論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Generalization Analysis on dependency of Model Complexity in Bayesian Deep Neural Network
著者(和文)	永安修也
Author(English)	Shuya Nagayasu
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12656号, 授与年月日:2024年3月26日, 学位の種別:課程博士, 審査員:渡邊 澄夫,金森 敬文,山下 真,中野 張,髙邉 賢史
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第12656号, Conferred date:2024/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Dissertation

Generalization Analysis on dependency of Model Complexity in Bayesian Deep Neural Network

Shuya Nagayasu

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

2023

Abstract

Deep Neural Networks (DNNs) are widely used as machine learning models for computer vision, natural language processing, robotics, and many other areas. The high performance of DNNs comes from the fact that the rules in machine learning models such as bias-variance tradeoff do not hold. Recent year studies revealed the relationship between Stochastic gradient descent and Stochastic gradient langevin dynamics which realize Bayesian learning. On the other hand, many studies try to realize the Bayesian Deep Neural Networks. From these viewpoints, Bayesian learning in DNNs gets more important.

In this thesis we theoretically show that the performance of deep neural networks is different from the classical statistical models in perspective of Bayesian learning through the statistical analysis of generalization error in two cases. One is the case that there exist multiple optimal probability distributions which are nearest to data generating process. In this case, the generalization error increases while the number of data increases. This case occurs when the model has relatively smaller complexity to the data. The other is the case that the model has larger number of parameters than necessary. In this case, the generalization error does not increase even if the number of the model parameters increases.

Contents

Chapter 1

Introduction

1.1 Deep Neural Networks

Deep Neural Networks(DNNs) are the most used as machine learning models in the last ten years. Though the original idea of Neural Network was shown in 1940s, the basis of modern Deep Learning arose in 1989 LeNet [23]. The structure of LeNet was the combination of convolutional layers and fully-connected layer same as modern Convolutional Neural Networks(CNNs) and the backpropagation was used. After twelve years of stagnation after success of LeNet, AlexNet [21] triggered modern using of Neural Networks. AlexNet is a development of LeNet and earlier architecture using GPU. Now many architectures of Neural Networks are used such as Residual Network (ResNet)[15] for computer vision, Generative Adversarial Network [11] for image processing, Transformer for natural language processing and many others for many areas. However, the clear theoretical explanation to the high performance of Neural Networks are unknown yet. In particular, it is well known that bias-variance trade-off shown in conventional learning models does not occur in deep neural networks. In present, there exist some different theories such as double descent [29] or Bayesian perspective of SGD [36].

1.2 Bayesian Learning

The name of Bayesian learning or Bayes inference comes from the Thomas Bayes who is an English statistician in 18th century. He discovered the basic probabilistic law of conditional probability distribution. For a long time, Bayesian inference is used only for analytically calculated cases such as exponential family and conjugate prior because the calculation of marginal likelihood needs the integration in Bayesian inference. The situation changed with the development of Metropolis–Hastings algorithm which is one of Markov chain Monte Carlo methods (MCMC) in statistical physics [13]. Metropolis-Hastings algorithm has the problem of rejection rate, Hamiltonian Monte Carlo(HMC) which comes from the Hamiltonian dynamics in phase space of statistical mechanics solved this problem [7]. In modern Bayesian inference, Hamiltonian Monte Carlo or No-U-Turn sampler [17] which is an improvement of HMC is mainly used for practical way, but the situation changed in Bayesian learning for DNNs. The name of Bayesian learning or Bayes inference comes from the Thomas Bayes who is an English statistician in 18th century. He discovered the basic probabilistic law of conditional probability

distribution. For a long time, Bayesian inference is used only for analytically calculated cases such as exponential family and conjugate prior because the calculation of marginal likelihood needs the integration in Bayesian inference. The situation changed with the development of Metropolis–Hastings algorithm which is one of Markov chain Monte Carlo methods (MCMC) in statistical physics [13]. Metropolis-Hastings algorithm has the problem of rejection rate, Hamiltonian Monte Carlo(HMC) which comes from the Hamiltonian dynamics in phase space of statistical mechanics solved this problem [7]. In modern Bayesian inference, Hamiltonian Monte Carlo or No-U-Turn sampler [17] which is an improvement of HMC is mainly used for practical way, but the situation changed in Bayesian learning for DNNs.

1.3 Neural Networks in Bayesian Learning

Earlier ideas of Neural Networks in Bayesian Learning were discussed by MacKay [24, 25]. In these papers, the gaussian approximation of posteriors which is the variational inference was mainly used. Neal [30] showed the idea of applying MCMC for Bayesian Neural Networks. Variational Approximation and MCMC are the basic methods of realizing Bayesian posterior for neural networks. Variational Autoencoder [20] and Monte Carlo dropout [9] are representative variational methods in practical use. On the other hand, Hamiltonian Monte Carlo or Langevin Dynamics are mainly used among MCMC. In particular, the development of Stochastic Gradient Langevin Dynamics(SGLD) [49] makes it easier to conduct MCMC in Neural Network because of the similarity to Stochastic Gradient Descent. In addition, recent studies showed that the randomness of Stochastic Gradient Descent(SGD) makes them closer to random walk and SGLD [36]. These studies connected the behavior of SGD with the flatness [16] through Bayesian learning and the explanation of generalization ability of Deep Learning. From both viewpoints of direct use of Bayesian Neural Networks and the theory of SGD in Neural Networks, the understanding of the behavior of Bayesian Neural Networks has become important.

1.4 Goals of studies

In this thesis we show the two cases of theoretical analysis in Bayesian generalization error about neural networks [28, 26]. In the first case, we clarify the generalization error of Bayesian learning when optimal probability distributions are not unique. Optimal probability distribution is the probability distribution nearest to the data generating process in learning model. In conventional learning theory, such distributions are assumed to be unique. However, in case the data generating process is more complicated than model approximation ability, this assumption cannot always be applied. In particular, if there exists the symmetry in data generating process which is mismatched to the learning model, this case occurs. In the second case, we reveal the generalization error of Convolutional Neural Network with and without skip connection in Bayesian learning. Skip connection is used for many architectures of CNNs. With skip connection case, we show that the variance term of generalization error only depends on the complexity of data generating process. As far as currently known, the models which have such property in Bayesian learning are only fully-connected neural network with ReLU activation function and CNNs with skip connection. Through analyzing these two cases we revealed the characteristics of DNNs in Bayesian learning.

Chapter 2

Bayesian learning theory

In this chapter, we explain the framework of Bayesian Learning theory.

2.1 Notations

Table2.1 show the definitions and notations used in this chapter.

Table 2.1: Notation

Notation	Definition	Name
$\mathbb{E}[\cdots]$	$\int \cdots \prod_{i=1}^{n} q(X_i) \mathrm{d}X^n$	average of generating of samples
$\mathbb{E}_w[\cdots]$	$\int \cdots p(w\|X^n) \mathrm{d}w$	average of posterior
$\mathbb{E}_X[\cdots]$	$\int \cdots q(x) \mathrm{d}x$	average of true distribution
S	$-\mathbb{E}_x[\log q(x)]$	entropy
F_n	$-\log Z_n$	the free energy
G_n	$\mathbb{E}_x[\log q(x)/p(x\|X^n)]$	the generalization error

2.2 Bayesian learning

First, we explain the basic calculation of Bayesian learning. In supervised learning, Let $X^n = (X_1, \cdots X_n)$ and $Y^n = (Y_1, \cdots Y_n)$ be training data and labels or output data. The natural number n is the number of the data. These data and labels are generated from a true joint distribution $q(x, y) = q(y|x)q(x)$. The prior distribution $\varphi(w)$, the learning model $p(y|x, w)$ is given on the bounded parameter set W. Then the posterior distribution is defined by

$$p(w|X^n, Y^n) = \frac{1}{Z(Y^n|X^n)} \varphi(w) \prod_{i=1}^{n} p(Y_i|X_i, w) \tag{2.1}$$

where $Z_n = Z(Y^n|X^n)$ is normalizing constant denoted as marginal likelihood:

$$Z_n = \int \varphi(w) \prod_{i=1}^{n} p(Y_i|X_i, w) \mathrm{d}w. \tag{2.2}$$

The posterior distribution estimates the parameters probabilistically. The posterior predictive distribution is defined as the average of the model by posterior:

$$p^*(y|x) = p(y|x, X^n, Y^n) = \int p(y|x, w)p(w|X^n, Y^n)\mathrm{d}w. \tag{2.3}$$

The posterior predictive distribution $p^*(y|x)$ is an estimation of true conditional distribution $q(y|x)$. In case unsupervised learning, the training data is $X^n = (X_1, \cdots X_n)$ and the true distribution is $q(x)$. The learning model is $p(x|w)$. In this case, the posterior distribution and the posterior predictive distribution are defined by

$$p(w|X^n) = \frac{1}{Z(X^n)}\varphi(w)\prod_{i=1}^{n} p(X_i|w) \tag{2.4}$$

$$p^*(x) = p(x|X^n) = \int p(x|w)p(w|X^n)\mathrm{d}w. \tag{2.5}$$

Hereinafter in this chapter, we explain about the case unsupervised learning. The same things hold also in case supervised learning.

2.3 Asymptotic analysis

2.3.1 Kullback-Leibler divergence

Kullback-Leibler(KL) divergence is a statistical distance measuring the probability distributions. For probability distributions $p(x), q(x)$ which supports P, Q satisfy the $Q \subset P$, KL divergence is defined by

$$D_{\mathrm{KL}}(q(x)|p(x)) = \int q(x)\log\frac{q(x)}{p(x)}\mathrm{d}x. \tag{2.6}$$

KL divergence has the following properties.

- $D_{\mathrm{KL}}(q(x)|p(x)) \geq 0$
- $D_{\mathrm{KL}}(q(x)|p(x)) = 0$ if and only if $p(x) = q(x)$
- $D_{\mathrm{KL}}(q(x)|p(x)) \neq D_{\mathrm{KL}}(p(x)|q(x))$ in general.

2.3.2 Free energy and Generalization Error

The generalization error in Bayesian learning is KL divergence between true distribution and estimated distribution.

$$G_n = D_{\mathrm{KL}}(q(x)|p^*(x)) \tag{2.7}$$

The free energy F_n is a negative log marginal likelihood also as known as stochastic complexity

$$F_n = -\log Z_n. \tag{2.8}$$

Minimizing the average of free energy in data generating is equivalent to minimizing true joint distribution of data and marginal distribution

$$\mathbb{E}[F_n] = -\mathbb{E}[\log Z_n] \tag{2.9}$$

$$= -\mathbb{E}[\log \prod_{i=1}^{n} q(X_i)] + \mathbb{E}\left[\log \frac{q(X^n)}{p(X^n)}\right] \tag{2.10}$$

$$= nS + D_{\mathrm{KL}}(q(X^n)|p(X^n)). \tag{2.11}$$

Average of the generalization error is difference between the average of Free energy of n and $n+1$:

Lemma 2.3.1.

$$\mathbb{E}[G_n] - S = \mathbb{E}[F_{n+1}] - \mathbb{E}[F_n]. \tag{2.12}$$

Proof. For $p(x|X^n) = p(X_{n+1}|X^n)$, the following equation holds

$$p(x|X^n) = \frac{1}{Z_n} \int p(X_{n+1}|w)\varphi(w) \prod_{i=1}^{n} p(X_i|w) \mathrm{d}w \tag{2.13}$$

$$= \frac{Z_{n+1}}{Z_n}. \tag{2.14}$$

The average of a sample of negative log of equation(2.14) completes the lemma. □

2.3.3 Regular Case

The set of optimal parameters W_0 is defined as the set of all parameters that minimize the KL divergence of $q(x)$ and $p(x|w)$,

$$W_0 = \{w \in W | \int q(x) \log \frac{q(x)}{p(x|w)} \mathrm{d}x \text{ is minimized.}\} \tag{2.15}$$

The log density ratio function for $w_0 \in W_0$ and $w \in W$ is defined as

$$f(x, w_0, w) = \log \frac{p(x|w_0)}{p(x|w)}. \tag{2.16}$$

The empirical error function $K_n(w)$ is defined as

$$K_n(w) = \frac{1}{n} \sum_{j=1}^{n} f(X_j, w_0, w). \tag{2.17}$$

The log loss function $L(w)$ and the average error function $K(w)$ are defined by

$$L(w) = -\mathbb{E}_X[\log p(X|w)], \tag{2.18}$$

$$K(w) = \mathbb{E}_X[f(X, w_0, w)] = L(w) - L(w_0). \tag{2.19}$$

K(w) satisfies

$$K(w_0) = 0, K(w) \geq 0. \tag{2.20}$$

If $K(w)$can be approximated by a quadratic form, in other words, the Laplace approximation can be applied to the posterior distribution, average of Free energy has the following asymptotic expansion with the number of parameters of the learning model d [35, 32]

$$E[F_n] = n(S + \text{Bias}) + \frac{d}{2}\log n + O(1) \tag{2.21}$$

where S is entropy of true distribution and Bias is

$$D_{\text{KL}}(q(x)|p(x|w_0)). \tag{2.22}$$

The generalization error is calculated from Free energy by using equation(2.3.1) [2]:

$$E[G_n] = \text{Bias} + \frac{d}{2n} + o\left(\frac{1}{n}\right). \tag{2.23}$$

2.4 Singular learning theory

Laplace approximation cannot be applied to the average Kullback-Leibler divergence of hierarchical model such as Gaussian Mixture or neural networks because of the degeneration of Fisher information matrix. In this section, we briefly review the theory of asymptotic behavior of Bayesian generalization error and free energy in such singular cases.[42]

Theorem 2.4.1 (Resolution of singularities)**.** *Let W be a bounded closed set in R^d. Assume that $K(w)$ is nonnegative analytic function on W and that the set $\{w \in W : K(w) = 0\}$ is not empty. Then there exists $\epsilon > 0$, $\{W_i : W_i \in W\}$, and $\{U_i : U_i \in R^d\}$ which satisfy*

$$\{w \in W : K(w) \leq \epsilon\} = \bigcup_i W_i \tag{2.24}$$

and, in each pair W_i and U_i, there exists an analytic map $g : U_i \to W_i$ which satisfies

$$K(g(u)) = u^{2k}, \tag{2.25}$$

$$|\det g'(u)| = b(u)|u^h| \tag{2.26}$$

where $k, h (k > 0, h \geq 0)$ are d-dimensional multi-indexes, and $b(u) > 0$ and $g'(u)$ is Jacobian matrix.

Definition 2.4.2 (Real Log Canonical Threshold)**.** *Let W be a bounded closed set in R^d. Assume that $K(w)$ is an analytic function of $w \in R^d$ and $\varphi(w)$ is C^∞ function with compact support $W \subset R^d$. Then, the zeta function is defined by following with a complex variable z*

$$\zeta(z) = \int_W K(w)^z \varphi(w) \mathrm{d}w.$$

This function is holomorphic in $\text{Re}(z) > 0$. A real log canonical threshold (RLCT) is defined by the negative maximum pole of ζ and its multiplicity is defined by the order of the maximum pole.

The normalized marginal likelihood is defined by

$$Z_n^0 = \int_{w \in W} \exp(-nK_n(w))\varphi(w)\mathrm{d}w.$$

The normalized marginal likelihood $Z_n^{(0)}$ can be divided into $Z_n^{(1)}$ and $Z_n^{(2)}$ as

$$Z_n^{(0)} = Z_n^{(1)} + Z_n^{(2)}, \tag{2.27}$$

$$Z_n^{(1)} = \int_{w \in W_i, K(w) < \epsilon} \exp(-nK_{ni}(w))\varphi(w)\mathrm{d}w, \tag{2.28}$$

$$Z_n^{(2)} = \int_{w \in W_i, K(w) \geq \epsilon} \exp(-nK_{ni}(w))\varphi(w)\mathrm{d}w. \tag{2.29}$$

For $Z_n^{(1)}$, $Z_n^{(2)}$,

Lemma 2.4.3. *Let $\epsilon > 0$ be a monotonically decreasing function of n which satisfies*

$$\lim_{n \to \infty} \epsilon = 0, \tag{2.30}$$

$$\lim_{n \to \infty} \sqrt{n}\epsilon = \infty. \tag{2.31}$$

If ϵ satisfies these conditions, then

$$Z_n^{(2)} = O_p(\exp(-\sqrt{n})). \tag{2.32}$$

Lemma 2.4.4. *If the log density ratio function has a relatively finite variance,*

$$Z_n^{(1)} = \frac{(\log n)^{m-1}}{n^\lambda} \int \mathrm{d}u_i^* \int t^{\lambda-1} \exp(-t + \sqrt{t}\xi_n(u))\mathrm{d}t + o_p\left(\frac{(\log n)^{m-1}}{n^\lambda}\right) \tag{2.33}$$

holds. In this equation, λ and m are respectively the real log canonical threshold and the multiplicity of zeta functions, and $\xi_n(u)$ is an empirical process that converges in distribution to a Gaussian process. $\mathrm{d}u^*$ is a measure represented by u, k, h of U_i on which there are the real log canonical threshold and the multiplicity.

The realizable case of this lemma is proven in [41] and non-realizable case is in [42]. From Lemma2.4.4 and 2.3.1 we can get the asymptotic form of free energy and the generalization error in singular case that

$$E[F_n] = n(S + \mathrm{Bias}) + \lambda \log n - (m-1)\log\log n + O(1), \tag{2.34}$$

$$E[G_n] = \mathrm{Bias} + \frac{\lambda}{n} + o\left(\frac{1}{n}\right). \tag{2.35}$$

Chapter 3

Learning theory for non-identifiable optimal probability distributions

3.1 Motivation

In statistical learning theory, a probability distribution which generates a sample is called a true distribution and one with a parameter is called a statistical model or a learning machine. A probability distribution is estimated by applying a training algorithm to a statistical model. Then, the difference between the true distribution and the estimated one is defined by some measure, for example, the Kullback-Leibler (KL) divergence. In practical applications, the true distribution is unknown, hence the free energy and the generalization loss, which give the relative difference of KL divergence, are used to evaluate the estimated one.

The theoretical values of the free energy and the generalization loss strongly depend on the geometrical situations of the true distribution and a statistical model. A statistical model is called regular if the parameter which minimizes the KL divergence of a true distribution and the statistical model is unique and Hessian matrix of the KL divergence at the minimum point is regular. For the regular case, the asymptotic behavior of the generalization loss was revealed by Akaike[1], while that of the free energy was revealed by Schwarz[35]. These results have been applied to statistical model selection criteria, i.e. Akaike(AIC), Bayesian(BIC), Deviance(DIC)[37], and Adjusted Bayesian(ABIC)[2].

If a statistical model is not regular, then it is called singular. Many practical probabilistic models such that neural network, normal mixture model, Bayesian network are singular. On some singular models, asymptotic behavior of generalization loss and free energy are revealed when statistical model includes the true distribution in other words bias equals to 0[51, 33]. These studies based on the algebraic geometrical methods[41]. The model selection criterion for singular cases:WAIC[44], WBIC[46], sBIC[6] are also suggested. In former studies in singular case, there exists an assumption which holds when bias equals to 0. However, it is not clear whether the assumption holds or not with larger bias. In particular if probability distribution on optimal parameter is not unique, this assumption does not always hold. In such cases asymptotic behavior of generalization loss and free energy are unknown. Whether WAIC and WBIC select appropriate models or not is also unclear.

In this chapter, we show the asymptotic behavior of the generalization error and the free energy when the optimal probability distributions are not unique. We show that the

generalization error gets larger as the sample size increases in this case. This behavior is not shown in conventional studies of singular model in Bayesian statistics. This behavior is near to Deep Double Descent [29]. In addition, we show that with small sample size, the bias of generalization is smaller than apparent bias. This situation occurs when complex true distribution is estimated by singular statistical model such as neural networks.

3.2 Theoretical conditions

In this section, we summarize some theoretical conditions in the previous researches and in this paper. Many studies required the following condition.

Definition 3.2.1 (Regular)**.** *Let $J(w_0)$ be Hessian matrix of KL divergence of $q(x)$ and $p(x|w)$ on optimal parameter $w = w_0$. If w_0 is a single point and $J(w_0)$ is a regular matrix, the statistical model is regular.*

This is one of the conditions of asymptotic normality. It is known that if the statistical model is regular, asymptotic behaviors of the free energy and the generalization loss have asymptotic expansion[35][2],

$$\mathbb{E}[F_n] = nL(w_0) + \frac{d}{2}\log n + O(1), \tag{3.1}$$

$$\mathbb{E}[G_n] = L(w_0) + \frac{d}{2n} + o\left(\frac{1}{n}\right). \tag{3.2}$$

If Hessian matrix is not regular, the statistical model is singular. In this case, asymptotic behaviors of the free energy and the generalization loss are also known[41][42] as,

$$\mathbb{E}[F_n] = nL(w_0) + \lambda \log n - (m-1)\log\log n + O(1), \tag{3.3}$$

$$\mathbb{E}[G_n] = L(w_0) + \frac{\lambda}{n} - \frac{m-1}{n\log n} + o\left(\frac{1}{n\log n}\right), \tag{3.4}$$

where $\lambda > 0$ is a rational number called the real log canonical threshold (RLCT) and $m \geq 1$ is a natural number called the multiplicity. Definitions of these numbers are described in Definition 2.4.2. These equations require an assumption(see Assumption 3.2.3).

The following condition is also set in many studies.

Definition 3.2.2 (Realizable)**.** *If a statistical model satisfies $q(x) = p(x|w_0)$, $q(x)$ is said to be realizable by $p(x|w)$.*

This condition means a statistical model includes true distribution. In other words, bias equals to 0 in this condition.

To focus on model selection criteria, AIC[1] and DIC[37] require both of regular and realizable for asymptotic equality to generalization loss. So, these criteria are useful when bias is small and eigenvalues of Hessian matrix is large. On the other hand, BIC[35] and ABIC[1] don't require realizable condition for asymptotic equality to free energy. sBIC[6] don't require regular condition for asymptotic equality to free energy. WAIC[44] and WBIC[46] don't require both of regular and realizable conditions. Thus, these conditions are applicable conditions of Information criteria.

In singular and realizable case, asymptotic behaviors of generalization errors are revealed in some statistical models in Bayesian estimation :Normal mixture[51], reduced rank regression[5], naive Bayesian network[33], Markov model[55], and latent Dirichlet allocation[14]. These researches reveal variance of each statistical model when bias equals to 0 by calculating RLCT and the multiplicity. By studying not only RLCT and the multiplicity but also higher order random variable terms, [19] constructs the Bayesian test of normal mixture. Asymptotic behavior of free energy of Bayesian network when the data are not simple independent identical random variables is also studied[50]. This is one of the studies removing conventional assumption in singular case.

Those singular case studies require realizable condition. On the other hand some regular case studies don't require realizable condition. To treat these conditions simultaneously the following condition is suggested.

Assumption 3.2.3. *If the following condition is satisfied, it is said that the log density ratio function has a relatively finite variance.*

$$\exists c_0 > 0, \forall w \in W, \forall w_0 \in W_0, \ \ \mathbb{E}_X[f(X, w_0, w)] \geq c_0 \mathbb{E}_X[f(X, w_0, w)^2], \tag{3.5}$$

The following theorem holds for this assumption.

Theorem 3.2.4. *If a statistical model is regular or realizable, the log density ration function has a relatively finite variance.*

The proof of this theorem is written in [47]. From this theorem it follows that the theorems under Assumption3.2.3 also hold with regular case or realizable case. Asymptotic equivalence of generalization loss and WAIC, free energy and WBIC, equation 3.3, 3.4 required this assumption. Thus, these information criteria and equations hold in regular or realizable case. For example, it is known that if a statistical model is regular, $\lambda = d/2$ and $m = 1$ therefore equation3.3 is same as equation3.1 and therefore equation3.4 is same as equation3.2.

This assumption holds when bias is near to 0 or the statistical model is regular, but other than those conditions, it is not unclear whether this assumption holds or not. Thus it is not also unclear that whether WAIC and WBIC are useful or not and equation3.3, equation3.4 hold. In particular, when a statistical model does not satisfy the following condition that model does not always satisfy Assumption3.2.3.

Definition 3.2.5 (Essentially Unique)**.** *If $p(x|w_0)$ is unique probability distribution for all w_0, the statistical model is essentially unique.*

The following theorem holds for this condition.

Theorem 3.2.6. *If the log density ratio function has a relatively finite variance, the statistical model is essentially unique.*

Proof If a statistical model satisfies Assumption3.2.3, there exists a positive real number $c_0 > 0$ such that for any $w_{01}, w_{02} \in W_0$,

$$0 = L(w_{01}) - L(w_{02}) = \mathbb{E}_X[f(X, w_{01}, w_{02})] \geq c_0 E_X[f(X, w_{01}, w_{02})^2], \tag{3.6}$$

and it follows that $f(x, w_{01}, w_{02}) = 0$ for all x, which means $p(x|w_{01}) = p(x|w_{02})$.

From contraposition of this theorem, if a statistical model is not essentially unique, the statistical model does not satisfy Assumption3.2.3. In this paper, for constructing the theory in non-essentially unique case, we discuss with following relaxed version of Assumption3.2.3.

Assumption 3.2.7.

$$\exists c_0 > 0, \forall w \in W, \exists w_0 \in W_0, \ \ \mathbb{E}_X[f(X, w_0, w)] \geq c_0 \mathbb{E}_X[f(X, w_0, w)^2], \tag{3.7}$$

From definition, Assumption3.2.7 includes the assumption3.2.3. Assumption3.2.7 holds if there exists even one w_0 for all w satisfying the inequality even though Assumption3.2.3 required all w_0 for all w.

Example. The following is an example for the case that the optimal probability distribution is not unique. We suppose supervised learning of $q(y|x)$ using a statistical model $p(y|x, a, b)$. The variables a, b are parameters. We suppose true distribution $q(y|x), q(x)$ is

$$q(y|x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - f(x))^2}{2}\right),$$
$$q(x) = \begin{cases} \frac{1}{4} & (-2 \leq x \leq 2) \\ 0 & (\text{otherwise}) \end{cases},$$
$$f(x) = \begin{cases} x + 2 & (-2 \leq x < -1) \\ 1 & (-1 \leq x < 1) \\ -x + 2 & (1 \leq x \leq 2) \end{cases}.$$

We also suppose a statistical model $p(y|x, a, b)$,

$$p(y|x, a, b) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - \sigma(ax + b))^2}{2}\right) \tag{3.8}$$
$$\sigma(x) = \frac{1}{1 + \exp(-x)}. \tag{3.9}$$

In this situation, the KL divergence between $q(y|x)$ and $p(y|x, a, b)$ can be calculated as

$$KL(q(y, x)|p(y, x|a, b)) = \int q(y, x) \log \frac{q(y|x)}{p(y|x, a, b)} = \int q(y|x)q(x) \log \frac{q(y|x)}{p(y|x, a, b)} \mathrm{d}x\mathrm{d}y, \tag{3.10}$$
$$= \frac{1}{2} \int q(x)(f(x) - \sigma(ax + b))^2 \mathrm{d}x. \tag{3.11}$$

Note that $q(x)$ and $f(x)$ are even functions. In addition, $\sigma(ax + b)$ and $\sigma(-ax + b)$ are line symmetric on $x = 0$. Therefore, $\mathrm{KL}(q(y|x)|p(y|x, a, b)) = \mathrm{KL}(q(y|x)|p(y|x, -a, b))$ holds. Thus, we can find two optimal parameters as (a_0, b_0) and $(-a_0, b_0)$. These two points satisfy $p(y, x|a_0, b_0) \neq p(y, x| - a_0, b_0)$. A numerical calculation shows that the optimal parameters in this case are $(5.13, 7.71)$, $(-5.13, 7.71)$. Note that each neighbor of optimal parameter $(5.13, 7.71)$, $(-5.13, 7.71)$ is same as regular condition, thus in this case statistical model satisfies the Assumption3.2.7.

3.3 Main theorem

We estimate a probability distribution $q(x)$ by Bayesian inference using a statistical model $p(x|w)$ and prior distribution $\varphi(w)$. It is assumed that the set of parameters W is compact and that $p(x|w)$ is continuous for w.

Problem treated in this paper. We assume about the set W_0.

Assumption 3.3.1. *We assume that the set W_0 can be represented by a disjoint union of $W_{0i}(i \in I)$ which satisfies*

$$\cup_{i \in I} W_{0i} = W_0, \tag{3.12}$$

$$W_{0i} \cap W_{0j} = \varnothing (i \neq j), \tag{3.13}$$

where in each subset W_{0i} for $w_{0i} \in W_{0i}$ optimal probability distribution $p(x|w_{0i})$ is unique on the other hand if $i \neq j, w_{0i} \in W_{0i}, w_{0j} \in W_{0j}$, $p(x|w_{0i}) \neq p(x|w_{0j})$

In particular, if I has only a single element, it is same as essentially unique case.
The main result of this paper is the following Theorem3.3.2.

Theorem 3.3.2. *If a statistical model satisfies Assumption3.2.7 and Assumption3.3.1, index set I are natural numbers the free energy and generalization loss has an asymptotic expansion that*

$$\mathbb{E}[F_n] = nL(w_0) - \sqrt{n}\mu + \hat{\lambda} \log n - (\hat{m} - 1) \log \log n + O(1), \tag{3.14}$$

$$\mathbb{E}[G_n] = L(w_0) - \frac{\mu}{2\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right) \tag{3.15}$$

where $\mu, \hat{\lambda}, \hat{m}$ are real numbers satisfying $\mu \geq 0$, $\hat{\lambda} > 0$, and $\hat{m} \geq 1$.

3.4 Proof of main theorem

In this section, we discuss the proof of Theorem3.3.2 in case the number of optimal probability distributions is finite.

To show Theorem3.3.2, let us represent the set of parameters as

$$W = \cup_{i=1}^{m} W_i,$$

where $W_{0i} \subset W_i$. We define the empirical log function as

$$L_n(w) = -\frac{1}{n} \sum_{j=1}^{n} \log p(X_j|w). \tag{3.16}$$

We also define the index empirical log function $L_{ni}(w)$ and the index empirical error function $K_{ni}(w)$ as

$$L_{ni} = L_n(w_{0i}). \tag{3.17}$$

$$K_{ni}(w) = \frac{1}{n} \sum_{j=1}^{n} f(X_j, w_{0i}, w). \tag{3.18}$$

The marginal likelihood of each domain $Z_n^{(i)}$ is also defined by

$$Z_n^{(i)} = \int_{w \in W_i} \exp(-nL_n(w)) \varphi(w) \mathrm{d}w. \tag{3.19}$$

By using $L_n(w) = K_{ni}(w) + L_{ni}$, we have

$$Z_n^{(i)} = \exp(-nL_{ni}) \int_{w \in W_i} \exp(-nK_{ni}(w))\varphi(w)\mathrm{d}w. \tag{3.20}$$

The index normalized marginal likelihood of each domain Z_n^{0i} is defined by

$$Z_n^{0i} = \int_{w \in W_i} \exp(-nK_{ni}(w))\varphi(w)dw. \tag{3.21}$$

In accordance with these definitions, the marginal likelihood is given by

$$Z_n = \int_{w \in W} \exp(-nL_n(w))\varphi(w)dw = \sum_{i=1}^{m} \exp(-n\beta L_{ni}) Z_n^{0i}. \tag{3.22}$$

The normalized marginal likelihood Z_n^{0i} can be divided into Z_n^{1i} and Z_n^{2i} as

$$Z_n^{(0i)} = Z_n^{(1i)} + Z_n^{(2i)}, \tag{37}$$

$$Z_n^{(1i)} = \int_{w \in W_i, K(w) < \epsilon} \exp(-nK_{ni}(w))\varphi(w)\mathrm{d}w, \tag{3.23}$$

$$Z_n^{(2i)} = \int_{w \in W_i, K(w) \geq \epsilon} \exp(-nK_{ni}(w))\varphi(w)\mathrm{d}w. \tag{3.24}$$

From Lemma2.4.3, with appropriate ϵ,$Z_n^{(2i)} = O_p(\exp(-\sqrt{n}))$ holds. The optimal parameter set in each W_i is W_{0i}; therefore, in this set, the optimal probability distribution is only $p(x|w_{0i})$. In Assumption3.2.7, w_0 corresponding to $w \in W_{0i}$ is only w_{0i}. Thus

$$\exists c_0 \forall w \in W_i \forall w_{0i} \;\; E_X[f(x, w_{0i}, w_i)] \geq c_0 E_X[f(x, w_{0i}, w_i)^2]. \tag{3.25}$$

holds. This condition is corresponding to Assumption3.2.3. Then, applying the Lemma2.4.4, there exists a measure du_i^* such that

$$Z_n^{(1i)} = \frac{(\log n)^{m_i - 1}}{n^{\lambda_i}} \int \mathrm{d}u_i^* \int t^{\lambda_i - 1} \exp(-\beta t + \beta\sqrt{t}\xi_{ni}(u_i))\mathrm{d}t + o_p\left(\frac{(\log n)^{m_i-1}}{n^{\lambda_i}}\right). \tag{3.26}$$

Thus, the marginal likelihood is

$$Z_n = \sum_{i=1}^{m} \exp(-nL_{ni}) Z_n^{(0i)} \tag{3.27}$$

$$= \sum_{i=1}^{m} \exp(-nL_{ni}) \frac{(\log n)^{m_i-1}}{n^{\lambda_i}} \left(\int \mathrm{d}u_i^* \int t^{\lambda_i - 1} \exp(-t + \sqrt{t}\xi_{ni}(u_i))\mathrm{d}t + o_p(1) \right). \tag{3.28}$$

From (4), the free energy is given by

$$\begin{aligned} F_n &= -\log Z_n \\ &= f_1 - f_1 + g_1 - g_1 + g_2 - g_2 - \log Z_n, \\ &= f_1 + f_2 + f_3 + o_p(1), \end{aligned} \tag{3.29}$$

where

$$g_1 = -\log\left(\sum_{i=1}^{m} e^{(-nL_{ni}-\lambda_i \log n+(m_i-1)\log\log n)}\right), \tag{3.30}$$

$$g_2 = -\log\left(\sum_{i=1}^{m} e^{(-\Theta(\beta,\xi_{ni})-nL_{ni}-\lambda_i \log n+(m_i-1)\log\log n)}\right), \tag{3.31}$$

$$f_1 = -\log\left(\sum_{k=1}^{m} e^{(-nL_{ni})}\right), \tag{3.32}$$

$$f_2 = -f_1 + g_1 = -\log\left(\frac{\sum_{i=1}^{m} e^{(-nL_{ni}-\lambda_i \log n+(m_i-1)\log\log n)}}{\sum_{k=1}^{m} e^{(-n\ L_{ni})}}\right), \tag{3.33}$$

$$f_3 = -g_1 + g_2 = -\log\left(\frac{\sum_{i=1}^{m} e^{(-\Theta(\xi_{ni})-nL_{ni}-\lambda_i \log n+(m_i-1)\log\log n)}}{\sum_{i=1}^{m} e^{(-nL_{ni}-\lambda_i \log n+(m_i-1)\log\log n)}}\right). \tag{3.34}$$

In the above equations, we have used the notation,

$$\Theta(\xi_{ni}) \tag{3.35}$$

$$= -\log\left(\int \mathrm{d}u_i^* \int_0^\infty \mathrm{d}t\, t^{\lambda_i-1} \exp(-t+\sqrt{t}\xi_{ni}(u_i))\right). \tag{3.36}$$

In the following, we examine the asymptotic behaviors of the three terms eq3.29. First, to study f_1, we define i_{max} and Y by

$$i_{\mathbf{max}} = \underset{i}{\mathrm{argmax}}\ (-L_{ni}), \tag{3.37}$$

$$Y = -nL_{ni_{max}}. \tag{3.38}$$

From the definition,

$$\log\left(\sum_{i=1}^{m} e^{(-nL_{ni})}\right) - Y = \log\left(1 + \sum_{i\neq i_{max}}^{m} e^{(-nL_{ni}-Y)}\right). \tag{3.39}$$

Since $Y + nL_{ni} \geq 0$,

$$0 < \log\left(\sum_{i=1}^{m} e^{(-nL_{ni})}\right) - Y \leq m\log 2. \tag{3.40}$$

Therefore,

$$f_1 = Y + O_p(1) = -nL_{ni_{\mathbf{max}}} + O_p(1).$$

Note that the average of L_{ni} is $L(w_{0i})$, which does not depend on i. Let us define a random variable,

$$\mathscr{L}_n(w_{0i}) \equiv \sqrt{n}(-L_n(w_{0i}) + L(w_{0i})). \tag{3.41}$$

By using the central limit theorem, $\mathscr{L}_n(w_{0i})$ $(i = 1, 2, ..., m)$ converges in distribution to an m-dimensional Gaussian random variable $\mathscr{L}(w_{0i})$ on $w_{0i} \in W_0$ whose average is zero and variance-covariance matrix $V = (V_{ij})$ is

$$V_{ij} = \mathbb{E}_X[(\log p(X|w_{0i}) + L(w_0))(\log p(X|w_{0j}) + L(w_0))]. \tag{3.42}$$

Then we obtain

$$f_1 = nL(w_0) - \sqrt{n} \max_{w_0 \in W_0} \mathscr{L}_n(w_0) + O_p(1). \tag{3.43}$$

Using $\mathscr{L}(w_0)$, the asymptotic behavior of its average is given by

$$\mathbb{E}[f_1] = nL(w_0) - \mathbb{E}[\sqrt{n} \max_{w_0 \in W_0} \mathscr{L}(w_0)] + O(1). \tag{3.44}$$

Now, let us examine the second term f_2 in eq3.29. If there exist multiple i_{max}, we define i_{max} as the i whose RLCT is smallest, and if the RLCTs are the same, we define i_{max} as the i whose multiplicity is biggest. Using $i_{\mathbf{max}}$ so defined, the asymptotic behavior of the second term f_2 is given by

$$-f_2 = \log \left(\frac{\sum_{i=1}^m e^{(-nL_{ni} - \lambda_i \log n + (m_i - 1) \log \log n)}}{\sum_{i=1}^m e^{(-nL_{ni})})} \right) \tag{3.45}$$

$$= -\lambda_{i_{\mathbf{max}}} \log n + (m_{k_{\mathbf{max}}} - 1) \log \log n \tag{3.46}$$

$$+ \log \left(\frac{1 + \sum_{k \neq i_{\mathbf{max}}}^m e^{(-nL_{nk} - Y - (\Delta\lambda_k) \log n + (\Delta m_k) \log \log n)}}{1 + \sum_{k \neq i_{max}}^m e^{(-nL_{ni} - Y)})} \right) \tag{3.47}$$

$$= -\lambda_{i_{\mathbf{max}}} \log n + (m_{i_{\mathbf{max}}} - 1) \log \log n + \log \left(\frac{1 + o_p(e^{-na})}{1 + o_p(e^{-na})} \right) \tag{3.48}$$

$$= -\lambda_{i_{\mathbf{max}}} \log n + (m_{i_{\mathbf{max}}} - 1) \log \log n + o_p(e^{-na}), \tag{3.49}$$

where $\Delta\lambda_i$ and Δm_i are $\lambda_i - \lambda_{i_{\mathbf{max}}}$ and $m_i - m_{i_{\mathbf{max}}}$ respectively, and $a \geq 0$ is the difference between Y and the second biggest $-nL_{ni}$.

$$\mathbb{E}[f_2] = \lambda_{i_{max}} \log n - (m_{i_{\mathbf{max}}} - 1) \log \log n + O(1). \tag{3.50}$$

Next, let us study the third term f_3 in eq3.29. We define a random variable a_i as follows.

$$a_i = \frac{e^{-nL_{ni} - \lambda_i \log n + (m_i - 1) \log \log n}}{\sum_{i=1}^m e^{(-nL_{ni} - \lambda_i \log n + (m_i - 1) \log \log n)}}. \tag{3.51}$$

The sum of a_i over i is 1. Using a_i is 1. We can describe the third term using a_i, so we can describe the free energy as follows.

$$F_n = nL(w_0) - \sqrt{n} \max_{w_0 \in W_0} \mathscr{L}_n(w_0) + \lambda_{i_{\mathbf{max}}} \log n - (m_{i_{\mathbf{max}}} - 1) \log \log n \tag{3.52}$$

$$- \log \left(\sum_{i=1}^m a_i e^{-\Theta(\xi_{ni})} \right) + O_p(1). \tag{3.53}$$

Lastly, in order to derive the asymptotic behavior of the average $\mathbb{E}[F_n]$, we show that the average of the random variable

$$f_4 \equiv \log\left(\sum_{i=1}^{m} a_i e^{-\Theta(\xi_{ni})}\right)$$

is finite. By the Cauchy-Schwarz inequality,

$$-\frac{t+\sup_u|\xi_{ni}(u_i)|^2}{2} \leq \sqrt{t}\xi_{ni}(u_i) \leq \frac{t+\sup_u|\xi_{ni}(u_i)|^2}{2} \tag{3.54}$$

holds. In the integral range of u_i,$[0,1]^d$, we have

$$-\log\int du_i^* - \log\int dt\, t^{\lambda_i-1}\exp(-\frac{1}{2}t) - \frac{1}{2}\sup_{u_i\in[0,1]^d}|\xi_{ni}(u_i)|^2 \tag{3.55}$$

$$\leq \Theta(\xi_{ni}(u_i)) \tag{3.56}$$

$$\leq -\log\int du_i^* - \log\int dt\, t^{\lambda_i-1}\exp(-\frac{3}{2}t) + \frac{1}{2}\sup_{u_i\in[0,1]^d}|\xi_{ni}(u_i)|^2. \tag{3.57}$$

Using Jensen's inequality,

$$f_4 \quad \geq \sum_{i=1}^{m} a_i(-\Theta(\xi_{ni}(u))) \tag{3.58}$$

$$\geq \sum_{i=1}^{m} a_i\left(\log\int du_i^* + \log\int dt\, t^{\lambda_i-1}\exp^{-\frac{3\beta}{2}t} - \frac{1}{2}\sup_{u_i\in[0,1]^d}|\xi_{ni}(u_i)|^2\right) \tag{3.59}$$

holds. From the nature of empirical processes, $\xi_{ni}(u_i)$ converge in low to Gaussian processes $\xi_i(u_i)$ and

$$\lim_{n\to\infty} E[\sup_{u_i\in[0,1]^d}|\xi_{ni}(u_i)|^2] = \lim_{n\to\infty} E[\sup_{u_i\in[0,1]^d}|\xi_i(u_i)|^2] \tag{3.60}$$

holds[40]. In regard to the lower bound of $\log\left(\sum_{i=1}^m a_i e^{-\Theta(\xi_{ni}(u))}\right)$, the only term that may diverge as a random variable is $\xi_{ni}(u_i)$, so $E[\log\left(\sum_{i=1}^m a_i e^{-\Theta(\xi_{ni}(u))}\right)]$ can be shown to be bounded below. In addition, because $\sup_{u_i\in[0,1]^d}|\xi_{ni}(u_i)|^2 \geq 0$ holds, we have

$$f_4 \quad \leq \log\left(\max_i \int dt\, t^{\lambda_i-1}e^{-\frac{1}{2}t}\int du_i^*\right)\left(me^{\frac{1}{2}\sum_{i=1}^m \sup_u|\xi_{ni}(u_i)|^2}\right) \tag{3.61}$$

$$= \log\left(\max_i \int dt\, t^{\lambda_i-1}e^{-\frac{1}{2}t}\int du_i^*\right) + \log m + \frac{1}{2}\sum_{i=1}^{m}\sup_u|\xi_{nk}(u_i)|^2. \tag{3.62}$$

Hence, as we did for the lower bound it can be shown that $\mathbb{E}[\log\left(\sum_{i=1}^m a_i e^{-\Theta(\xi_{ni}(u))}\right)]$ is bounded from above. By summing up the above equations, the asymptotic behavior of $\mathbb{E}[F_n]$

can be described as

$$\mathbb{E}[F_n] = nL(w_0) - \sqrt{n}\mathbb{E}[\max_{w_0 \in W_0} \mathscr{L}(w_0)] + \sum_{k=1}^{m} \alpha_i(\lambda_i \log n - (m_i - 1)\log\log n) + O(1), \tag{3.63}$$

where α_i is the probability that $i = i_{max}$. By putting $\hat{\lambda} = \sum_{k=1}^{m} \alpha_i \lambda_i$ and $\hat{m} = \sum_{k=1}^{m} \alpha_i m_i$, we obtain

$$\mathbb{E}[F_n] = nL(w_0) - \sqrt{n}\mathbb{E}[\max_{w_0 \in W_0} \mathscr{L}(w_0)] + \hat{\lambda}\log n - (\hat{m} - 1)\log\log n + O(1) \tag{3.64}$$

holds.
Using Theorem2.3.1, and assuming that $\mathbb{E}[G_n(1)]$ has an asymptotic expansion, we find that

$$\mathbb{E}[G_n] = L(w_0) - \frac{1}{2\sqrt{n}}\mathbb{E}[\max_{w_0 \in W_0} \mathscr{L}(w_0)] + o\left(\frac{1}{\sqrt{n}}\right). \tag{3.65}$$

From this Theorem3.3.2 is proved.

3.5 Experiment

In this section, we show the results of an experiment for the case when the optimal probability distribution is not unique.

3.5.1 Experiment1

We set the true distribution as

$$q(y|x) = \frac{1}{\sqrt{0.08\pi}} \exp\left(-\frac{(y - f(x))^2}{0.08}\right).$$

$$f(x) = \begin{cases} x + 2 & (-2 \leq x < -1) \\ 1 & (-1 \leq x < 1) \\ -x + 2 & (1 \leq x \leq 2) \end{cases}$$

$$q(x) = \begin{cases} \frac{1}{4} & (-2 \leq x \leq 2) \\ 0 & \text{otherwise.} \end{cases}$$

We use following statistical model and prior distributions.

$$p(y|x, a, b) = \frac{1}{\sqrt{0.08}} \exp\left(-\frac{(y - \sigma(ax + b))^2}{0.08}\right).$$

$$\sigma(x) = \frac{1}{1 + \exp(-x)}.$$

$$a \sim \textbf{Unifrom}(0, 20)$$

$$b \sim \textbf{Unifrom}(-20, 20)$$

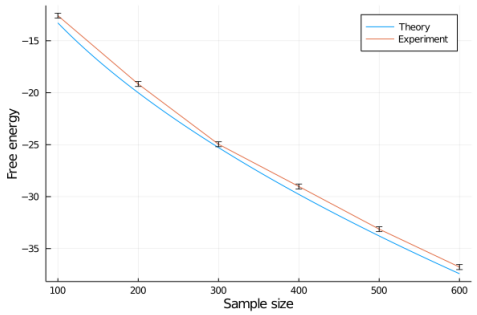


Figure 3.1: Experimental value and theoretical free energy depending on the sample size. The error bar is the SE of the average of free energy. The theoretical value of log likelihood($nL(w_0)$) is subtracted from each value.

This statistical model has two optimal parameters

$$w_{01} = (5.13, 7.71), w_{02} = (-5.13, 7.71).$$

At these points,

$$p(x|w_{01}) \neq p(x|w_{02})$$

holds. In this case, eq3.64 gives the theoretical asymptotic behavior of the free energy versus inverse temperature for $\beta = 1$. Note that the KL-divergence between $q(y, x)$ and $p(y, x|a, b)$ in each neighborhood of the optimal parameter is regular, so $\lambda = 1$ and $m = 1$. The expectation of the maximum value of a 2-dimensional Gaussian distribution follows a 1-dimensional Gaussian distribution (see the appendix). We will show that the theoretical behavior of free energy obeys

$$\mathbb{E}[F_n(1)] = nL(w_0) - \sqrt{n}\sqrt{\frac{\mathbb{V}[\log(p(y|x, a_0, b_0)) - \log(p(y|x, -a_0, b_0))]}{2\pi}} + \log n + O(1). \tag{3.66}$$

In eq3.66, $L(w_0)$ and the coefficient of $\sqrt{n}$ can be calculated by numerical integration. We used the average of F_n calculated from the true distribution $q(y|x), q(x)$ as the experimental value of $\mathbb{E}[F_n]$. The prior distribution $p(a)$, $p(b)$ does not have an effect on the asymptotic behavior. For this reason, we used equally spaced fixed values for integration. We compared this experimental values and theoretical values, except for the $O(1)$ term.

We calculated the experimental values of $\mathbb{E}[F_n]$ whose sample size were $n = 100$ to 600 every 100. For making variance the same size, we repeat to calculate F_n from 10000 times to 60000 times in steps of 10000 for each sample sizes. This means $\mathbb{E}[F_n]$ of $n = 100$ is average of 10000 samples of F_100. Figure1 compares the experimental and theoretical values. The experimental behavior of the free energy depending on the sample size is similar to the theoretical behavior. Figure2 shows the difference between the theoretical value and experimental value. This difference corresponds to $O(1)$ term. This difference is remains on this order regardless of the sample size. The experimental results support the theoretical formula, Theorem3.3.2.

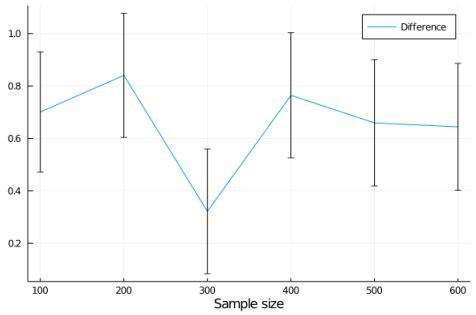


Figure 3.2: Difference between experimental value and theoretical value in Figure1. The error bar is the same as in Figure1.

3.5.2 Experiment2

We experimentally show that if optimal probability distribution is not unique, the generalization loss gets larger as the number of data becomes larger. We set the true distribution as 2-dimensional 4-component Gaussian mixture where

$$q(\boldsymbol{x}) = \sum_{i=1}^{4} \frac{1}{4} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}||\boldsymbol{x} - \overline{\boldsymbol{b}}^{(\boldsymbol{i})}||^2\right)$$
$$\overline{\boldsymbol{b}}^{(\boldsymbol{1})}, \overline{\boldsymbol{b}}^{(\boldsymbol{2})}, \overline{\boldsymbol{b}}^{(\boldsymbol{3})}, \overline{\boldsymbol{b}}^{(\boldsymbol{4})} = (1,1)^T, (1,-1)^T, (-1,1)^T, (-1,-1)^T.$$

We use 2-dimensional 2-component Gaussian mixture model and prior distribution where

$$p(\boldsymbol{x}|a, \boldsymbol{b}^{(\boldsymbol{1})}, \boldsymbol{b}^{(\boldsymbol{2})}) = \frac{a}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}||\boldsymbol{x} - \boldsymbol{b}^{(\boldsymbol{1})}||^2\right) + \frac{1-a}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}||\boldsymbol{x} - \boldsymbol{b}^{(\boldsymbol{2})}||^2\right)$$
$$\hat{a} \sim \textbf{Normal}(0, 4)$$
$$a = \frac{1}{1 + \exp(-\hat{a})}.$$
$$b_j^{(i)}, \sim \textbf{Unifrom}(-10, 10)\ (i = 1, 2\ j = 1, 2).$$

In this statistical model, mixture weight a is reparameterized for Markov chain Monte Carlo method(MCMC). The average of each Gaussian in true distribution $\overline{b}^{(1)}, \overline{b}^{(2)}, \overline{b}^{(3)}, \overline{b}^{(4)}$ is rotationally symmetric in $\pi/2$ for coordinate system of $\boldsymbol{x}$. We set the optimal parameter of statistical model as $(\hat{a}, \hat{\boldsymbol{b}}_{\boldsymbol{1}}, \hat{\boldsymbol{b}}_{\boldsymbol{2}})$. We also set the $R(\theta)$ as the rotation matrix where

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{3.67}$$

By using the symmetric property of $q(x)$ and the property of $R(\pi/2)$ the following equation holds.

$$\begin{aligned}\int q(\boldsymbol{x}) \log p(\boldsymbol{x}|\hat{a}, \hat{\boldsymbol{b}}^{(\mathbf{1})}, \hat{\boldsymbol{b}}^{(\mathbf{2})}) \mathbf{dx} &= \int q(R(-\pi/2)R(\pi/2)\boldsymbol{x}) \log p(R(\pi/2)\boldsymbol{x}|\hat{a}, R(\pi/2)\hat{\boldsymbol{b}}^{(\mathbf{1})}, R(\pi/2)\hat{\boldsymbol{b}}^{(\mathbf{2})}) \mathbf{d}\boldsymbol{x} \\ &= \int q(R(-\pi/2)\boldsymbol{x'}) \log p(\boldsymbol{x'}|\hat{a}, R(\pi/2)\hat{\boldsymbol{b}}^{(\mathbf{1})}, R(\pi/2)\hat{\boldsymbol{b}}^{(\mathbf{2})})|R(-\pi/2)| \mathbf{d}\boldsymbol{x'} \\ &= \int q(\boldsymbol{x'}) \log p(\boldsymbol{x'}|\hat{a}, R(\pi/2)\hat{\boldsymbol{b}}^{(\mathbf{1})}, R(\pi/2)\hat{\boldsymbol{b}}^{(\mathbf{2})}) \mathbf{d}\boldsymbol{x'} \quad (\boldsymbol{x'} = R(\pi/2)\boldsymbol{x}) \end{aligned} \tag{3.68}$$

From eq3.68, $(\hat{a}, R(\pi/2)\hat{\boldsymbol{b}}^{(\mathbf{1})}, R(\pi/2)\hat{\boldsymbol{b}}^{(\mathbf{2})})$ is also optimal parameter. Moreover,

$$p(x|\hat{a}, \hat{\boldsymbol{b}}^{(\mathbf{1})}, \hat{\boldsymbol{b}}^{(\mathbf{2})}) \neq p(x|\hat{a}, R(\pi/2)\hat{\boldsymbol{b}}^{(\mathbf{1})}, R(\pi/2)\hat{\boldsymbol{b}}^{(\mathbf{2})}) \tag{3.69}$$

holds in general, therefore optimal probability distribution is not unique.

We experimentally calculate the generalization loss$\mathbb{E}[G_n]$ in this case. From n-dataX^n generating from $q(\boldsymbol{x})$, we calculate posterior distribution $p(a, \boldsymbol{b}^{(\mathbf{1})}, \boldsymbol{b}^{(\mathbf{2})}|X^n)$ and predictive distribution $p(\boldsymbol{x}|X^n)$. In this experiment we use the No-U-Turn Sampler(NUTS)[17] to realize the posterior distribution. We calculate the experimental value of G_n by the average of $\log p(\boldsymbol{x}|X^n)$ of 10000 test sample. We estimate $\mathbb{E}[G_n]$ by the average of 100 times of this calculation.

Table 3.1: Experimental value of $\mathbb{E}[G_n]$

sample size	$\mathbb{E}[G_n]$	Standard error
100	2.56864	0.01091
200	2.58427	0.00991
300	2.58111	0.01013
400	2.58471	0.01082
500	2.58640	0.01105

Table1 shows the results of the experiment. $\mathbb{E}[G_n]$ increases by the sample size increase. There is a reversal of change $n = 200$, that may be caused by the effect of lower order term in $o(1/\sqrt{n})$. This result supports the main result.

3.6 Discussion

We found that if the optimal probability distribution is not unique, the apparent bias or the variance gets smaller for a finite sample number n corresponding to a Gaussian process determined by the log loss of the optimal parameter set, and the reduction converges to 0 asymptotically. This behavior can be explained qualitatively as follows: when there are two or more optimal probability distributions, the posterior distribution can be selected to be the nearest optimal probability distribution by bias of data, and this makes the generalization loss smaller than the average generation of data. As the number of samples and the bias increase, data generates averagely and generalization loss gets larger.

Both experiment1 and experiment2 are the case while the true distribution is symmetric, the statistical model is smaller than sufficiently realizing them. In real data analysis, if the generalization error gets larger by the sample size larger, it can be detected that the statistical model is smaller than appropriate size or the model does not fit to the unknown symmetric property of true distribution. In particular, such situations may occur depending on the relationship between dimensions in high dimensional data.

The behavior of generalization error shown in this paper is near to phenomena in Deep Double Descent[29]. The section "Sample-wise non-monotonicity" in [29] shows the behavior generalization error gets larger with increasing sample size. The test error shown in the experiment of [29] does not converge to 0 by increasing the sample size. Therefore, Deep Double Descent of sample-wise case is also the case that bias does not equal to 0. Although the relation between Deep Double Descent and the term discovered in this paper is unclear in the range of this paper, to develop the study of multiple optimal probability distribution case in particular the unfitness between true and statistical model in high dimensional distributions, it is possible to contribute clarifying the irregular behavior of neural network. The specification of unfitness between true and statistical model can help the efficient expansion of statistical model to estimate true distribution with lower bias and generalization error.

In this paper, we showed that the asymptotic behavior of the free energy and generalization loss are determined by $n^{\frac{1}{2}}$ and $n^{-\frac{1}{2}}$ order. Previous research[45] provides concrete example in which there is a unique optimal probability distribution but Assumption3.2.3 does not hold. In that paper, the asymptotic behavior of the free energy and the generalization loss are determined by $n^{\frac{1}{3}}$ and $n^{-\frac{2}{3}}$ order, so we predict that the lowest order determining the asymptotic behavior of the free energy and generalization loss are $n^{\frac{1}{2}}$ and $n^{-\frac{1}{2}}$.

We showed that the asymptotic behaviors of the free energy and generalization loss are determined by the maximum value of a Gaussian process. The probability distribution of the maximum value of a Gaussian process or multivariate normal distribution can not be calculated analytically, but an approximate calculation, called the "tube method"[22] exists. There is also a method for calculating the upper and lower bounds of the expectation of the maximum value of a Gaussian process, called "chaining"[39]. This maximum value is what determines the free energy and generalization loss in this paper.

3.7 Conclusion

We examined the case of when an important assumption in singular learning theory about the log density ration function is loosened. In this case there is a new term that is determined by a Gaussian process, whereby the generalization loss asymptotically increases as the size of the dataset increases. In the future, we should examine the asymptotic behavior of the generalization loss as a random variable, in particular the asymptotic equivalence of WAIC [44] and WBIC [46] in this case, and in the case in which the assumption is completely removed.

Appendix 3.A Maximum value of 2-dimensional Gaussian

We will derive the following equation.

$$\mathbb{E}[\max \mathscr{L}(w_0)] = \sqrt{\frac{\mathbb{V}[\log(p(x|w_{01})) - \log(p(x|w_{02}))]}{2\pi}} \tag{A1}$$

In this equation, $\mathscr{L}(w_0)$ is a 2-dimensional Gaussian distribution which average is 0 and variance-covariance matrix is

$$V_{ij} = \mathbb{E}[(\log p(x|w_{0i}) + L(w_0))(\log p(x|w_{0j}) + L(w_0))] \tag{A2}$$

We define two random variables as

$$z_1 = \mathscr{L}(w_{01}) - \mathscr{L}(w_{02}).$$
$$z_2 = \mathscr{L}(w_{01}) + \mathscr{L}(w_{02})$$

The random variables (z_1, z_2) are also from 2-dimensional Gaussian distribution whose average is 0 and variance-covariance matrix is

$$\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} V_{11} + V_{22} - V_{12} - V_{21} & V_{11} - V_{22} + V_{12} - V_{21} \\ V_{11} - V_{22} - V_{12} + V_{21} & V_{11} + V_{22} + V_{12} + V_{21} \end{pmatrix}.$$

The marginal distribution about z_1 is a 1-dimensional Gaussian distribution whose average is 0 and the variance is

$$V_{11} + V_{22} - V_{12} - V_{21}.$$

According to eq.(16) and eq.(34), we have

$$\begin{aligned} & V_{11} + V_{22} - V_{12} - V_{21} \\ & \quad = \mathbb{E}[(\log p(x|w_{01})(\log p(x|w_{01})] - L(w_0)^2 + \mathbb{E}[(\log p(x|w_{02})(\log p(x|w_{02})] \\ & \qquad - L(w_0)^2 - 2(\mathbb{E}[(\log p(x|w_{01})(\log p(x|w_{02})] - L(w_0)^2) \\ & \quad = \mathbb{E}[(\log p(x|w_{01}) - \log p(x|w_{02}))^2] \\ & \quad = \mathbb{V}[(\log p(x|w_{01}) - \log p(x|w_{02}))] \end{aligned} \tag{A3}$$

We define a random variable z_3

$$z_3 = \begin{cases} z_1 & z_1 \geq 0 \\ 0 & z_1 < 0 \end{cases}$$

By using z_3 we can describe the maximum value of $\mathscr{L}(w_0)$ in the following way,

$$\max\mathscr{L}(w_0) = \mathscr{L}(w_{02}) + z_3. \tag{A4}$$

Considering the average of $\mathscr{L}(w_{02})$ is 0, we find that

$$\mathbb{E}[\max\mathscr{L}(w_0)] = \mathbb{E}[z_3]. \tag{A5}$$

$\mathbb{E}[z_3]$ is the expectation of a positive value in a Gaussian distribution. This integration of a Gaussian whose variance is σ^2 can be calculated as

$$\begin{aligned} \int_0^\infty \frac{x}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) &= \left[-\frac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)\right]_0^\infty \\ &= \frac{\sigma}{\sqrt{2\pi}}. \end{aligned} \tag{A6}$$

From (A3), (A5), and(A6), we have

$$\mathbb{E}[\max \mathscr{L}(w_0)] = \mathbb{E}[z_3] = \sqrt{\frac{\mathbb{V}[\log(p(x|w_{01})) - \log(p(x|w_{02}))]}{2\pi}}.$$

Therefore, (A1) holds.

Chapter 4

Free Energy of Bayesian Convolutional Neural Networks

4.1 Motivation

Convolutional Neural Networks (CNNs) are a type of Neural Networks mainly used for computer vision. CNNs have been shown high performance with deep layers[38, 21]. Residual Network(ResNet)[15] adopted the skip connection for addressing the problem that the loss function of CNN with deep layers does not decrease well through optimization. After success of ResNet, the CNNs with more than 100 layers are realized. The high performance of ResNet has been explained by similarity to the ensemble learning [18, 31, 10]. On the other hand, there is a common issue in neural networks that the reason why the overparametrized deep neural network generalized has been unknown yet.

In conventional learning theory, if the Fisher information matrix of a learning machine is positive definite, and the data size is sufficient large, the generalization error of the learning machine is determined from the number of its parameter in maximum likelihood estimator[1]. The similar property is shown in free energy and generalization error in Bayesian learning[35, 32, 2]. From these characteristics of generalization error and free energy some information criteria such as AIC, BIC and MDL are proposed. However, most of the hierarchical models such as neural networks have degenerated Fisher information matrix. In such models, the Bayesian generalization error and free energy are determined by a rational number called Real Log Canonical Threshold(RLCT) and that is smaller than the number of parameters [42, 43]. In particular, RLCTs are revealed in some concrete models such as three layered neural networks[41, 4], normal mixtures [12, 51], Poisson mixtures[34], Boltzmann machine[52, 3], reduced rank regression[5], Latent Dirichlet allocation[14], matrix factorization, and Bayesian Network[53]. While RLCTs of many hierarchical models are revealed, that of neural networks with multiple layer of nonlinear transformation has not been clarified. Yet the possibility of that is shown in [48], the RLCT of Deep Neural Network is revealed[27]. On the other hand the RLCT of neural networks other than DNN was not explored.

In Bayesian learning for neural networks, how to realize the posterior is important. There exist approaches for generating posterior, Variational Approximation and Markov chain Monte Carlo(MCMC) methods. For Variational Approximation methods for neural networks, Variational Autoencoder[20] and Monte Carlo dropout[9] are practically used. Also for CNNs,

variational approach for Bayesian inference was proposed [8]. MCMC for neural networks, Hamiltonian Monte Carlo and Langevin Dynamics are useful for sampling from posterior. Stochastic Gradient Langevin Dynamics(SGLD)[49] which is a method applying Stochastic Gradient Descent instead of Gradient Descent to Langevin Dynamics is popular MCMC for Bayesian Neural Networks. [54] used SGLD for generating posterior of CNNs.

In this chapter we clarify the free energy and generalization error of Bayesian CNNs with and without skip connection. In both case the free energy and generalization error don't depend on the number of parameters in redundant filters. Then, in case without skip connection, the redundant layers affect the free energy and generalization error whereas they don't affect in case with skip connection. This chapter consists of seven main sections and one appendix. In section4.2, we describe the setting of Convolutional Neural Network analyzed in this chapter. In section4.3, we show the main theorem of this paper. In section4.4, we prove the main theorem of this chapter. In section4.5, we conduct the experiment of synthetic data. In section4.6 and section4.7, we discuss the theorem in this chapter and conclusion. In appendix4.A we explain former study about fully connected deep neural network case.

4.2 Convolutional Neural Network

In this section we describe the function of Convolutional Neural Network. First, we explain CNN without skip connection. The kernel size is 3×3 with zero padding and 1-stride. The activation function is ReLU. The numbers of the layers of the CNN are $K_1(\geq 3)$ for Convolutional Layers and $K_2(\geq 3)$ for Fully Connected Layers.

Let $x \in \mathbb{R}^{L_1 \times L_2 \times H_1}$ be an input vector generated from $q(x)$ with bounded support and $y \in \{0,1\}^{H_{K_1+K_2}}$ be an output vector with $q(y|x)$. We define $w^{(k)} \in \mathbb{R}^{3 \times 3 \times H_{k-1} \times H_k}$, $b^{(k)} \in \mathbb{R}^{H_k}$ as weight and bias parameters in each Convolutional Layer $(2 \leq k \leq K_1)$. $f^{(k)} \in \mathbb{R}^{L_1 \times L_2 \times H_k}$ is output of each layer for $1 \leq k \leq K_1$. $\mathrm{Conv}(f, w)$ is the convolution operation with zero padding and 1-stride:

$$\mathrm{Conv}(f^{k-1}, w^k)_{l_1,l_2,h_k} = \sum_{h_{k-1}} \sum_{p=1,q=1}^{p=3,q=3} f_{l_1+p-1,l_2+p-1,h_{k-1}} w_{p,q,h_{k-1},h_k}. \tag{4.1}$$

We define $g(b^k) : \mathbb{R}^{H_k} \to \mathbb{R}^{L_1 \times L_2 \times H_k}$ as

$$g(b^{(k)})_{l_1,l_2} = b^{(k)} \tag{4.2}$$

for $1 \leq l_1 \leq L_1, 1 \leq l_2 \leq L_2$. By using $w^{(k)}$, $g(b^{(k)})$, and $f^{(k-1)}$, $f^{(k)}$ is described by

$$f^{(k)}(w, b, x) = \sigma(\mathrm{Conv}(f^{(k-1)}(w, b, x), w^{(k)}) + g(b^{(k)})) \tag{4.3}$$

where w, b are the set of all weight and bias parameters. $\sigma()$ is a function that applies the ReLU to all the elements of the input tensor.

The output of $k = K_1 + 1$ layer is result of Global Average Pooling on $k = K_1$ layer:

$$f^{(K_1+1)}(w, b, x) = \frac{1}{L_1 L_2} \sum_{l_1=1}^{l_1=L_1} \sum_{l_2=1}^{l_2=L_2} f^{(K_1)}(w, b, x)_{l_1,l_2}. \tag{4.4}$$

Let $w^{(k)} \in \mathbb{R}^{H_k} \times \mathbb{R}^{H_{k-1}}$, $b^{(k)} \in \mathbb{R}^{H_k}$ be weight and bias parameters in each Fully Connected Layer ($K_1 + 2 \leq k \leq K_1 + K_2$). For $K_1 + 2 \leq k \leq K_1 + K_2 - 1$, $f^{(k)}$ is defined by

$$f^{(k)}(w, b, x) = \sigma(w^{(k)} f^{(k-1)}(w, b, x) + b^{(k)}), \tag{4.5}$$

and for $k = K_1 + K_2$,

$$f^{(K_1+K_2)}(w, b, x) = \text{softmax}(w^{(k)} f^{(k-1)}(w, b, x) + b^{(k)}), \tag{4.6}$$

where softmax() is a softmax function

$$\text{softmax}(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{J} e^{z_j}}. \tag{4.7}$$

The output of the model is represented stochastically

$$y \sim \text{Categorical}(f^{(K_1+K_2)}(w, b, x)) \tag{4.8}$$

where Categorical() is a categorical distribution.

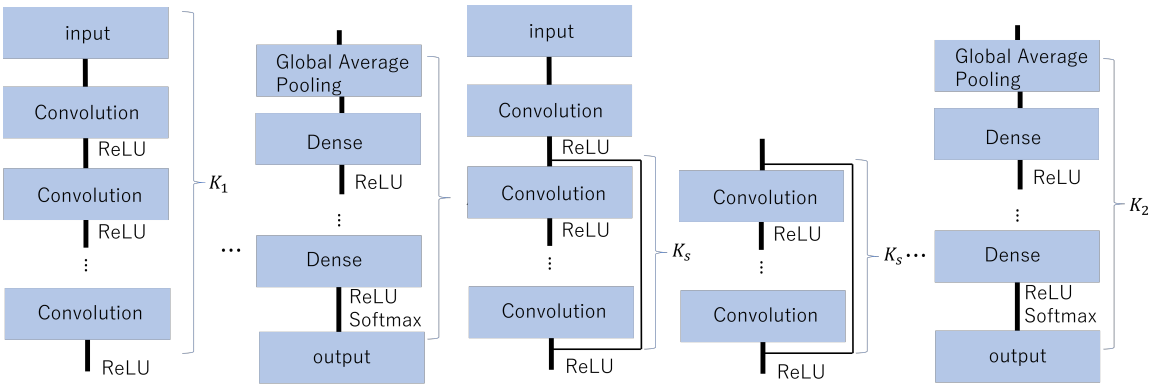
Then we describe CNN with skip connection. The number of layers within the skip connection is K_s and the number of skip connection is M. The output of the layer with skipped connection is described by

$$\begin{aligned} f^{(mK_s+2)}(w, b, x) = \sigma(&\text{Conv}(f^{(mK_s+1)}(w, b, x), w^{(mK_s+2)}) \\ &+ B^{(mK_s+2)} + f^{((m-1)K_s+2)}(w, b, x)). \end{aligned} \tag{4.9}$$

In this case, CNN satisfies the following conditions

$$\begin{aligned} K_1 &= MK_s + 2 \\ H_{mK_s+2} &= \text{const}(1 \leq m \leq M). \end{aligned} \tag{4.10}$$

The other conditions are the same as the case without skip connection.



Without Skip Connection With Skip Connection

Figure 4.1: The structure of Convolutional Neural Network with and without Skip Connection

Figure4.1 shows the configuration of Convolutional Neural Network analyzed in this paper.

4.3 Main Theorem

In this subsection, the main result of this paper is introduced. We assume that the model CNN has enough complexity to approximate the data generating process. In such situation, the data generating process is described by a CNN which is smaller than model. We define the data generating network. Both in case with and without skip connection, the data generating network satisfies the following conditions about the number of layers and filters,

$$K_1^* \leq K_1, K_2^* \leq K_2, H_1^* = H_1, H_{K_1}^* = H_{K_1+K_2} \tag{4.11}$$

and

$$H_k \geq \begin{cases} H_{K_1^*}^* & (K_1^* + 1 \leq k \leq K_1) \\ H_{K_1+K_2^*}^* & (K_1 + K_2^* + 1 \leq k \leq K_1 + K_2 - 1) \\ H_k^* & (others) \end{cases}$$

where K_1^*, K_2^*, H_k^* are the corresponding sizes of K_1, K_2, H_k on data generating network. These conditions indicate that the data generating network has smaller number of layers and width on each layer. Then, we show the main theorem.

Theorem 4.3.1. *(No Skip connection) Assume that the learning machine and the data generating distribution are given by $p(y|x,w,b)$ and $q(y|x) = p(y|x,w^*,b^*)$ in case without skip connection which satisfy the conditions (4.11) and (4.12), and that a training data $\{(X_i, Y_i)\ \ i = 1, 2, ..., n\}$ is independently taken from $q(x)q(y|x)$. Then the average free energy satisfies the inequality,*

$$\mathbb{E}[F_n] \leq nS + \lambda_{CNN} \log n + C \tag{4.12}$$

where

$$\lambda_{CNN} = \frac{1}{2}\left(|w^*|_0 + |b^*|_0 + (K_1 - K_1^*)(9H_{K_1^*}^* + 1)H_{K_1^*}^*\right) \tag{4.13}$$

where $|w^*|_0, |b^*|_0$ are the numbers of parameters of weights and biases in data generating network.

Theorem 4.3.2. *(Skip connection) Assume that the learning machine and the data generating distribution are given by $p(y|x,w,b)$ and $q(y|x) = p(y|x,w^*,b^*)$ in case with skip connection which satisfy the conditions (4.10), (4.11) and (4.12), and that a training data $\{(X_i, Y_i)\ \ i = 1, 2, ..., n\}$ is independently taken from $q(x)q(y|x)$. Then*

$$\lambda_{CNN} = \frac{1}{2}(|w^*|_0 + |b^*|_0). \tag{4.14}$$

If there exists asymptotic expansion of the generalization error $\mathbb{E}[G_n]$ in theorem4.3.1 and theorem4.3.2, that satisfies the following inequality

$$\mathbb{E}[G_n] \leq \frac{\lambda_{CNN}}{n} + o\left(\frac{1}{n}\right), \tag{4.15}$$

where

$$G_n = \int q(x) \sum_{i=1}^{H_{K_1+K_2}} f_i^{(K_1^*+K_2^*)}(w^*, b^*, x) \log \frac{f_i^{(K_1^*+K_2^*)}(w^*, b^*, x)}{\mathbb{E}_{w,b}[f_i^{(K_1+K_2)}(w, b, x)]} \mathrm{d}x \tag{4.16}$$

which corresponds to categorical cross entropy.

4.4 Proof of main theorem

In this section, we show the proof of main theorem.

4.4.1 Inequalities

Note that we describe the Frobenius norm of any order of tensor as $\|\cdots\|$. We denote the Kullback-Leibler divergence of a data-generating distribution $q(y|x) = p(y|x, w^*, b^*)$ and a model $p(y|x)$ that

$$K(w,b) = \int q(x)q(y|x)\log\frac{q(y|x)}{p(y|x,w,b)}\mathrm{d}x\mathrm{d}y. \tag{4.17}$$

Lemma 4.4.1. *[27] Assume that a set W is contained in the set determined by the prior distribution $\{(w,b);\varphi(w,b)>0\}$. Then for an arbitrary positive integer n,*

$$\mathbb{E}[F_n] \leq nS - \log\int_W \exp(-nK(w,b))\varphi(w,b)\mathrm{d}w\mathrm{d}b. \tag{4.18}$$

Lemma 4.4.2. *[27] For arbitrary vectors s,t,*

$$\|\sigma(s)-\sigma(t)\| \leq \|s-t\|. \tag{4.19}$$

Lemma 4.4.3. *[27] For arbitrary w,w', b, b', and $K_1+1 \leq k \leq K_1+K_2$, the following inequality holds,*

$$\begin{aligned}
&\|f^{(k)}(w,b,x) - f^{(k)}(w',b',x)\| \\
&\leq \|w^{(k)} - w'^{(k)}\|\|f^{(k-1)}(w,b,x)\| + \|b^{(k)} - b'^{(k)}\| \\
&+ \|w^{(k)}\|\|f^{(k-1)}(w,b,x) - f^{(k-1)}(w',b',x)\|.
\end{aligned} \tag{4.20}$$

Corollary 4.4.4. *For arbitrary w,w', b, b', and $1 \leq k \leq K_1$, the following inequality holds,*

$$\begin{aligned}
&\|f^{(k)}(w,b,x) - f^{(k)}(w',b',x)\| \\
&\leq 9\|w^{(k)} - w'^{(k)}\|\|f^{(k-1)}(w,b,x)\| + L_1L_2\|b^{(k)} - b'^{(k)}\| \\
&+ 9\|w^{(k)}\|\|f^{(k-1)}(w,b,x) - f^{(k-1)}(w',b',x)\| \\
&+ \delta^{(k)}\|w^{(k)}\|\|f^{(k-K_2-1)}(w,b,x) - f^{(k-K_2-1)}(w',b',x)\|.
\end{aligned} \tag{4.21}$$

where $\delta^{(k)}$ equals to 1 if the network has Skip connection and $k = mK_2+2$, otherwise it equals to 0

Proof.

$$\begin{aligned}
&f^{(k)}(w,b,x) - f^{(k)}(w',b',x) \\
&= \sigma(\mathrm{Conv}(f^{(k-1)}(w,b,x), w^{(k)}) + g(b^{(k)})) - \sigma(\mathrm{Conv}(f^{(k-1)}(w,b,x), w'^{(k)})) + g(b'^{(k)})) \\
&+ \sigma(\mathrm{Conv}(f^{(k-1)}(w,b,x), w'^{(k)}) + g(b'^{(k)})) - \sigma(\mathrm{Conv}(f^{(k-1)}(w',b',x), w'^{(k)}) + g(b'^{(k)})).
\end{aligned} \tag{4.22}$$

From definition of Conv(), the following equation holds.

$$\|\mathrm{Conv}(f^{(k-1)}(w,b,x),w^{(k)})_j\| \leq \sum_{i=1}^{H^{k-1}} \|f^{(k-1)}(w,b,x)_i\||(\sum_{p=1}^{3}\sum_{q=1}^{3} w^{(k)}_{pqij})|_1$$
$$\leq 9\|f^{(k-1)}(w,b,x)\|\|w_{:,:,:,j}\| \tag{4.23}$$

By using lemma4.A.2, (4.22) and (4.23), corollary4.4.4 is proved. □

Lemma 4.4.5. *For arbitrary w, b, x,*

$$\|f^{(k)}(w,b,x)\| \leq \mathcal{D}_k\|w^{(k)}\|\|w^{(k-1)}\|\cdots\|w^{(2)}\|\|x\| \tag{4.24}$$
$$+ \mathcal{D}_0\|b^{(k)}\| + \sum_{j=1}^{k-2}\mathcal{D}_j\|w^{(k)}\|\|w^{(k-1)}\|\cdots\|w^{(k-j)}\|\|b^{(k-j)}\|.$$

where $\mathcal{D}_j, 0 \leq j \leq k$ is constant.

Proof. By considering the case all the parameters of w' and b' are 0, in Lemma 4.A.3, it follows that

$$\|f^{(k)}(w,b,x)\| \leq 9\|w^{(k)}\|\|f^{(k-1)}(w,b,x)\| + L_1L_2\|b^{(k)}\| \tag{4.25}$$
$$+ \delta^{(k)}\|w^{(k)}\|\|f^{(k-K_2-1)}(w,b,x) - f^{(k-K_2-1)}(w',b',x)\|.$$

Then mathematical induction gives the Lemma. □

4.4.2 Notations of parameters

In order to prove the main theorem, we need several notations. We divide the filters of learning model in each convolutional layer $1 \leq h_k \leq H_k$ into the $1 \leq h_k \leq H_k^*$ and $H_k^* + 1 \leq h_k \leq H_k$. The former is denoted as A and the latter is denoted as B. The convergent tensor $\mathcal{E}^{(k)} \in \mathbb{R}^{3\times3\times H_{k-1}\times H_k}$ and vector $\mathcal{E}_0^{(k)} \in \mathbb{R}^{H_k}$ where the absolute values of all elements are smaller than $1/\sqrt{n}$ are denoted by

$$\mathcal{E}^{(k)}_{pq} = \begin{pmatrix} \mathcal{E}^{(k)}_{pqAA} & \mathcal{E}^{(k)}_{pqAB} \\ \mathcal{E}^{(k)}_{pqBA} & \mathcal{E}^{(k)}_{pqBB} \end{pmatrix}, \quad (1 \leq p \leq 3, 1 \leq q \leq 3), \tag{4.26}$$

$$\mathcal{E}^{(k)}_0 = \begin{pmatrix} \mathcal{E}^{(k)}_{A0} \\ \mathcal{E}^{(k)}_{B0} \end{pmatrix}. \tag{4.27}$$

The positive constant tensor $\mathcal{M}^{(k)}$ and vector $\mathcal{M}_0^{(k)}$ are defined by the condition that all elements are in the interval $[A, B]$,

$$\mathcal{M}^{(k)}_{pq} = \begin{pmatrix} \mathcal{M}^{(k)}_{pqAA} & \mathcal{M}^{(k)}_{pqAB} \\ \mathcal{M}^{(k)}_{pqBA} & \mathcal{M}^{(k)}_{pqBB} \end{pmatrix}, \quad (1 \leq p \leq 3, 1 \leq q \leq 3), \tag{4.28}$$

$$\mathcal{M}^{(k)}_0 = \begin{pmatrix} \mathcal{M}^{(k)}_{A0} \\ \mathcal{M}^{(k)}_{B0} \end{pmatrix}. \tag{4.29}$$

To prove Theorem 4.3.1 4.3.2, we show an upper bound of $\mathbb{E}[F_n]$ is given by choosing a set W_E which consists of essential weight and bias parameters in convolutional layers and fully connected layers.

4.4.3 No Skip Connection Case

Definition. (Essential parameter set W_E without Skip Connection). A parameter (w, b) is said to be in an essential parameter set W_E if it satisfies the following conditions (1),(2) for $2 \leq k \leq K_1$,

(1) For $2 \leq k \leq K_1^*$

$$w_{pq}^{(k)} = \begin{pmatrix} (w^*)^{(k)} + \mathcal{E}_{pqAA}^{(k)} & \mathcal{M}_{pqAB}^{(k)} \\ -\mathcal{M}_{pqBA}^{(k)} & -\mathcal{M}_{pqBB}^{(k)} \end{pmatrix}, \tag{4.30}$$

$$b^{(k)} = \begin{pmatrix} (b^*)^{(k)} + \mathcal{E}_{A0}^{(k)} \\ -\mathcal{M}_{B0}^{(k)} \end{pmatrix}, \tag{4.31}$$

for $1 \leq p \leq 3, 1 \leq q \leq 3$

(2) For $K_1^* + 1 \leq k \leq K_1$

$$w_{pq}^{(k)} = \begin{pmatrix} \mathcal{Z}_{pqAA}^{(k)} & \mathcal{M}_{pqAB}^{(k)} \\ -\mathcal{M}_{pqBA}^{(k)} & -\mathcal{M}_{pqBB}^{(k)} \end{pmatrix}, \tag{4.32}$$

$$b^{(k)} = \begin{pmatrix} (b^*)^{(k)} + \mathcal{E}_{A0}^{(k)} \\ -\mathcal{M}_{B0}^{(k)} \end{pmatrix}, \tag{4.33}$$

where

$$\mathcal{Z}_{pqAA}^{(k)} = \begin{cases} I_{22AA} + \mathcal{E}_{22AA}^{(k)} & (p = q = 2) \\ \mathcal{E}_{pqAA}^{(k)} & (\text{others}) \end{cases}. \tag{4.34}$$

where $I_{22AA} \in \mathbb{R}^{(H^*)^{(k)}} \times \mathbb{R}^{(H^*)^{(k)}}$ is an identity matrix.

Lemma 4.4.6. *Assume that the weight and bias parameters of convolutional layers are in the essential set W_E in case without Skip Connection. Then there exist constants $c_1, c_2 > 0$ such that*

$$\|f_{:,:,A}^{(K_1)}(w, b, x) - f^{(K_1^*)}(w^*, b^*, x)\| \leq \frac{c_1}{\sqrt{n}}(\|x\| + 1), \tag{4.35}$$

$$\|f_{:,:,A}^{(K_1)}(w, b, x)\| \leq c_2(\|x\| + 1). \tag{4.36}$$

Proof. Eq.(4.79) is derived from Lemma 4.A.4. By the definitions (4.71), (4.72), for $2 \leq k \leq K^*$

$$f_A^{(2)}(w, b, x) = \sigma(\mathrm{Conv}(f_{:,:,A}^{(1)}(w, b, x), (w^*)^{(2)} + \mathcal{E}_{:,:,AA}^{(2)}) + g((b^*)^{(2)} + \mathcal{E}_{A0}^{(2)})), \tag{4.37}$$

$$\begin{aligned} f_A^{(k)}(w, b, x) = \sigma(&\mathrm{Conv}(f_{:,:,A}^{(k-1)}(w, b, x), (w^*)^{(k)} + \mathcal{E}_{:,:,AA}^{(k)}) \\ &+ \mathrm{Conv}(f_{:,:,B}^{(k-1)}(w, b, x), \mathcal{M}_{:,:,AB}^{(k)}) + g((b^*)^{(k)} + \mathcal{E}_{A0}^{(k)})). \end{aligned} \tag{4.38}$$

In $k = 2$, $|x|$ is bounded and $\mathcal{M}_{:,:,AB}^{(k)}$ is a constant tensor, $\mathcal{M}_{B0}^{(k)}$ is large sufficiently, $f_{:,:,B}^{(2)}(w, b, x) = 0$ because all the elements of the output of ReLU function $f^{(2)}(w, b, x)$ are nonnegative. For

$3 \leq k \leq K_1$, $f^{(k)}_{:,:,B}(w,b,x) = 0$, since all elements of $w^{(k)}_{:,:,BA}$, $w^{(k)}_{:,:,BB}$, and $w^{(k)}_{B0}$ are negative. Hence, by Lemma 4.A.3, for $2 \leq k \leq K_1^*$,

$$\begin{aligned}
&\|f^{(k)}_{:,:,A}(w,b,x) - f^{(k)}(w^*,b^*,x)\| \\
&\leq 9\|\mathcal{E}^{(k)}_{:,:,AA}\|\|f^{(k-1)}(w,b,x)\| + L_1L_2\|\mathcal{E}^{(k)}_{A0}\| \\
&+ 9\|(w^*)^{(k)}\|\|f^{(k-1)}_{:,:,A}(w,b,x) - f^{(k-1)}(w^*,b^*,x)\|. \qquad (4.39)
\end{aligned}$$

and for $K_1^* + 1 \leq k \leq K_1$, by using $f^{(K_1^*)}(w^*,b^*,x)$ as $f^{(k)}(w^*,b^*,x)$,

$$\begin{aligned}
&\|f^{(k)}_{:,:,A}(w,b,x) - f^{(K_1^*)}(w^*,b^*,x)\| \\
&\leq 9\|\mathcal{E}^{(k)}_{:,:,AA}\|\|f^{(k-1)}(w,b,x)\| + L_1L_2\|\mathcal{E}^{(k)}_{A0}\| \\
&+ 9\|(w^*)^{(k)}\|\|f^{(k-1)}_{:,:,A}(w,b,x) - f^{(K_1^*)}(w^*,b^*,x)\|. \qquad (4.40)
\end{aligned}$$

The elements of tensors and vectors in $\mathcal{E}^{(k-1)}_{:,:,AA}$ and $\mathcal{E}^{(k)}_{:,:,A0}$ are bounded by $1/\sqrt{n}$ order term, hence $\|\mathcal{E}^{(k-1)}_{AA}\|$ and $\|\mathcal{E}^{(k)}_{A0}\|$ are bounded by $1/\sqrt{n}$ order term. Moreover, $\|(w^*)^{(k)}\|$ is a constant term. For $k = 2$, $f^{(k-1)}_{:,:,A}(w,b,x) - f^{(k-1)}(w^*,b^*,x) = x - x = 0$. Then, by using mathematical induction for (4.39) and (4.40) , the all terms can be bounded by $1/\sqrt{n}$ terms, hence we obtained the Lemma. □

From [27], because of the output in $k = K_1 + 1$ is nonnegative there exist the essential parameters for fully connected layers such that the number of the convergent parameters $\mathcal{E}$ equals to that of data generating network. From these lemmas, the main theorem can be proved.

(Proof of Theorem 4.3.1). By Lemma 4.A.1, it is sufficient to prove that there exists a constant $C > 0$ such that

$$\int_{W_E} \exp(-nK(w,b))\varphi(w,b)\mathrm{d}w\mathrm{d}b \geq \frac{C}{n^\lambda} \qquad (4.41)$$

From the property of KL-divergence, there exists the positive constant c_4

$$K(w,b) \leq \frac{c_4}{2}\int \|f^{(K_1+K_2)}(w,b,x) - f^{(K_1^*+K_2^*)}(w^*,b^*,x)\|^2 q(x)\mathrm{d}x. \qquad (4.42)$$

By using Lemma 4.A.5, if $(w,b) \in W_E$,

$$K(w,b) \leq \frac{c_4c_3^2}{2n}\int (\|x\|+1)^2 q(x)\mathrm{d}x = \frac{c_5}{n} < \infty. \qquad (4.43)$$

It follows that

$$\begin{aligned}
&\int_{W_E} \exp(-nK(w,b))\varphi(w,b)\mathrm{d}w\mathrm{d}b \\
&\geq \exp(-c_5)\left(\min_{(w,b)\in W_E}\varphi(w,b)\right)\mathrm{Vol}(W_E). \qquad (4.44)
\end{aligned}$$

where $c_5 > 0$, $\min_{(w,b)\in W_E} \varphi(w,b) > 0$, and $\mathrm{Vol}(W_E)$ is the volume of the set W_E by the Lebesgue measure. The convergent scale of $\mathrm{Vol}(W_E)$ is determined from the number of convergent parameter $\mathcal{E}$ in W_E. Then,

$$\mathrm{Vol}(W_E) \geq \frac{C_1}{n^\lambda}, \tag{4.45}$$

where

$$\begin{aligned}\lambda &= \frac{1}{2}\left(\sum_{k=2}^{k=K_1}(9H^*_{k-1}+1)H^*_k + \sum_{k=K_1+1}^{k=K_1+K_2}(9H^*_{k-1}+1)H^*_k\right) \\ &= \frac{1}{2}\left(|w^*|_0 + |b^*|_0 + (K_1 - K_1^*)(9H^*_{K_1^*}+1)H^*_{K_1^*}\right).\end{aligned} \tag{4.46}$$

We obtained theorem4.3.1.

4.4.4 Skip Connection Case

Definition. (Essential parameter set W_E with Skip Connection). An essential parameter set W_E with Skip Connection satisfies the following conditions (1),(2) for $2 \leq k \leq K_1$,
(1) For $2 \leq k \leq K_1^*$, the same conditions as (4.71) and (4.72).
(2) For $K_1^* + 1 \leq k \leq K_1$

$$w^{(k)}_{pq} = \begin{pmatrix} -\mathcal{M}^{(k)}_{pqAA} & -\mathcal{M}^{(k)}_{pqAB} \\ -\mathcal{M}^{(k)}_{pqBA} & -\mathcal{M}^{(k)}_{pqBB} \end{pmatrix}, \tag{4.47}$$

$$b^{(k)} = \begin{pmatrix} -\mathcal{M}^{(k)}_{A0} \\ -\mathcal{M}^{(k)}_{B0} \end{pmatrix}, \tag{4.48}$$

Lemma 4.4.7. *Assume that the weight and bias parameters of convolutional layers are in the essential set W_E in case with Skip Connection. Then there exist constants $c_1, c_2 > 0$ such that*

$$\|f^{(K_1)}_{:,:,A}(w,b,x) - f^{(K_1^*)}(w^*,b^*,x)\| \leq \frac{c_1}{\sqrt{n}}(\|x\|+1), \tag{4.49}$$

$$\|f^{(K_1)}_{:,:,A}(w,b,x)\| \leq c_2(\|x\|+1). \tag{4.50}$$

Proof. Because of similar reason to lemma4.A.5, holds. By Lemma 4.A.3, for $k = mK_s + 1$,

$$\begin{aligned}&\|f^{(k)}_{:,:,A}(w,b,x) - f^{(k)}(w^*,b^*,x)\| \\ &\leq 9\|\mathcal{E}^{(k)}_{:,:,AA}\|\|f^{(k-1)}(w,b,x)\| + L_1L_2\|\mathcal{E}^{(k)}_{A0}\| \\ &+ 9\|(w^*)^{(k)}\|\|f^{(k-1)}_{:,:,A}(w,b,x) - f^{(k-1)}(w^*,b^*,x)\| \\ &+ \|w^{(k)}\|\|f^{(k-K_2-1)}(w,b,x) - f^{(k-K_2-1)}(w',b',x)\|.\end{aligned} \tag{4.51}$$

If $k \neq mK_s + 1$ and $2 \leq k \leq K_1^*$, inequality (4.39) holds. Same as the lemma4.A.5, from mathematical induction, $\|f^{(K_1^*)}_{:,:,A}(w,b,x) - f^{(K_1^*)}(w^*,b^*,x)\|$ is bounded by $1/\sqrt{n}$ terms. For

$2 \leq k \leq K_1$, $f^{(k)}_{:,:,B}(w,b,x) = 0$ same reason as lemma4.A.5. For $K_1^* + 1 \leq k \leq K_1$, since all elements of $w^{(k)}$ and $b^{(k)}$ are negative, the following equations are given.

$$f^{(k)}_{:,:,A}(w,b,x) = \begin{cases} f^{(K_1^*)}(w,b,x) & (k = nK_s + 1) \\ 0 & (\text{others}) \end{cases} . \tag{4.52}$$

Hence, we obtained the Lemma. □

Same as without Skip connection case, by using the result of [27] for fully connected layer and inequality(4.44),(4.45), we obtained theorem4.3.2.

4.5 Experiment

In this section, we show the result of experiment of synthetic data.

4.5.1 Methods

We prepared 2-class labeled simple data shown in fig4.2. The data is $x \in R^{4\times 4}$ and the values of each element are in $(-1, 1)$. The average of each element is 0.5 or -0.5 and added the truncated normal distribution noise within the interval$(-0.5, 0.5)$. The probability of each label of data is 0.5. We trained CNN whose number of convolutional layer $K_1 = 2$ and fully connected layers $K_2 = 2$ with SGD. The number of filter is $H_2 = 2$ and the parameters are L_2 regularized. We use the trained CNN named "true model" as a data generating distribution. Note that the label of original data fig4.2 is deterministic, but the label of true model is probabilistic. We prepare three learning CNN models. Each number of convolutional layers is $K_1 = 2, 3, 4$. Each model has skip connection every one layers or does not have skip connection. The number of filters in each layer is $H_k = 4$. They have $K_2 = 2$ fully connected layers. The prior distribution is the Gaussian distribution which covariance matrix is $10^4 I$ for weight parameter and $10^2 I$ for bias parameter. We trained the learning CNN models by using the Langevin dynamics. The learning rate is 10^{-2} and the interval of sampling is 100. We use the average of 1000 samples of learning CNN models as the average of posterior. We estimated the generalization error by the test error of 10000 test data from true model and trained each learning model 10 times and estimated the $\mathbb{E}[G_n]$ from the average of test error.

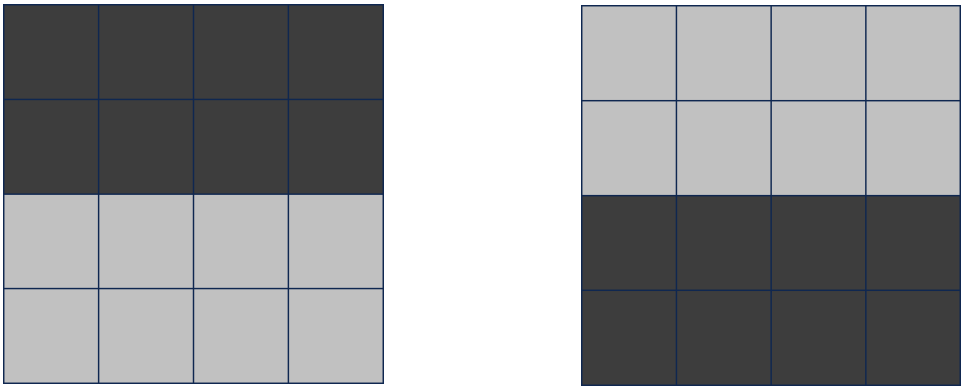


Figure 4.2: The average of input x of each label

4.5.2 Result of experiments

Table 4.1: Experimental value and theoretical upper bound of the generalization error

model	$n\times$ Test Error	λ_{CNN}	$d_{\mathrm{model}}/2$
$K_1 = 2$	16.0(1.9)	13	25
$K_1 = 3$ no skip	10.0(0.9)	32	99
$K_1 = 4$ no skip	58.4(2.3)	51	173
$K_1 = 3$ with skip	11.4(2.3)	13	99
$K_1 = 4$ with skip	15.6(1.2)	13	173

Table4.1 shows the result of the experiment. Test Error shows n times of the average of 10 test error in each model and the standard error of them. d_{model} is a number of parameters of each model. All the CNN models include the true model, hence the bias is 0. Then from equation(4.15), theoretical upper bound of the generalization error is λ_{CNN}/n. In table4.1, the experimental values of all models are smaller than $d_{\mathrm{model}}/2$. Moreover, in case with skip connection, the experimental value did not so increase as the increase of the number of layer. Then, in case $K_1 = 4$ without skip connection, the experimental value increased from the case $K_1 = 2$. In case $K_1 = 3$ without skip connection, the experimental value is smaller than that of $K_1 = 2$. Behavior of MCMC is considered to be the cause of this result. Since MCMC in high dimensional model needs the long series for convergence in general, the result is deviated from theoretical predict.

4.6 Discussion

4.6.1 Difference with or without Skip Connection

In this paper for analyzing the overparametrized CNN, the data generating network is smaller than learning network both case of Skip Connection. Nevertheless, two cases of the data generating network is different, if the learning model network has double filter $H^{(k)}$ to the data generating network in each convolutional layer, the model network can represent the generating network in different case. The output of each layer is nonnegative hence the model can represent the skip connection or the negative of that. If the model network doesn't have larger layer to the data generating network, the free energy of CNN with skip connection can be both larger or smaller than that without skip connection by the data generating network. Then, the layer of model network gets larger, the free energy of CNN with skip connection does not change but that without skip connection gets larger and the free energy of CNN with skip connection comes to have smaller free energy for all data generating network.

4.6.2 Comparison to Deep Neural Network

Firstly we compare the result of this paper to that of DNN in [27]. In case of DNN, the free energy depends on the layers of the model and only on that of the data generating network. This stands to the reason that mapping of the linear transformation in lower layer can be represented in higher layer. On the other hand, convolution operation doesn't have such property, hence the free energy of CNN without skip connection depends on the layer of

learning model network. However, with skip connection, there exists the essential parameter which doesn't depend on overparametrized layers and the free energy does not also depend on the layer of learning model network.

4.7 Conclusion

In this chapter, we studied Free energy of Bayesian Convolutional Neural Network with Skip Connection and compared to the case without Skip Connection. Free energy of Bayesian CNN with Skip Connection doesn't depend on the layer of the model unlike the case without Skip Connection. In Bayesian learning, the increase of Free energy is equivalent to generalization error, hence the generalization error has same property about the Skip Connection. In particular, Free energy of CNN with skip connection does not depend on the number of parameters in learning network but depends only on that in data generating network. This feature shows the generalization ability of CNN with skip connection does not decrease with respect to any overparameterization in Bayesian learning.

Appendix 4.A Fully connected case

In this appendix, we introduce the main theorem and proof of a paper[27] which is about the free energy of full connected Bayesian Deep Neural Network.

4.A.1 Main theorem

Assume that the learning machine and the data generating distribution are given by $p(y|x,w,b)$ and $q(y|x) = p(y|x,w^*,b^*)$ which satisfy the conditions eq.(4.11), eq.(4.12), and eq.(4.10), and that a sample $\{(X_i, Y_i) \;\; i = 1,2,...,n\}$ is independently subject to $q(x)q(y|x)$. Then the average free energy satisfies the inequality,

$$\mathbb{E}[F_n] \leq nS + \lambda_{ReLU} \log n + C.$$

For general cases,

$$\lambda_{ReLU} = \frac{1}{2}\left(d^* + H_3^*(H_2 - H_2^*)\right), \tag{4.53}$$

where d^* is sum of the number of parameters in w^* and b^*. If the support of the input distribution is bounded or contained in nonnegative region,

$$\lambda_{ReLU} = \frac{d^*}{2} \tag{4.54}$$

These results show that the average free energy is bounded, even if the number of layers are larger than necessary to estimate the data-generation network. In particular eq.(4.53) is equal to the half the number of parameters in the data-generating network.

4.A.2 Lemmas

In this section, we prepare several lemmas which are necessary to prove the main theorem.

Let the Kullback-Leibler divergence of a data-generating network $q(y|x) = p(y|x, w^*, b^*)$ and a learning machine $p(y|x)$ be

$$K(w,b) = \int q(x)q(y|x)\log\frac{q(y|x)}{p(y|x,w,b)}dxdy.$$

It is well-known that $K(w,b) \geq 0$ for an arbitrary (w,b) and $K(w,b) = 0$ if and only if $q(y|x) = p(y|x,w,b)$.

Lemma 4.A.1. *Assume that a set W is contained in the set determined by the prior distribution $\{(w,b); \varphi(w,b) > 0\}$. Then for an arbitrary positive integer n,*

$$\mathbb{E}[F_n] \leq nS - \log\int_W \exp(-nK(w,b))\varphi(w,b)dwdb.$$

Proof. An empirical Kullback-Leibler divergence is defined by

$$K_n(w,b) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{p(Y_i|X_i,w^*,b^*)}{p(Y_i|X_i,w,b)},$$

which satisfies $\mathbb{E}[K_n](w,b)] = K(w,b)$.

$$\frac{q(y^n|x^n)}{p(y^n|x^n)} = \exp(-\sum_{i=1}^{n}\log\frac{q(Y_i|X_i)}{p(Y_i|X_i,w,b)}) \tag{4.55}$$

$$= \exp(-nK_n(w,b)). \tag{4.56}$$

From the definition of free energy,

$$\mathbb{E}[F_n] = -\mathbb{E}[\log\frac{q(y^n|x^n)}{p(y^n|x^n)}] + nS \tag{4.57}$$

$$= -\mathbb{E}[\log\int\varphi(w,b)\exp(-nK_n(w,b))dwdb] + nS. \tag{4.58}$$

By applying Lemma.1 in [42],

$$\mathbb{E}[F_n] \leq -\log\int\varphi(w,b)\exp(-\mathbb{E}[nK_n(w,b)])dwdb + nS \tag{4.59}$$

$$\leq -\log\int\varphi(w,b)\exp(-nK(w,b))dwdb + nS \tag{4.60}$$

$$\leq -\log\int_W\varphi(w,b)\exp(-nK(w,b))dwdb + nS, \tag{4.61}$$

where the last inequality is derived the fact that the restriction of integrated region makes the integration not larger. □

Lemma 4.A.2. *For arbitrary vectors s, t,*

$$\|\sigma(s) - \sigma(t)\| \leq \|s - t\|.$$

Proof. If $s_i, t_i \geq 0$ or $s_i, t_i \leq 0$, then $|\sigma_i(s) - \sigma_i(t)| = |s_i - t_i|$. If $s_i \geq 0, t_i < 0$, then $|\sigma_i(s) - \sigma_i(t)| = |s_i| \leq |s_i - t_i|$. If $s_i < 0, t_i \geq 0$, then $|\sigma_i(s) - \sigma_i(t)| = |t_i| \leq |s_i - t_i|$. Hence,

$$\|\sigma(s) - \sigma(t)\|^2 = \sum_i |\sigma_i(s) - \sigma_i(t)|^2 \leq \sum_i |s_i - t_i|^2 = \|s - t\|^2.$$

□

Lemma 4.A.3. *For arbitrary w,w', b, b', the following inequality holds,*

$$\begin{aligned}
&\|f^{(k)}(w, b, x) - f^{(k)}(w', b', x)\| \\
&\leq \|w^{(k)} - w'^{(k)}\| \|f^{(k-1)}(w, b, x)\| + \|b^{(k)} - b'^{(k)}\| \\
&+ \|w^{(k)}\| \|f^{(k-1)}(w, b, x) - f^{(k-1)}(w', b', x)\|,
\end{aligned} \tag{4.62}$$

where $\|w^{(k)}\|$ is the operator norm of a matrix $w^{(k)}$.

Proof.

$$\begin{aligned}
&f^{(k)}(w, b, x) - f^{(k)}(w', b', x) \\
&= \sigma(w^{(k)} f^{(k-1)}(w, b, x) + b^{(k)}) - \sigma(w'^{(k)} f^{(k-1)}(w, b, x) + b'^{(k)}) \\
&+ \sigma(w'^{(k)} f^{(k-1)}(w, b, x) + b'^{(k)}) - \sigma(w'^{(k)} f^{(k-1)}(w', b', x) + b'^{(k)}).
\end{aligned} \tag{4.63}$$

Hence, by using Lemma 4.A.2,

$$\begin{aligned}
&\|f^{(k)}(w, b, x) - f^{(k)}(w', b', x)\| \\
&\leq \|\sigma(w^{(k)} f^{(k-1)}(w, b, x) + b^{(k)}) - \sigma(w'^{(k)} f^{(k-1)}(w, b, x) + b'^{(k)})\| \\
&+ \|\sigma(w'^{(k)} f^{(k-1)}(w, b, x) + b'^{k}) - \sigma(w'^{(k)} f^{(k-1)}(w', b', x) + b'^{(k)})\| \\
&\leq \|w^{(k)} - w'^{(k)}\| \|f^{(k-1)}(w, b, x)\| + \|b^{(k)} - b'^{(k)}\| \\
&+ \|w'^{(k)}\| \|f^{(k-1)}(w, b, x) - f^{(k-1)}(w', b', x)\|.
\end{aligned} \tag{4.64}$$

Hence, lemma is proved. □

Lemma 4.A.4. *For arbitrary w, b, x,*

$$\|f^{(k)}(w, b, x)\| \leq \|w^{(k)}\| \|w^{(k-1)}\| \cdots \|w^{(2)}\| \|x\| \tag{4.65}$$

$$+ \|b^{(k)}\| + \sum_{j=1}^{k-2} \|w^{(k)}\| \|w^{(k-1)}\| \cdots \|w^{(k-j)}\| \|b^{(k-j)}\|. \tag{4.66}$$

Proof. By substituting $w' := 0$ and $b' = 0$, in Lemma 4.A.3, it follows that

$$\|f^{(k)}(w, b, x)\| \leq \|w^{(k)}\| \|f^{(k-1)}(w, b, x)\| + \|b^{(k)}\|. \tag{4.67}$$

Then mathematical induction gives the Lemma. □

In order to prove the main theorem, we need several notations. The convergent matrix $\mathcal{E}^{(k)}$ and vector $\mathcal{E}_0^{(k)}$ defined by the condition that the absolute values of all entries are smaller than $1/\sqrt{n}$, which is denoted by

$$\mathcal{E}^{(k)} = \begin{pmatrix} \mathcal{E}_{AA}^{(k)} & \mathcal{E}_{AB}^{(k)} \\ \mathcal{E}_{BA}^{(k)} & \mathcal{E}_{BB}^{(k)} \end{pmatrix}, \quad \mathcal{E}_0^{(k)} = \begin{pmatrix} \mathcal{E}_{A0}^{(k)} \\ \mathcal{E}_{B0}^{(k)} \end{pmatrix}. \tag{4.68}$$

The positive-small-constant matrix $\mathcal{D}^{(k)}$ and vector $\mathcal{D}_0^{(k)}$ are defined by the condition that all entries are positive and smaller than $\delta > 0$ where δ does not depend on n, which is denoted by

$$\mathcal{D}^{(k)} = \begin{pmatrix} \mathcal{D}_{AA}^{(k)} & \mathcal{D}_{AB}^{(k)} \\ \mathcal{D}_{BA}^{(k)} & \mathcal{D}_{BB}^{(k)} \end{pmatrix}, \quad \mathcal{D}_0^{(k)} = \begin{pmatrix} \mathcal{D}_{A0}^{(k)} \\ \mathcal{D}_{B0}^{(k)} \end{pmatrix}. \tag{4.69}$$

The positive constant matrix $\mathcal{M}^{(k)}$ and vector $\mathcal{M}_0^{(k)}$ are defined by the condition that all entries are in the interval $[1, 2]$,

$$\mathcal{M}^{(k)} = \begin{pmatrix} \mathcal{M}_{AA}^{(k)} & \mathcal{M}_{AB}^{(k)} \\ \mathcal{M}_{BA}^{(k)} & \mathcal{M}_{BB}^{(k)} \end{pmatrix}, \quad \mathcal{M}_0^{(k)} = \begin{pmatrix} \mathcal{M}_{A0}^{(k)} \\ \mathcal{M}_{B0}^{(k)} \end{pmatrix}. \tag{4.70}$$

To prove Theorem 4.3.1, we show an upper bound of $\mathbb{E}[F_n]$ is given by choosing a set W_E which consists of essential weight and bias parameters.

Definition. (Essential parameter set W_E). A parameter (w, b) is said to be in an essential parameter set W_E if it satisfies the following conditions, (1), (2), and (3).

(1) For $2 \leq k \leq N^* - 1$, there exist convergent matrices $\mathcal{E}^{(k)}$ and positive constant matrices $\mathcal{M}^{(k)}$ such that

$$w^{(k)} = \begin{pmatrix} (w^*)^{(k)} + \mathcal{E}_{AA}^{(k)} & \mathcal{Z}_{AB}^{(k)} \\ -\mathcal{M}_{BA}^{(k)} & -\mathcal{M}_{BB}^{(k)} \end{pmatrix}, \tag{4.71}$$

$$b^{(k)} = \begin{pmatrix} (b^*)^{(k)} + \mathcal{E}_{A0}^{(k)} \\ -\mathcal{M}_{B0}^{(k)} \end{pmatrix}, \tag{4.72}$$

where

$$\mathcal{Z}_{AB}^{(k)} = \begin{cases} \mathcal{E}_{AB}^{(3)} & (k = 3) \\ \mathcal{M}_{AB}^{(k)} & (k \neq 3) \end{cases}. \tag{4.73}$$

Note that, for $k = 2$, $\mathcal{Z}_{AB}^{(k)}$, $\mathcal{M}_{BB}^{(k)}$, and $\mathcal{M}_{B0}^{(k)}$ are the empty matrix.

(2) For $N^* \leq k \leq N-1$, there exist positive-small-constant matrix $\mathcal{D}^{(k)}$ and positive constant matrix $\mathcal{M}^{(k)}$

$$w^{(k)} = \begin{pmatrix} I_{N^*-1} + \mathcal{D}_{AA}^{(k)} & \mathcal{M}_{AB}^{(k)} \\ -\mathcal{M}_{BA}^{(k)} & -\mathcal{M}_{BB}^{(k)} \end{pmatrix}, \tag{4.74}$$

$$b^{(k)} = \begin{pmatrix} \mathcal{M}_{A0}^{(k)} \\ -\mathcal{M}_{B0}^{(k)} \end{pmatrix}, \tag{4.75}$$

where I_{N^*-1} is the identity matrix of $H_{N^*-1} \times H_{N^*-1}$.
(3) For $k = N$, there exist convergent matrix $\mathcal{E}^{(N)}$ and vector $\mathcal{E}_0^{(N)}$ such that

$$w^{(N)} = \begin{pmatrix} (w^*)^{(N^*)}P^{-1} + \mathcal{E}_{AA}^{(N)}, & \mathcal{M}_{AB}^{(N)} \end{pmatrix} \tag{4.76}$$

$$b^{(N)} = (b^*)^{(N^*)} - \sum_{k=N^*}^{N-1} w^{(N-1)}w^{(N-2)}\cdots w^{(k)}b^{(k-1)} + \mathcal{E}_{B0}^{(N)}, \tag{4.77}$$

where $P \in \mathbb{R}^{(H^*_{N^*-1})\times(H^*_{N^*-1})}$ is defined by matrices in eq.(4.74)

$$P = w_{AA}^{(N-1)}w_{AA}^{(N-2)}\cdots w_{AA}^{(N^*)}.$$

Note that a positive constant $\delta > 0$ is taken sufficiently small such that arbitrary $w_{AA}^{(k)}$ ($N^* \le k \le N-1$) is invertible.

Lemma 4.A.5. *Assume that the weight and bias parameters are in the essential set W_E. Then there exist constants $c_1, c_2 > 0$ such that*

$$\|f_A^{(N^*-1)}(w,b,x) - f^{(N^*-1)}(w^*,b^*,x)\| \le \frac{c_1}{\sqrt{n}}(\|x\|+1), \tag{4.78}$$

$$\|f_B^{(N^*-1)}(w,b,x)\| \le c_2(\|x\|+1). \tag{4.79}$$

Proof. Eq (4.79) is derived from Lemma 4.A.4. By the definitions (4.71), (4.72), for $4 \le k \le N^*-1$

$$f_A^{(2)}(w,b,x) = \sigma(((w^*)^{(2)} + \mathcal{E}_{AA}^{(2)})f_A^{(1)}(w,b,x) + (b^*)^{(2)} + \mathcal{E}_{A0}^{(2)}), \tag{4.80}$$

$$\begin{aligned} f_A^{(3)}(w,b,x) = \sigma(&((w^*)^{(3)} + \mathcal{E}_{AA}^{(3)})f_A^{(2)}(w,b,x) \\ &+ \mathcal{E}_{AB}^{(3)}f_B^{(2)}(w,b,x) + (b^*)^{(3)} + \mathcal{E}_{A0}^{(3)}), \end{aligned} \tag{4.81}$$

$$\begin{aligned} f_A^{(k)}(w,b,x) = \sigma(&((w^*)^{(k)} + \mathcal{E}_{AA}^{(k)})f_A^{(k-1)}(w,b,x) \\ &+ \mathcal{M}_{AB}^{(k)}f_B^{(k-1)}(w,b,x) + (b^*)^{(k)} + \mathcal{E}_{A0}^{(k)}). \end{aligned} \tag{4.82}$$

Here, for $4 \le k \le N^*-1$, $f_B^{(k-1)}(w,b,x) = 0$, since all entries of $w_{BA}^{(k-1)}$, $w_{BB}^{(k-1)}$, and $w_{B0}^{(k-1)}$ are negative and the output of ReLU function $f_B^{(k-2)}(w,b,x)$ is nonnegative. On the other hand,

$$f^{(k)}(w^*,b^*,x) = \sigma((w^*)^{(k)}f^{(k-1)}(w^*,b^*,x) + (b^*)^{(k)}). \tag{4.83}$$

Hence, by Lemma 4.A.3, $2 \le k \le N^*-1$,

$$\|f_1^{(k)}(w,b,x) - f^{(k)}(w^*,b^*,x)\| \tag{4.84}$$

$$\le \|\mathcal{E}_{AA}^{(k-1)}f_A^{(k-1)}(w,b,x) + \mathcal{E}_{01}^{(k)}\| + \delta_{k,3}\|\mathcal{E}_{AB}^{(3)}f_B^{(2)}(w,b,x)\| \tag{4.85}$$

$$+ \|(w^*)^{(k)}(f_A^{(k-1)}(w,b,x) - f^{(k-1)}(w^*,b^*,x))\| \tag{4.86}$$

$$\le \|\mathcal{E}_{AA}^{(k-1)}\|\|f_A^{(k-1)}(w,b,x)\| + \|\mathcal{E}_{A0}^{(k)}\| + \delta_{k,3}\|\mathcal{E}_{AB}^{(3)}\|\|f_B^{(2)}(w,b,x)\| \tag{4.87}$$

$$+ \|(w^*)^{(k)}\|\|f_A^{(k-1)}(w,b,x) - f^{(k-1)}(w^*,b^*,x)\|, \tag{4.88}$$

where $\delta_{k,3} = 1$ if $k = 1$ or 0 otherwise. The entries of matrices in $\mathcal{E}_{AA}^{(k-1)}$, $\mathcal{E}_{AB}^{(3)}$, and $\mathcal{E}_{A0}^{(k)}$ are bounded by $1/\sqrt{n}$ order term and the operator norm is bounded by the Frobenius norm, hence $\|\mathcal{E}_{AA}^{(k-1)}\|$, $\|\mathcal{E}_{AB}^{(3)}\|$, and $\|\mathcal{E}_{A0}^{(k)}\|$ are bounded by $1/\sqrt{n}$ order term. Moreover, $\|(w^*)^{(k)}\|$ is a constant term. For $k = 2$, $f_A^{(k-1)}(w, b, x) - f^{(k-1)}(w^*, b^*, x) = x - x = 0$. Then by using mathematical induction we obtain the Lemma. □

Lemma 4.A.6. *Assume that the weight and bias parameters are in the set W_E. Then there exists a constant $c_3 > 0$ such that*

$$\|f^{(N)}(w, b, x) - f^{(N^*)}(w^*, b^*, x)\| \leq \frac{c_3}{\sqrt{n}}(\|x\| + 1). \tag{4.89}$$

Proof. Let $h \in \mathbb{R}^{H_N}$ and $h^* \in \mathbb{R}^{H_{N^*}^*}$ ($H_N = H_{N^*}^*$) be input vectors into the output layers of the learning and data-generating machines respectively. In other words, h and h^* is defined such that $f^{(N)}(w, b, x) = \sigma(h)$ and $f^{(N^*)}(w^*, b^*, x) = \sigma(h^*)$. By the definition of the essential parameter set (2), for $N^* - 1 \leq k \leq N - 1$, all entries of $w_{BA}^{(k)}$, $w_{BB}^{(k)}$ and $b_{B0}^{(k)}$ are negative. Hence, for $N^* \leq k \leq N - 1$, $f_2^{(k)}(w, b, x) = 0$. For $N^* \leq k \leq N - 1$, all entries of $w_{AA}^{(k)}$, $w_{AB}^{(k)}$ and $b_{A0}^{(k)}$ are positive. Hence, by using $\sigma(t) = t$ for $t \geq 0$,

$$h = w_{AA}^{(N)} w_{AA}^{(N-1)} \cdots w_{AA}^{(N^*)} f_A^{(N^*-1)}(w, b, x) \tag{4.90}$$

$$+ b_{A0}^{(N)} + \sum_{k=N^*}^{N-1} w_{AA}^{(N-1)} \cdots w_{AA}^{(k)} b_{A0}^{(k-1)}. \tag{4.91}$$

On the other hand,

$$h^* = (w^*)^{(N^*)} f^{(N^*-1)}(w^*, b^*, x) + (b^*)^{(N)}. \tag{4.92}$$

If w is in the essential set of parameters,

$$w_{AA}^{(N)} w_{AA}^{(N-1)} \cdots w_{AA}^{(N^*)} = ((w^*)^{(N^*)} B^{-1} + \mathcal{E}_{AA}^{(N)}) w_{AA}^{(N-1)} \cdots w_{AA}^{(N^*)} \tag{4.93}$$

$$= (w^*)^{(N^*)} + \mathcal{E}_{AA}^{(N)} w_{AA}^{(N-1)} \cdots w_{AA}^{(N^*)}. \tag{4.94}$$

It follows that

$$\|w^{(N)} w^{(N-1)} \cdots w^{(N^*)} f^{(N^*-1)}(w, b, x) - w^{(N*)} f^{(N^*-1)}(w^*, b^*, x)\| \tag{4.95}$$

$$\leq \|w_{AA}^{(N)} w_{AA}^{(N-1)} \cdots w_{AA}^{(N^*)} f_A^{(N^*-1)}(w, b, x) - w^{(N*)} f^{(N^*-1)}(w^*, b^*, x)\| \tag{4.96}$$

$$\leq \|(w^*)^{(N^*)}(f_A^{(N^*-1)}(w, b, x) - f^{(N^*-1)}(w^*, b^*, x))\| \tag{4.97}$$

$$+ \|\mathcal{E}_{AB}^{(N)}\| \|w_{AA}^{(N-1)}\| \cdots \|w_{AA}^{(N^*)}\| \|f_A^{(N^*-1)}(w^*, b^*, x)\| \tag{4.98}$$

$$\leq \frac{c_4}{\sqrt{n}}(\|x\| + 1), \tag{4.99}$$

where the last inequality is derived by Lemma 4.A.5. Also by the definition,

$$\|b^{(N)} + \sum_{k=N^*}^{N-1} w^{(N-1)} \cdots w^{(k)} b^{(k-1)} - (b^*)^{(N^*)}\| \leq \frac{c_4}{\sqrt{n}}, \tag{4.100}$$

it follows that

$$\|h - h^*\| \leq \frac{c_5}{\sqrt{n}}(\|x\| + 1).$$

Then applying Lemma 4.A.2 completes the lemma. □

Lemma 4.A.7. *(1) If the support of $q(x)$ is contained in a positive region, the same conclusion as Lemma 4.A.5 holds by replacing $\mathcal{Z}_{AB}^{(3)}$ in (4.73) with $\mathcal{M}_{AB}^{(3)}$.*
(2) If the support of $q(x)$ is contained in a bounded region, the same conclusion as Lemma 4.A.5 holds by replacing $\mathcal{Z}_{AB}^{(3)}$ in (4.73) with $\mathcal{M}_{AB}^{(3)}$ and by replacing $-\mathcal{M}_{B0}^{(3)}$ in (4.72) with a matrix in a sufficiently small region.

Proof. In both cases, $f_B^{(2)}(w, b, x) = 0$ in eq.(4.81) holds. Hence, the same conclusion of Lemma 4.A.5 holds. □

4.A.3 Proof of Main Theorem

In this section we prove the main theorem.

Proof. (Main theorem). By Lemma 4.A.1, it is sufficient to prove that there exists a constant $C > 0$ such that

$$\int_{W_E} \exp(-nK(w, b))\varphi(w, b)dwdb \geq \frac{C}{n^\lambda}$$

where

$$K(w, b) = \frac{1}{2}\int \|f^{(N)}(w, b, x) - f^{(N^*)}(w^*, b^*, x)\|^2 q(x)dx.$$

By using Lemma 4.A.6, if $(w, b) \in W_E$,

$$K(w, b) \leq \frac{c_3^2}{2n}\int (\|x\| + 1)^2 q(x)dx = \frac{c_4}{n} < \infty.$$

It follows that

$$\int_{W_E} \exp(-nK(w, b))\varphi(w, b)dwdb \tag{4.101}$$

$$\geq \exp(-c_4)\left(\min_{(w,b)\in W_E} \varphi(w, b)\right)\mathrm{Vol}(W_E). \tag{4.102}$$

where $c_4 > 0$, $\min_{(w,b)\in W_E} \varphi(w, b) > 0$, and $\mathrm{Vol}(W_E)$ is the volume of the set W_E by the Lebesgue measure. By the definition of the essential parameter set W_E, its volume is determined by the dimension of the convergent matrices and vectors. Let 2λ be the number of parameters in convergent matrices and vectors. Then

$$\mathrm{Vol}(W_E) \geq \frac{C_1}{n^\lambda},$$

where in general cases,

$$\lambda = \frac{1}{2}\left(H_{N^*}^*(H_{N-1}^* + 1) + H_3^*(H_2 - H_2^*) + \sum_{k=2}^{N^*-1} H_k^*(H_{k-1}^* + 1)\right) \tag{4.103}$$

$$= \frac{1}{2}\left(d^* + H_3^*(H_2 - H_2^*)\right). \tag{4.104}$$

If the support of the input distribution is contained in a positive region or a bounded region,

$$\lambda = \frac{1}{2}\left(H^*_{N^*}(H^*_{N-1}+1) + \sum_{k=2}^{N^*-1} H^*_k(H^*_{k-1}+1)\right) \tag{4.105}$$

$$= \frac{d^*}{2}, \tag{4.106}$$

which completes the main theorem. □

Chapter 5

Conclusion

This thesis shows the free energy and the generalization error of Bayesian learning in cases that model has multiple optimal probability distributions and CNNs are overparamtetrized. The first study revealed that if there exist multiple optimal probability distributions, the generalization error gets larger if the number of data increases. This situation occurs when complex learning models are used for complex data such as learning of deep neural networks. The second study revealed that even if CNNs are overparamtetrized, the generalization error only depends on the complexity of data generating process with skip connection. Among the models which are revealed the asymptotic behavior in Bayesian learning, only fully connected DNNs and CNNs with skip connection have such property. We clarified the peculiarities of DNNs on Bayesian learning in these two studies.

Acknowledgment

I would like to thank the following people for helping with this work. My supervisor Prof Sumio Watanabe supported me in many ways from Master course to PhD. I have learned many from his approach to mathematical science. Colleagues in Watanabe laboratory helps me with the discussions and advices about my work and daily life. Dr Naoki Hayashi in particular, advised me after completing his PhD. Dr Susan Wei, University Melbourne invited me to IMS-APRM. I had a good opportunity to discuss singular learning theory as statistics. Edmund Lau, University Melbourne discussed with me and show me around Melbourne city. Prof Makoto Yamashita, Prof Satoshi Takabe, Prof Takafumi Kanamori and Yumiharu Nakano reviewed this work and providing many helpful comments. Finally, I would like to express my strong gratitude to my father and relatives for their support to me for years after my mother's death.

Bibliography

[1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[2] Hirotugu Akaike. Likelihood and the bayes procedure. In *Springer Series in Statistics*, pages 309–332. Springer New York, 1998.

[3] Miki Aoyagi. A bayesian learning coefficient of generalization error and vandermonde matrix-type singularities. *Communications in Statistics - Theory and Methods*, 39(15):2667–2687, jul 2010.

[4] Miki Aoyagi and Kenji Nagata. Learning coefficient of generalization error in bayesian estimation and vandermonde matrix-type singularity. *Neural Computation*, 24(6):1569–1610, jun 2012.

[5] Miki Aoyagi and Sumio Watanabe. Stochastic complexities of reduced rank regression in bayesian estimation. *Neural Networks*, 18(7):924–933, sep 2005.

[6] Mathias Drton and Martyn Plummer. A bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):323–380, feb 2017.

[7] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[8] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

[9] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[10] Mudasir A Ganaie, Minghui Hu, AK Malik, M Tanveer, and PN Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[12] J. A. Hartigan. A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Barkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, 1985*, volume 2, pages 807–810, 1985.

[13] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[14] Naoki Hayashi. The exact asymptotic form of bayesian generalization error in latent dirichlet allocation. *Neural Networks*, 137:127–137, may 2021.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

[17] Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[18] Furong Huang, Jordan Ash, John Langford, and Robert Schapire. Learning deep resnet blocks sequentially using boosting theory. In *International Conference on Machine Learning*, pages 2058–2067. PMLR, 2018.

[19] Natsuki Kariya and Sumio Watanabe. Asymptotic analysis of singular likelihood ratio of normal mixture by bayesian learning theory for testing homogeneity. *Communications in Statistics - Theory and Methods*, pages 1–18, nov 2020.

[20] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[22] Satoshi Kuriki and Akimichi Takemura. The tube method for the moment index in projection pursuit. *Journal of Statistical Planning and Inference*, 138(9):2749–2762, sep 2008.

[23] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[24] David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

[25] David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

[26] Shuya Nagayasu and Sumio Watanabe. Free energy of convolutional neural network with skip connection. In *Proceedings of Asian Conference on Machine Learning.*

[27] Shuya Nagayasu and Sumio Watanabe. Bayesian free energy of deep relu neural network in overparametrized cases. *arXiv preprint arXiv:2303.15739*, 2023.

[28] Shuya Nagayasu and Sumio Watanbe. Asymptotic behavior of free energy when optimal probability distribution is not unique. *Neurocomputing*, 500:528–536, aug 2022.

[29] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

[30] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 1995.

[31] Atsushi Nitanda and Taiji Suzuki. Functional gradient boosting for learning residual-like networks with statistical guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 2981–2991. PMLR, 2020.

[32] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, sep 1978.

[33] Dmitry Rusakov and Dan Geiger. Asymptotic model selection for naive bayesian networks. *Journal of Machine Learning Research*, 6(Jan):1–35, 2005.

[34] Kenichiro Sato and Sumio Watanabe. Bayesian generalization error of poisson mixture and simplex vandermonde matrix type singularity. *arXiv preprint arXiv:1912.13289*, 2019.

[35] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[36] Sam Smith and Quoc V. Le. A bayesian perspective on generalization and stochastic gradient descent.

[37] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639, 2002.

[38] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[39] Michel Talagrand. Gaussian processes and the generic chaining. In *Upper and Lower Bounds for Stochastic Processes*, pages 13–73. Springer Berlin Heidelberg, 2014.

[40] A. W. van der Vaart and J. A. Wellner. Weak convergence. In *Weak Convergence and Empirical Processes*, pages 16–28. Springer New York, 1996.

[41] Sumio Watanabe. Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, 13(4):899–933, apr 2001.

[42] Sumio Watanabe. Algebraic geometrical methods for hierarchical learning machines. *Neural Networks*, 14(8):1049–1060, 2001.

[43] Sumio Watanabe. Almost all learning machines are singular. In *2007 IEEE Symposium on Foundations of Computational Intelligence*, pages 383–388. IEEE, 2007.

[44] Sumio Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12), 2010.

[45] Sumio Watanabe. Asymptotic learning curve and renormalizable condition in statistical-learning theory. *Journal of Physics: Conference Series*, 233:012014, jun 2010.

[46] Sumio Watanabe. A widely applicable bayesian information criterion. *The Journal of Machine Learning Research*, 14(1):867–897, 2013.

[47] Sumio Watanabe. *Mathematical theory of Bayesian statistics*. CRC Press, 2018.

[48] Susan Wei, Daniel Murfet, Mingming Gong, Hui Li, Jesse Gell-Redman, and Thomas Quella. Deep learning is singular, and that's good. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2022.

[49] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[50] K. Yamazaki and Y. Motomura. Hidden node detection between observable nodes based on bayesian clustering. *Entropy*, 21(1):32, jan 2019.

[51] Keisuke Yamazaki and Sumio Watanabe. Singularities in mixture models and upper bounds of stochastic complexity. *Neural Networks*, 16(7):1029–1038, sep 2003.

[52] Keisuke Yamazaki and Sumio Watanabe. Singularities in complete bipartite graph-type boltzmann machines and upper bounds of stochastic complexities. *IEEE Transactions on Neural Networks*, 16(2):312–324, mar 2005.

[53] Keisuke Yamazaki and Sumio Watanbe. Stochastic complexity of bayesian networks. *arXiv preprint arXiv:1212.2511*, 2012.

[54] Yingxue Zhang, Soumyasundar Pal, Mark Coates, and Deniz Ustebay. Bayesian graph convolutional neural networks for semi-supervised classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5829–5836, 2019.

[55] Piotr Zwiernik. An asymptotic behaviour of the marginal likelihood for general markov models. *The Journal of Machine Learning Research*, 12:3283–3310, 2011.

Publications

Peer-reviewed Journal Papers

1. Shuya Nagayasu, Sumio Watanbe. "Asymptotic behavior of free energy when optimal probability distribution is not unique." Neurocomputing, Vol.500, pp. 528-536, August 2022.

Peer-reviewed International Conference

1. Shuya Nagayasu, Sumio Watanabe "Free Energy of Bayesian Convolutional Neural Network with Skip Connection." The 15th Asian Conference on Machine Learning, November, 11-14, 2023, Istanbul, Turkey. to appear in Proceedings of Machine Learning Research.

Not-Peer-reviewed International Conference

1. Shuya Nagayasu, "Generalization Error of Bayesian Deep Neural Network with non analytic activation function" The Institute for Mathematical Statistics Asia-Pacific Rim Meeting January, 4-7, 2024.Melbourne, Australia

Domestic Conference

1. 永安修也, 渡辺澄夫. "最適な確率分布が一意でないときのベイズ学習曲線", 信学技報, vol. 119, no. 453, NC2019-94, pp.107-112, 2020 年 3 月. 電気通信大学
2. 永安修也, 渡辺澄夫. "最適パラメータの学習モデルが一意でないときのベイズ学習の漸近解析", 第 25 回情報論的学習理論ワークショップ, 20221 年 11 月 22 日（火）, つくば国際会議場 茨城

Preprint

1. Shuya Nagayasu, Sumio Watanabe "Bayesian Free Energy of Deep ReLU Neural Network in Overparametrized Cases" https://doi.org/10.48550/arXiv.2303.15739