is an unbiased estimator of Bayes generalization loss for both cases when the posterior distribution is regular and singular. We show the theoretical support of the proposed information criterion, and its effectiveness is illustrated by numerical experiments.

This thesis consists of six chapters. In chapter 1, we overview the background, purpose and the main contribution of this research. In chapter 2, we introduce the framework of Bayes learning and review the essence of singular learning theory. In chapter 3. we describe the statistical learning theory in quasi-regular cases. In chapter 4, we propose an information criterion WAIC-VB for variational Bayes learning. In chapter 5, we discuss some related topics and left problems for a future study. In chapter 6, we conclude this thesis.

Acknowledgement

This thesis integrates the result of my work during the doctoral course at Watanabe laboratory, Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology. This thesis would not have been possible without the help of many people.

First and foremost, I would like to express my deep gratitude to my supervisor, Professor Sumio Watanabe. His support and encouragement have always been invaluable to me. I have really enjoyed discussion with him on wide range of topics, such as mathematics, physics, statistics, computer science and sometimes social problems. His profound knowledge and insight always gives an inspiration for me.

I would also like to express my gratitude to my thesis committee, Professor Yoshiyuki Kabashima, Professor Osamu Hasegawa, Professor Isao Ono, Professor Toshiaki Murofushi for taking time to read this thesis and giving me valuable comments on my presentation and draft of this thesis.

I would also like to thank present and former members of Watanabe laboratory. Especially, I am very grateful to Assistant Professor Keisuke Yamazaki for his willingness to support me both in academic and daily life. I am also grateful to all other members of Watanabe laboratory, which includes Dr.Shi-ichi Nakajima, Dr.Kazuho Watanabe, Dr.Kenji Nagata, Dr.Yu Nishiyama, Dr.Kaori Fujiwara, Dr.Daisuke Kaji, Dr.Fumiaki Saitoh, Takuto Naito, Shinji Oyama, Masahiro Kohjima, Tetsutaro Yamada, Takashi Kato, Koichi Kobayashi, Takushi Miki, Shigehiro Ohara, Kazumi Dakujaku and Fumito Nakamura. Thank you all for constructive discussion and having enjoyable conversation.

Finally, I wish to thank my parents for their patient support and encouragement throughout my study.

Contents

Chapter 1

Introduction

In this chapter, we briefly introduce the contents of this research. In section 1.1, the background of this research is explained. In section 1.2, the purpose of this research is explained. In section 1.3, the main contribution of this research is summarized before discussing in detail in the subsequent chapters.

1.1 Overview of Background

Statistical learning is now being used in broad engineering fields, such as pattern recognition, bioinformatics, robotic control, artificial intelligence and so on. When samples $X^n := \{X_1, X_2 \cdots X_n\}$ are obtained independent and identically, it is one of the fundamental problems in statistical learning to estimate the probability distribution $q(x)$ from which samples are generated. The probability distribution $q(x)$ is called the true distribution.

Evaluating the result of learning is an important problem from both theoretical and practical points of view. Two concepts are essential in evaluation of the learning result, one is generalization error $G_n^{(0)}$ and the other is training error $T_n^{(0)}$. Generalization error is defined by Kullback-Leibler distance from the true distribution $q(x)$ to the estimated distribution $p(x|X^n)$. Training error is defined by empirical Kullback-Leibler distance from the true distribution $q(x)$ to the estimated distribution $p(x|X^n)$. That is,

$$G_n^{(0)} = \int q(x) \log \frac{q(x)}{p(x|X^n)} dx, \quad T_n^{(0)} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i|X^n)}.$$

In practice, generalization loss G_n and training loss T_n are often used, which are defined by adding entropy terms to generalization and training errors.

$$G_n = -\int q(x) \log p(x|X^n) dx, \quad T_n = -\frac{1}{n} \sum_{i=1}^{n} \log p(X_i|X^n).$$

From the viewpoint of prediction for unknown samples, it can be said that the goal of statistical learning is to construct estimated distribution $p(x|X^n)$ so that generalization error or loss gets smaller. However, we cannot calculate the generalization loss G_n from samples directly, since generalization loss includes the unknown true distribution $q(x)$ in its definition. Hence, if we can estimate the generalization loss only using samples, it would be very useful in evaluation of learning result.

The classical and the first result for this problem was given by Akaike's seminal work [1] in 1974. Let us assume the true distribution is realizable by a statistical model. Then, Akaike showed that, if the model satisfies regularity condition, which are called *regular* model, and we construct $p(x|X^n)$ by maximum likelihood estimation, then the following relation holds.

$$\mathbb{E}[G_n] = \mathbb{E}[T_n] + \frac{d}{n} + o(\frac{1}{n}),$$

where $\mathbb{E}[\ ]$ represents expectation over samples X^n, d is the dimension of the parameter space and n is the number of samples. This result is the theoretical foundation of the famous *information criterion* AIC, which is an asymptotically unbiased estimator of the generalization loss under regularity condition. In fact, under regularity condition, it can be proved that this result also holds when we construct $p(x|X^n)$ by Bayes method or maximum a *posteriori* estimation.

On the other hand, many learning machines used in information engineering fields, such as layered neural networks, normal mixture, hidden Markov model, Bayes network, do not satisfy this regularity condition. That is, the parameter of a learning machine does not correspond to probability distribution one-to-one and Fisher infomation matrix is degenerate in their parameter space. These learning machines or statistical models are called *singular*. For singular learning machines, Akaike's AIC does not have theoretical foundation as an unbiased estimator of generalization loss.

Recently, Watanabe has constructed a learning theory which can be also applied to singular learning machines based on algebraic geometrical method [25, 26, 28, 29]. Watanabe has clarified that, if we construct $p(x|X^n)$ by Bayes method, Bayes generalization loss and Bayes training loss have the following asymptotic expansion.

$$\begin{aligned} \mathbb{E}[G_n] &= S + (\frac{\lambda - \nu}{\beta} + \nu)\frac{1}{n} + o(\frac{1}{n}), \\ \mathbb{E}[T_n] &= S + (\frac{\lambda - \nu}{\beta} - \nu)\frac{1}{n} + o(\frac{1}{n}), \end{aligned}$$

where S is the entropy of the true distribution $q(x)$, β is the inverse temparature and λ and ν are birational invariants called *real log canonical threshold* (RLCT) and *singular fluctuation* (SF), respectively. From this result, the following relation holds,

$$\mathbb{E}[G_n] = \mathbb{E}[T_n] + \frac{2\nu}{n} + o(\frac{1}{n}).$$

Then it is shown that the functional variance V_n defined by

$$V_n = \sum_{i=1}^{n} \{\mathbb{E}_w[(\log p(X_i|w))^2] - \mathbb{E}_w[\log p(X_i|w)]^2\}$$

can be calculated only from samples and satisfies

$$\beta \mathbb{E}[V_n] \to 2\nu.$$

Hence the following relation holds,

$$\mathbb{E}[G_n] = \mathbb{E}[T_n + \frac{\beta V_n}{n}] + o(\frac{1}{n}).$$

This asymptotically unbiased estimator of G_n has been proposed as an information criterion WAIC (A Widely Applicable Information Criterion) and could be considered as a generalized AIC for singular learning machines. It should be emphasized that Watanabe's theory is not a specialized theory for singular learning machines but is a generalized theory which even holds for regular models. That is, for regular models, $\lambda = \nu = \frac{d}{2}$ and $\beta \mathbb{E}[V_n] \to d$ hold as a special case. (In fact, βV_n converges to d in probability in regular cases.)

1.2 Purpose of the research

The purpose of this research is to investigate the property of unbiased estimator of Bayes generalization loss from both theoretical and practical perspective.

Firstly, we explain the theoretical purpose of research. In section 1.1, we stated that asymptotic behavior of Bayes generalization and training losses could be described by two birational invariants, λ and ν. In particular, the asymptotic difference between two losses was determined by singular fluctuation ν. In the literature, many researches [2, 3, 16, 19, 20, 32, 33, 34, 35] have been done on the the property of RLCT λ, because λ appears also as the leading term of the asymptotic Bayes free energy. In singular learning theory, the concrete value of λ has been discussed for many learning machines, such as layered neural networks [2], normal mixture [32], Bayes network [20, 33], hidden Markov model [34], reduced rank regression [3] and so on. In addition, the concept of log canonical threshold, complexification of RLCT, plays an important role also in the field of high-dimensional algebraic geometry [21]. On the other hand, it is considered that singular fluctuation, SF, is a birational invariant firstly recognized in singular learning theory and the mathematical property of SF has not been clarified yet totally. In particular, the concrete value of singular fluctuation has not been known for any singular learning machines. Therefore, one of the purpose of this research is to gain insight in singular fluctuation theoretically.

Secondly, we explain the practical purpose of research. In section 1.1, we stated that WAIC has been proposed as an asymptotically unbiased estimator of Bayes generalization loss even in singular cases. To compute the value of WAIC,

we have to realize Bayes posterior distribution using, for example, Markov chain Monte Carlo (MCMC) method. However, for singular learning machines, it requires higher computational cost than the regular model case because Bayes posterior distribution is also singular for such learning machines. Therefore, approximate inference, such as variational Bayes learning, is also important in practice. For model selection in variational Bayes learning, minimization of variational free energy is often employed, however, it is desired to select a model based on generalization performance from the viewpoint of prediction for unknown samples. It has been difficult to construct an unbiased estimator of Bayes generalization loss in variational Bayes learning, since a learning machine to which variational Bayes learning is applied often has hidden variables, resulting that it is a singular learning machine. Therefore, another purpose of this research is to propose an unbiased estimator of Bayes generalization loss in variational Bayes learning based on WAIC.

1.3 Overview of Contribution

In section 1.2, we stated that this research has both theoretical and practical purposes.

Research for theoretical purpose will be discussed in chapter 3. Through investigating the property of singular fluctuation ν, we propose a new concept in singular learning theory, *a quasi-regular case*. Quasi-regular case is generally included in singular cases but has similar properties as regular cases. That is, in quasi-regular case, it will be proved that two birational invariants, λ and ν, are equal to each other as they are in regular cases, but they does not take a value of $\frac{d}{2}$ in general. Quasi-regular case provides the first example of singular cases for which the concrete value of singular fluctuation is clarified.

Research for practical purpose will be discussed in chapter 4. For the practical purpose, we propose a computational method of WAIC in variational Bayes learning, which method is called WAIC-VB. In WAIC-VB, WAIC is computed by the importance sampling from a proposal distribution constructed based on variational posterior distribution. Since it is easier to sample from such a proposal distribution than from the original Bayes posterior distribution, we can compute WAIC at lower computational cost compared to the method using MCMC. To validate the proposed method, we conduct some numerical experiments on WAIC-VB and compare its performance with generalized DIC proposed in [18].

In summary, through the research conducted in this paper, we will gain new insight in unbiased estimator of generalization loss in statistical learning. Details on the research will be discussed in the subsequent chapters.

Chapter 2

Framework of Bayes learning

In this chapter, we introduce fundamental concepts in Bayes learning and review its basic theory. In section 2.1, we introduce some fundamental concepts in Bayes learning theory and fundamental assumptions in this research. In section 2.2, we review the general theoretical framework of Bayes learning based on cumulant generating functions of two losses. In section 2.3, we introduce the concepts of regular case and singular case and briefly review its learning theory for each case. Lastly, in section 2.4, we review "Equation of States in Statistical Learning" and introduce a widely applicable information criterion, WAIC, as an unbiased estimator of Bayes generalization loss in general singular case.

2.1 Fundamental concepts in Bayes learning

2.1.1 Fundamental concepts and assumptions

Here we introduce some fundamental concepts and assumptions. Let N, n and d be natural numbers. Let $X_1, X_2, ..., X_n$ be random variables on $\mathbb{R}^N$ which are independently subject to the same probability density function as $q(x)$, which is called the true distribution below. Let $p(x|w)$ be a probability density function of x for a parameter $w \in W \subset \mathbb{R}^d$, which is called a learning machine or a parametric model below. The prior distribution is represented by the probability density function $\varphi(w)$ on W. In this research, we assume

- W is compact and its open kernel W^o, where open kernel is the maximal open subset included in W, is not empty.
- There exists a parameter $w_0 \in W^o$ which satisfies $q(x) = p(x|w_0)$.
- $\varphi(w) > 0$ for $\forall w \in W$.

Definition 1 *For a given pair of the true distribution $q(x)$ and a parametric model $p(x|w)$, the log density ratio function and Kullback-Leibler distance are respectively defined by*

$$f(x,w) = \log \frac{q(x)}{p(x|w)},$$

$$K(w) = \int q(x) \log \frac{q(x)}{p(x|w)} dx.$$

In this research, the log density ratio function $f(x,w)$ and Kullback-Leibler distance $K(w)$ are required to satisfy the following fundamental assumptions to develop the theory rigorously.

Fundamental assumptions

(1) It is said that $q(x)$ and $p(x|w)$ satisfy the fundamental assumption (1) with index s if the following conditions are satisfied. In this research, we assume $s = 6$.

For $f(x,w)$, there exists an open set $W^{(C)} \subset \mathbb{C}^d$ such that:
(1-a) $W \subset W^{(C)}$,
(1-b) $W^{(C)} \ni w \mapsto f(\cdot, w)$ is a $L^s(q)$-valued complex analytic function,
(1-c) $M(x) \equiv \sup_{w \in W^{(C)}} |f(x,w)|$ is contained in $L^s(q)$,

where $\mathbb{C}^d$ is the set of all d-dimensional complex numbers and $L^s(q)$ is the set of functions whose s-th powers are integrable with probability distribution $q(x)dx$.

(2) There exists $\epsilon > 0$ such that, for

$$Q(x) \equiv \sup_{K(w) \leq \epsilon} p(x|w),$$

the following integral is finite,

$$\int M(x)^2 Q(x) dx < \infty.$$

(3) $K(w)$ is an analytic function of w.

Definition 2 *The empirical loss function is defined by*

$$K_n(w) = \frac{1}{n} \sum_{i=1}^{n} f(X_i, w) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i|w)}.$$

Remark. It is clear that $\mathbb{E}[K_n(w)] = K(w)$ holds, where $\mathbb{E}[\ ]$ represents the expectation over X^n.

2.1.2 Bayes learning

Here we introduce Bayes learning. For a given training set

$$X^n = \{X_1, X_2, ..., X_n\},$$

the Bayes posterior distribution is defined by

$$p(w|X^n) = \frac{1}{Z(X^n)} \prod_{i=1}^{n} p(X_i|w)^\beta \varphi(w),$$

where $0 < \beta < \infty$ is the inverse temperature and $Z(X^n)$ is the normalizing constant. The case $\beta = 1$ is most important because it corresponds to the strict Bayes estimation.

Remark. The posterior distribution can be represented using empirical loss function as follows.

$$p(w|X^n) = \frac{1}{Z_0(X^n)} \exp(-n\beta K_n(w))\varphi(w),$$

where $Z_0(X^n)$ is also the normalizing constant.

The expectation value over the posterior distribution is denoted by

$$\mathbb{E}_w[f(w)] = \int f(w)p(w|X^n)dw,$$

where $f(w)$ is an arbitrary function of w. The Bayes predictive distribution is defined by

$$p(x|X^n) = \mathbb{E}_w[p(x|w)].$$

2.1.3 Generalization and training losses

From both theoretical and practical viewpoints, it is important to evaluate how well the predictive distribution $p(x|X^n)$ approximates the true distribution $q(x)$ as a result of learning. In general, Kullback-Leibler distance plays a key role as a measure of discrepancy of two probability distributions in statistical learning theory. Generalization error is defined by Kullback-Leibler distance from $q(x)$ to $p(x|X^n)$ and training error is defined by empirical Kullback-Leibler distance from $q(x)$ to $p(x|X^n)$.

Definition 3 *Generalization error* $G_n^{(0)}$ *and Training error* $T_n^{(0)}$ *are defined by*

$$G_n^{(0)} = \int q(x) \log \frac{q(x)}{p(x|X^n)} dx, \quad T_n^{(0)} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i|X^n)}$$

respectively.

In practice, the concepts of generalization loss and training loss are often employed, which are obtained by adding entropy term or empirical entropy term of the true distribution $q(x)$ to generalization error or training error respectively.

Definition 4 *Generalization loss G_n and Training loss T_n are defined by*

$$G_n = -\int q(x) \log p(x|X^n) dx = S + G_n^{(0)},$$

$$T_n = -\frac{1}{n} \sum_{i=1}^{n} \log p(X_i|X^n) = S_n + T_n^{(0)},$$

where S and S_n are entropy terms defined by

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

From the definition, it can be said that G_n is a loss for unknown samples and T_n is a loss for obtained samples. In statistical learning, it is desired that G_n gets smaller from the viewpoint of prediction, however, we cannot calculate G_n directly from samples. Hence, it is very useful if we can estimate G_n using quantities which can be calculated from samples, such as T_n.

2.2 Theoretical framework of Bayes learning

One of the main goals of Bayes learning theory is to clarify the asymptotic behavior of G_n and T_n. Here we introduce the common theoretical framework for deriving the asymptotic behavior of two losses.

2.2.1 Cumulant generating functions for G_n and T_n

Firstly, we introduce cumulant generating functions for G_n and T_n.

Definition 5 *Let α be a real number. Then cumulant generating functions for generalization loss and training loss, which is denoted by $\mathcal{G}_n(\alpha)$ and $\mathcal{T}_n(\alpha)$, are respectively defined by*

$$\mathcal{G}_n(\alpha) = \mathbb{E}_X[\log \mathbb{E}_w[p(X|w)^\alpha]], \quad \mathcal{T}_n(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \log \mathbb{E}_w[p(X_i|w)^\alpha].$$

Then, k-th cumulants are defined by their k-th derivatives at $\alpha = 0$.

From the definition,

$$G_n = -\mathcal{G}_n(1), \quad T_n = -\mathcal{T}_n(1)$$

clearly holds. Hence, it can be said that the behavior of G_n and T_n is clarified by investigating the cumulant generating funtion $\mathcal{G}_n(\alpha)$ and $\mathcal{T}_n(\alpha)$. To describe $\mathcal{G}_n^{(k)}(\alpha)$ and $\mathcal{T}_n^{(k)}(\alpha)$ in a unified way, we introduce the following notation.

Definition 6 *For a random variable A, $l_k(A)$ $(k \in \mathbb{N})$ is defined by*

$$l_k(A) = \frac{\mathbb{E}_w[(\log p(A|w))^k p(A|w)^\alpha]}{\mathbb{E}_w[p(A|w)^\alpha]}.$$

Then, following lemma holds.

Lemma 1

$$\mathcal{G}_n^{(1)}(\alpha) = \mathbb{E}_X[l_1(X)], \quad \mathcal{G}_n^{(2)}(\alpha) = \mathbb{E}_X[l_2(X) - l_1^2(X)]$$

$$\mathcal{T}_n^{(1)}(\alpha) = \frac{1}{n}\sum_{i=1}^n l_1(X_i), \quad \mathcal{T}_n^{(2)}(\alpha) = \frac{1}{n}\sum_{i=1}^n [l_2(X_i) - l_1^2(X_i)]$$

Proof.

Since the first derivative of $\mathcal{G}_n(\alpha)$ is calculated as

$$\mathcal{G}_n^{(1)}(\alpha) = \mathbb{E}_X\left[\frac{\frac{d}{d\alpha}(\mathbb{E}_w[p(X|w)^\alpha])}{\mathbb{E}_w[p(A|w)^\alpha]}\right] = \mathbb{E}_X\left[\frac{\mathbb{E}_w[(\log p(X|w))p(X|w)^\alpha]}{\mathbb{E}_w[p(A|w)^\alpha]}\right],$$

$\mathcal{G}_n^{(1)}(\alpha) = \mathbb{E}_X[l_1(X)]$ holds. Note that, in general, for an arbitrary differentiable function $f(\alpha)$ and a natural number k, the following relation holds.

$$\frac{d}{d\alpha}\left(\frac{f^{(k)}(\alpha)}{f(\alpha)}\right) = \frac{f^{(k+1)}(\alpha)}{f(\alpha)} - \frac{f^{(k)}(\alpha)}{f(\alpha)}\frac{f^{(1)}(\alpha)}{f(\alpha)}$$

From the above relation, $\mathcal{G}_n^{(2)}(\alpha) = \mathbb{E}_X[l_2(X) - l_1^2(X)]$ clearly holds.
The second half of the lemma can be proved in the same way. (Q.E.D.)

2.2.2 Fundamental theorem for Bayes learning

Here we state the fundamental theorem for Bayes learning, which provides the common theoretical framework for deriving asymptotic behavior of generalization and training losses.

Theorem 1 *Assume that*

$$\left|(\frac{d}{d\alpha})^3\mathcal{G}_n(\alpha)\right| = o_p(\frac{1}{n}), \quad \left|(\frac{d}{d\alpha})^3\mathcal{T}_n(\alpha)\right| = o_p(\frac{1}{n})$$

holds. Then generalization loss G_n and training loss T_n can be calculated by the following formula.

$$G_n = -\mathcal{G}_n(1) = -\mathcal{G}_n'(0) - \frac{1}{2}\mathcal{G}_n''(0) + o_p(\frac{1}{n}),$$
$$T_n = -\mathcal{T}_n(1) = -\mathcal{T}_n'(0) - \frac{1}{2}\mathcal{T}_n''(0) + o_p(\frac{1}{n}),$$

where

$$\mathcal{G}_n^{'}(0) = -S - \mathbb{E}_w[K(w)],$$
$$\mathcal{G}_n^{''}(0) = \mathbb{E}_X[\mathbb{E}_w[f(X,w)^2] - \mathbb{E}_w[f(X,w)]^2],$$
$$\mathcal{T}_n^{'}(0) = -S_n - \mathbb{E}_w[K_n(w)],$$
$$\mathcal{T}_n^{''}(0) = \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_w[f(X_i,w)^2] - \mathbb{E}_w[f(X_i,w)]^2\right\}.$$

Proof.

From the definition of $\mathcal{G}_n(\alpha)$, $\mathcal{G}_n(0) = 0$ holds. Hence, from the mean-value theorem, there exists $0 < |\alpha^*| < |\alpha|$ such that

$$\mathcal{G}_n(\alpha) = \alpha\mathcal{G}_n^{'}(0) + \frac{1}{2}\alpha^2\mathcal{G}_n^{''}(0) + \frac{1}{6}\alpha^3\mathcal{G}_n^{'''}(\alpha^*).$$

By substituting $\alpha = 1$ to the above equation, we obtain

$$G_n = -\mathcal{G}_n(1) = -\mathcal{G}_n^{'}(0) - \frac{1}{2}\mathcal{G}_n^{''}(0) + o_p(\frac{1}{n})$$

from the assumption of the theorem.

$$T_n = -\mathcal{T}_n(1) = -\mathcal{T}_n^{'}(0) - \frac{1}{2}\mathcal{T}_n^{''}(0) + o_p(\frac{1}{n})$$

can be proved in the same way.
From Lemma 1, the first order cumulants $\mathcal{G}_n^{'}(0)$ and $\mathcal{T}_n^{'}(0)$ are given by

$$\mathcal{G}_n^{'}(0) = \mathbb{E}_X[l_1(X)|_{\alpha=0}] = \mathbb{E}_X[\mathbb{E}_w[\log p(X|w)]],$$

$$\mathcal{T}_n^{'}(0) = \frac{1}{n}\sum_{i=1}^{n} l_1(X_i)|_{\alpha=0} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_w[\log p(X_i|w)].$$

On the other hand,

$$\begin{aligned}\mathbb{E}_w[K(w)] &= \mathbb{E}_w[\mathbb{E}_X[f(X,w)]] = \mathbb{E}_w[\mathbb{E}_X[\log\frac{q(X)}{p(X|w)}]] \\ &= -S - \mathbb{E}_w[\mathbb{E}_X[\log p(X|w)]] = -S - \mathbb{E}_X[\mathbb{E}_w[\log p(X|w)]],\end{aligned}$$

$$\begin{aligned}\mathbb{E}_w[K_n(w)] &= \mathbb{E}_w[\frac{1}{n}\sum_{i=1}^{n} f(X_i,w)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_w[\log\frac{q(X_i)}{p(X_i|w)}] \\ &= -S_n - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_w[\log p(X_i|w)].\end{aligned}$$

Hence,

$$\mathcal{G}_n^{'}(0) = -S - \mathbb{E}_w[K(w)], \quad \mathcal{T}_n^{'}(0) = -S_n - \mathbb{E}_w[K_n(w)]$$

holds. Also from the Lemma 1, the second order cumulants $\mathcal{G}_n''(0)$ and $\mathcal{T}_n''(0)$ are given by

$$\mathcal{G}_n''(0) = \mathbb{E}_X[l_2(X)|_{\alpha=0} - l_1^2(X)|_{\alpha=0}] = \mathbb{E}_X[\mathbb{E}_w[\log p(X|w)^2] - \mathbb{E}_w[\log p(X|w)]^2],$$

$$\mathcal{T}_n''(0) = \frac{1}{n}\sum_{i=1}^{n}\{l_2(X_i)|_{\alpha=0} - l_1^2(X_i)|_{\alpha=0}\} = \frac{1}{n}\sum_{i=1}^{n}\{\mathbb{E}_w[\log p(X_i|w)^2] - \mathbb{E}_w[\log p(X_i|w)]^2\}.$$

Then, since

$$\begin{aligned}
& \mathbb{E}_w[f(X,w)^2] - \mathbb{E}_w[f(X,w)]^2 \\
= \; & \mathbb{E}_w[\{\log q(X) - \log p(X|w)\}^2] - \mathbb{E}_w[\{\log q(X) - \log p(X|w)\}]^2 \\
= \; & \log q(X)^2 - 2\log q(X)\mathbb{E}_w[\log p(X|w)] + \mathbb{E}_w[\log p(X|w)^2] \\
- \; & \{\log q(X)^2 - 2\log q(X)\mathbb{E}_w[\log p(X|w)] + \mathbb{E}_w[\log p(X|w)]^2\} \\
= \; & \mathbb{E}_w[\log p(X|w)^2] - \mathbb{E}_w[\log p(X|w)]^2
\end{aligned}$$

holds, we completes the proof. (Q.E.D.)

Remark. In Theorem 1, it is assumed that

$$\left|(\frac{d}{d\alpha})^3\mathcal{G}_n(\alpha)\right| = o_p(\frac{1}{n}), \quad \left|(\frac{d}{d\alpha})^3\mathcal{T}_n(\alpha)\right| = o_p(\frac{1}{n})$$

holds. It is shown that these assumptions are really satisfied in Bayes learning theory [29].

The next theorem holds only for Bayes learning and plays an important role in Bayes learning theory.

Theorem 2 *Let* $\mathbb{E}[\ ]$ *represents the expectation over samples* X^n. *Then*

$$\mathbb{E}[\mathcal{G}_{n-1}(\beta)] = -\mathbb{E}[\mathcal{T}_n(-\beta)].$$

Hence, when third and over order cumulants go to zero faster than $\frac{1}{n}$, *the following relation holds.*

$$\mathbb{E}[\mathcal{G}_{n-1}'(0) + \frac{\beta}{2}\mathcal{G}_{n-1}''(0)] = \mathbb{E}[\mathcal{T}_n'(0) - \frac{\beta}{2}\mathcal{T}_n''(0)] + o(\frac{1}{n}).$$

Proof.

Here we denote a normalizing constant $Z(X^n)$ as $Z_n(\beta)$, where $0 < \beta < \infty$ represents the inverse temparature. Then, from the definition,

$$Z_n(\beta) = \int \prod_{i=1}^{n} p(X_i|w)^\beta \varphi(w) dw.$$

Besides, we denote expectation over the posterior distribution $p(w|X^n)$ as $\mathbb{E}_w^n[\,]$ in this proof. Then, the following relation holds,

$$\mathbb{E}_w^{n-1}[p(X_n|w)^\beta] = \frac{Z_n(\beta)}{Z_{n-1}(\beta)}.$$

Then,

$$\begin{aligned}
\mathbb{E}[\mathcal{G}_{n-1}(\beta)] &= \mathbb{E}_{X^{n-1}}[\mathbb{E}_{X_n}[\log \mathbb{E}_w^{n-1}[p(X_n|w)^\beta]]] \\
&= \mathbb{E}_{X^n}[\log \frac{Z_n(\beta)}{Z_{n-1}(\beta)}] = -\mathbb{E}[\log \frac{Z_{n-1}(\beta)}{Z_n(\beta)}] \\
&= -\mathbb{E}[\log \frac{1}{Z_n(\beta)} \int p(X_n|w)^{-\beta} \prod_{i=1}^n p(X_i|w)^\beta \varphi(w) dw] \\
&= -\mathbb{E}[\log \mathbb{E}_w^n[p(X_n|w)^{-\beta}]]
\end{aligned}$$

Since

$$\mathbb{E}[\log \mathbb{E}_w^n[p(X_1|w)^{-\beta}]] = \mathbb{E}[\log \mathbb{E}_w^n[p(X_2|w)^{-\beta}]] = \cdots = \mathbb{E}[\log \mathbb{E}_w^n[p(X_n|w)^{-\beta}]],$$

$$\mathbb{E}[\mathcal{G}_{n-1}(\beta)] = -\mathbb{E}[\frac{1}{n}\sum_{i=1}^n \log \mathbb{E}_w^n[p(X_i|w)^{-\beta}]] = -\mathbb{E}[\mathcal{T}_n(-\beta)]$$

holds, which completes the proof of the first half of Theorem 2. Then, from the mean-value theorem, there exists $0 < |\alpha^*| < |\alpha|$ such that

$$\mathcal{G}_{n-1}(\alpha) = \alpha \mathcal{G}_n^{'}(0) + \frac{1}{2}\alpha^2 \mathcal{G}_n^{''}(0) + \frac{1}{6}\alpha^3 \mathcal{G}_n^{'''}(\alpha^*).$$

By substituting $\alpha = \beta$, we obtain

$$\mathcal{G}_{n-1}(\beta) = \beta \mathcal{G}_{n-1}^{'}(0) + \frac{1}{2}\beta^2 \mathcal{G}_{n-1}^{''}(0) + o_p(\frac{1}{n}).$$

In the same way, we obtain

$$\mathcal{T}_n(-\beta) = -\beta \mathcal{T}_n^{'}(0) + \frac{1}{2}\beta^2 \mathcal{T}_n^{''}(0) + o_p(\frac{1}{n}).$$

Hence, from the first half of Theorem 2,

$$\mathbb{E}[\beta \mathcal{G}_{n-1}^{'}(0) + \frac{1}{2}\beta^2 \mathcal{G}_{n-1}^{''}(0)] = \mathbb{E}[\beta \mathcal{T}_n^{'}(0) - \frac{1}{2}\beta^2 \mathcal{T}_n^{''}(0)] + o(\frac{1}{n})$$

Then, by dividing by $\beta \neq 0$, we obtain the second half of the Theorem. (Q.E.D.)

2.3 Regular and Singular cases

Here we introduce two important concepts, *regular* and *singular*, for the pair of $(q(x), p(x|w))$. After that, we briefly summarize the fundamental result of statistical learning theory in both regular case and singular case.

2.3.1 Definition of regular and singular cases

Here we state the definition of *regular* and *singular* cases for the pair of $(q(x), p(x|w))$. Recall that we assume the true distribution $q(x)$ is realizable by a parametric model $p(x|w)$.

Definition 7 *Let $W \subset \mathbb{R}^d$ be a parameter set and W_0 be*

$$W_0 = \{w \in W; q(x) = p(x|w)\},$$

which is called a set of true parameters. Then, a pair of the true distribution and a parametric model $(q(x), p(x|w))$ is said to be in a regular case if and only if the set W_0 consists of a single element $w_0 \in W^o$ and Hessian matrix of $K(w)$ at w_0

$$J := \nabla^2 K(w_0),$$

is positive definite, where $\nabla^2 K(w_0)$ represents a $d \times d$ matrix whose (i, j) component is given by

$$(\nabla^2 K(w_0))_{i,j} = (\frac{\partial^2 K}{\partial w_i \partial w_j})(w_0).$$

Otherwise, it is said to be in a singular case.

Remark. (1) Under the assumption that the true distribution is realizable by a parametric model, it can be proved that Hessian matrix of $K(w)$ at w_0 is equal to the following Fisher information matrix I at w_0. (cf. Lemma 5)

$$I := \int \nabla \log p(x|w_0)(\nabla \log p(x|w_0))^T q(x) dx,$$

where

$$\nabla = \frac{\partial}{\partial w} = \Big(\frac{\partial}{\partial w_1}, ..., \frac{\partial}{\partial w_d}\Big)^T$$

and $(\ )^T$ represents its transpose. Hence, on that assumption, we can use Fisher information matrix I instead of Hessian matrix J in the Definition 7. However, when the true distribution is unrealizable by a parametric model $p(x|w)$, it turns out that regularity of Hessian matrix J is more essential than that of Fisher information matrix I for the regularity condition of $(q(x), p(x|w))$.
(2) The terms such as 'regular model', 'singular learning machines', are often used. That is, the concepts of *regular* and *singular* are employed as a term not for representing the property of the pair of $(q(x), p(x|w))$ but for the model $p(x|w)$. Roughly speaking, a parametric model or a learning machine is said to be regular if and only if a parameter corresponds to a parametric model one-to-one and there does not exist a parameter in the parameter space at which Fisher information matrix is degenerate. Otherwise, it is said to be singular.

2.3.2 Overview of Statistical Learning Theory

The goal in this section is to summarize how to derive asymptotic behavior of G_n and T_n in both regular case and singular case. Before discussing in detail, we explain our basic strategy which is common in both cases.
From Theorem 1, for deriving asymptotic behavior of G_n and T_n, we have to clarify the asymptotic behavior of the first and second order cumulants, that is,

$$\mathcal{G}_n^{'}(0) = -S - \mathbb{E}_w[K(w)],$$

$$\mathcal{G}_n^{''}(0) = \mathbb{E}_X[\mathbb{E}_w[f(X,w)^2] - \mathbb{E}_w[f(X,w)]^2],$$

$$\mathcal{T}_n^{'}(0) = -S_n - \mathbb{E}_w[K_n(w)],$$

$$\mathcal{T}_n^{''}(0) = \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_w[f(X_i,w)^2] - \mathbb{E}_w[f(X_i,w)]^2\right\},$$

where

$$\mathbb{E}_w[\ ] = \frac{\int(\ )\Omega(w)dw}{\int\Omega(w)dw}, \qquad \Omega(w)dw := \exp(-n\beta K_n(w))\varphi(w)dw.$$

Hence, the following is our basic strategy for deriving asymptotic behavior of G_n and T_n.

Step 1: Clarify the asymptotic behavior of $\Omega(w)dw$ and $\mathbb{E}_w[\ ]$.

Step 2: Clarify the asymptotic behavior of the first and second order cumulants $\mathcal{G}_n^{'}(0), \mathcal{G}_n^{''}(0), \mathcal{T}_n^{'}(0), \mathcal{T}_n^{''}(0)$.

Step 3: Clarify the asymptotic behavior of G_n and T_n based on Theorem 1.

To derive asymptotic behavior of $\Omega(w)dw$, it is sufficient to consider the behavior of $K_n(w)$ in the neighborhood of $K(w) = 0$ [28, 29].

2.3.3 Statistical Learning Theory in regular cases

Here we briefly summarize the statistical learning theory in regular cases.
We begin with **Step 1**, that is, clarifying the asymptotic behavior of $\Omega(w)dw$ and $\mathbb{E}_w[\ ]$. Firstly, we investigate the behavior of $K_n(w)$ in the neighborhood of $w = w_0$.

Definition 8 *A stochastic process $\eta_n(w)$ is defined by*

$$\eta_n(w) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(K(w) - f(X_i,w)).$$

Lemma 2 *The mean function of $\eta_n(w)$ is the zero function and the corelation function of $\eta_n(w)$ is given by*

$$E_X[f(X,w)f(X,w')] - K(w)K(w').$$

When $n \to \infty$, a stachastic process $\eta_n(w)$ converges to a gaussian process $\eta(w)$ in law, which has the same mean and corelation function with $\eta_n(w)$.

From Definition 8, empirical loss function can be described as

$$K_n(w) = K(w) - \frac{1}{\sqrt{n}}\eta_n(w).$$

Note that

$$K(w_0) = 0, \quad \nabla K(w_0) = 0, \quad \eta_n(w_0) = 0$$

holds. Then, from mean-value theorem, there exist w^*, w^{**} such that

$$K(w) = \frac{1}{2}(w - w_0) \cdot J(w^*)(w - w_0), \ \eta_n(w) = (w - w_0) \cdot \nabla\eta_n(w^{**}),$$

where $J(w^*)$ represents Hessian matrix at $w = w^*$. Therefore, following equation holds.

$$\begin{aligned} nK_n(w) &= \frac{n}{2}(w - w_0) \cdot J(w^*)(w - w_0) - \sqrt{n}(w - w_0) \cdot \nabla\eta_n(w^{**}) \\ &= \frac{n}{2}||J(w^*)^{\frac{1}{2}}(w - w_0 - \frac{1}{\sqrt{n}}J(w^*)^{-1}\nabla\eta_n(w^{**}))||^2 \\ &\quad -\frac{1}{2}||J(w^*)^{-\frac{1}{2}}\nabla\eta_n(w^{**})||^2 \end{aligned}$$

Hence, in the neighborhood of $w = w_0$,

$$nK_n(w) \cong \frac{n}{2}||J^{\frac{1}{2}}(w - w_0 - \frac{1}{\sqrt{n}}J^{-1}\nabla\eta_n(w_0))||^2 - \frac{1}{2}||J^{-\frac{1}{2}}\nabla\eta_n(w_0)||^2$$

Here we introduce the following notation ξ_n and $\hat{\xi}_n$.

Definition 9 *Random variables ξ_n and $\hat{\xi}_n$ are defined by*

$$\xi_n = J^{-1/2}\nabla\eta_n(w_0),$$

$$\hat{\xi}_n = J^{-1}\nabla\eta_n(w_0) = J^{-1/2}\xi_n.$$

Hence, from Definition 9,

$$nK_n(w) \cong \frac{n}{2}||J^{\frac{1}{2}}(w - w_0 - \frac{\hat{\xi}_n}{\sqrt{n}})||^2 - \frac{1}{2}||\xi_n||^2$$

holds in the neighborhood of $w = w_0$.

Then, we can directly obtain the following lemma.

Lemma 3 *Assume that $(q(x), p(x|w))$ is in regular cases. Then, $\mathbb{E}_w[\ ]$ has the following asymptotic behavior.*

$$\mathbb{E}_w[\ ] = \frac{\int (\ ) \exp(-\frac{n\beta}{2}||J^{1/2}(w - w_0 - \frac{\hat{\xi}_n}{\sqrt{n}})||^2) dw}{\int \exp(-\frac{n\beta}{2}||J^{1/2}(w - w_0 - \frac{\hat{\xi}_n}{\sqrt{n}})||^2) dw}(1 + o_p(1)).$$

Next, as **Step 2**, we investigate the asymptotic behavior of the first and second order cumulants.

We begin with introducing an important concept, Fisher information matrix and showing some lemmas.

Definition 10 *$I(w)$ and I, $d \times d$ matrices, are defined by*

$$I(w) = \mathbb{E}_X[\nabla f(X, w)(\nabla f(X, w))^T] - \nabla K(w)(\nabla K(w))^T,$$

$$I = I(w_0) = \mathbb{E}_X[\nabla f(X, w)(\nabla f(X, w))^T].$$

A matrix I is called Fisher information matrix at $w = w_0$.

Lemma 4 *Assume that $I(w)$ is positive definite. Then, following convergence in law holds.*
(1) For each $w \in W$, a random variable $\nabla \eta_n(w)$ converges to $N(0, I(w))$ in law.
(2) A random variable ξ_n converges to $N(0, J^{-1/2} I J^{-1/2})$ in law.
(3) A random variable $\hat{\xi}_n$ converges to $N(0, J^{-1} I J^{-1})$ in law.

Proof.

(1) For each $w \in W$,

$$\nabla \eta_n(w) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\nabla K(w) - \nabla f(X_i, w))$$

holds. Hence, from the central limit theorem, $\nabla \eta_n(w)$ converges to $N(0, I(w))$ in law.

(2),(3) It clearly holds from the definition of ξ_n and $\hat{\xi}_n$. (Q.E.D.)

Following lemma is an important property of realizable and regular case.

Lemma 5 *Assume that the true distribution $q(x)$ is realizable by a parametric model $p(x|w)$ and $(q(x), p(x|w))$ is in regular cases. Then $I = J$ holds.*

Proof.

From the definition,

$$I = \int q(x) \nabla \log p(x|w_0) (\nabla \log p(x|w_0))^T dx, \quad J = -\int q(x) \nabla^2 \log p(x|w_0) dx$$

holds. Then,

$$\nabla \log p(x|w_0) = \frac{\nabla p(x|w_0)}{p(x|w_0)},$$

$$\begin{aligned}\nabla^2 \log p(x|w_0) &= \frac{\nabla^2 p(x|w_0)p(x|w_0) - \nabla p(x|w_0)(\nabla p(x|w_0))^T}{p(x|w_0)^2} \\ &= \frac{\nabla^2 p(x|w_0)}{p(x|w_0)} - \frac{\nabla p(x|w_0)(\nabla p(x|w_0))^T}{p(x|w_0)^2}.\end{aligned}$$

Therefore, from the assumption of $q(x) = p(x|w_0)$,

$$\begin{aligned}J &= -\int q(x)\nabla^2 \log p(x|w_0)dx = -\int p(x|w_0)\nabla^2 \log p(x|w_0)dx, \\ &= -\int \nabla^2 p(x|w_0)dx + \int q(x)\frac{\nabla p(x|w_0)(\nabla p(x|w_0))^T}{p(x|w_0)^2}dx, \\ &= \int q(x)\nabla \log p(x|w_0)(\nabla \log p(x|w_0))^T dx = I,\end{aligned}$$

which completes the proof. (Q.E.D.)

Now we prepare some formulae to derive asymptotic behavior of the first and second order cumulants.

Lemma 6 *Following equations hold.*

$$\mathbb{E}_w[w] = w_0 + \frac{1}{\sqrt{n}}\hat{\xi}_n + o_p(\frac{1}{\sqrt{n}}),$$

$$\mathbb{E}_w[(w-w_0)(w-w_0)^T] = \frac{J^{-1}}{n\beta} + \frac{\hat{\xi}_n\hat{\xi}_n^T}{n} + o_p(\frac{1}{n}),$$

$$\mathbb{E}_w[f(x,w)] = \frac{\hat{\xi}_n}{\sqrt{n}} \cdot \nabla f(x,w_0) + o_p(\frac{1}{\sqrt{n}})$$

and

$$\begin{aligned}&\mathbb{E}_w[f(x,w)^2] - \mathbb{E}_w[f(x,w)]^2 \\ =\ &\frac{1}{n\beta}\mathrm{tr}(J^{-1}(\nabla f(x,w_0))(\nabla f(x,w_0))^T) + o_p(\frac{1}{n}).\end{aligned}$$

Proof.

Firstly, from Lemma 3,

$$\begin{aligned}\mathbb{E}_w[w] &= \frac{\int w \exp(-\frac{n\beta}{2}||J^{1/2}(w-w_0-\frac{\hat{\xi}_n}{\sqrt{n}})||^2)dw}{\int \exp(-\frac{n\beta}{2}||J^{1/2}(w-w_0-\frac{\hat{\xi}_n}{\sqrt{n}})||^2)dw}(1+o_p(1)) \\ &= w_0 + \frac{1}{\sqrt{n}}\hat{\xi}_n + o_p(\frac{1}{\sqrt{n}}).\end{aligned}$$

Note that the following equation holds,

$$-\frac{n\beta}{2}||J^{1/2}(w-w_0-\frac{\hat{\xi}_n}{\sqrt{n}})||^2 = -\frac{1}{2}(w-w_0-\frac{\hat{\xi}_n}{\sqrt{n}})\cdot(\frac{J^{-1}}{n\beta})^{-1}(w-w_0-\frac{\hat{\xi}_n}{\sqrt{n}}).$$

Hence,

$$\mathbb{E}_w[(w-w_0-\frac{\hat{\xi}_n}{\sqrt{n}})(w-w_0-\frac{\hat{\xi}_n}{\sqrt{n}})^T] = \frac{J^{-1}}{n\beta}+o_p(\frac{1}{n}).$$

That is,

$$\mathbb{E}_w[(w-w_0)(w-w_0)^T]-\mathbb{E}_w[w-w_0]\frac{\hat{\xi}_n^T}{\sqrt{n}}-\frac{\hat{\xi}_n}{\sqrt{n}}\mathbb{E}_w[(w-w_0)^T]+\frac{\hat{\xi}_n\hat{\xi}_n^T}{n} = \frac{J^{-1}}{n\beta}+o_p(\frac{1}{n}).$$

Then, since

$$\mathbb{E}_w[w-w_0] = \frac{1}{\sqrt{n}}\hat{\xi}_n+o_p(\frac{1}{\sqrt{n}}),$$

we obtain

$$\mathbb{E}_w[(w-w_0)(w-w_0)^T]-\frac{\hat{\xi}_n\hat{\xi}_n^T}{n} = \frac{J^{-1}}{n\beta}+o_p(\frac{1}{n}),$$

that is,

$$\mathbb{E}_w[(w-w_0)(w-w_0)^T] = \frac{J^{-1}}{n\beta}+\frac{\hat{\xi}_n\hat{\xi}_n^T}{n}+o_p(\frac{1}{n}).$$

Next, from mean-value theorem, there exists w^+ such that

$$f(x,w) = (w-w_0)\cdot\nabla f(x,w^+).$$

Therefore,

$$\mathbb{E}_w[f(x,w)] = \mathbb{E}_w[(w-w_0)\cdot(\nabla f(x,w_0)+o_p(1))] = \frac{\hat{\xi}_n}{\sqrt{n}}\cdot\nabla f(x,w_0)+o_p(\frac{1}{\sqrt{n}}).$$

Lastly, since

$$\begin{aligned}
\mathbb{E}_w[f(x,w)^2] &= \mathbb{E}_w[((w-w_0)\cdot\nabla f(x,w^+))^2] \\
&= \mathbb{E}_w[\mathrm{tr}((w-w_0)(w-w_0)^T(\nabla f(x,w^+)(\nabla f(x,w^+))^T)] \\
&= \mathrm{tr}(\mathbb{E}_w[(w-w_0)(w-w_0)^T](\nabla f(x,w_0))(\nabla f(x,w_0))^T)+o_p(\frac{1}{n}) \\
&= \mathrm{tr}((\frac{J^{-1}}{n\beta}+\frac{\hat{\xi}_n\hat{\xi}_n^T}{n})(\nabla f(x,w_0))(\nabla f(x,w_0))^T)+o_p(\frac{1}{n}),
\end{aligned}$$

we obtain

$$\begin{aligned}
&\mathbb{E}_w[f(x,w)^2]-\mathbb{E}_w[f(x,w)]^2 \\
&= \frac{1}{n\beta}\mathrm{tr}(J^{-1}(\nabla f(x,w_0))(\nabla f(x,w_0))^T)+o_p(\frac{1}{n}),
\end{aligned}$$

which completes the proof. (Q.E.D.)

Based on Lemma 6, we can derive the asymptotic behavior of the first and second order cumulants and finally the asymptotic behavior of G_n and T_n as **Step 3**.

Theorem 3 *Assume that the true distribution $q(x)$ is realizable by a parametric model $p(x|w)$ and $(q(x), p(x|w))$ is in regular cases. Then, asymptotic behavior of G_n and T_n can be described as*

$$G_n = S + \frac{1}{2n}||\xi_n||^2 + o_p(\frac{1}{n}), \quad T_n = S_n - \frac{1}{2n}||\xi_n||^2 + o_p(\frac{1}{n}).$$

Proof.

From Theorem 1, we can derive asymptotic behavior of G_n and T_n by calculating the first and second order cumulants $\mathcal{G}_n^{'}(0), \mathcal{G}_n^{''}(0), \mathcal{T}_n^{'}(0), \mathcal{T}_n^{''}(0)$ using Lemma 5 and 6.

At first, we derive the asymptotic behavior of G_n in regular cases.
Since there exists w^* such that

$$K(w) = \frac{1}{2}(w - w_0) \cdot J(w^*)(w - w_0) = \frac{1}{2}\mathrm{tr}(J(w^*)(w - w_0)(w - w_0)^T),$$

the first order cumulant $\mathcal{G}_n^{'}(0)$ can be calculated as

$$\begin{aligned}
-\mathcal{G}_n^{'}(0) &= S + \mathbb{E}_w[K(w)] = S + \frac{1}{2}\mathbb{E}_w[\mathrm{tr}(J(w^*)(w - w_0)(w - w_0)^T)] \\
&= S + \frac{1}{2}\mathrm{tr}(J(w^*)\mathbb{E}_w[(w - w_0)(w - w_0)^T]) \\
&= S + \frac{1}{2}\mathrm{tr}(J(\frac{J^{-1}}{n\beta} + \frac{\hat{\xi}_n\hat{\xi}_n^T}{n})) + o_p(\frac{1}{n}) \\
&= S + \frac{d}{2n\beta} + \frac{1}{2n}\mathrm{tr}((J^{1/2}\hat{\xi}_n)(J^{1/2}\hat{\xi}_n)^T) + o_p(\frac{1}{n}) \\
&= S + \frac{d}{2n\beta} + \frac{||\xi_n||^2}{2n} + o_p(\frac{1}{n}).
\end{aligned}$$

Similarly, from Lemma 5 and 6,

$$\begin{aligned}
\mathcal{G}_n^{''}(0) &= \mathbb{E}_X[\mathbb{E}_w[f(X,w)^2] - \mathbb{E}_w[f(X,w)]^2], \\
&= \frac{1}{n\beta}\mathbb{E}_X[\mathrm{tr}(J^{-1}(\nabla f(X,w_0))(\nabla f(X,w_0))^T)] + o_p(\frac{1}{n}), \\
&= \frac{1}{n\beta}\mathrm{tr}(IJ^{-1}) + o_p(\frac{1}{n}) = \frac{d}{n\beta} + o_p(\frac{1}{n}).
\end{aligned}$$

Therefore, from Theorem 1,

$$\begin{aligned} G_n &= -\mathcal{G}_n(1) = -\mathcal{G}_n'(0) - \frac{1}{2}\mathcal{G}_n''(0) + o_p(\frac{1}{n}), \\ &= S + \frac{d}{2n\beta} + \frac{||\xi_n||^2}{2n} - \frac{d}{2n\beta} + o_p(\frac{1}{n}) = S + \frac{||\xi_n||^2}{2n} + o_p(\frac{1}{n}). \end{aligned}$$

Next, we derive the asymptotic behavior of T_n in regular cases.
Then, since

$$K_n(w) = \frac{1}{2}||J^{1/2}(w - w_0 - \frac{1}{\sqrt{n}}\hat{\xi}_n)||^2 - \frac{1}{2n}||\xi_n||^2 + o_p(\frac{1}{n})$$

and the integral formula

$$\frac{\int ||J^{1/2}(w - w_0 - \frac{1}{\sqrt{n}}\hat{\xi}_n)||^2 \exp(-\frac{n\beta}{2}||J^{1/2}(w - w_0 - \frac{1}{\sqrt{n}}\hat{\xi}_n)||^2)du}{\int \exp(-\frac{n\beta}{2}||J^{1/2}(w - w_0 - \frac{1}{\sqrt{n}}\hat{\xi}_n)||^2)du} = \frac{d}{n\beta},$$

the first order cumulant $\mathcal{T}_n'(0)$ can be calculated as

$$-\mathcal{T}_n'(0) = S_n + \mathbb{E}_w[K_n(w)] = S_n + \frac{d}{2n\beta} - \frac{||\xi_n||^2}{2n} + o_p(\frac{1}{n}).$$

Besides, from the law of large numbers,

$$\mathcal{T}_n''(0) = \mathcal{G}_n''(0) + o_p(\frac{1}{n}) = \frac{d}{n\beta} + o_p(\frac{1}{n}).$$

Hence, from Theorem 1,

$$\begin{aligned} T_n &= -\mathcal{T}_n(1) = -\mathcal{T}_n'(0) - \frac{1}{2}\mathcal{T}_n''(0) + o_p(\frac{1}{n}), \\ &= S_n + \frac{d}{2n\beta} - \frac{||\xi_n||^2}{2n} - \frac{d}{2n\beta} + o_p(\frac{1}{n}) = S_n - \frac{||\xi_n||^2}{2n} + o_p(\frac{1}{n}), \end{aligned}$$

which completes the proof. (Q.E.D.)

Lemma 7 *Assume that the true distribution $q(x)$ is realizable by a parametric model $p(x|w)$ and $(q(x), p(x|w))$ is in regular cases. Then, following holds.*

$$\mathbb{E}[||\xi_n||^2] = d + o(1)$$

Proof.

From Lemma 4, ξ_n converges to $N(0, J^{-1/2}IJ^{-1/2})$ in law. Hence, when $q(x)$ is realizable by $p(x|w)$ and $(q(x), p(x|w))$ is in regular cases, $\mathbb{E}[||\xi_n||^2]$ converges to $\mathrm{tr}(IJ^{-1})$, which is equal to d from Lemma 5. (Q.E.D.)

From Theorem 3 and Lemma 7, we obtain the following theorem.

Theorem 4 *Assume that the true distribution $q(x)$ is realizable by a parametric model $p(x|w)$ and $(q(x), p(x|w))$ is in regular cases. Then, expected value of G_n and T_n can be asymptotically described as*

$$\mathbb{E}[G_n] = S + \frac{d}{2n} + o(\frac{1}{n}), \quad \mathbb{E}[T_n] = S - \frac{d}{2n} + o(\frac{1}{n}).$$

From this theorem, it is shown that

$$\mathbb{E}[G_n] = \mathbb{E}[T_n] + \frac{d}{n} + o(\frac{1}{n})$$

holds. Hence, when $q(x)$ is realizable by $p(x|w)$ and $(q(x), p(x|w))$ is in regular cases, $T_n + \frac{d}{n}$ can be used as an asymptotically unbiased estimator of generalization loss G_n.

2.3.4 Statistical Learning Theory in singular cases

Here we briefly summarize the statistical learning theory in singular cases. Firstly, we state three fundamental theorems in singular learning theory without proof.

Theorem 5 (Hironaka's theorem) *Let $K(w) \geq 0$ be an nonnegative analytic function on an open set $W \subset \mathbb{R}^d$ and assume that there exists $w \in W$ which satisfies $K(w) = 0$. Then there exists a d-dimentional manifold $\mathcal{M}$ and an analytic map $g : \mathcal{M} \to W$ such that, on each local coordinate,*

$$K(g(u)) = u^{2k} = u_1^{2k_1} u_2^{2k_2} \cdots u_d^{2k_d},$$

$$|g^{'}(u)| = c(u)|u^h| = c(u)|u_1^{h_1} u_2^{h_2} \cdots u_d^{h_d}|,$$

where $|g^{'}(u)|$ represents Jacobian of $w = g(u)$, that is,

$$|g^{'}(u)| = |\det(\frac{\partial w_i}{\partial u_j})|$$

and $c(u) > 0$ is a positive analytic function.

Remark. (1) A pair of a manifold $\mathcal{M}$ and a mapping g, $(\mathcal{M},g)$, is called **resolution of singularities**.
(2) In general, resolution of singularities $(\mathcal{M},g)$ is not unique.
(3) In singular learning theory, Hironaka's theorem is applied to Kullback-Leibler distance $K(w)$.
(4) A map $g : \mathcal{M} \to W$ and multi-indices k, h depend on each local coordinate. When we emphasize such dependence on each local coordinate α, the notation such as k_α, h_α is employed.

The next theorem shows that, on each local coordinate in $\mathcal{M}$, the log density ratio function $f(x, g(u))$ can be divided by u^k.

Theorem 6 *Assume that the log density ratio function $f(x, g(u))$ is an analytic function of u. Then, there exists an analytic function $a(x, u)$ such that*

$$f(x, g(u)) = u^k a(x, u).$$

Remark. Note that $f(x, w)$ is assumed to be a $L^s(q)$-valued complex analytic function of w in the fundamental assumption in this research.

From this theorem, following empirical process is well-defined, which plays an important role in singular learning theory.

Definition 11 *An empirical process $\xi_n(u)$ is defined by*

$$\xi_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{u^k - a(X_i, u)\}$$

Remark. This empirical process is a stochastic process determined by a sample X^n. Its mean function is zero function and covariance function is

$$\mathbb{E}[\xi_n(u)\xi_n(v)] = \mathbb{E}_X[a(X, u)a(X, v)] - u^k v^k.$$

When $n \to \infty$, $\xi_n(u)$ converges to an gaussian process $\xi(u)$ in law which has the same mean function and covariance function with $\xi_n(u)$. We denote averaging operation with respect to $\xi(u)$ as $\mathbb{E}_\xi[\ ]$ below.

The third fundamental theorem is on the asymptotic expansion of Schwartz distribution, in which one of the important birational invariants, real log canonical threshold (RLCT) firstly appears. Before discussing the theorem, we introduce the concept of real log canonical threshold.

Definition 12 (Real log canonical threshold) *Let us consider the values*

$$\left(\frac{h_{\alpha,j} + 1}{2k_{\alpha,j}} \right) \quad (j = 1, 2, \cdots, d)$$

with respect to multi-indices $k_\alpha = (k_{\alpha,j}), h_\alpha = (h_{\alpha,j})$ for each local coordinate α. When $k_{\alpha,j} = 0$, we define the above value as $+\infty$. Then, we define real log canonical threshold λ as

$$\lambda = \min_{\alpha} \min_{j=1}^{d} \left\{ \frac{h_{\alpha,j} + 1}{2k_{\alpha,j}} \right\}$$

and define its multiplicity m as

$$m = \max_{\alpha} \sharp \left\{ j : \lambda = \frac{h_{\alpha,j} + 1}{2k_{\alpha,j}} \right\},$$

where $\sharp$ represents the number of elements included in the set.

Remark. (1) In the above, the concept of RLCT is defined by resolution of singularities of Kullback-Leibler distance $K(w)$. RLCT also can be defined by the zeta function of statistical learning as follows. The zeta function of statistical learning is defined by

$$\zeta(z) = \int K(w)^z \varphi(w) dw \quad (z \in \mathbb{C}),$$

where $\mathbb{C}$ is a set of the complex variables. Then $\zeta(z)$ is a holomorphic function on the region $Re(z) > 0$, which can be analytically continued to the unique meromorphic function on the entire complex plane [25]. All poles of the zeta function are real, negative, and rational numbers [25].

Definition 13 (Real Log Canonical Threshold) *If the largest pole of $\zeta(z)$ is $(-\lambda)$ $(\lambda > 0)$, then the real log canonical threshold is defined by λ. The order of the pole $z = -\lambda$ is referred to as the multiplicity m.*

(2) The real log canonical threshold is invariant under a birational transform

$$\begin{aligned} w &= g(w'), \\ p(x|w) &\mapsto p(x|g(w')), \\ \varphi(w) &\mapsto \varphi(g(w'))|g'(w')|, \end{aligned}$$

where $|g'(w')|$ is the Jacobian determinant. Such constants are called birational invariants.
(3) The real log canonical thresholds for several learning machines were clarified [2, 3, 16, 19, 20, 32, 33, 34, 35] using resolution of singularities.
(4) The real log canonical threshold is a well known birational invariant in algberaic geometry, which plays an important role in higher dimensional algberaic geometry [21].

Based on the concept of RLCT, we obtain the following theorem, which is the asymptotic expansion of Schwartz distribution.

Theorem 7 *When $t \to 0$, there exists an integral elements du^* such that*

$$\delta(t - u^{2k})|u^h|b(u)du = t^{\lambda-1}(-\log t)^{m-1}du^* + o(t^{\lambda-1}(-\log t)^{m-1}),$$

where we define du^* as

$$du^* = \left(\frac{1}{(m-1)!2^m \prod_{j=1}^m k_j} \right) \cdot \delta(u_a) u_b^{\mu} b(u) du$$

$\mu = \{\mu_j; j = m+1, \cdots, d\}$ is a multi-index which is defined by $\mu_j = -2\lambda k_j + h_j$.

Remark. (1) $(u_a, u_b) \in \mathbb{R}^m \times \mathbb{R}^{d-m}$, where u_a corresponds to variables u_j whose suffix j satisfies

$$\frac{h_j + 1}{2k_j} = \min_{j=1}^{d} \left\{ \frac{h_j + 1}{2k_j} \right\}$$

and u_b corresponds to other variables, respectively.
(2) In application, the term $|u^h|b(u)du$ corresponds to transformed prior distribution, that is, $|u^h|b(u)du = \varphi(g(u))|g^{'}(u)|du = |u^h|c(u)\varphi(g(u))du$.
(3) The proof of this theorem is conducted by using Mellin transform and inverse Mellin transform, which is one of the most basic techniques in singular learning theory [28, 29].

Based on the three fundamental theorems, we can clarify the asymptotic behavior of

$$\Omega(w)dw := \exp(-n\beta K_n(w))\varphi(w)dw.$$

Firstly, we introduce the standard form of empirical loss function $K_n(w)$.

Theorem 8 *For Kullback-Leibler distance $K(w)$, there exists a resolution of singularities $(\mathcal{M}, g)$, that is, on each local coordinate in $\mathcal{M}$, there exists a mapping $w = g(u)$ such that*

$$K_n(g(u)) = u^{2k} - \frac{1}{\sqrt{n}}u^k\xi_n(u),$$

$$\varphi(w)dw = \varphi(g(u))|g^{'}(u)|dw = |u^h|b(u)du.$$

Proof.

$$\begin{aligned} K_n(w) &= K(w) - (K(w) - K_n(w)) \\ &= K(w) - \frac{1}{n}\sum_{i=1}^{n}(K(w) - f(X_i, w)) \\ &= K(w) - \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(K(w) - f(X_i, w)) \end{aligned}$$

Then, from Theorem 5, there exists an analytic manifold $\mathcal{M}$ and a mapping $w = g(u)$ such that, on each local coordinate in $\mathcal{M}$,

$$\begin{aligned} K_n(g(u)) &= u^{2k} - \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{u^{2k} - u^k a(X_i, u)\} \\ &= u^{2k} - \frac{1}{\sqrt{n}} \cdot u^k \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{u^k - a(X_i, u)\} \\ &= u^{2k} - \frac{1}{\sqrt{n}} \cdot u^k\xi_n(u), \end{aligned}$$

which completes the proof. (Q.E.D.)

From Theorem 8, integral element $\Omega(w)dw$ can be represented as

$$\Omega(w)dw = \sum_{\alpha} \exp(-n\beta u^{2k_\alpha} + \sqrt{n}\beta u^{k_\alpha}\xi_n(u))|u^{h_\alpha}|\phi_\alpha(u)du,$$

where $\sum_\alpha$ represents the summation over local coordinates in $\mathcal{M}$, and k_α, h_α are multi-indices on each local coordinate in $\mathcal{M}$ and

$$\phi_\alpha(u) := b_\alpha(u)\varphi_\alpha(u),$$

where $\varphi_\alpha(u)$ represents partition of unity. Then, from Theorem 7, it turns out that integral element $\Omega(w)dw$ has the following asymptotic behavior.

Theorem 9

$$\Omega(w)dw = \frac{(\log n)^{m-1}}{n^\lambda} \sum_{\alpha^*} \int dt t^{\lambda-1} \exp(-\beta t + \beta\sqrt{t}\xi_n(u))du^* + o_p\left(\frac{(\log n)^{m-1}}{n^\lambda}\right),$$

where du^* represents

$$du^* = \left(\frac{1}{(m-1)!2^m \prod_{j=1}^m k_j}\right) \cdot \delta(u_a) u_b^\mu \phi_\alpha(u) du$$

and α^* is local coordinates which satisfy

$$\lambda = \min_\alpha \min_{j=1}^d \left\{\frac{h_{\alpha,j}+1}{2k_{\alpha,j}}\right\}, \quad m = \max_\alpha \sharp\left\{j : \lambda = \frac{h_{\alpha,j}+1}{2k_{\alpha,j}}\right\}.$$

Proof.

$$\begin{aligned}
\Omega(w)dw &= \sum_\alpha \exp(-n\beta u^{2k_\alpha} + \sqrt{n}\beta u^{k_\alpha}\xi_n(u))|u^{h_\alpha}|\phi_\alpha(u)du \\
&= \sum_\alpha \int_0^\infty d\tau\, \delta(\tau - u^{2k_\alpha}) u^{h_\alpha} \exp(-n\beta\tau + \sqrt{n\tau}\beta\xi_n(u))\phi_\alpha(u)du \\
&= \sum_\alpha \int_0^\infty \frac{dt}{n}\, \delta(\frac{t}{n} - u^{2k_\alpha}) u^{h_\alpha}\phi_\alpha(u)du\, \exp(-\beta t + \beta\sqrt{t}\xi_n(u))
\end{aligned}$$

Then, from Theorem 7,

$$\delta(\frac{t}{n} - u^{2k_\alpha})u^{h_\alpha}\phi_\alpha(u)du = (\frac{t}{n})^{\lambda-1}(-\log\frac{t}{n})^{m-1}du^* + \cdots = \frac{(\log n)^{m-1}}{n^{\lambda-1}} t^{\lambda-1}du^* + \cdots.$$

Hence,

$$\Omega(w)dw = \frac{(\log n)^{m-1}}{n^\lambda} \sum_{\alpha^*} \int dt t^{\lambda-1} \exp(-\beta t + \beta\sqrt{t}\xi_n(u))du^* + o_p\left(\frac{(\log n)^{m-1}}{n^\lambda}\right),$$

which completes the proof. (Q.E.D.)

From Theorem 9, we can derive the asymptotic behavior of $\mathbb{E}_w[\ ]$.

Definition 14 *For a given function $F(t,u)$, an averaging operator $\langle\ \rangle$ is defined by*

$$\langle F(t,u)\rangle = \frac{\sum_{\alpha^*}\int_{[0,1]^d} du^* \int_0^\infty dt F(t,u)t^{\lambda-1}\exp(-\beta t+\beta\sqrt{t}\xi_n(u))}{\sum_{\alpha^*}\int_{[0,1]^d} du^* \int_0^\infty dt t^{\lambda-1}\exp(-\beta t+\beta\sqrt{t}\xi_n(u))}.$$

Remark. $\mathbb{E}_w[\ ] = \langle\ \rangle(1+o_p(1))$ holds.

Between the parameter w and the variables (u,t), the following relation holds.

$$w = g(u), \quad K(w) = u^{2k} = \frac{t}{n}.$$

From this relation, it turns out that

$$f(x,g(u)) = u^k a(x,u) = \sqrt{\frac{t}{n}}a(x,u)$$

$$K_n(g(u)) = u^{2k} - \frac{1}{\sqrt{n}}u^k\xi_n(u) = \frac{1}{n}(t-\sqrt{t}\xi_n(u))$$

holds.

Theorem 10 *When $n\to\infty$, for an arbitrary positive real number $s\geq 0$,*

$$\mathbb{E}_w[f(x,w)^s] = \frac{1}{n^{\frac{s}{2}}}\left\langle(\sqrt{t}a(x,u))^s\right\rangle + o_p(\frac{1}{n^{\frac{s}{2}}}), \quad \langle t\rangle = \frac{\lambda}{\beta}+\frac{1}{2}\left\langle\sqrt{t}\xi_n(u)\right\rangle$$

holds.

Proof.

We can directly obtain the first equation from the relation

$$f(x,g(u)) = u^k a(x,u) = \sqrt{\frac{t}{n}}a(x,u).$$

Let us prove the second equation. From the definition,

$$\langle t\rangle = \frac{\sum_{\alpha^*}\int_{[0,1]^d} du^* \int_0^\infty dt\, t^{\lambda}\exp(-\beta t+\beta\sqrt{t}\xi_n(u))}{\sum_{\alpha^*}\int_{[0,1]^d} du^* \int_0^\infty dt\, t^{\lambda-1}\exp(-\beta t+\beta\sqrt{t}\xi_n(u))}.$$

Then, since

$$\begin{aligned}
&\int_0^\infty e^{-\beta t}t^\lambda e^{\beta\sqrt{t}\xi_n(u)} \\
&= -\frac{1}{\beta}\left[e^{-\beta t}t^\lambda e^{\beta\sqrt{t}\xi_n(u)}\right]_0^\infty + \frac{1}{\beta}\int_0^\infty e^{-\beta t}\frac{d}{dt}(t^\lambda e^{\beta\sqrt{t}\xi_n(u)}) \\
&= \frac{\lambda}{\beta}\int_0^\infty e^{-\beta t}t^{\lambda-1}e^{\beta\sqrt{t}\xi_n(u)} + \int_0^\infty e^{-\beta t}t^\lambda e^{\beta\sqrt{t}\xi_n(u)}\frac{\xi_n(u)}{2\sqrt{t}},
\end{aligned}$$

we obtain the second equation. (Q.E.D.)

Before discussing the asymptotic behavior of G_n and T_n, we introduce the concept of functional variance.

Definition 15 *For an empirical process $\xi_n(u)$, functional variance $V(\xi_n(u))$ is defined by*

$$V(\xi_n(u)) := \mathbb{E}_X[\langle ta(X,u)^2\rangle - \langle \sqrt{t}a(X,u)\rangle^2].$$

Finally, we have completed the preparation for deriving the asymptotic behavior of G_n and T_n based on Theorem 1.

Theorem 11 *Asymptotic behavior of G_n and T_n can be described as follows.*

$$G_n = S + \frac{1}{n}(\frac{\lambda}{\beta} + \frac{1}{2}\left\langle \sqrt{t}\xi_n(u)\right\rangle - \frac{1}{2}V(\xi_n)) + o_p(\frac{1}{n}),$$

$$T_n = S_n + \frac{1}{n}(\frac{\lambda}{\beta} - \frac{1}{2}\left\langle \sqrt{t}\xi_n(u)\right\rangle - \frac{1}{2}V(\xi_n)) + o_p(\frac{1}{n}).$$

Proof.

Firstly, we derive the asymptotic behavior of Bayes generalization loss G_n. According to Theorem 1, it is sufficient to clarify the asymptotic behavior of the first and second order cumulants. Then, from Theorem 10, following holds.

$$\begin{aligned}
\mathcal{G}_n^{'}(0) &= -S - \mathbb{E}_w[K(w)] = -S - \frac{1}{n}\langle t\rangle + o_p(\frac{1}{n}) \\
&= -S - \frac{1}{n}(\frac{\lambda}{\beta} + \frac{1}{2}\left\langle \sqrt{t}\xi_n(u)\right\rangle) + o_p(\frac{1}{n}) \\
\mathcal{G}_n^{''}(0) &= \mathbb{E}_X[\mathbb{E}_w[f(X,w)^2] - \mathbb{E}_w[f(X,w)]^2] \\
&= \frac{1}{n}\mathbb{E}_X[\langle ta(X,u)^2\rangle - \langle \sqrt{t}a(X,u)\rangle^2] + o_p(\frac{1}{n}) = \frac{1}{n}V(\xi_n) + o_p(\frac{1}{n})
\end{aligned}$$

Hence, from Theorem 1,

$$\begin{aligned}
G_n &= -\mathcal{G}_n^{'}(0) - \frac{1}{2}\mathcal{G}_n^{''}(0) + o_p(\frac{1}{n}) \\
&= S + \frac{1}{n}(\frac{\lambda}{\beta} + \frac{1}{2}\left\langle \sqrt{t}\xi_n(u)\right\rangle - \frac{1}{2}V(\xi_n)) + o_p(\frac{1}{n}).
\end{aligned}$$

Secondly, we derive the asymptotic behavior of Bayes training loss T_n. Also from Theorem 10, following holds.

$$\begin{aligned}
\mathcal{T}_n'(0) &= -S_n - \mathbb{E}_w[K_n(w)] = -S_n - \frac{1}{n}\left\langle t - \sqrt{t}\xi_n(u)\right\rangle + o_p(\frac{1}{n}) \\
&= -S_n - \frac{1}{n}(\frac{\lambda}{\beta} - \frac{1}{2}\left\langle \sqrt{t}\xi_n(u)\right\rangle) + o_p(\frac{1}{n}) \\
\mathcal{T}_n''(0) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_w[f(X_i,w)^2] - \mathbb{E}_w[f(X_i,w)]^2\right\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left\{\langle ta(X,u)^2\rangle - \langle\sqrt{t}a(X,u)\rangle^2\right\} + o_p(\frac{1}{n})
\end{aligned}$$

Then, since

$$\frac{1}{n}\sum_{i=1}^{n} a(X_i,u)a(X_i,v) = \mathbb{E}_X[a(X,u)a(X,v)] + o_p(1)$$

holds, the difference between $n\mathcal{G}_n''(0)$ and $n\mathcal{T}_n''(0)$ converges to zero in probability. Hence, from Theorem 1,

$$\begin{aligned}
T_n &= -\mathcal{T}_n'(0) - \frac{1}{2}\mathcal{T}_n''(0) + o_p(\frac{1}{n}) \\
&= S_n + \frac{1}{n}(\frac{\lambda}{\beta} - \frac{1}{2}\left\langle\sqrt{t}\xi_n(u)\right\rangle - \frac{1}{2}V(\xi_n)) + o_p(\frac{1}{n}),
\end{aligned}$$

which completes the proof. (Q.E.D.)

Then, from Theorem 2, we obtain the following lemma.

Lemma 8 *For an gaussian process $\xi(u)$, the following equation holds.*

$$\mathbb{E}_\xi[\langle\sqrt{t}\xi(u)\rangle] = \beta\mathbb{E}_\xi[V(\xi)]$$

Proof.

From Theorem 2,

$$\mathbb{E}[\mathcal{G}_{n-1}'(0) + \frac{\beta}{2}\mathcal{G}_{n-1}''(0)] = \mathbb{E}[\mathcal{T}_n'(0) - \frac{\beta}{2}\mathcal{T}_n''(0)] + o(\frac{1}{n}).$$

holds. Then, since the differences between $\mathcal{G}_{n-1}'(0)$ and $\mathcal{G}_n'(0)$ and between $\mathcal{G}_{n-1}''(0)$ and $\mathcal{G}_n''(0)$ are smaller than $o_p(\frac{1}{n})$,

$$\mathbb{E}[\mathcal{G}_n'(0) + \frac{\beta}{2}\mathcal{G}_n''(0)] = \mathbb{E}[\mathcal{T}_n'(0) - \frac{\beta}{2}\mathcal{T}_n''(0)] + o(\frac{1}{n}).$$

holds. Hence,

$$-S - \frac{1}{n}\frac{\lambda}{\beta} - \frac{1}{2n}\mathbb{E}_{\xi_n}[\langle\sqrt{t}\xi_n(u)\rangle] + \frac{\beta}{2n}\mathbb{E}_{\xi_n}[V(\xi_n)] =$$
$$-S - \frac{1}{n}\frac{\lambda}{\beta} + \frac{1}{2n}\mathbb{E}_{\xi_n}[\langle\sqrt{t}\xi_n(u)\rangle] - \frac{\beta}{2n}\mathbb{E}_{\xi_n}[V(\xi_n)] + o(\frac{1}{n}),$$

that is,

$$\mathbb{E}_{\xi_n}[\langle\sqrt{t}\xi_n(u)\rangle] = \beta\mathbb{E}_{\xi_n}[V(\xi_n)] + o(1).$$

Then, since $\xi_n(u) \to \xi(u)$ in law,

$$\mathbb{E}_{\xi}[\langle\sqrt{t}\xi(u)\rangle] = \beta\mathbb{E}_{\xi}[V(\xi)],$$

which completes the proof. (Q.E.D.)

From the above lemma, we can introduce the concept of singular fluctuation.

Definition 16 *The constant ν defined by*

$$2\nu := \mathbb{E}_{\xi}[\langle\sqrt{t}\xi(u)\rangle] = \beta\mathbb{E}_{\xi}[V(\xi)]$$

is called **singular fluctuation.**

Remark. (1) In general, singular fluctuation ν depends on β.
(2) It is known that singular fluctuation ν is also a birational invariant.

By introducing the constant ν, we can describe the expected Bayes generalization loss and training loss using two birational invariants λ and ν as follows.

Theorem 12 *Expectation values of G_n and T_n can be asymptotically described as follows.*

$$\mathbb{E}[G_n] = S + \frac{1}{n}(\frac{\lambda-\nu}{\beta} + \nu) + o(\frac{1}{n}),$$
$$\mathbb{E}[T_n] = S + \frac{1}{n}(\frac{\lambda-\nu}{\beta} - \nu) + o(\frac{1}{n}).$$

2.4 Unbiased estimator of Bayes Generalization Loss

2.4.1 Equation of States in Statistical Learning

Definition 17 *We define a random variable V_n as*

$$V_n := \sum_{i=1}^{n}\{\mathbb{E}_w[(\log p(X_i|w))^2] - \mathbb{E}_w[\log p(X_i|w)]^2\}$$

and call it also as functional variance.

Lemma 9 *When $n \to \infty$, the following holds.*

$$\lim_{n\to\infty} \beta\mathbb{E}[V_n] = 2\nu$$

From Lemma 9, the following theorem holds.

Theorem 13 *For an arbitrary triple $(q(x), p(x|w), \varphi(w))$, the following relation holds among G_n, T_n and V_n.*

$$\mathbb{E}[G_n] = \mathbb{E}[T_n + \frac{\beta V_n}{n}] + o(\frac{1}{n}).$$

Remark. The above equation is called *Equation of States in Statistical Learning*, which always holds in both regular and singular cases.

2.4.2 A Widely Applicable Information Criterion : WAIC

From the *Equation of States in Statistical Learning*, the following asymptotically unbiased estimator of Bayes generalization loss is proposed [26].

Definition 18 *A widely applicable information criterion ($WAIC$) is defined by*

$$WAIC = T_n + \frac{\beta V_n}{n}.$$

Remark. From the *Equation of States in Statistical Learning*,

$$\mathbb{E}[G_n] = \mathbb{E}[WAIC] + o(\frac{1}{n})$$

holds.

Chapter 3

Statistical Learning Theory of Quasi-Regular Cases

In this chapter, we introduce a new concept in statistical learning theory, *quasi-regular case* and explain its basic theory. Quasi-regular case is included in singular cases in general but has similar properties as regular cases. The concept of quasi-regular case was discovered through investigating singular fluctuation defined in chapter 2. In section 3.1, we explain our motivation for developing the concept of quasi-regular case. In section 3.2, we define quasi-regular case and show some examples. In section 3.3, we describe theoretical foundation of quasi-regular case. In section 3.4, we prove the main theorem in this chapter, which provides the first example of singular cases for which the exact value of singular fluctuation is clarified.

3.1 Motivation for Quasi-Regular Cases

In chapter 2, we explained that asymptotic behavior of expected Bayes generalization loss and training loss was determined by two birational invariants, real log canonical threshold λ and singular fluctuation ν. As stated in chapter 2, real log canonical threshold, RLCT, is a well known birational invariant in algebraic geometry. Also in singular learning theory, the exact value of RLCT has been discussed for many learning machines. However, in contrast, singular fluctuation, SF, is considered to be firstly discovered in singular learning theory and its mathematical properties are left totally unknown. It is no doubt that SF is an important birational invariant for singular learning theory, because it determines the asymptotic difference between expected Bayes generalization loss and training loss, that is,

$$\mathbb{E}[G_n] = \mathbb{E}[T_n] + \frac{2\nu}{n} + o(\frac{1}{n}).$$

Hence, it is a challenging problem in singular learning theory to investigate the mathematical properties of singular fluctuation, in other words, to clarify what the singular fluctuation is mathematically. In this thesis, we focus on clarifying the exact value of singular fluctuation in singular cases. Since it seems difficult to consider on singular fluctuation $\nu = \nu(\beta)$ in general singular cases, we begin with a restricted class of singular cases. As the first step, we propose the concept of quasi-regular cases.

3.2 Definition of quasi-regular cases and its examples

3.2.1 Definition of quasi-regular cases

Here we state the definition of quasi-regular cases and show some examples. Note that we assume the true distribution $q(x)$ is realizable by a parametric model $p(x|w)$. Then, we can assume $q(x) = p(x|0)$ without loss of generality.

Definition 19 (Quasi-Regular Case) *Let g be a natural number and $d_0, d_1, \cdots, d_g$ be integers satisfying*

$$0 = d_0 < d_1 < \cdots < d_g = d.$$

We define a function $u = (u_1, u_2, ..., u_g) \in \mathbb{R}^g$ of the paramater $w \in \mathbb{R}^d$ by

$$\begin{aligned} u_1 &= \prod_{j=1}^{d_1} w_j, \\ u_2 &= \prod_{j=d_1+1}^{d_2} w_j, \\ \cdots &= \cdots, \\ u_g &= \prod_{j=d_{g-1}+1}^{d_g} w_j. \end{aligned}$$

If there exist constants $c_1, c_2 > 0$ such that, for arbitrary $w \in W$,

$$c_1(u_1^2 + \cdots + u_g^2) \leq K(w) \leq c_2(u_1^2 + \cdots + u_g^2),$$

then the pair $(q(x), p(x|w))$ is said to be in a quasi-regular case.

Remark. (1) If $g = d$, then

$$\{w; q(x) = p(x|w)\} = \{0\}$$

and Fisher information matrix at 0, which is equal to $\nabla^2 K(0) = (\frac{\partial^2 K(0)}{\partial w_i \partial w_j})_{1 \leq i,j \leq d}$, is positive definite, hence the quasi-regular case corresponds to the regular case.

We can prove positive-definiteness of $\nabla^2 K(0)$ as follows. Firstly, from the assumption of $q(x) = p(x|0)$, $K(0) = 0$ holds. In addition, since $K(w)$ is non-negative in the neighborhood of $w = 0$, $\nabla K(0) = 0$ holds and $\nabla^2 K(0)$ is positive semi-definite. Then, from the Taylor's theorem, there exists $0 < \theta < 1$ such that

$$K(w) = \frac{1}{2} w^T \nabla^2 K(0) w + \frac{1}{6} \sum_{1 \leq i,j,k \leq d} \frac{\partial^3 K(\theta w)}{\partial w_i \partial w_j \partial w_k} w_i w_j w_k.$$

Assume $\nabla^2 K(0)$ is not positive definite, that is, $\nabla^2 K(0)$ has eigenvalue 0 and corresponding eigenvector $v = (v_i)_{i=1}^d \neq 0$. Then, for arbitrary $t \neq 0$,

$$K(tv) = \frac{t^3}{6} \sum_{1 \leq i,j,k \leq d} \frac{\partial^3 K(\theta(tv))}{\partial w_i \partial w_j \partial w_k} v_i v_j v_k.$$

Since $(q(x), p(x|w))$ is in a quasi-regular case,

$$c_1 t^2 ||v||^2 \leq \frac{t^3}{6} \sum_{1 \leq i,j,k \leq d} \frac{\partial^3 K(\theta(tv))}{\partial w_i \partial w_j \partial w_k} v_i v_j v_k.$$

By dividing both sides by $t^2 \neq 0$, we obtain

$$c_1 ||v||^2 \leq \frac{t}{6} \sum_{1 \leq i,j,k \leq d} \frac{\partial^3 K(\theta(tv))}{\partial w_i \partial w_j \partial w_k} v_i v_j v_k.$$

When $t \to 0$, we obtain $c_1 ||v||^2 \leq 0$. Since $v \neq 0$, this is a contradiction. Therefore, a quasi-regular case contains a regular case as a special one.

(2) If $g \neq d$, then "$K(w) = 0 \Longleftrightarrow w = 0$" does not hold, because, for at least one variable w_j, $K(0, 0, .., w_j, 0, .., 0) = 0$. Hence a quasi-regular case with $g \neq d$ is not a regular case but a singular case.

(3) There are singular cases which are not contained in quasi-regular cases. Therefore,

$$\text{Regular} \subsetneq \text{Quasi-Regular} \subsetneq \text{Singular}$$

holds. We show in Theorem 14 that a quasi-regular case is not a regular case, however, it has the same property as a regular case.

3.2.2 Examples of quasi-regular cases

In this section, we show some examples of $(q(x), p(x|w))$ included in quasi-regular cases.

Example.1 Let a parametric model

$$p(x, y|w) = \frac{r(x)}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(y - a \tanh(bx))^2\},$$

where $w = (a, b)$ is the parameter and $r(x)$ is the probability density function of x. If the true distribution is given by $q(x, y) = p(x, y|0, 0)$, then

$$K(w) = \frac{1}{2}\int \{a \tanh(bx)\}^2 r(x) dx$$

holds and there exist constants $c_1, c_2 > 0$ such that

$$c_1(ab)^2 \leq K(w) \leq c_2(ab)^2.$$

Hence, this case satisfies the condition for a quasi-regular case with $g = 1$.

Example.2 Let a parametric model be

$$p(x, y|w) = \frac{r(x)}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(y - ax^2 - b\tanh(cx))^2\},$$

where $w = (a, b, c)$ is the parameter and $r(x)$ is the probability density function of x. If the true distribution is given by $q(x, y) = p(x, y|0, 0, 0)$, then by using

$$u_1 = a, \quad u_2 = bc,$$

it follows that

$$K(w) = \frac{1}{2}\int \{ax^2 + b\tanh(cx)\}^2 r(x) dx$$

satisfies the condition for a quasi-regular case with $g = 2$, because x^2 and $\tanh(cx)/c$ is linearly independent. In fact there exist $c_1, c_2 > 0$ such that

$$c_1(a^2 + (bc)^2) \leq K(w) \leq c_2(a^2 + (bc)^2).$$

Hence the set of true parameters consists of the union of two lines,

$$\{w; q(x, y) = p(x, y|w)\} = \{a = 0, bc = 0\}.$$

Example.3 Let a parametric model be

$$p(x, y|w) = \frac{r(x)}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(y - ax - b\tanh(cx))^2\},$$

where $w = (a, b, c)$ is the parameter and the true distribution is given by $q(x, y) = p(x, y|0, 0, 0)$. Then because x and $\tanh(cx)/c$ is not linearly independent as $c \to 0$, hence this case does not satisfies the quasi-regular condition. In this case

$$c_1\{(a + bc)^2 + b^2c^6\} \leq K(w) \leq c_2\{(a + bc)^2 + b^2c^6\}.$$

Example.3 resembles Example.2, however, from the viewpoint of statistical learning theory, they are different.

Example.4 Let $X = (X_1, X_2)$ and Y are random variables which take values in {-1,1}. Then, a parametric model $p(x, y|w)$ is given by

$$p(x, y|w) = \frac{1}{Z(a, b)} \exp(ax_1 y + bx_2 y),$$

where $w = (a, b, c)$ and $Z(a, b)$ is a normalizing constant. The true distribution is given by $q(x, y) = p(x, y|0, 0) = \frac{1}{4}$. Such a model is often called naive Bayes network model. In this case, it is shown that there exist $c_1, c_2 > 0$ such that

$$c_1(ab)^2 \leq K(w) \leq c_2(ab)^2,$$

Hence, this case satisfies the condition for a quasi-regular case with $g = 1$.

Example.5 Let a parametric model be

$$p(x, y, z|w) = \frac{r(x, y)}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(z - f(x, y, w))^2\},$$

where

$$\begin{aligned} f(x, y, w) &= a_1 \sin(b_1 x) + a_2 x \sin(b_2 x) \\ &\quad + a_3 \sin(b_3 y) + a_4 y \sin(b_4 y), \end{aligned}$$

and $w = \{(a_i, b_i)\}_{i=1}^4$ is the parameter and the true distribution is given by $q(x, y, z) = p(x, y, z|0)$. Then $(q(x, y, z), p(x, y, z|w))$ is in a quasi-regular case with $g = 4$.

3.3 Theoretical foudation of quasi-regular cases

In this section, we explain the theoretical foundation of quasi-regular cases.

Note that we assume the true distribution $q(x)$ is realizable by a parametric model $p(x|w)$. The following lemma shows that the log density ratio function of the quasi-regular case is represented by g linearly independent functions.

Lemma 10 *Assume that the pair $(q(x), p(x|w))$ is in a quasi-regular case. Then there exists a set of functions $\{e_j(x, u); j = 1, 2, ..., g\}$ which are analytic functions of u and*

$$f(x, w) = \sum_{j=1}^{g} u_j e_j(x, u)$$

in an open neighborhood of $u = 0$.

(Proof) Let us define a function

$$F(t) = t + e^{-t} - 1$$

for $t \in \mathbb{R}^1$. Then $F(0) = 0$, $F'(0) = 0$, and $F''(0) = 1$, resulting that $F(t) \geq 0$ and that $F(t) = 0$ if and only if $t = 0$. Then,

$$\begin{aligned} K(w') &= \int q(x) F\Big(\log \frac{q(x)}{p(x|w')}\Big) dx \\ &= \int q(x) F(f(x,w')) dx \\ &= \frac{1}{2} \int f(x,w')^2 e^{-t^* f(x,w')} q(x) dx, \end{aligned} \tag{3.1}$$

where $0 < t^* < 1$ holds. Moreover,

$$\begin{aligned} & \frac{1}{2} \int f(x,w')^2 e^{-t^* f(x,w')} q(x) dx \\ \leq & \frac{1}{2} \int M(x)^2 \max_{w'} \{1, \frac{p(x|w')}{q(x)}\} q(x) dx \\ = & \frac{1}{2} \int M(x)^2 \max_{w'} \{q(x), p(x|w')\} dx \\ \leq & \frac{1}{2} \int M(x)^2 Q(x) dx \end{aligned}$$

holds in the neighborhood of $w' = w$ such that $K(w) = 0$. From the fundamental assumption (2), $M(x)^2 Q(x)$ is an integrable function, hence eq.(3.1) is bounded by the integrable function. By using Lebesgue's convergence theorem for $w' \to w$ such that $K(w) = 0$, we obtain

$$0 = \frac{1}{2} \int f(x,w)^2 q(x) dx. \tag{3.2}$$

Hence, by the assumption of the quasi-regular case, $K(w) = 0$ if and only if $u_1 = u_2 = \cdots = u_g = 0$, which is equivalent to $f(x,w) \equiv 0$. That is to say, $f(x,w)$ is contained in the ideal of analytic functions generated by $u_1, u_2, ..., u_g$. Hence there exist a set $\{e_j(x,u)\}$ of analytic functions of u, which satisfies

$$f(x,w) = \sum_{j=1}^{g} u_j e_j(x,u).$$

Therefore, we obtained the Lemma. (Q.E.D.)

In the following lemma, we show that the quasi-regular case has the generalized Fisher information matrix. $\cdot$ or $(\ \cdot\ )$ denote the inner product of vectors below.

Lemma 11 *The $g \times g$ matrix $I(u)$ is defined by*

$$I_{ij}(u) \equiv \int q(x) e_i(x,u) e_j(x,u) dx.$$

Then $I(u)$ is positive definite in an open neighborhood of $u = 0$.

(Proof) By Lemma 10 and eq.(3.2), in the neighborhood of $u = 0$,

$$K(w) = \frac{1}{2}(u \cdot I(u)u).$$

By the condition of the quasi-regular case,

$$c_1 \sum_{j=1}^{g} u_j^2 \leq K(w).$$

Hence the minimum eigenvalue of $I(u)$ is positive, which shows $I(u)$ is positive definite. (Q.E.D.)

Remark. When $(q(x), p(x|w))$ is in regular cases, analytic functions $\{e_j(x,u)\}$ corresponds to $\{-\frac{\partial}{\partial w_j} \log p(x|w)\}$, hence $I(u)$ is equal to Fisher information matrix defined in Definition 10.

The following definition and lemma show that the empirical loss function of the quasi-regular case has the same decomposition as that of the regular and realizable case.

Definition 20 *A random process $\xi_n(u) \in \mathbb{R}^g$ is defined by*

$$\xi_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\frac{1}{2} I(u)u - e(X_i, u), \}$$

where

$$e(x,u) = (e_1(x,u), e_2(x,u), ..., e_g(x,u))^T.$$

Lemma 12 *The empirical loss function defined by*

$$K_n(w) = \frac{1}{n} \sum_{i=1}^{n} f(X_i, w)$$

is represented by

$$K_n(w) = \frac{1}{2}(u \cdot I(u)u) - \frac{1}{\sqrt{n}}\, u \cdot \xi_n(u)$$

in the neighborhood of $u = 0$. Moreover, the random process $\xi_n(u)$ converges to the gaussian process $\xi(u)$ that satisfies

$$\mathbb{E}[\xi(0) \cdot I(0)^{-1} \xi(0)] = g.$$

(Proof) The empirical loss function is given by

$$K_n(w) = K(w) - \frac{1}{n} \sum_{i=1}^{n} \{K(w) - f(X_i, w)\}.$$

By combining this equation with the definition of $\xi_n(u)$, the first half of the Lemma is obtained. For the second half, the convergence of $\xi_n(u)$ is derived from Theorem 5.9 and 5.10 in [28] under the fundamental assumption (1). Moreover,

$$\begin{aligned}&\mathbb{E}[\xi_n(0) \cdot I(0)^{-1}\xi_n(0)] \\ &= \mathbb{E}[\mathrm{tr}(I(0)^{-1}\xi_n(0)\xi_n(0)^T)] = g,\end{aligned}$$

where we used the covariance matrix of $\xi_n(0)$

$$\mathbb{E}[\xi_n(0)\xi_n(0)^T] = \int q(x)e(x,0)e(x,0)^T dx = I(0),$$

which completes the proof of the Lemma. (Q.E.D.)

In the quasi-regular case, the relation between $w = (w_1, w_2, ..., w_d)$ and $u = (u_1, u_2, ..., u_g)$ is important. The following lemma shows the essential property of the quasi-regular case. This lemma does not hold in general singular cases.

Lemma 13 *When n tends to infinity,*

$$\prod_{j=1}^{g} \delta\Big(\frac{u_j}{\sqrt{n}} - \prod_{k=d_{j-1}+1}^{d_j} w_k\Big) \cong c_3(\log n)^{m-1} \prod_{j=1}^{d} \delta(w_j)$$

where $m = d - g + 1$ and $c_3 > 0$ is a constant.

(Proof) Firstly, we prove that the delta function with variables $\mathbf{x} = (x_1, x_2, ..., x_d)$ in $M \equiv [0,1]^d$

$$D(t, \mathbf{x}) = \delta(t - x_1 x_2 \cdots x_d)$$

has asymptotic expansion for $t \to 0$,

$$D(t, \mathbf{x}) = \frac{(-\log t)^{d-1}}{(d-1)!} \prod_{k=1}^{d} \delta(x_k) + o((-\log t)^{d-2}). \tag{3.3}$$

Let $\phi(\mathbf{x})$ be an arbitrary C^∞-class function of $\mathbf{x}$ whose support is contained in M.

$$D_t(\phi) \equiv \int_M D(t, \mathbf{x})\phi(\mathbf{x})d\mathbf{x}.$$

Then its Mellin transform is

$$\int D_t(\phi)t^z dt = \int_M \prod_{i=1}^{d} (x_i)^z \phi(\mathbf{x})d\mathbf{x}.$$

From the mean-value theorem, there exists $\mathbf{x}^*$ such that

$$\phi(\mathbf{x}) = \phi(0) + \mathbf{x} \cdot \nabla\phi(\mathbf{x}^*),$$

where $\mathbf{x}^* = \theta\mathbf{x}$ for some $0 < \theta < 1$. Then,

$$\begin{aligned} & \int_M \prod_{i=1}^d (x_i)^z \phi(\mathbf{x}) d\mathbf{x} \\ = & \int_M \prod_{i=1}^d (x_i)^z (\phi(0) + \mathbf{x} \cdot \nabla\phi(\mathbf{x}^*)) d\mathbf{x} \\ = & \frac{\phi(0)}{(z+1)^d} + \sum_{j=1}^d \int_M \prod_{i=1}^d (x_i)^z x_j \frac{\partial \phi}{\partial x_j}(\mathbf{x}^*) d\mathbf{x}. \end{aligned} \tag{3.4}$$

Let us describe the second term as $\Phi(z)$. Then we can prove that the maximal pole of $\Phi(z)$ is (-1) and its multiplicity is not larger than $d-1$. Note that all poles of $\Phi(z)$ is on the real axis from the property of zeta function. Then, for real $z > -1$,

$$|\Phi(z)| \leq d \max_{x^* \in M, 1 \leq j \leq d} \left| \frac{\partial \phi}{\partial x_j}(\mathbf{x}^*) \right| \times \frac{1}{(z+1)^{d-1}(z+2)}.$$

Hence, the maximal pole of the function (3.4) is (-1) with multiplicity d. Therefore, we have the asymptotic expansion,

$$\int D_t(t, \mathbf{x}) t^z dt = \frac{1}{(z+1)^d} \prod_{k=1}^d \delta(x_k) + \cdots$$

for $\mathbf{x} \in [0,1]^d$. By using inverse Mellin transform, we obtain eq.(3.3). Secondly, let us prove the Lemma. By using eq.(3.3), for each u_j,

$$\delta(\frac{u_j}{\sqrt{n}} - \prod_{k=d_{j-1}+1}^{d_j} w_k) \propto (\log n)^{d_j - d_{j-1} - 1} \prod_{j=d_{j-1}+1}^{d_j} \delta(w_j)$$

when $n \to \infty$. By summing up these relations for $j = 1, 2, ..., g$, Lemma is obtained. (Q.E.D.)

Remark. In the proof of asymptotic expansion of $D(t, \mathbf{x})$ as $t \to 0$, we constrained the set M to be $[0,1]^d$. It is enough to prove the asymptotic expansion when $M \equiv [0,1]^d$ from the following reason. Firstly, since $\varphi(w)$ has the support which is compact, we can assume that there exists a posive constant $M^{'} > 0$ such that the support of $\phi(\mathbf{x})$ is contained in $[-M^{'}, M^{'}]^d$. Then

$$D_t(\phi) = \left\{ \int_{[-1,1]^d} + \int_{[-M',M']^d \setminus [-1,1]^d} \right\} D(t, \mathbf{x}) \phi(\mathbf{x}) d\mathbf{x}$$

holds and the second term has the lower order value than the first term asymptotically (Example 4.7 in [28]). Secondly, since $\delta(\mathbf{x}) = \delta(-\mathbf{x})$ holds, it is enough to consider the asymptotic expansion of $D(t, \mathbf{x})$ on $[0,1]^d$.

3.4 Two Birational invariants in quasi-regular cases

The following is the main theorem in this chapter.

Theorem 14 (Main Theorem). *Assume that the pair $(q(x), p(x|w))$ is in a quasi-regular case and that $\varphi(w) > 0$ on W. Then the real log canonical threshold and the singular fluctuation are given by*

$$\lambda = \nu = \frac{g}{2}$$

and

$$m = d - g + 1.$$

Corollary 1 *Assume that the pair $(q(x), p(x|w))$ is in a quasi-regular case and that $\varphi(w) > 0$ on W. For arbitrary $0 < \beta < \infty$ the symmetry of the generalization and training errors holds,*

$$\begin{aligned} \mathbb{E}[G_n^{(0)}] &= \frac{g}{2n} + o(\frac{1}{n}), \\ \mathbb{E}[T_n^{(0)}] &= -\frac{g}{2n} + o(\frac{1}{n}). \end{aligned}$$

Remarks.(1) The above theorem shows the generalization and training errors for Bayes estimation. In the quasi-regular case, they have the same property as those in regular cases, however, the generalization and training errors of the maximum likelihood estimation is different from regular case in general.

(2) In the maximum likelihood method, the training error of a singular case is far smaller than that of a regular case, whereas the generalization error of a singular case is far larger than that of a regular case. From the viewpoint of the maximum likelihood method, the quasi-regualr case is contained in the singular case. Here we prove that the quasi-regular case has the same property as the regular case from the viewpoint of the Bayes estimation.

In this section, we prove Theorem 14. At first, we derive the real log canonical threshold of the quasi-regular case.

Lemma 14 *Assume that the pair $(q(x), p(x|w))$ is in a quasi-regular case. The real log canonical threshold and its multiplicity are given by $\lambda = g/2$ and $m = d - g + 1$ respectively.*

(Proof) Firstly, since $(q(x), p(x|w))$ is in a quasi-regular case, in other words,

$$c_1(u_1^2 + \cdots + u_g^2) \leq K(w) \leq c_2(u_1^2 + \cdots + u_g^2),$$

holds for some $c_1, c_2 > 0$, it is enough to consider the function

$$\int (u_1^2 + \cdots + u_g^2)^z \varphi(w) dw$$

as zeta function $\zeta(z)$ (Remark 7.2 in [28]). Hence we consider $u_1^2 + \cdots + u_g^2$ instead of the original $K(w)$. We assumed that $\varphi(w) > 0$ $\forall w \in W$ and that W is compact, therefore, $\varphi(w)$ can be bounded by constants. Hence we can replace $\varphi(w)$ in the definition of zeta function by a constant without changing its maximal pole and multiplicity. Then, in general, if $w = (w_a, w_b)$ and $K(w)$ can be represented by

$$K(w) = K(w_a) + K(w_b),$$

then the maximal pole $(-\lambda)$ of $\zeta(z)$ and its multiplicity m satisfy

$$\lambda = \lambda_a + \lambda_b,$$

$$m = m_a + m_b - 1,$$

where $(-\lambda_a, m_a)$ and $(-\lambda_b, m_b)$ are the maximal poles and multiplicities of zeta function corresponding to $K(w_a)$ and $K(w_b)$ respectively (Remark 7.2 in [28]). Since each function $\{u_j; j = 1, 2, ..., g\}$ does not have common variable w_k, the real log canonical threshold is given by the sum of individual real log canonical thresholds for zeta function $\zeta_j(z)$ defined by

$$\zeta_j(z) = \int \prod_{i=d_{j-1}+1}^{d_j} (w_i)^{2z} dw_{d_{j-1}+1} \cdots dw_{d_j}.$$

The maximal pole of each $\zeta_j(z)$ and its multiplicity are $-1/2$ and $d_j - d_{j-1}$. Hence λ is equal to g times $1/2$, that is, $\lambda = g/2$. The multiplicity is also given by

$$\begin{aligned} m &= d_1 + d_2 - d_1 + \cdots + d_g - d_{g-1} - (g-1) \\ &= d - g + 1, \end{aligned}$$

which shows the Lemma. (Q.E.D.)

Let us return to the proof of the Main theorem.

(Proof of Main Theorem) It was proved by eq.(6.4) in [28] that the expectation value of $K_n(w)$ is given by two birational invariants,

$$\mathbb{E}[\mathbb{E}_w[K_n(w)]] = \frac{\lambda}{n\beta} - \frac{\nu}{n} + o(\frac{1}{n}).$$

Since we have already obtained the value of λ in Lemma 14, that is to say, $\lambda = g/2$, we can derive the value of ν by calculating $\mathbb{E}[\mathbb{E}_w[K_n(w)]]$. The posterior distribution is represented by the empirical loss function by

$$p(w|X^n) \propto \exp(-n\beta K_n(w))\varphi(w)dw.$$

The integration of the outside of the neighborhood of $u = 0$ with respect to the posterior distribution goes to zero with the smaller order than $\exp(-\sqrt{n})$ as Lemma 6.3 in [28], hence we can restrict the integrated region to the neighborhood of $u = 0$. The empirical loss function is rewritten as

$$\begin{aligned} K_n(w) &= \frac{1}{2}\|I(u)^{\frac{1}{2}}\Big(u - I(u)^{-1}\frac{\xi_n(u)}{\sqrt{n}}\Big)\|^2 \\ &\quad -\frac{1}{2n}(\xi_n(u)\cdot I(u)^{-1}\xi_n(u)). \end{aligned}$$

In the neighborhood of $u = 0$, we obtain

$$\begin{aligned} K_n(w) &\cong \frac{1}{2}\|I(0)^{\frac{1}{2}}\Big(u - I(0)^{-1}\frac{\xi_n(0)}{\sqrt{n}}\Big)\|^2 \\ &\quad -\frac{1}{2n}(\xi_n(0)\cdot I(0)^{-1}\xi_n(0)). \end{aligned}$$

By using Lemma 13, for an arbitrary function $F(\ )$,

$$\begin{aligned} &\int F(\sqrt{n}\,u)dw \\ &= \int F(\sqrt{n}\,u)\prod_{j=1}^{g}\delta\Big(u - \prod_{k=d_{j-1}+1}^{d_j} w_k\Big)dwdu \\ &= \int F(u)\prod_{j=1}^{g}\delta\Big(\frac{u}{\sqrt{n}} - \prod_{k=d_{j-1}+1}^{d_j} w_k\Big)dw\frac{du}{n^{g/2}} \\ &= \frac{c_3(\log n)^{m-1}}{n^{g/2}}\int F(u)du + o\left(\frac{(\log n)^{m-1}}{n^{g/2}}\right). \end{aligned} \tag{3.5}$$

On the other hand,

$$\begin{aligned} nK_n(w) &= \frac{1}{2}\|I(0)^{1/2}\Big(\sqrt{n}\,u - I(0)^{-1}\xi_n(0)\Big)\|^2 \\ &\quad -\frac{1}{2}(\xi_n(0)\cdot I(0)^{-1}\xi_n(0)) \\ &\equiv \hat{K}_n(\sqrt{n}\,u). \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}_w[K_n(w)] &= \frac{\int K_n(w)\exp(-n\beta K_n(w))\varphi(w)dw}{\int \exp(-n\beta K_n(w))\varphi(w)dw} \\ &= \frac{1}{n}\frac{\int \hat{K}_n(\sqrt{n}\,u)\exp(-\beta\hat{K}_n(\sqrt{n}\,u))\varphi(w)dw}{\int \exp(-\beta\hat{K}_n(\sqrt{n}\,u))\varphi(w)dw} \end{aligned}$$

Then, by using eq.(3.5),

$$= \frac{1}{n} \frac{\int \hat{K}_n(u) \exp(-\beta \hat{K}_n(u)) du}{\int \exp(-\beta \hat{K}_n(u)) du}$$
$$= \frac{1}{2n} \frac{\int \|I(0)^{\frac{1}{2}}(u - \xi_n^*)\|^2 \exp(-\beta \hat{K}_n(u)) du}{\int \exp(-\beta \hat{K}_n(u)) du}$$
$$- \frac{1}{2n} (\xi_n(0) \cdot I(0)^{-1} \xi_n(0)),$$

where the notation

$$\xi_n^* = I(0)^{-1} \xi_n(0)$$

is used. Finally, by the integral formlula

$$\frac{\int \|I(0)^{1/2} u\|^2 \exp(-\frac{\beta}{2} \|I(0)^{1/2} u\|^2) du}{\int \exp(-\frac{\beta}{2} \|I(0)^{1/2} u\|^2) du} = \frac{g}{\beta}$$

and by Lemma 4, we have

$$\mathbb{E}[\mathbb{E}_w[K_n(w)]] = \frac{g}{2\beta n} - \frac{g}{2n} + o(\frac{1}{n}),$$

Then, because $\lambda = g/2$ holds from Lemma 14, we obtain the Theorem. (Q.E.D.)

3.5 Conclusion

A new concept in statistical learning theory, quasi-regular case, has been proposed. Quasi-regular case is included in singular cases in general but has similar properties as regular cases, that is, two birational invariants, λ and ν are equal to each other.

Chapter 4

WAIC-VB : an information criterion for variational Bayes learning

In this chapter, we propose an information criterion for variational Bayes learning based on WAIC. We call the proposed information criterion as WAIC-VB. In section 4.1, we explain our motivation for developing WAIC-VB. In section 4.2, we briefly review variational Bayes learning for Gaussian mixture and a conventional method for model selection. In section 4.3, we propose WAIC-VB and explain its algorithm. In section 4.4, we conduct some numerical experiments on WAIC-VB. Lastly, in section 4.5, we discuss and conclude this chapter.

4.1 Motivation for WAIC-VB

Variational Bayes learning was proposed by the mean field approximation of the Bayes posterior distribution [5, 6, 8, 17, 23]. It was shown that accurate estimation can be realized with small computational cost as the expectation and maximization algorithm. It is important to design a learning machine also in variational Bayes learning, so it is desired to estimate generalization loss of learning result. However, it has been difficult to estimate generalization loss for variational Bayes, because learning machines used in variational Bayes are not statistically regular but singular.

A lot of methods have been proposed for model selection in variational Bayes learning. In practice, minimization of variational free energy is often employed, however, this method does not evaluate generalization performance of learning result. Besides, for variational Bayes learning, the generalized DIC was proposed [18], however, DIC [22] is not an unbiased estimator of the generalization loss in singular cases [27]. In fact, the effective number of parameters defined in DIC does not correspond to the difference between the generalization loss and the

training loss. For singular statistical models, neither AIC, BIC, nor DIC can be applied because the posterior distribution does not satisfy the asymptotic normality condition.

Recently, as stated in chapter 2, a new information criterion, a widely applicable information criterion (WAIC) has been introduced, which is an unbiased estimator of Bayes generalization loss in both cases when the posterior distributions are regular and singular [26]. It was also proved that WAIC is asymptotically equivalent to the leave-one-out Bayes cross validation [27], even if the true distribution is not realizable by the learning machine. In other words, the expectation value of WAIC is equal to that of Bayes generalization loss for an arbitrary set of a true distribution, a statistical model, and a prior distribution.

In this chapter, we propose a new information criterion for variational Bayes learning by combining the importance sampling with WAIC, and show its effectiveness experimentally. The proposed information criterion, WAIC-VB, is calculated by the following procedures.

1. The variational Bayes learning is performed and the variational posterior distribution is realized.

2. The Bayes posterior distribution is approximated by the importance sampling method based on the variational posterior distribution.

3. WAIC-VB is numerically calculated.

The proposed information criterion, WAIC-VB, has two advantages compared to other criteria.

1. WAIC-VB is an unbiased estimator of Bayes generalization loss even if the posterior distribution is singular and even if the true distrbution is not realizable by the learning machine.

2. WAIC-VB can be calculated with smaller computational cost than WAIC.

In general, computing WAIC requires to realize Bayes posterior distribution by Markov chain Monte Carlo (MCMC) method, so it takes considerable computational cost. By contrast, we can compute WAIC-VB with lower computational cost because we can sample from variational posterior distribution easily.

4.2 Variational Bayes learning

In this section, we briefly introduce variational Bayes learning for Gaussian mixture.

4.2.1 Gaussian Mixture Model

In this research, Gaussian mixture is defined by a probability density function of $x \in \mathbb{R}^M$ for a parameter $w = (a, b) = \{(a_k, b_k)\}_{k=1}^K$,

$$p(x|w) = \sum_{k=1}^{K} \frac{a_k}{\sqrt{2\pi\sigma^2}^M} \exp\Big(-\frac{||x - b_k||^2}{2\sigma^2}\Big),$$

where $||\cdot||$ represents M-dimensional Euclidean norm, $\sigma > 0$ is a constant, $a_k \geq 0$, $b_k \in \mathbb{R}^M$, and

$$a_1 + a_2 + \cdots + a_K = 1.$$

Note that σ is not a parameter but a hyperparameter. In variational Bayes learning, we employ the conjugate prior

$$\begin{aligned}
\varphi(w) &= \varphi_1(a)\varphi_2(b), \\
\varphi_1(a) &= \frac{\Gamma(K\phi_0)}{\Gamma(\phi_0)^K} \prod_{k=1}^{K} (a_k)^{\phi_0 - 1}, \\
\varphi_2(b) &= (\frac{\beta_0}{2\pi\sigma^2})^{KM/2} \prod_{k=1}^{K} \exp(-\frac{\beta_0}{2\sigma^2}||b_k||^2),
\end{aligned}$$

where $\phi_0 > 0$ and $\beta_0 > 0$ are hyperparameters. Note that the Fisher information matrix of a Gaussian mixture has zero eigenvalue, to which we can not apply information criteria AIC, BIC, or DIC. It is convenient to represent Gaussian mixture model by using a competitive random variable $y = (y^1, y^2, \cdots, y^K)$ which takes value in

$$C = \{(1, 0, ..., 0), (0, 1, ..., 0), (0, 0, ..., 1)\}.$$

Then, by introducing a probability distribution of (x, y),

$$p(x, y|w) = \prod_{k=1}^{K} \Big(\frac{a_k}{\sqrt{2\pi\sigma^2}^M} \exp(-\frac{||x - b_k||^2}{2\sigma^2})\Big)^{y^k},$$

we can represent $p(x|w)$ as the marginal distribution,

$$p(x|w) = \sum_{y \in C} p(x, y|w).$$

4.2.2 Algorithm of variational Bayes learning

For a given set of training samples X^n, Bayes posterior distribution $p(y^n, w|X^n)$ is given by

$$p(y^n, w|X^n) = \frac{1}{Z_n} \varphi(w) \prod_{i=1}^{n} p(X_i, y_i|w),$$

where Z_n is the normalizing constant. In variational Bayes learning, we approximate Bayes posterior distribution $p(y^n, w|X^n)$ by assuming $y^n \in C^n$ and w are independent of each other. For a given set of training samples X^n, the variational posterior distribution $q(y^n)r(w)$ is given by

$$\begin{aligned} q(y^n) &= \prod_{i=1}^{n}\prod_{k=1}^{K}(\hat{y}_{ik})^{y_{ik}}, \\ r(w) &= \varphi_1(a|\hat{\phi})\varphi_2(b|\hat{\eta}) \\ &= \frac{1}{Z(\hat{\phi})}\prod_{k=1}^{K}\frac{(a_k)^{\hat{\phi}_k-1}}{z(\hat{\eta})}\exp(-\frac{\hat{\eta}_{k2}}{2\sigma^2}\left\|b_k-\frac{\hat{\eta}_{k1}}{\hat{\eta}_{k2}}\right\|^2), \end{aligned}$$

where $\hat{y}_{ik}$ and $(\hat{\phi}, \hat{\eta}) = (\hat{\phi}, \hat{\eta}_1, \hat{\eta}_2)$ are optimized so that Kullback-Leibler distance from $q(y^n)r(w)$ to $p(y^n, w|X^n)$,

$$K(q,r) = \sum_{y^n\in C^n}\int q(y^n)r(w)\log\frac{q(y^n)r(w)}{p(y^n,w|X^n)}dw,$$

is minimized. Such minimization is realized by the following recursive formula.

$$\begin{aligned} \hat{\phi}_k &= \sum_{i=1}^{n}\hat{y}_{ik}+\phi_0, \\ \hat{\eta}_{k1} &= \sum_{i=1}^{n}\hat{y}_{ik}x_i, \\ \hat{\eta}_{k2} &= \sum_{i=1}^{n}\hat{y}_{ik}+\beta_0, \\ L_{ik} &= \psi(\hat{\phi}_k)-\psi(n+K\phi_0)-\frac{M}{2\hat{\eta}_{k2}}-\frac{1}{2\sigma^2}\left\|x_i-\frac{\hat{\eta}_{k1}}{\hat{\eta}_{k2}}\right\|^2, \\ \hat{y}_{ik} &= \frac{\exp(L_{ik})}{\sum_{k=1}^{K}\exp(L_{ik})}, \end{aligned}$$

where $\psi(\ )$ represents the digamma function. In this research, we set the initial values as follows.

$$\begin{aligned} \hat{\phi}_0 &= \frac{n}{K}+\phi_0, \\ \hat{\eta}_{01} &= \frac{1}{n}\sum_{i=1}^{n}x_i+N(0,I_N), \\ \hat{\eta}_{02} &= \frac{n}{K}+\beta_0, \end{aligned}$$

where I_N represents $N\times N$ unit matrix. By using these iterative equations, $\hat{y}_{ik}$ and $(\hat{\phi}, \hat{\eta})$ converge to some values and then the variational posterior distribution $q(y^n)r(w)$ is obtained.

4.2.3 Method of model selection

Here we introduce a method for model selection in variational Bayes learning. That is a method which minimizes generalized DIC proposed in [18]. In this thesis, we call this generalized DIC as DIC-VB for comparison. In section 4.4, we experimentally compare DIC-VB with a proposed method.

Deviance information criterion (DIC) was proposed in [22] for model selection in Bayes learning.

$$\begin{aligned} DIC &= -2\sum_{i=1}^{n}\log p(X_i|\mathbb{E}_w[w]) + 2D_{eff}, \\ D_{eff} &= 2\sum_{i=1}^{n}\{-\mathbb{E}_w[\log p(X_i|w)] + \log p(X_i|\mathbb{E}_w[w])\}, \end{aligned}$$

where D_{eff} represents effective number of parameters. In fact, it was shown that $\mathrm{DIC}/2n$ provided asymptotically unbiased estimator of Bayes generalization loss in regular cases. In [22], DIC was applied to exponential-family models and little was said about other scenarios such as models for incomplete data. For this problem, Mcgrory et.al. [18] derived an analytical formula which approximately calculate DIC for mixture model by approximating Bayes posterior distribution using variational posterior distribution.

The analytical formula is given as follows in the setting of this research.

Lemma 15

$$\sum_{i=1}^{n}\log p(X_i|\mathbb{E}_w[w]) \cong \sum_{i=1}^{n}\log\left[\sum_{k=1}^{K}\left(\frac{\hat{\phi}_k}{\sum_{k=1}^{K}\hat{\phi}_k}\right)\frac{1}{\sqrt{2\pi}^M}\exp\{-\frac{1}{2}||X_i - \frac{\hat{\eta}_{k1}}{\hat{\eta}_{k2}}||^2\}\right],$$

$$D_{eff} \cong -2\left[\sum_{k=1}^{K}(\sum_{i=1}^{n}\hat{y}_{ik})\{(\psi(\hat{\phi}_k) - \psi(\sum_{k=1}^{K}\hat{\phi}_k)) - \frac{1}{2\hat{\eta}_{k2}}\}\right] + 2\left[\sum_{k=1}^{K}(\sum_{i=1}^{n}\hat{y}_{ik})\log(\frac{\hat{\phi}_k}{\sum_{k=1}^{K}\hat{\phi}_k})\right],$$

where $\psi(\ )$ represents the digamma function.

In this research, we calculate

$$-\frac{1}{n}\sum_{i=1}^{n}\log p(X_i|\mathbb{E}_w[w]) + \frac{1}{n}D_{eff}$$

for comparison with other criteria and we call this criterion as DIC-VB.

4.3 Proposed method

In this section, we propose a new method to estimate Bayes generalization loss in variational Bayes learning based on the importance sampling technique and

WAIC. To estimate Bayes generalization loss, that is, to calculate WAIC, we have to calculate the expectation over Bayes posterior distribution $\mathbb{E}_w[\ ]$, which requires in general to sample from $p(w|X^n)$ using the Markov Chain Monte Carlo method. In the proposed method, we use a technique proposed in [6] to approximately calculate $\mathbb{E}_w[\ ]$ in variational Bayes learning.

Firstly, we prepare $\tilde{p}_B(w)$ and $\tilde{p}_{VB}(w)$ which satisfy

$$\begin{aligned} p(w|X^n) &= \frac{1}{Z_B}\tilde{p}_B(w), \\ r(w) &= \frac{1}{Z_{VB}}\tilde{p}_{VB}(w), \end{aligned}$$

where Z_B and Z_{VB} are the normalizing constants. For instance, we can employ

$$\begin{aligned} \tilde{p}_B(w) &= \varphi(w)\prod_{i=1}^{n} p(X_i|w), \\ \tilde{p}_{VB}(w) &= \prod_{k=1}^{K}(a_k)^{\hat{\phi}_k - 1}\exp\Big(-\frac{\hat{\eta}_{k2}}{2\sigma^2}\Big\|b_k - \frac{\hat{\eta}_{k1}}{\hat{\eta}_{k2}}\Big\|^2\Big). \end{aligned}$$

In what follows, we represent $p(w|X^n)$ and $r(w)$ as $p_B(w|X^n)$ and $p_{VB}(w|X^n)$, respectively. Let $f(w)$ be an arbitrary function of w and $\{w^{(1)}, ..., w^{(L)}\}$ be independent samples from variational posterior distribution $p_{VB}(w|X^n)$. Then, since

$$\begin{aligned} \mathbb{E}_w[f(w)] &= \int f(w)p_B(w|X^n)dw \\ &= \int f(w)\frac{p_B(w|X^n)}{p_{VB}(w|X^n)}p_{VB}(w|X^n)dw \end{aligned}$$

holds, we can approximately calculate $\mathbb{E}_w[f(w)]$ by

$$\mathbb{E}_w[f(w)] \cong \sum_{l=1}^{L} w_l f(w^{(l)}),$$

where w_l represents an importance weight

$$w_l = \frac{\tilde{p}_B(w^{(l)})/\tilde{p}_{VB}(w^{(l)})}{\sum_m \tilde{p}_B(w^{(m)})/\tilde{p}_{VB}(w^{(m)})}.$$

Note that all parameters in variational posterior distribution are independent of each other though they are not independent in the original Bayes posterior distribution. Hence, we can easily generate samples $\{w^{(1)}, ..., w^{(L)}\}$ from variational posterior distribution.

Now, we propose WAIC-VB algorithm by which we can estimate Bayes generalization loss.

WAIC-VB algorithm

Input: $X^n = \{X_1, X_2, ..., X_n\}$, hyperparameters: σ, ϕ_0, β_0.
Output: WAIC-VB.

Step1: Calculate $(\hat{\phi}, \hat{\eta})$ in variational Bayes learning.
Step2: Sample $\{w^{(1)}, ..., w^{(L)}\}$ from a proposal distribution constructed based on variational posterior distribution.
Step3: Calculate importance weights w_l $(l = 1, ..., L)$ for $\{w^{(1)}, ..., w^{(L)}\}$.
Step4: Calculate

$$WAIC\text{-}VB = T_n + \frac{V_n}{n},$$

where

$$\begin{aligned} T_n &\cong -\frac{1}{n}\sum_{i=1}^{n} \log\Big(\sum_{l=1}^{L} w_l p(X_i|w^{(l)})\Big), \\ V_n &\cong \sum_{i=1}^{n}\Big\{\sum_{l=1}^{L} w_l (\log p(X_i|w^{(l)}))^2 - \{\sum_{l=1}^{L} w_l \log p(X_i|w^{(l)})\}^2\Big\}. \end{aligned}$$

Based on Theorem 13, the expectation value of WAIC-VB is asymptotically equal to that of the Bayes generalization loss.

Remark. Although WAIC-VB is an asymptotically unbiased estimator of Bayes generalization loss theoretically, there are several problems in application. In Step 1, variational Bayes learning may converge to local minima, resulting that variational posterior distribution would be a poor approximation of the original Bayes posterior distribution. Besides, in Step 2, it is known that sampling from variational posterior distribution tends to be localized compared to the original Bayes posterior distribution. For this problem, we construct another proposal distribution for importance sampling based on variational posterior distribution obtained in Step 1 so that we can sample from broader region. In this research, we make a proposal distribution by converting $(\hat{\phi}, \hat{\eta}_1, \hat{\eta}_2)$ to $(\hat{\phi}/50, \hat{\eta}_1, \hat{\eta}_2/50)$ after Step 1.

4.4 Numerical Experiments

In this section, we evaluate the WAIC-VB algorithm through numerical experiments. We conduct three types of experiments, from Experiment 1 to Experiment 3. In Experiment 1, we compare $\mathbb{E}[WAIC\text{-}VB]$ with $\mathbb{E}[G_n]$ for Gaussian mixture. In Experiment 2, a model selection task for Gaussian mixture with $\sigma = 1$ is studied, in which the number of components that minimizes WAIC-VB is selected. We perform model selection under two different situations. One is under the same setting as Experiment 1 where each component of Gaussian

mixture is well separated, the other is under a 'delicate' situation where each component of Gaussian mixture is not clearly separated. In Experiment 1 and 2, the value of DIC-VB is also calculated for comparison. In Experiment 3, hyperparameter selection for Gaussian mixture is analyzed. The hyperparameter σ which minimizes WAIC-VB is chosen. In all experiments, hyperparameters, ϕ_0 and β_0, are set to be $\phi_0 = 1$ and $\beta_0 = 1$.

Experiment 1: Comparison of $\mathbb{E}[\mathit{WAIC\text{-}VB}]$ with $\mathbb{E}[G_n]$

In Experiment 1, we set the true distribution $q(x)$ as follows.

$$q(x) = \sum_{k=1}^{K_0} \frac{a_k^0}{\sqrt{2\pi}^M} \exp(-\frac{||x - b_k^0||^2}{2}),$$

where $K_0 = 3$ and the true parameters are

$$\begin{aligned} a^0 &= (0.3, 0.3, 0.4), \\ b_1^0 &= (3, 3), \quad b_2^0 = (-3, 1), \quad b_3^0 = (2, -2). \end{aligned}$$

A sample set from the true distribution $q(x)$ is shown below.

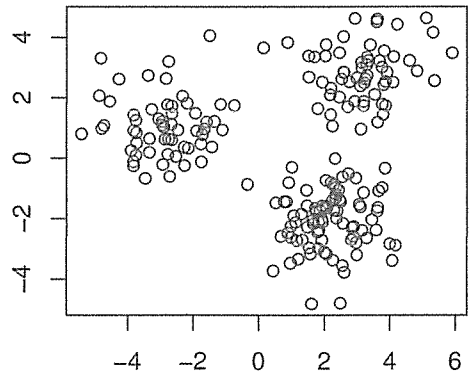

Figure 4.1: A sample set from $q(x)$

The learning machines were set to be a Gaussian mixture $p(x|w)$ with $\sigma = 1$ and $K = 3, 4$. When $K = 3$, the learning machine had the same number of

components as the true distribution $q(x)$, hence the posterior distribution was regular, whereas, when $K = 4$, it had a redundant component compared to the true distribution, hence the posterior distribution was singular. In a regular case, the posterior distribution can be approximated by a normal distribution and the number of effective parameters is equal to the dimension of the parameter set. However, in a singular case, the posterior distribution cannot be approximated by any normal distribution and the number of effective parameters is different from the dimension of the parameter set. Training samples ($n = 200$) were taken from $q(x)$ and the values of WAIC-VB and DIC-VB were calculated using these samples, and the Bayes generalization loss G_n was calculated by using 5000 test samples. For calculating WAIC-VB, the number of samples L from variational posterior distribution was 2000 in this experiment. The expectation value $\mathbb{E}[\ ]$ was approximated by the average over 100 sets of training samples.

	G_n	T_n	V_n	*WAIC-VB*	*DIC-VB*
MEAN	3.9326	3.8902	7.7793	3.9291	3.9290
STD	0.0178	0.0736	0.4723	0.0758	0.0747

Table 4.1: comparison of $\mathbb{E}[$*WAIC-VB*$]$ with $\mathbb{E}[G_n]$: Case $K = 3$

	G_n	T_n	V_n	*WAIC-VB*	*DIC-VB*
MEAN	3.9389	3.8978	8.0642	3.9381	3.9479
STD	0.0161	0.0648	0.5653	0.0673	0.0661

Table 4.2: comparison of $\mathbb{E}[$*WAIC-VB*$]$ with $\mathbb{E}[G_n]$: Case $K = 4$

The results of the experiment are shown in Table 4.1 and 4.2, where 'MEAN' and 'STD' are respectively the expectation value and the standard deviation.

Experiment 2: Model selection

In Experiment 2, we conducted two kinds of simulations of model selection for Gaussian mixture. In the first simulation, we employed the same true distribution $q(x)$ and a learning machine $p(x|w)$ as Experiment 1, respectively. The values of WAIC-VB and DIC-VB were calculated for 100 fixed datasets for each learning machine with $K = 1, 2, \cdots, 10$. The results of the experiment are shown in Fig.4.2 and Fig.4.3. The solid line represents the mean value of WAIC-VB and dotted line does that of DIC-VB, respectively. The error bar represents the standard deviation of each value. In Fig.4.3, we investigated the values of WAIC-VB and DIC-VB for $K = 3, \cdots, 10$ in detail.

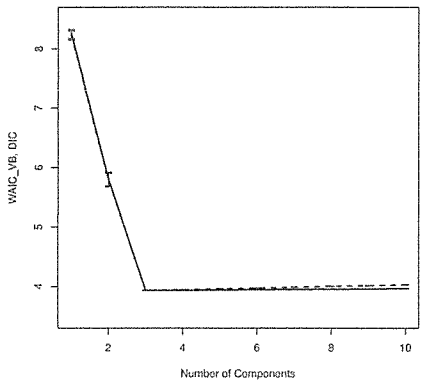

Figure 4.2: Experiment 2

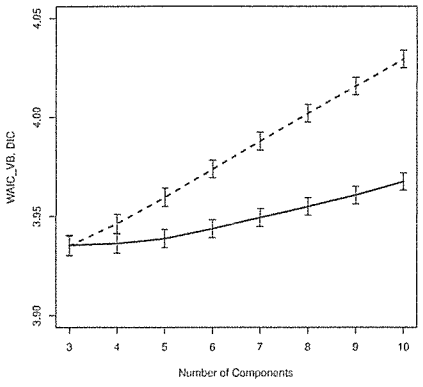

Figure 4.3: Experiment 2 (in detail)

In the second simulation, we performed model section for Gaussian mixture under 'delicate' situation. Let the true distribution $q(x)$ and parametric models $p_1(x|w), p_2(x|w)$ be defined as follows.

$$\text{True distribution} : q(x) = \frac{0.5}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\} + \frac{0.5}{\sqrt{2\pi}} \exp\{-\frac{(x-k)^2}{2}\}$$

$$\text{parametric model-1} : p_1(x|w) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(x-b)^2}{2}\}$$

$$\text{parametric model-2} : p_2(x|w) = \frac{1-a}{\sqrt{2\pi}} \exp\{-\frac{(x-b_1)^2}{2}\} + \frac{a}{\sqrt{2\pi}} \exp\{-\frac{(x-b_2)^2}{2}\}$$

The value of k determines the distance between the mean of each component. In this simulation, we increase the value of k from 0 to 2 by 0.2. When $k = 0$, the true distribution consists of a single Gaussian distribution. As the value of k increases, the true deistribution will be clearly separated into two components. For both parametric models, $p_1(x|w)$ and $p_2(x|w)$, we investigated the value of Bayes generalization loss G_n, WAIC-VB and DIC-VB for each k from 0 to 2. For calculating WAIC-VB, the number of samples L from variational posterior distribution was 5000 in this experiment. They were calculated for 100 fixed datasets for each k.

The results of the experiment are shown in Fig.4.4 and Fig.4.5. In Fig.4.4, we compare the values of WAIC-VB and Bayes generalization loss G_n. The solid line represents mean values of WAIC-VB and G_n for $(q(x), p_1(x|w))$ and dotted line does for $(q(x), p_2(x|w))$, respectively. The error bar represents the standard deviation of each value. In Fig.4.5, we compare the values of DIC-VB and Bayes

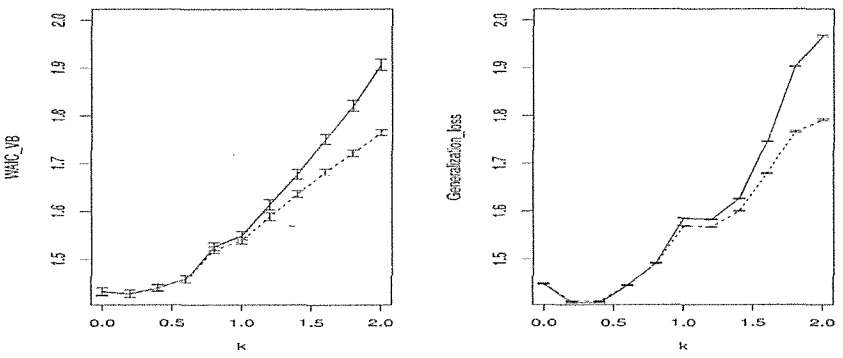

Figure 4.4: WAIC-VB and Generalization loss

generalization loss G_n. The solid line represents mean values of DIC-VB and G_n for $(q(x), p_1(x|w))$ and dotted line does for $(q(x), p_2(x|w))$, respectively. The error bar represents the standard deviation of each value.

Experiment 3: Hyperparameter selection

In Experiment 3, we employed the same true distribution $q(x)$ as Experiment 1 and learning machines $p(x|w)$ with $K = K_0 = 3$ and $\sigma = 0.6, 0.7, ..., 1.5$ were compared.

$$p(x|w) = \sum_{k=1}^{K} \frac{a_k}{\sqrt{2\pi\sigma^2}^M} \exp\Big(-\frac{||x - b_k||^2}{2\sigma^2}\Big),$$

The values of WAIC-VB were calculated for 100 fixed detasets and a learning machine with each σ. The results of the experiment are shown in Fig.4.6.

4.5 Discussion and Conclusion

In this section, we discuss the results of numerical experiment conducted in section 4.4 and conclude this chapter.

Experiment 1: Comparison of $\mathbb{E}[WAIC\text{-}VB]$ with $\mathbb{E}[G_n]$

Table 4.1 and 4.2 show that WAIC-VB works well as an unbiased estimator both for regular and singular cases. As stated in chapter 2, it is theoretically proved that $\mathbb{E}[V_n]$ converges to 2ν, where ν represents singular fluctuation. When $K = 3$, since $(q(x), p(x|w))$ is in regular cases, 2ν is equal to the number of parameter, which is 8 in this situation. This almost coinsides to the experimental result. On the other hand, when $K = 4$, since $(q(x), p(x|w))$ is in singular

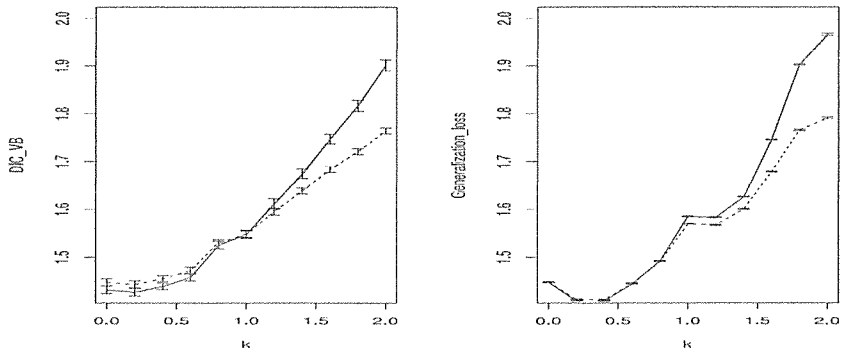

Figure 4.5: DIC-VB and Generalization loss

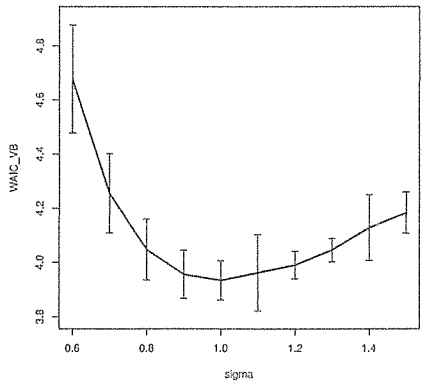

Figure 4.6: Experiment 3

cases, 2ν is not equal to the number of parameter, which is 10. Although the theoretical value of 2ν in this case is still unclarified, the experimental result of $\mathbb{E}[V_n]$ is explicitly smaller than 10. DIC-VB works well in regular cases ($K = 3$), but it seems to overestimate Bayes generalization loss in singular cases ($K = 4$).

Experiment 2: Model selection

In the first simulation, the true distribution $q(x)$ consisted of three components and they were well separated. Fig.4.2 and Fig.4.3 showed that the model that has the true number of components was chosen by both WAIC-VB and DIC-VB. However, DIC-VB tended to take a bigger value than WAIC-VB and the more singular learning is, the bigger the difference is. From this result, it may be considered that DIC-VB works well enough in practical model selection tasks even though it does not provide an unbiased estimator of Bayes generalization loss in singular cases. However, the difference between WAIC-VB and DIC-VB becomes more important in delicate situation, where each component of Gaussian mixture is not clearly separated and hence model selection task gets more difficult.

In the second simulation, the true distribution $q(x)$ depended on the value of k. Hence, which model, $p_1(x|w)$ or $p_2(x|w)$, was selected by WAIC-VB and DIC-VB also depended on the value of k. Fig.4.4 shows that the values of WAIC-VB for $p_1(x|w)$ and $p_2(x|w)$ indicates similar behavior as Bayes generalization loss. On the other hand, Fig.4.5 showes that the values of DIC-VB for $p_1(x|w)$ and $p_2(x|w)$ indicates different behavior from Bayes generalization loss. Specifically, when k is small, the value of DIC-VB for $p_1(x|w)$ takes explicitly smaller value than that for $p_2(x|w)$. It is considered that this is because DIC-VB overestimates Bayes generalization loss in singular cases.

To compare the difference between WAIC-VB and DIC-VB in delicate situation more precisely, we investigate the values of WAIC-VB-S_n, DIC-VB-S_n and Bayes generalization error $G_n^{(0)}$ for small k. Note that the entropy term S_n which is included in the values of WAIC-VB, DIC-VB and G_n depends on k, since the true distribution $q(x)$ is different for each k. In this situation, when the value of k is small, the model $p_1(x|w)$ is more appropriate than $p_2(x|w)$ to attain good generalization error, but the appropriate model changes to $p_2(x|w)$ as the value of k increases. Fig.4.7 shows that the model selected WAIC-VB switches from $p_1(x|w)$ to $p_2(x|w)$ around $k = 0.6$ and Bayes generalization error $G_n^{(0)}$ for $p_1(x|w)$ and $p_2(x|w)$ behaves similarly. On the other hand, Fig.4.8 shows that the model selected DIC-VB does not switch from $p_1(x|w)$ to $p_2(x|w)$ yet even when $k = 1.0$. In other words, DIC-VB selects the model $p_1(x|w)$ even though the model $p_2(x|w)$ attains better Bayes generalization error. This result means that, in delicate situation, WAIC-VB can select better model than DIC-VB on average from the viewpoint of Bayes generalization error.

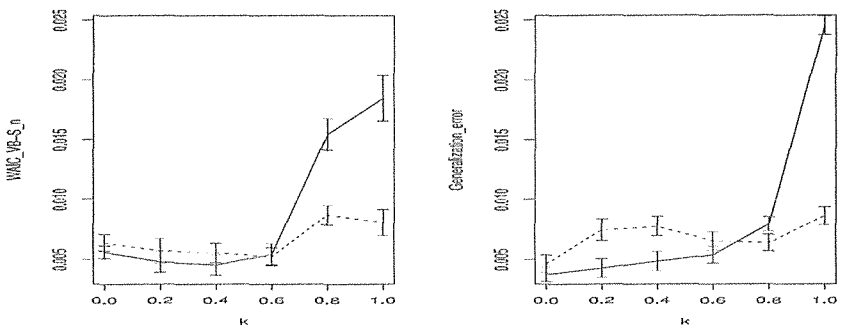

Figure 4.7: WAIC-VB-S_n and Generalization error

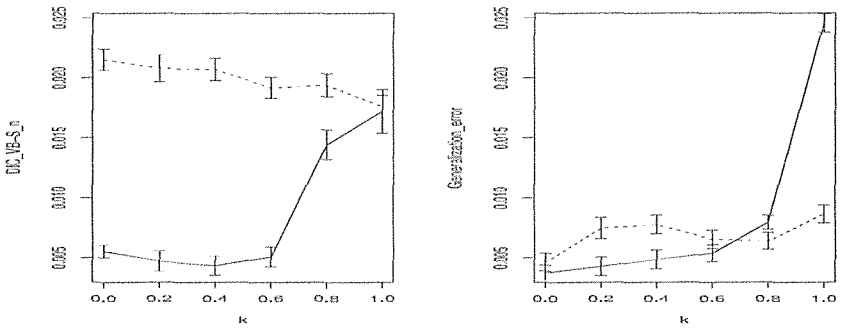

Figure 4.8: DIC-VB-S_n and Generalization error

Experiment 3: Hyperparameter selection

This result showed that the true standard deviation was chosen by minimization of WAIC-VB.

From the results of numerical experiment, it can be said that we can utilize WAIC-VB as an information criterion for variational Bayes learning. As a conclusion, in this chapter, we proposed a new information criterion, WAIC-VB, for variational Bayes learning and investigated its property by numerical experiments. The effectiveness of WAIC-VB was shown through numerical experiments although there are several problems in its implementation. It will be a problem for a future study.

Chapter 5

Discussion

In this section, we discuss some related topics to the results in chapter 3,4 and left problems for a future study.

In chapter 3, we have proposed a new concept in statistical learning theory, a quasi-regular case, and have constructed its basic theory. It has been shown that, in general, a quasi-regular case is included in singular cases, but it has similar properties as regular cases, that is, two birational invariants, λ and ν, are equal to each other. In other words, symmetry between Bayes generalization loss and Bayes training loss holds even in quasi-regular cases. In addition, although singular fluctuation ν depends on inverse temparature β in general singular cases, $\nu = \nu(\beta)$ takes the constant value independent of β in quasi-regular cases.

Here we discuss some properties of quasi-regular cases and left problems for a future study. From theoretical perspective, it can be said that singularities which appears in quasi-regular cases are relatively easy to handle. Lemma.13 is a key property in quasi-regular cases, but it is not so easy to extend this lemma to more general singular cases. It will be an important and interesting problem for a future study to clarify singular fluctuation ν in more general singular cases. From practical perspective, we can utilize the theory of quasi-regular cases in two ways. For one thing, we can use the exact value of singular fluctuation in the reserch of Markov Chain Monte Carlo method. In general, it is difficult to check whether we successfuly realize Bayes posterior distribution or not, especially in singular cases. For this problem, it would be helpful to monitor the difference between the value of functional variance and the theoretical value of 2ν, because $\mathbb{E}[\beta V_n]$ converges to 2ν asymptotically. Another is an application in Bayes hypothesis test. In Bayes hypothesis test, it is essential to clarify the asymptotic distribution of Bayes free energy. In fact, we can clarify the asymptotic distribution of Bayes free energy in quasi-regular cases based on the basic theory developed in chapter 3.

In chapter 4, we have proposed an information criterion for variational Bayes learning, WAIC-VB, and have conducted some numerical experiments. It has been shown that WAIC-VB is an asymptotically unbiased estimator of Bayes generalization loss and it is effective especially in 'delicate' cases compared to other imformation criteria such as generalized DIC.

Here we discuss some properties of WAIC-VB and left problems for a future study. Firstly, from theoretical perspective, WAIC-VB estimates Bayes generalization loss in variational Bayes learning. In general, in variational Bayes learning, we construct variational predictive distribution instead of Bayes predictive distribution. In a strict sense, Bayes generalization loss differs from the generalization loss of variational predictive distribution. To develop an asymptotically unbiased estimator of generalization loss for variational predictive distribution is an important theoretical problem for a future study. Secondly, we discuss the variance of WAIC and WAIC-VB. In chapter 2, we have stated that expected value of WAIC is asymptotically equal to the expected value of Bayes generalization loss G_n, but nothing has been stated on the variance of WAIC and WAIC-VB. In fact, it is known that WAIC-S_n has the same variance as Bayes generalization error asymptotically [29]. It is considered that the variance of WAIC-VB has the same property theoretically. However, in singular cases, exact value of the variance has not been clarified yet. It will be an important probrem for a future study. Thirdly, we discuss the computational cost for WAIC and WAIC-VB. As stated in chapter 4, computing WAIC needs to realize Bayes posterior distribution by Markov chain Monte Carlo method in general. On the other hand, computing WAIC-VB utilizes samples from a proposal distribution which is based on variational posterior distribution and sampling from such a distribution is easy, since each parameter is conditionally independent.

Lastly, we discuss some problems in implementation of WAIC-VB.

(1) Variational Bayes learning may converge to local minima, resulting that variational posterior distribution would be a poor approximation of the original Bayes posterior distribution.
(2) Samples from variational posterior distribution tend to be localized compared to the original Bayes posterior distribution. For this problem, in this research, we constructed a proposal distribution for importance sampling based on variational posterior distribution by converting $(\hat{\phi}, \hat{\eta_1}, \hat{\eta_2})$ to $(\hat{\phi}/50, \hat{\eta_1}, \hat{\eta_2}/50)$. However, it is not clear how to construct this transformation for making an appropriate proposal distribution. In practice, it is recommended to make a proposal distribution through monitoring the distribution of imporatance weights.
(3) WAIC-VB tends to fluctuate depending on samples from proposal distribution. Especially, V_n tends to fluctuate much compared to T_n.

Chapter 6

Conclusion

In this thesis, we have discussed an unbiased estimator of Bayes generalization loss, WAIC, from theoretical and practical viewpoints.

From theoretical perspective, we proposed a quasi-regular case, which is a new concept in statistical learning theory. In singular learning theory, it has been clarified that the asymptotic Bayes generalization loss and Bayes training loss are determined by two birational invariants, real log canonical threshold λ and singular fluctuation ν for a triple $(q(x), p(x|w), \varphi(w))$. Especially, singular fluctuation ν determines the asymptotic difference of Bayes generalization loss and Bayes training loss, however, the properties of singular fluctuation ν was left totally unknown except for regular cases. In this thesis, we clarified the value of singular fluctuation ν in quasi-regular cases, which was equal to the value of real log canical threshold λ as it was in regular cases. In other words, quasi-regular cases have the similar properties as regular cases from the viewpoint of singular learning theory, although they are generally not in regular cases but in singular cases.

From practical perspective, we proposed WAIC-VB, which is a new information criterion for variational Bayes learning. WAIC-VB provides an unbiased estimator of Bayes generalization loss theoretically and it can be calculated with small computational cost because sampling from variational posterior distribution is easier than that from Bayes posterior distribution. The effectiveness of WAIC-VB was shown by numerical experiments.

These results will be a basis for more thorough understanding of unbiased estimator of generalization loss in statistical learning from both theoretical and practical viewpoints.

Bibliography

[1] H.Akaike. "A new look at the statistical model identification," IEEE Transactions on Automatic Control, Vol.19, pp.716-723, 1974.

[2] M.Aoyagi,S.Watanabe,"Resolution of singularities and generalization error with Bayesian estimation for layered neural network," Vol.J88-D-II, No.10, pp.2112-2124, 2005.

[3] M.Aoyagi, S.Watanabe,"Stochastic complexities of reduced rank regression in Bayesian estimation," Neural Networks, Vol.18, No.7, pp.924-933, 2005.

[4] M.F.Atiyah,"Resolution of singularities and division of distributions," Comm. Pure Appl. Math., Vol.13, pp.145-150, 1970.

[5] H. Attias, Inferring parameters and structure of latent variable models by variational Bayes, in: Proceedings of Uncertainty in Artificial Intelligence (UAI ' 99), 1999.

[6] M.J. Beal, Variational algorithms for approximate Bayesian inference, Ph.D. Thesis, University Colledge London, 2003.

[7] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[8] Z. Ghahramani, M.J. Beal, Graphical models and variational methods, in: D. Saad, M. Opper (Eds.), Advanced Mean Field Methods: Theory and Practice, MIT Press, 2000.

[9] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian data analysis. Chapman & Hall CRC, Boca Raton, 2004.

[10] K. Hagiwara, "On the Problem in Model Selection of Neural Network Regression in Overrealizable Scenario," Neural Comput., Vol.14,Vol.8, pp.1979 - 2002, 2002.

[11] J.A.Hartigan,"A failure of likelihood asymptotics for normal mixture," Proc. of Barkeley Conf. in honor of Jerzy Neyman and Jack Keifer, Vol.2, pp.807-810,1985.

[12] T. Hayasaka, M. Kitahara, and S. Usui, "On the Asymptotic Distribution of the Least-Squares Estimators in Unidentifiable Models," Neural Comput., Vol.16 ,No.1, pp.99 - 114, 2004.

[13] H. Hironaka, "Resolution of singularities of an algebraic variety over a field of characteristic zero," Ann. of Math., Vol.79, 109-326,1964.

[14] D. Kaji, S. Watanabe, "Two design methods of hyperparameters in variational Bayes learning for Bernoulli mixtures," Neurocomputing, Vol.74, pp.2002-2007, 2011.

[15] M. Kashiwara, "B-functions and holonomic systems," Inventions Math., 38, 33-53.1976.

[16] Shaowei Lin, "Algebraic Methods for Evaluating Integrals in Bayesian Statistics," Ph.D thesis, Ph.D. dissertation, University of California, Berkeley, 2011

[17] D.J.C. MacKay, Developments in probabilistic modeling with neural networks ensemble learning, in: Proceedings of the Third Annual Symposium on Neural Networks, Springer, New York, 1995, pp. 191-198.

[18] C.A. McGrory and D.M. Titterington, Variational Approximations in Bayesian Model Selection for Finite Mixture Distributions, Computational Statistics & Data Analysis 51(11), pp.5352-5367, 2007.

[19] K. Nagata, S. Watanabe, "Asymptotic Behavior of Exchange Ratio in Exchange Monte Carlo Method," International Journal of Neural Networks, Vol. 21, No. 7, pp. 980-988, 2008.

[20] Dmitry Rusakov and Dan Geiger, "Asymptotic model selection for naive Bayesian network," Journal of Machine Learning Research, pp.1-35, 2005

[21] Morihiko Saito, "On real log canonical threhsolds," arXiv:0707.2308v1, 2007

[22] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, A. Linde, Bayesian measures of model complexity and fit. Journal of Royal Statistical Society B, 64(4), 583-639, 2002.

[23] K. Watanabe, S. Watanabe, Stochastic complexities of gaussian mixtures in variational bayesian approximation. Journal of Machine Learning Research, 7, 625-644, 2006.

[24] S. Watanabe, "Generalized Bayesian framework for neural networks with singular Fisher information matrices," Proc. of International Symposium on Nonlinear Theory and Its applications, (Las Vegas), pp.207-210, 1995.

[25] S. Watanabe, "Algebraic Analysis for Nonidentifiable Learning Machines," Neural Computation, Vol.13, No.4, pp.899-933, 2001.

[26] S. Watanabe, "Equations of states in singular statistical estimation", Neural Networks, Vol.23, No.1, pp.20-34, 2010.

[27] S. Watanabe, "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory," Journal of Machine Learning Research, Vol.11, (DEC), pp.3571-3591, 2010.

[28] S. Watanabe, "Algebraic geometry and statistical learning theory," Cambirdge University Press, 2009.

[29] S. Watanabe, "Theory and method of Bayes statistics (in Japanese)," Koronasha, 2012

[30] K. Yamada, S. Watanabe, "Statistical Learning Theory of Quasi-Regular Cases," IEICE Transactions, Vol.E95-A, No.12, pp.2479-2487, 2012

[31] Koshi Yamada and Sumio Watanabe, "Information Criterion for Variational Bayes Learning in Regular and Singular Cases," Proc. of The 6th International Conference on Soft Computing and Intelligent Systems, F2-55-3, 2012, Kobe

[32] K. Yamazaki, S. Watanabe,"Singularities in mixture models and upper bounds of stochastic complexity," International Journal of Neural Networks, Vol.16, No.7, pp.1029-1038,2003.

[33] K. Yamazaki,"Stochastic complexity of Bayesian networks," Proceeding of International Conference on Uncertainty in Artifitial Intelligence, 2003

[34] K. Yamazaki and S. Watanabe, "Algebraic geometry and stochastic complexity of hidden Markov models," Neurocomputing, Vol.69, pp.62-84, 2005

[35] Piotr Zwiernik, "An asymptotic behavior of the marginal likelihood for general markov models," Journal of Machine Learning Research, 12: 3283-3310, 2011