Free Energy of Bayesian Convolutional Neural Network with Skip Connection

Shuya Nagayasu NAGAYASU.S.AA@M.TITECH.AC.JP
Sumio Watanabe WATANABE.S.AE@M.TITECH.AC.JP
Department of Mathematical and Computing Science
Tokyo Institute of Technology

Editors: Berrin Yanıkoğlu and Wray Buntine

Abstract

Since the success of Residual Network(ResNet), many of architectures of Convolutional Neural Networks(CNNs) have adopted skip connection. While the generalization performance of CNN with skip connection has been explained within the framework of Ensemble Learning, the dependency on the number of parameters has not been revealed. In this paper, we show that Bayesian free energy of Convolutional Neural Network both with and without skip connection in Bayesian learning. Bayesian Free Energy is the negative log marginal likelihood which is equivalent to Stochastic Complexity or Minimum Description Length (MDL) used for evaluating model complexity. The upper bound of free energy of Bayesian CNN with skip connection does not depend on the oveparametrization and, the generalization error of Bayesian CNN has similar property.

Keywords: Learning theory; Convolutional Neural Network; Bayesian Learning; Free Energy

1. Introduction

Convolutional Neural Networks (CNNs) are a type of Neural Networks mainly used for computer vision. CNNs have been shown the high performance with deep layers Szegedy et al. (2015); Krizhevsky et al. (2017). Residual Network(ResNet) He et al. (2016) 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 Huang et al. (2018); Nitanda and Suzuki (2020); Ganaie et al. (2022). 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 Akaike (1974). The similar property is shown in free energy and generalization error in Bayesian learning Schwarz (1978); Rissanen (1978); Akaike (1998). 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 Watanabe (2001b, 2007). In particular, RLCTs are revealed in some concrete models such as three layered neural networks Watanabe (2001a); Aoyagi and Nagata (2012), normal mixtures Hartigan (1985); Yamazaki and Watanabe (2003), Poisson mixtures Sato and Watanabe (2019), Boltzmann machine Yamazaki and Watanabe (2005); Aoyagi (2010), reduced rank regression Aoyagi and Watanabe (2005), Latent Dirichlet allocation Hayashi (2021), matrix factorization, and Bayesian Network Yamazaki and Watanbe (2012). 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 Wei et al. (2022), the RLCT of Deep Neural Network is revealed Nagayasu and Watanabe (2023). 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 or Markov chain Monte Carlo(MCMC) methods. Variational Approximation for neural netowrks, Variational Autoencoder Kingma and Welling (2013) or Monte Carlo dropout Gal and Ghahramani (2016) are practically used. Also for CNNs, variational approach for Bayesian inference was proposed Gal and Ghahramani (2015). MCMC for neural networks, Hamiltonian Monte Carlo or Langevin Dynamics are useful for sampling from posterior. Stochastic Gradient Langevin Dynamics(SGLD) Welling and Teh (2011) is a MCMC method applying Stochastic Gradient Descent instead of Gradient Descent to Langevin Dynamics is popular MCMC for Bayesian Neural Networks. Zhang et al. (2019) used SGLD for generating posterior of CNNs.

In this paper 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 with skip connection, the redundant layers don't affect the free energy and generalization error whereas they affect in case without skip connection. This paper consists of seven main sections and one appendix. In section2, we describe the setting of Convolutional Neural Network analyzed in this paper. In section3, we explain the basic terms of the Bayesian learning. In section4, we note the main theorem of this paper. In section5, we conduct the experiment of synthetic data. In section6 and section7, we discuss about the theorem in this paper and conclusion. In appendixA, we prove the main theorem of this paper.

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:

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

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

$$g(b^{(k)})_{l_1,l_2} = b^{(k)} \tag{2}$$

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

$$f^{(k)}(w, b, x) = \sigma(\mathrm{Conv}(f^{(k-1)}(w, b, x), w^{(k)}) + g(b^{(k)})) \tag{3}$$

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:

$$f^{(K_1+1)}(w, b, x) = \frac{1}{L_1 L_2} \sum_{l_1=1}^{l_1=L_1} \sum_{l_2=1}^{l_2=L_2} f^{(K_1)}(w, b, x)_{l_1,l_2}. \tag{4}$$

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

$$f^{(k)}(w, b, x) = \sigma(w^{(k)} f^{(k-1)}(w, b, x) + b^{(k)}), \tag{5}$$

and for $k = K_1 + K_2$,

$$f^{(K_1+K_2)}(w, b, x) = \mathrm{softmax}(w^{(k)} f^{(k-1)}(w, b, x) + b^{(k)}), \tag{6}$$

where softmax() is a softmax function

$$\mathrm{softmax}(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{J} e^{z_j}}. \tag{7}$$

The output of the model is represented stochastically

$$y \sim \mathrm{Categorical}(f^{(K_1+K_2)}(w, b, x)) \tag{8}$$

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

$$\begin{aligned} f^{(mK_s+2)}(w, b, x) = \sigma(&\mathrm{Conv}(f^{(mK_s+1)}(w, b, x), w^{(mK_s+2)}) \\ &+ g(b^{(mK_s+2)}) + f^{((m-1)K_s+2)}(w, b, x)). \end{aligned} \tag{9}$$

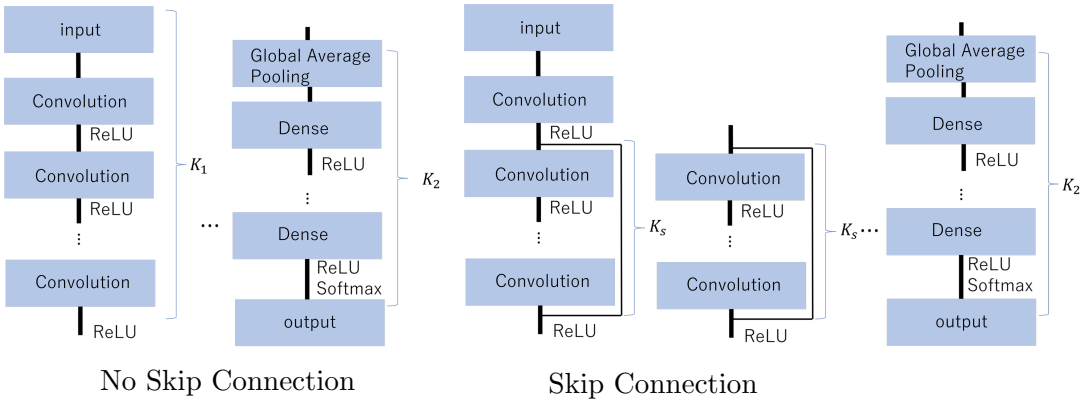

Figure 1: The structure of Convolutional Neural Network with and without Skip Connection

In this case, CNN satisfies the following conditions

$$\begin{aligned} K_1 &= MK_s + 2 \\ H^{mK_s+2} &= \text{const} \quad (1 \leq m \leq M). \end{aligned} \tag{10}$$

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

Figure1 shows the configuration of Convolutional Neural Network analyzed in this paper.

3. Free energy in Bayesian Learning

3.1. Bayesian Learning

Let $X^n = (X_1, \cdots X_n)$ and $Y^n = (Y_1, \cdots Y_n)$ be training data and labels. n is the number of the data. These data and labels are generated from a true 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

$$p(w|X^n, Y^n) = \frac{1}{Z(Y^n|X^n)} \varphi(w) \prod_{i=1}^{n} p(Y_i|X_i, w) \tag{11}$$

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

$$Z_n = \int \varphi(w) \prod_{i=1}^{n} p(Y_i|X_i, w) \mathrm{d}w. \tag{12}$$

The free energy is negative log value of marginal likelihood

$$F_n = -\log Z_n. \tag{13}$$

Free energy is equivalent to evidence and stochastic complexity.

The posterior predictive distribution is defined as the average of the model by posterior:

$$p^*(y|x) = p(y|x, X^n, Y^n) = \int p(y|x, w)p(w|X^n, Y^n)\mathrm{d}w. \tag{14}$$

Generalization error G_n is given by Kullback-Leibler divergence between the true distribution and posterior distribution as follows

$$G_n = \int q(y|x)q(x)\log\frac{q(y|x)}{p^*(y|x)}\mathrm{d}x\mathrm{d}y. \tag{15}$$

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

$$\mathbb{E}[G_n] - S = \mathbb{E}[F_{n+1}] - \mathbb{E}[F_n], \tag{16}$$

where $\mathbb{E}[f(X^n, Y^n)]$ is the average of the generation of n data $\mathbb{E}_{X^n,Y^n}[f(X^n, Y^n)]$ and S is the entropy of $q(y|x)$.

3.2. Asymptotic property of Free energy and Generalization error

It is well known that if the average Kullback-Leibler divergence

$$K(w) = \int q(y|x)q(x)\log\frac{q(y|x)}{p(y|x, w)}\mathrm{d}x\mathrm{d}y. \tag{17}$$

can be approximated by 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 learning model d Schwarz (1978); Rissanen (1978)

$$E[F_n] = n(S + \mathrm{Bias}) + \frac{d}{2}\log n + O(1) \tag{18}$$

where S is entropy of true distribution and Bias is the minimum value of $K(w)$ for $w \in W$. The generalization error is calculated from Free energy by using equation(16) Akaike (1998):

$$E[G_n] = \mathrm{Bias} + \frac{d}{2n} + o\left(\frac{1}{n}\right). \tag{19}$$

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 such models, the average of Free energy and Generalization error have the following asymptotic expansions Watanabe (2001b)

$$E[F_n] = n(S + \mathrm{Bias}) + \lambda\log n + o(\log n), \tag{20}$$

$$E[G_n] = \mathrm{Bias} + \frac{\lambda}{n} + o\left(\frac{1}{n}\right), \tag{21}$$

where λ is a rational number called Real Log Canonical Threshold(RLCT). In particular, Nagayasu and Watanabe (2023) showed that in case Bias $= 0$ and x is bounded, when the Deep Neural Network is trained from the data generated from smaller network,

$$\lambda \leq \frac{d^*}{2} \tag{22}$$

where $d^* \leq d$ is the number of parameter of data generating Network.

4. Main Theorem

In this subsection the main result of this paper is introduced. First, to state the theorem, we define the data generating network. Both in skip connection the data generating network satisfies the following conditions about the number of layers and filters,

$$K_1^* \leq K_1, K_2^* \leq K_2, (H^*)^{(1)} = H^{(1)} (H^*)^{(K_1)} = H^{(K_1+K_2)} \tag{23}$$

and

$$\begin{aligned} H^{(k)} &\geq (H^*)^{(K_1^*)} (K_1^* + 1 \leq k \leq K_1) \\ H^{(k)} &\geq (H^*)^{(K_1+K_2^*)} (K_1 + K_2^* + 1 \leq k \leq K_1 + K_2 - 1) \\ H^{(k)} &\geq (H^*)^{(k)} (others). \end{aligned} \tag{24}$$

Then, we show the main theorem.

Theorem 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 (23) and (24), 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,*

$$\mathbb{E}[F_n] \leq nS + \lambda_{CNN} \log n + C \tag{25}$$

where

$$\lambda_{CNN} = \frac{1}{2} \left(|w^*|_0 + |b^*|_0 + \sum_{k=K_1^*+1}^{K_1} (9H_k + 1) H_k \right) \tag{26}$$

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

Theorem 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 (10), (23) and (24), and that a training data $\{(X_i, Y_i) \ \ i = 1,2,...,n\}$ is independently taken from $q(x)q(y|x)$. Then*

$$\lambda_{CNN} = \frac{1}{2} (|w^*|_0 + |b^*|_0). \tag{27}$$

Proof of main theorems are shown in AppendixA.

If there exists asymptotic expansion of the generalization error $\mathbb{E}[G_n]$ in theorem1 and theorem2, that satisfies the following inequality

$$\mathbb{E}[G_n] \leq \frac{\lambda_{CNN}}{n} + o\left(\frac{1}{n}\right), \tag{28}$$

where

$$G_n = \int q(x) \sum_{i=1}^{H_{K_1+K_2}} f_i^{(K_1^*+K_2^*)}(w^*, b^*, x) \log \frac{f_i^{(K_1^*+K_2^*)}(w^*, b^*, x)}{\mathbb{E}_{w,b}[f_i^{(K_1+K_2)}(w, b, x)]} \mathrm{d}x \tag{29}$$

which corresponds to categorical cross entropy.

5. Experiment

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

5.1. Methods

We prepared the 2-class labeled simple data shown in fig2. The the data is $x \in R^{4\times4}$ and the values of each elements 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 fig2 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 layers 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 train 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 estimate the generalization error by the test error of 10000 test data from true model. We trained each learning model 10 times and estimated the $\mathbb{E}[G_n]$ from the average of test error.

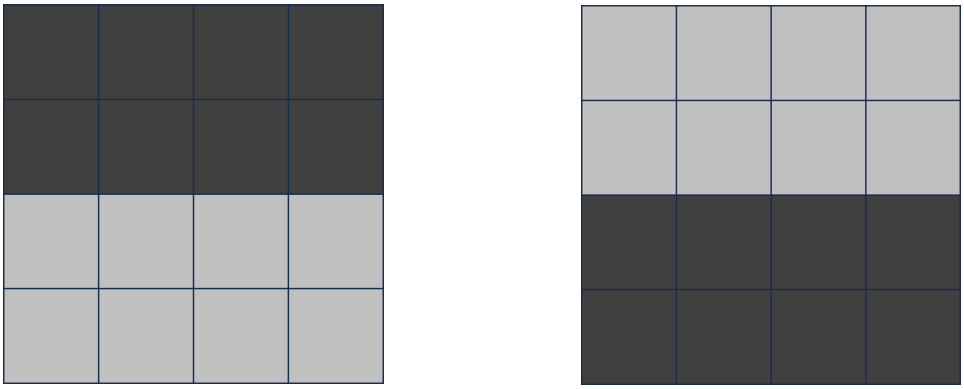

Figure 2: The average of input x of each label

5.2. Result of experiments

Table1 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(21), theoretical upper bound of the generalization error is λ_{CNN}/n. In table1, 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 with the increase of the number of layer. Then, in case $K_1 = 4$ without skip connection, the experimental value increased from

Table 1: Experimental value and theoretical upper bound of the generalization error

model	$n\times$ Test Error	$\lambda_{\rm CNN}$	$d_{\rm model}/2$
$K_1 = 2$	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

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.

6. Discussion

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.

6.2. Comparison to Deep Neural Network

Firstly we compare the result of this paper to that of DNN in Nagayasu and Watanabe (2023). 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.

7. Conclusion

In this paper, we studied Free energy of Bayesian Convolutinal 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 without skip connection does not depends 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.

References

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974. doi: 10.1109/tac.1974.1100705.

Hirotugu Akaike. Likelihood and the bayes procedure. In *Springer Series in Statistics*, pages 309–332. Springer New York, 1998. doi: 10.1007/978-1-4612-1694-0_24.

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. doi: 10.1080/03610920903094899.

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. doi: 10.1162/neco_a_00271.

Miki Aoyagi and Sumio Watanabe. Stochastic complexities of reduced rank regression in bayesian estimation. *Neural Networks*, 18(7):924–933, sep 2005. doi: 10.1016/j.neunet.2005.03.014.

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

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.

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.

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.

Naoki Hayashi. The exact asymptotic form of bayesian generalization error in latent dirichlet allocation. *Neural Networks*, 137:127–137, may 2021. doi: 10.1016/j.neunet.2021.01.024.

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.

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.

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

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

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

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.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, sep 1978. doi: 10.1016/0005-1098(78)90005-5.

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

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2): 461–464, 1978. doi: 10.1214/aos/1176344136.

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.

Sumio Watanabe. Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, 13(4):899–933, apr 2001a. doi: 10.1162/089976601300014402.

Sumio Watanabe. Algebraic geometrical methods for hierarchical learning machines. *Neural Networks*, 14(8):1049–1060, 2001b. doi: 10.1016/s0893-6080(01)00069-7.

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

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. doi: 10.1109/tnnls.2022.3167409.

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.

Keisuke Yamazaki and Sumio Watanabe. Singularities in mixture models and upper bounds of stochastic complexity. *Neural Networks*, 16(7):1029–1038, sep 2003. doi: 10.1016/s0893-6080(03)00005-4.

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. doi: 10.1109/tnn.2004.841792.

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

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.

Appendix A. Proof of main theorem

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

A.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

$$K(w,b) = \int q(x)q(y|x)\log\frac{q(y|x)}{p(y|x,w,b)}\mathrm{d}x\mathrm{d}y. \tag{30}$$

Lemma 3 *Nagayasu and Watanabe (2023) 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 postive integer n,*

$$\mathbb{E}[F_n] \leq nS - \log\int_W \exp(-nK(w,b))\varphi(w,b)\mathrm{d}w\mathrm{d}b. \tag{31}$$

Lemma 4 *Nagayasu and Watanabe (2023) For arbitrary vectors s, t,*

$$\|\sigma(s) - \sigma(t)\| \leq \|s - t\|. \tag{32}$$

Lemma 5 *Nagayasu and Watanabe (2023) For arbitrary w,w', b, b', and $K_1 + 1 \leq k \leq K_1 + K_2$, the following inequality holds,*

$$\begin{aligned}
&\|f^{(k)}(w,b,x) - f^{(k)}(w',b',x)\| \\
&\leq \|w^{(k)} - w'^{(k)}\|\|f^{(k-1)}(w,b,x)\| + \|b^{(k)} - b'^{(k)}\| \\
&+ \|w^{(k)}\|\|f^{(k-1)}(w,b,x) - f^{(k-1)}(w',b',x)\|.
\end{aligned} \tag{33}$$

Corollary 6 *For arbitrary w,w', b, b', and $1 \leq k \leq K_1$, the following inequality holds,*

$$\begin{aligned}
&\|f^{(k)}(w,b,x) - f^{(k)}(w',b',x)\| \\
&\leq 9\|w^{(k)} - w'^{(k)}\|\|f^{(k-1)}(w,b,x)\| + L_1L_2\|b^{(k)} - b'^{(k)}\| \\
&+ 9\|w^{(k)}\|\|f^{(k-1)}(w,b,x) - f^{(k-1)}(w',b',x)\| \\
&+ \delta^{(k)}\|w^{(k)}\|\|f^{(k-K_2-1)}(w,b,x) - f^{(k-K_2-1)}(w',b',x)\|.
\end{aligned} \tag{34}$$

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

Proof

$$
\begin{aligned}
&f^{(k)}(w,b,x) - f^{(k)}(w',b',x) \\
&= \sigma(\mathrm{Conv}(f^{(k-1)}(w,b,x), w^{(k)}) + g(b^{(k)})) - \sigma(\mathrm{Conv}(f^{(k-1)}(w,b,x), w'^{(k)})) + g(b'^{(k)})) \\
&+ \sigma(\mathrm{Conv}(f^{(k-1)}(w,b,x), w'^{(k)}) + g(b'^{(k)})) - \sigma(\mathrm{Conv}(f^{(k-1)}(w',b',x), w'^{(k)}) + g(b'^{(k)})).
\end{aligned} \tag{35}
$$

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

$$
\begin{aligned}
\|\mathrm{Conv}(f^{(k-1)}(w,b,x), w^{(k)})_j\| &\leq \sum_{i=1}^{H^{k-1}} \|f^{(k-1)}(w,b,x)_i\| |(\sum_{p=1}^{3}\sum_{q=1}^{3} w^{(k)}_{pqij})|_1 \\
&\leq 9\|f^{(k-1)}(w,b,x)\| \|w_{:,:,:,j}\|
\end{aligned} \tag{36}
$$

By using lemma4, (35) and (36), corollary6 is proved. ■

Lemma 7 *For arbitrary w, b, x,*

$$
\begin{aligned}
\|f^{(k)}(w,b,x)\| &\leq \mathcal{D}_k \|w^{(k)}\| \|w^{(k-1)}\| \cdots \|w^{(2)}\| \|x\| \\
&+ \mathcal{D}_0 \|b^{(k)}\| + \sum_{j=1}^{k-2} \mathcal{D}_j \|w^{(k)}\| \|w^{(k-1)}\| \cdots \|w^{(k-j)}\| \|b^{(k-j)}\|.
\end{aligned} \tag{37}
$$

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 5, it follows that

$$
\begin{aligned}
\|f^{(k)}(w,b,x)\| &\leq 9\|w^{(k)}\| \|f^{(k-1)}(w,b,x)\| + L_1 L_2 \|b^{(k)}\| \\
&+ \delta^{(k)} \|w^{(k)}\| \|f^{(k-K_2-1)}(w,b,x) - f^{(k-K_2-1)}(w',b',x)\|.
\end{aligned} \tag{38}
$$

Then mathematical induction gives the Lemma. ■

A.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 later is denoted as B. The

convergent tensor $\mathcal{E}^{(k)} \in \mathbb{R}^{3\times 3\times H^{(k-1)}\times H^{(k)}}$ and vector $\mathcal{E}_0^{(k)}\mathbb{R}^{H^{(k)}}$ where the absolute value of all elements are smaller than $1/\sqrt{n}$ are denoted by

$$\mathcal{E}_{pq}^{(k)} = \begin{pmatrix} \mathcal{E}_{pqAA}^{(k)} & \mathcal{E}_{pqAB}^{(k)} \\ \mathcal{E}_{pqBA}^{(k)} & \mathcal{E}_{pqBB}^{(k)} \end{pmatrix}, \quad (1 \le p \le 3, 1 \le q \le 3), \tag{39}$$

$$\mathcal{E}_0^{(k)} = \begin{pmatrix} \mathcal{E}_{A0}^{(k)} \\ \mathcal{E}_{B0}^{(k)} \end{pmatrix}. \tag{40}$$

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

$$\mathcal{M}_{pq}^{(k)} = \begin{pmatrix} \mathcal{M}_{pqAA}^{(k)} & \mathcal{M}_{pqAB}^{(k)} \\ \mathcal{M}_{pqBA}^{(k)} & \mathcal{M}_{pqBB}^{(k)} \end{pmatrix}, \quad (1 \le p \le 3, 1 \le q \le 3), \tag{41}$$

$$\mathcal{M}_0^{(k)} = \begin{pmatrix} \mathcal{M}_{A0}^{(k)} \\ \mathcal{M}_{B0}^{(k)} \end{pmatrix}. \tag{42}$$

To prove Theorem 1 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 Convlutional Layers and Fully connected layers.

A.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 $2 \le k \le K_1$,

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

$$w_{pq}^{(k)} = \begin{pmatrix} (w^*)^{(k)} + \mathcal{E}_{pqAA}^{(k)} & \mathcal{M}_{pqAB}^{(k)} \\ -\mathcal{M}_{pqBA}^{(k)} & -\mathcal{M}_{pqBB}^{(k)} \end{pmatrix}, \tag{43}$$

$$b^{(k)} = \begin{pmatrix} (b^*)^{(k)} + \mathcal{E}_{A0}^{(k)} \\ -\mathcal{M}_{B0}^{(k)} \end{pmatrix}, \tag{44}$$

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

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

$$w_{pq}^{(k)} = \begin{pmatrix} \mathcal{Z}_{pqAA}^{(k)} & \mathcal{M}_{pqAB}^{(k)} \\ -\mathcal{M}_{pqBA}^{(k)} & -\mathcal{M}_{pqBB}^{(k)} \end{pmatrix}, \tag{45}$$

$$b^{(k)} = \begin{pmatrix} (b^*)^{(k)} + \mathcal{E}_{A0}^{(k)} \\ -\mathcal{M}_{B0}^{(k)} \end{pmatrix}, \tag{46}$$

where

$$\mathcal{Z}_{pqAA}^{(k)} = \begin{cases} I_{22AA} + \mathcal{E}_{22AA}^{(k)} & (p = q = 2) \\ \mathcal{E}_{pqAA}^{(k)} & (\text{others}) \end{cases}. \tag{47}$$

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

Lemma 8 *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*

$$\|f^{(K_1)}_{:,:,A}(w,b,x) - f^{(K_1^*)}(w^*,b^*,x)\| \leq \frac{c_1}{\sqrt{n}}(\|x\|+1), \tag{48}$$

$$\|f^{(K_1)}_{:,:,A}(w,b,x)\| \leq c_2(\|x\|+1). \tag{49}$$

Proof Eq.(49) is derived from Lemma 7. By the definitions (43), (44), for $2 \leq k \leq K^*$

$$f^{(2)}_A(w,b,x) = \sigma(\mathrm{Conv}(f^{(1)}_{:,:,A}(w,b,x), (w^*)^{(2)} + \mathcal{E}^{(2)}_{:,:,AA}) + g((b^*)^{(2)} + \mathcal{E}^{(2)}_{A0})), \tag{50}$$

$$\begin{aligned} f^{(k)}_A(w,b,x) = \sigma(&\mathrm{Conv}(f^{(k-1)}_{:,:,A}(w,b,x), (w^*)^{(k)} + \mathcal{E}^{(k)}_{:,:,AA}) \\ &+ \mathrm{Conv}(f^{(k-1)}_{:,:,B}(w,b,x), \mathcal{M}^{(k)}_{:,:,AB}) + g((b^*)^{(k)} + \mathcal{E}^{(k)}_{A0})). \end{aligned} \tag{51}$$

In $k=2$, $|x|$ is bounded and $\mathcal{M}^{(k)}_{:,:,AB}$ is constant tensor, $\mathcal{M}^{(k)}_{B0}$ is large sufficiently, $f^{(2)}_{:,:,B}(w,b,x) = 0$ because all the elements of the output of ReLU function $f^{(2)}(w,b,x)$ is 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 5, for $2 \leq k \leq K_1^*$,

$$\begin{aligned} &\|f^{(k)}_{:,:,A}(w,b,x) - f^{(k)}(w^*,b^*,x)\| \\ &\leq 9\|\mathcal{E}^{(k)}_{:,:,AA}\|\|f^{(k-1)}(w,b,x)\| + L_1L_2\|\mathcal{E}^{(k)}_{A0}\| \\ &+ 9\|(w^*)^{(k)}\|\|f^{(k-1)}_{:,:,A}(w,b,x) - f^{(k-1)}(w^*,b^*,x)\|. \end{aligned} \tag{52}$$

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

$$\begin{aligned} &\|f^{(k)}_{:,:,A}(w,b,x) - f^{(K_1^*)}(w^*,b^*,x)\| \\ &\leq 9\|\mathcal{E}^{(k)}_{:,:,AA}\|\|f^{(k-1)}(w,b,x)\| + L_1L_2\|\mathcal{E}^{(k)}_{A0}\| \\ &+ 9\|(w^*)^{(k)}\|\|f^{(k-1)}_{:,:,A}(w,b,x) - f^{(K_1^*)}(w^*,b^*,x)\|. \end{aligned} \tag{53}$$

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 (52) and (53) , the all terms can be bounded by $1/\sqrt{n}$ terms, hence we obtained the Lemma. ■

From Nagayasu and Watanabe (2023), because of the output in $k = K_1 + 1$ is nonnegative there exists 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 1). By Lemma 3, it is sufficient to prove that there exists a constant $C > 0$ such that

$$\int_{W_E} \exp(-nK(w,b))\varphi(w,b)\mathrm{d}w\mathrm{d}b \geq \frac{C}{n^{\lambda}} \tag{54}$$

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

$$K(w,b) \leq \frac{c_4}{2}\int \|f^{(K_1+K_2)}(w,b,x) - f^{(K_1^*+K_2^*)}(w^*,b^*,x)\|^2 q(x)\mathrm{d}x. \tag{55}$$

By using Lemma 8, if $(w,b) \in W_E$,

$$K(w,b) \leq \frac{c_4 c_3^2}{2n}\int (\|x\|+1)^2 q(x)\mathrm{d}x = \frac{c_5}{n} < \infty. \tag{56}$$

It follows that

$$\begin{aligned}&\int_{W_E} \exp(-nK(w,b))\varphi(w,b)\mathrm{d}w\mathrm{d}b \\ &\geq \exp(-c_5)\left(\min_{(w,b)\in W_E}\varphi(w,b)\right)\mathrm{Vol}(W_E).\end{aligned} \tag{57}$$

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,

$$\mathrm{Vol}(W_E) \geq \frac{C_1}{n^{\lambda}}, \tag{58}$$

where

$$\begin{aligned}\lambda &= \frac{1}{2}\left(\sum_{k=2}^{k=K_1}(9H_{k-1}^*+1)H_k^* + \sum_{k=K_1+1}^{k=K_1+K_2}(H_{k-1}^*+1)H_k^*\right) \\ &= \frac{1}{2}\left(|w^*|_0 + |b^*|_0 + \sum_{k=K_1^*+1}^{K_1}(9H_{K_1^*}+1)H_{K_1^*}\right).\end{aligned} \tag{59}$$

We obtained theorem1.

A.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 (43) and (44).
(2) For $K_1^*+1 \leq k \leq K_1$

$$w_{pq}^{(k)} = \begin{pmatrix} -\mathcal{M}_{pqAA}^{(k)} & -\mathcal{M}_{pqAB}^{(k)} \\ -\mathcal{M}_{pqBA}^{(k)} & -\mathcal{M}_{pqBB}^{(k)} \end{pmatrix}, \tag{60}$$

$$b^{(k)} = \begin{pmatrix} -\mathcal{M}_{A0}^{(k)} \\ -\mathcal{M}_{B0}^{(k)} \end{pmatrix}, \tag{61}$$

Lemma 9 *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*

$$\|f^{(K_1)}_{:,:,A}(w,b,x) - f^{(K_1^*)}(w^*,b^*,x)\| \leq \frac{c_1}{\sqrt{n}}(\|x\|+1), \tag{62}$$

$$\|f^{(K_1)}_{:,:,A}(w,b,x)\| \leq c_2(\|x\|+1). \tag{63}$$

Proof Because of similar reason to lemma8, holds. By Lemma 5, for $k = mK_s + 1$,

$$\begin{aligned}
&\|f^{(k)}_{:,:,A}(w,b,x) - f^{(k)}(w^*,b^*,x)\| \\
&\leq 9\|\mathcal{E}^{(k)}_{:,:,AA}\|\|f^{(k-1)}(w,b,x)\| + L_1L_2\|\mathcal{E}^{(k)}_{A0}\| \\
&+ 9\|(w^*)^{(k)}\|\|f^{(k-1)}_{:,:,A}(w,b,x) - f^{(k-1)}(w^*,b^*,x)\| \\
&+ \|w^{(k)}\|\|f^{(k-K_2-1)}(w,b,x) - f^{(k-K_2-1)}(w',b',x)\|.
\end{aligned} \tag{64}$$

If $k \neq mK_s + 1$ and $2 \leq k \leq K_1^*$, inequality (52) holds. Same as the lemma8, 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 lemma8. 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.

$$f^{(k)}_{:,:,A}(w,b,x) = \begin{cases} f^{(K_1^*)}(w,b,x) & (k = nK_s + 1) \\ 0 & (\text{others}) \end{cases}. \tag{65}$$

Hence, we obtained the Lemma. ■

Same as without Skip connection case, by using the result of Nagayasu and Watanabe (2023) for fully connected layer and inequality(57),(58), we obtained theorem2.