ASYMPTOTIC THEORY OF EMPIRICAL AND VARIATIONAL BAYES LEARNING

Department of Computational Intelligence and Systems Science
Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology

Shinichi Nakajima

2006

Abstract

It is well-known that, in singular models, the Bayes estimation provides much better generalization performance than the maximum likelihood (ML) estimation. However, its accurate approximation by Markov chain Monte Carlo methods requires huge computational costs. As an alternative, a tractable approximation method, called the variational Bayes (VB) approach, has recently been proposed and been attracting people's attention. Its advantage over the expectation-maximization algorithm, which was proposed to realize the ML estimation, has been experimentally shown in many applications, however, has not been theoretically shown yet.

In this thesis, focusing on three-layer linear neural networks (LNNs), the simplest singular models, we analyze the VB approach and a subspace Bayes (SB) approach, the other approximation method considered to be an extension of the empirical Bayes approach. We derive their solutions, and then theoretically clarify their generalization properties, i.e., the generalization error, the training error, and the free energy. We thus discuss when and why the approximation methods provide good generalization performance. As a result, we find the following facts: that, in LNNs, the SB and the VB approaches are asymptotically equivalent to a positive-part James-Stein type shrinkage estimation and another one, respectively, which are similar to each other; that, in typical cases, they provide good generalization performance comparable to the Bayes estimation; that they have not only a property similar to the Bayes estimation but also another property similar to the ML estimation; and that the free energy and the generalization error in the VB approach are not simply related with each other, unlike in the Bayes estimation.

We also consider the *delicate* situations when the Kullback-Leibler divergence

of the true distribution from the singularities is comparable to the inverse of the number of training samples. This consideration is important in model selection problems and in statistical tests, and necessary to show the consistency of our results with the proved admissibility of the Bayes estimation.

Acknowledgement

This thesis is the summary of my work from 2003 to 2005 in the PhD course of the Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology. Many people have helped me conduct this thesis.

First and foremost, with a deep sense of gratitude, I would like to express my sincere thanks to Professor Sumio Watanabe for his supervision and suggestions, as well as his invitation to Watanabe laboratory. His suggestions often cleared obstacles from my way, and guided me to fruitful fields. It was a great pleasure to work under his supervision.

I must deeply thank Nikon corporation for the approval of my study in Watanabe laboratory and the financial support. I would like to express my gratitude to Michio Kariya, President, Chief Executive Officer, and Chief Operating Officer, Kazuo Ushida, Managing Director and Senior Executive Officer, Toshikazu Umatate, Executive Officer and General Manager of Development Headquarters, Precision Equipment Company, and Masahiro Nei, General Manager of First Development Department, Precision Equipment Company. I would especially like to thank Nobutaka Magome, Manager of Sixth Development Section, First Development Department, Precision Equipment Company, for his understanding and encouragement. I am obliged to all the other people working for Nikon as well.

I would like to thank the members of Watanabe laboratory for productive discussions. Professor Keisuke Yamazaki taught me the basics of the Bayes estimation and of the algebraic geometrical analysis. Kazuho Watanabe taught me the basics of the variational Bayes approach, and was very helpful during our discussions. Discussions with Tikara Hosino, Nobuhiro Nakano, Kenji Nagata, and

Shingo Takamatsu were very helpful for me to understand singularities, Bayesian learning methods, and so on.

I would also like to thank the members of my PhD committee, Professor Yoshiyuki Kabashima, Professor Toshiaki Murofushi, Professor Misako Takayasu, and Professor Toru Aonishi, for providing valuable comments on my presentation and on the draft of this thesis. I would also like to express my gratitude to Professor Miki Aoyagi of Sophia University for her lecture on the algebraic geometry, and Professor Kenji Fukumizu of Institute of Statistical Mathematics for his answer to my question concerned with his work. I wish to express my sincere thanks to anonymous reviewers of the journals, the conferences, and the workshops to which I submitted my papers. Their kind and meaningful comments polished this thesis as well as the papers.

I am obliged to my colleagues of Nikon corporation for their help. Discussions with Yuho Kanaya make my understanding clearer. His knowledge on TEX and other software also helped a lot. Yuji Kokumai, Taro Sugihara, Hiroki Okuno, and Ayako Sukegawa helped cover my other work for Nikon in my absence. Many other people helped me as well.

Last but not least, I would like to express my gratitude to my family. I am very grateful to my parents, Eiji and Tomoko, for the education they gave me. I remember that Eiji, who loved mathematics but was a teacher of Japanese language and literature, gave me mathematical quizzes in my childhood, which has made me who I am. Special thanks go to my wife, Yuki, for her support, encouragement, and patience during my PhD period.

Contents

Chapter 1

Introduction

This thesis contributes to the theory of supervised learning with singular models and Bayesian learning methods in the asymptotic limit, i.e., when the number of training samples goes to infinity. The first chapter is devoted to the background, the purpose, and the overview of this thesis. In Section 1.1, we explain a statistical learning machine, before we state the purpose in Section 1.2, and the overview in Section 1.3.

1.1 Learning Machine

A statistical learning machine consists of two fundamental elements (Fig. 1.1). The first one is a model, usually denoted by a probabilistic distribution with un-

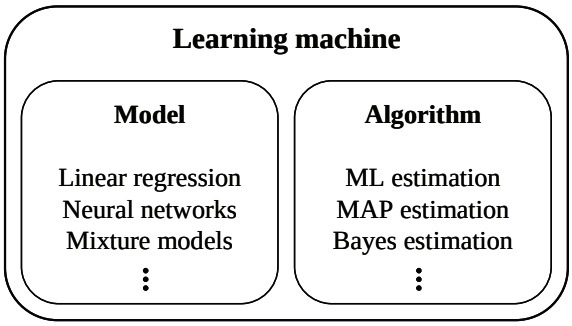

Figure 1.1: A statistical learning machine consists of a model and an algorithm.

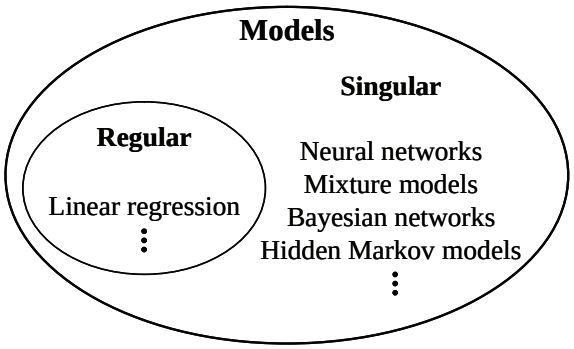

Figure 1.2: Models are classified into regular and singular.

known parameters. The other one is an algorithm or method to estimate the parameters from training samples.

1.1.1 Models — Regular and Singular

The statistical models can be classified into two classes, regular and singular (Fig. 1.2). Although the regularity condition is a little complicated [Fukumizu *et al.*, 2004], we can roughly classify by checking identifiability.[1] We say that a parametric model is identifiable if the map from the parameter to the probability distribution is one-to-one. Therefore, identifiable models do not have any singularity, on which the Fisher information matrix is singular, in their parameter spaces. If the true parameter is at a regular point, the distribution of the maximum likelihood (ML) estimator, as well as the Bayes posterior distribution, asymptotically converges to the normal distribution. Because the asymptotic normality makes theoretical analysis relatively easy, many of the properties of the regular models were clarified. Thus, the conventional learning theory (CLT) was established [Cramer, 1949; Sakamoto *et al.*, 1983; 1986].

The CLT describes the asymptotic behavior of the generalization error, that of the training error, that of the Bayes free energy, etc., which provides the mathemat-

[1] Boundary points of the parameter space can be singularities as well.

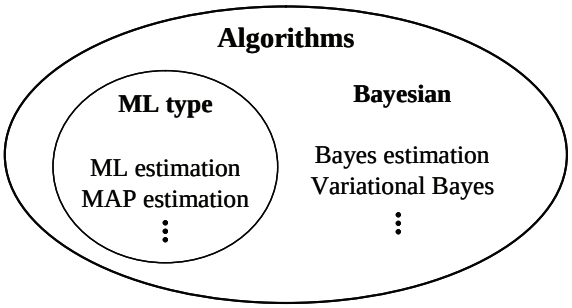

Figure 1.3: Algorithms are classified into ML type and Bayesian.

ical foundation of penalized likelihood type information criteria, such as Akaike's information criterion (AIC) [Akaike, 1974], Bayesian information criterion (BIC) [Schwarz, 1978], the minimum description length criterion (MDL) [Rissanen, 1986], etc.. However, many models recently used in the field of machine learning, such as neural networks, mixture models, hidden Markov models, Bayesian networks, etc., are unidentifiable and hence have singularities in their parameter spaces. Around the singularities, the log-likelihood cannot be approximated by any quadratic form of the parameter. Hence, neither the distribution of the ML estimator nor the Bayes posterior distribution converges to the normal distribution even in the asymptotic limit, which prevents the CLT to hold [Hartigan, 1985; Watanabe, 1995; Amari *et al.*, 2002; Hagiwara, 2002]. In addition, the information criteria, accordingly, have no theoretical foundation in such singular models. Some properties of learning in singular models have been theoretically clarified in the recent decades, however, many are yet to be clarified.

1.1.2 Algorithms — Maximum Likelihood type and Bayesian

On the other hand, the learning algorithms can also be classified into two types, maximum likelihood (ML) type and Bayesian (Fig. 1.3). In the ML estimation, which typifies the former type, we estimate the parameter value so as to maximize the likelihood function given training data. After training, we predict a new observation with the *one* model denoted by the estimated parameter. In contrast, in the

Bayes estimation, which typifies the latter type, we estimate the posterior distribution of the parameter given training data, and then predict a new observation with the *ensemble* of the models subject to the posterior distribution. The difference between the ML type and the Bayesian methods is whether the predictive distribution is denoted by *one* model or by *ensemble* of models. Note that the maximum a posterior (MAP) estimation is classified into the ML type because its predictive distribution is denoted by *one* model, although it utilizes a prior distribution based on the Bayesian framework.

In the regular models, the CLT states that asymptotic generalization performance does not depend on algorithm type, as will be mentioned in Section 1.1.3. However, we should emphasize that, in singular models, there is a significant difference between the algorithm types, of which the reason will be explained in Section 1.1.4.

1.1.3 Conventional Learning Theory

Now, we summarize some results, which are the most important and related with this thesis, of the CLT [Cramer, 1949; Sakamoto *et al.*, 1983; 1986], and based on which, we introduce some information criteria. First of all, remember that this theory is based on the regularity condition, so it does not hold in the singular models.

Suppose $p(y|x;w)$ is the distribution of a regular model, where x is an input, y is an output, and w is a parameter. Let $X^n = \{x_1, \ldots, x_n\}$ and $Y^n = \{y_1, \ldots, y_n\}$ be arbitrary n training samples independently and identically taken from the true distribution $q(x,y) = q(x)q(y|x)$. After training, we have the predictive distribution $p(y|x, X^n, Y^n)$ by using a learning algorithm. For example, in the ML estimation, we have the predictive distribution given by

$$p_{\text{MLE}}(y|x, X^n, Y^n) = p(y|x; \hat{w}_{\text{MLE}}), \tag{1.1}$$

where

$$\hat{w}_{\text{MLE}} = \underset{w}{\text{argmax}} \left(\prod_{i=1}^{n} p(y_i|x_i;w) \right) \tag{1.2}$$

is the ML estimator. The generalization error, a criterion of generalization performance, is defined as the Kullback-Leibler (KL) divergence of the predictive distribution from the true distribution:

$$G(X^n, Y^n) = \int q(x)q(y|x) \log \frac{q(y|x)}{p(y|x, X^n, Y^n)} dxdy, \tag{1.3}$$

and the training error, often utilized as an estimator of the generalization error, is defined as the empirical KL divergence:

$$T(X^n, Y^n) = n^{-1} \sum_{i=n}^{n} \log \frac{q(y|x)}{p(y|x, X^n, Y^n)}. \tag{1.4}$$

We write the average of them as follows:

$$G(n) = \langle G(X^n, Y^n) \rangle_{q(X^n, Y^n)}, \tag{1.5}$$

$$T(n) = \langle T(X^n, Y^n) \rangle_{q(X^n, Y^n)}, \tag{1.6}$$

where $\langle \cdot \rangle_{q(X^n, Y^n)}$ denotes the expectation value over all sets of n training samples. The average generalization error, Eq.(1.5), represents the generalization performance of the learning machine consisting of the model and the learning algorithm. Consider a learning machine whose predictive distribution converges to the true model in the asymptotic limit. Then, the generalization error and the training error can, commonly, be asymptotically expanded as follows:

$$G(n) = \lambda n^{-1} + o(n^{-1}), \tag{1.7}$$

$$T(n) = \nu n^{-1} + o(n^{-1}), \tag{1.8}$$

where the coefficients of the leading terms, λ and ν, are called the generalization and the training coefficients, respectively, in this thesis. We denote the variables with respect to the regular models by the ones subscripted with 'Regular', for example, G_{Regular}, λ_{Regular}, etc.. In the regular models, the asymptotic normality of the ML estimator leads to the following simple and important relation:

$$2\lambda_{\text{Regular}} = -2\nu_{\text{Regular}} = K, \tag{1.9}$$

where K is the parameter dimension of the model. In addition, this relation holds not only in the ML estimation but also in the Bayes estimation.

Based on the simple relation, Eq.(1.9), Akaike's information criterion (AIC) for model selection was proposed [Akaike, 1974], as shown below. We first explain the model selection problem. Consider the polynomial regression model, for example:

$$y = \sum_{k=1}^{K} w_k x^{k-1} + \varepsilon,$$

where ε denotes a noise. Suppose that the true distribution can be realized just with K^* terms:

$$y = \sum_{k=1}^{K^*} w_k^* x^{k-1} + \varepsilon,$$

where $\{w_k^*; 1 \leq k \leq K^*\}$ denotes the set of the true parameters. If we adopt the model with $K < K^*$, we expect poor generalization performance because of lack of expression ability. On the other hand, if we adopt the model with $K > K^*$, we also expect no good generalization performance because of the estimation error of the redundant parameters. Accordingly, adopting the model with $K = K^*$ naturally provides the best generalization performance. Model selection is to determine the best model from observation in such a situation.

Since we can ignore the terms independent of the model for model selection, we rewrite the generalization error, Eq.(1.3), as follows:

$$G(X^n, Y^n) = -\int q(x)q(y|x) \log p(y|x, X^n, Y^n) dxdy + \text{const.}. \tag{1.10}$$

Because we do not know the true distribution $q(y|x)$, we cannot, unfortunately, calculate Eq.(1.10). Therefore, we often utilize the training error as an estimator of the generalization error,

$$T(X^n, Y^n) = -n^{-1} \sum_{i=1}^{n} \log p(y|x, X^n, Y^n) + \text{const.}, \tag{1.11}$$

which can be calculated from the observed finite training samples. It might seem that the problem had been solved. However, the training error, Eq.(1.11), is not a good estimator of the generalization error, Eq.(1.10), because of its bias. Actually,

the training error does not reflect the estimation error of the redundant parameters, and is a non-increasing function of the parameter dimension, K. We can expect that correcting the bias provides a better model selection criterion, and find from Eq.(1.9) that, in the asymptotic limit, the average of the bias between the generalization error and the training error is given by

$$\begin{aligned} G_{\text{Regular}}(n) - T_{\text{Regular}}(n) &= \frac{(\lambda_{\text{Regular}} - \nu_{\text{Regular}})}{n} + o(n^{-1}) \\ &= \frac{K}{n} + o(n^{-1}). \end{aligned} \tag{1.12}$$

Based on Eq.(1.12), the model selection method minimizing Akaike's information criterion (AIC) was proposed [Akaike, 1974]:

$$\text{AIC} = -2\sum_{i=1}^{n} \log p(y_i|x_i, X^n, Y^n) + 2K, \tag{1.13}$$

which can be considered to be an unbiased estimator of the generalization error, Eq.(1.10), multiplied by $2n$.[2]

Another information criterion, called Bayesian information criterion (BIC), was proposed based on the framework of the Bayes estimation[3] [Schwarz, 1978]. Let $D^n = (X^n, Y^n)$. The Bayes estimation is based on the following equation, called the Bayes theorem:

$$p(w|D^n) = \frac{p(D^n|w)p(w)}{p(D^n)}, \tag{1.14}$$

where $p(w|D^n)$ is the Bayes posterior distribution of the parameter, $p(D^n|w)$ is the likelihood of the model, $p(w)$ is the prior distribution, and $p(D^n)$ is the normalization factor, given by

$$p(D^n) = \int p(D^n|w)p(w)dw. \tag{1.15}$$

The normalization factor can be regarded as the likelihood of the *ensemble* of the models, and is called the marginal likelihood.[4] Following the concept of *like-*

[2]Note that the minimized AIC, however, is biased, i.e., $\langle \min \text{AIC} \rangle_{q(X^n,Y^n)} \neq \min \langle \text{AIC} \rangle_{q(X^n,Y^n)}$.

[3]The Bayes estimation will be described in detail in Section 2.1.1.

[4]The normalization factor $p(D^n)$ corresponds to the partition function in statistical physics. (See the first paragraph of Section 1.1.5, as well as Section 6.1.4.)

lihood in statistics, we can accept that the *ensemble* of models with the maximum marginal likelihood is the most likely. Thus, the model selection method maximizing the marginal likelihood was proposed [Efron and Morris, 1973; Schwarz, 1978; Akaike, 1980; MacKay, 1992]. Note that this method is appropriate not only in the regular models but also in the singular models.

The free energy, evidence, or stochastic complexity is defined as the negative logarithm of the marginal likelihood:

$$F = -\log p(D^n). \tag{1.16}$$

Consider the free energy subtracted by the entropy, which we call the normalized free energy in this thesis. Commonly, the average normalized free energy

$$F(n) = \langle F(D^n) + \log q(D^n) \rangle_{q(D^n)} \tag{1.17}$$

can be asymptotically expanded as follows:

$$F(n) = \lambda' \log n + o(\log n), \tag{1.18}$$

where the coefficient of the leading term, λ', is called the free energy coefficient in this thesis. In the regular models, the asymptotic normality of the Bayes posterior distribution leads to the following simple relation:

$$2\lambda'_{\text{Regular}} = K, \tag{1.19}$$

and based on which, Bayesian information criterion (BIC) was proposed [Schwarz, 1978]. Interestingly, the minimum description length criterion (MDL), is equivalent to BIC, although it was derived from the context of information theory in communication [Rissanen, 1986]. Thus, BIC, as well as the MDL, is given by

$$\text{BIC} = \text{MDL} = -2\sum_{i=1}^{n} \log p(y_i|x_i, X^n, Y^n) + K \log n. \tag{1.20}$$

The second term of Eq.(1.13), as well as that of Eq.(1.20), is called the penalty term, which penalizes the complexity of the model. AIC, BIC, and the MDL

are easily calculated and very useful in many applications. However, the simple relations, Eqs.(1.9) and (1.19), do not hold in singular models, because of the singularities, as will be explained in Section 1.1.4.

According to recent works, it has been known that the generalization coefficients in singular models depend not only on the parameter dimension, K, but also on the true distribution. So, it seemed to be difficult to propose any information criterion in singular models. Nevertheless, an information criterion, called the singular information criterion (SingIC), has just been proposed rather by utilizing the dependence on the true distribution [Yamazaki *et al.*, 2005]. Clarifying the coefficients, which, of course, is very important to evaluate the performance of a learning machine, is also useful for such kind of work.

1.1.4 Effects of Singularities

Generally, a model with hierarchical parameter structure has singularities. Considering neural networks, for example, we explain the singularities in detail. Let $x \in \mathbb{R}^M$ be an input vector, $y \in \mathbb{R}^N$ an output vector, and w a parameter vector. A neural network model can be described as a parametric family of maps $\{f(\cdot; w) : \mathbb{R}^M \mapsto \mathbb{R}^N\}$. A three-layer neural network with H hidden units is defined by

$$f(x; w) = \sum_{h=1}^{H} b_h \psi \left(a_h^t x\right), \tag{1.21}$$

where $w = \{(a_h, b_h) \in \mathbb{R}^M \times \mathbb{R}^N; h = 1, \ldots, H\}$ summarizes all the parameters, $\psi(\cdot)$ is an activation function, which is usually a bounded, non-decreasing, antisymmetric, nonlinear function like $\tanh(\cdot)$, and ^{t} denotes the transpose of a matrix or vector. The model, Eq.(1.21), is unidentifiable because it is invariant for any a_h if $b_h = 0$, or vice versa. The continuous points denoting the same distribution are called the singularities, because the Fisher information matrix on them is singular. The shadowed locations in Fig. 1.4 indicate the singularities. We can see in Fig. 1.4 that the model denoted by the singularities has a larger number of neighborhoods and a larger state density than any other model denoted by only one point each.

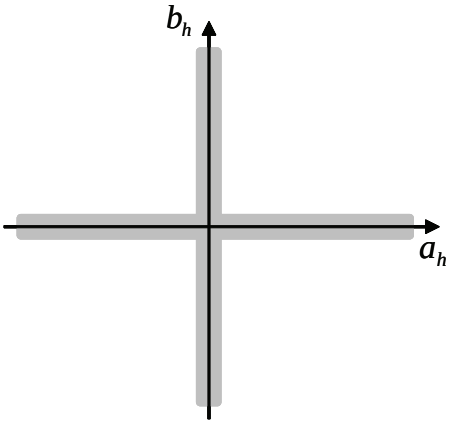

Figure 1.4: Singularities of a neural network model.

When the true model is off the singularities, they asymptotically do not affect prediction, and therefore, the CLT of the regular models holds. However, when the true model is on the singularities, they significantly affect generalization performance as follows:

***ML type effect*:** In the ML type learning methods, increase of the neighborhoods of the true distribution leads to increase of the flexibility of imitating noises, and therefore, accelerates overfitting.

***Bayesian effect*:** In the Bayesian learning methods, the large state density of the true distribution increases its weight, and therefore, suppresses overfitting.

According to the effects above, we can expect the Bayesian learning methods to provide better generalization performance than the ML type learning methods. Remember that we call the former effect the *ML type effect* and the latter one the *Bayesian effect* of singularities in this thesis.

In the recent decades, some of the generalization properties in the ML estimation and in the Bayes estimation have been theoretically clarified. In the ML estimation, the asymptotic behavior of the log-likelihood ratio in some singular models was analyzed [Hartigan, 1985; Bickel and Chernoff, 1993; Takemura and Kuriki, 1997; Kuriki and Takemura, 2001; Fukumizu, 2003], with

facilitation by the idea of the locally conic parameterization [Dacunha-Castelle and Gassiat, 1997]. It has, thus, been known that the ML estimation, in general, provides poor generalization performance, and in the worst cases, the ML estimator diverges. We denote the variables with respect to a learning method in singular models by the ones subscripted with the abbreviation of the method, for example, G_{MLE}, λ_{Bayes}, etc.. Then, we can say that

$$2\lambda_{\text{MLE}} \geq 2\lambda_{\text{Regular}} \ (= K). \tag{1.22}$$

On the other hand, for the analysis of the Bayes estimation in singular models, the algebraic geometrical analysis (AGA) was established [Watanabe, 2001a]. By using the AGA, the asymptotic behavior of the free energies or their upper bounds was clarified in some singular models [Watanabe, 2001a; Yamazaki and Watanabe, 2003a; 2003b; 2003c; Rusakov and Geiger, 2002; Aoyagi and Watanabe, 2005]. In the Bayes estimation, the following relation holds even in singular models:

$$\lambda'_{\text{Bayes}} = \lambda_{\text{Bayes}}, \tag{1.23}$$

which can be proved by using the following well-known relation [Levin *et al.*, 1990]:

$$G_{\text{Bayes}}(n) = F_{\text{Bayes}}(n+1) - F_{\text{Bayes}}(n). \tag{1.24}$$

Therefore, clarifying the asymptotic behavior of the free energy immediately informs us of that of the generalization error. It has, thus, been proved that the generalization error of any singular model is no greater than that of the regular models when we use a prior distribution having positive values on the singularities [Watanabe, 2001b]:

$$2\lambda'_{\text{Bayes}} = 2\lambda_{\text{Bayes}} \leq 2\lambda_{\text{Regular}} \ (= K). \tag{1.25}$$

However, we should note that suppression of overfitting accompanies insensitivity to the true components with small amplitude. There is a trade-off, which would, however, be ignored in asymptotic analysis if we would consider only the

situations when the true model is *distinctly* on or off the singularities. Therefore, we should consider the situations, called *delicate* situations in this thesis, when the KL divergence of the true distribution from the singularities is comparable to n^{-1} [Watanabe and Amari, 2003]. The *delicate* situations are important in model selection problems and in statistical tests with a finite number of training samples for the following reasons: first, that there naturally exist a few true components with amplitude comparable to $n^{-1/2}$ when neither the smallest nor the largest model is selected; and secondly, that whether the selected model involves such components essentially affects generalization performance. In addition, we can design an algorithm, which must not perform well in real situations, with arbitrary small generalization error for redundant components if we ignore the *delicate* situations. In Section 2.5, we will describe the Bayes generalization coefficient in the *delicate* situations, clarified in [Watanabe and Amari, 2003], and will consider the *delicate* situations in our cases in Chapter 5.

1.1.5 Approximation Methods of Bayes Estimation

As discussed in Section 1.1.4, it can be said that, in singular models, the Bayes estimation has the advantage of generalization performance over the ML estimation. However, the Bayes posterior distribution can seldom be exactly realized. Markov chain Monte Carlo (MCMC) methods, originally proposed in the field of statistical physics to obtain samples subject to the Boltzmann distribution, are often used for approximation of the Bayes posterior distribution. The Bayes posterior distribution, Eq.(1.14), can be written in the following form of the Boltzmann distribution:

$$p(w|D^n) = \frac{1}{Z(D^n)} \exp\left(-nE(w; D^n)\right), \tag{1.26}$$

where the factor

$$E(w; D^n) = -\frac{1}{n} \log\left(p(D^n|w)p(w)\right) \tag{1.27}$$

corresponds to the Hamiltonian or energy, the marginal likelihood, $Z(D^n) = p(D^n)$, corresponds to the partition function, and the number of training sam-

ples, n, corresponds to the inverse temperature.[5] The procedure of the Metropolis method, which is typical and the most widely applicable, is the following:

Step 1. Set the initial value $w^{(s)} \in \mathbb{R}^K$, where $s = 0$.

Step 2. Randomly select the candidate of the next sample w', usually among the neighborhoods of $w^{(s)}$, for example,

$$w' = w^{(s)} + \varepsilon_1, \tag{1.28}$$

where ε_1 is a random variable subject to the K-dimensional normal distribution with zero average and sufficiently small variance.

Step 3. Probabilistically select *stay* or *transition*, using another random variable ε_2 subject to the uniform distribution from zero to one, as follows:

$$w^{(s+1)} = \begin{cases} w' & \text{if } \exp\left\{-n\left(E(w';D^n) - E(w^{(s)};D^n)\right)\right\} > \varepsilon_2 \\ w^{(s)} & \text{otherwise} \end{cases}, \tag{1.29}$$

Step 4. Let $s = s + 1$. Repeat Step 2 to Step 4 until we have a sufficient number of samples after the series of the samples seem to be in equilibrium.

The Metropolis method can be applied to arbitrary distribution, however, they require huge computational costs,[6] especially in singular models [Nakano *et al.*, 2005]. As an alternative, the variational Bayes (VB) approach, which provides computational tractability, was proposed [Hinton and van Camp, 1993; MacKay, 1995a; Attias, 1999; Jaakkola and Jordan, 2000; Ghahramani and Beal, 2001]. The original VB approach was proposed for learning of neural networks. Recently, for learning of models with hidden variables, an iterative algorithm, which has a similar procedure to the expectation-maximization (EM) algorithm to provide the ML estimator [Dempster *et al.*, 1977], was derived in the framework of the VB approach [Attias, 1999; Ghahramani and Beal, 2001]. Since then, the

[5] The relation between Bayesian learning and statistical physics is well-known. (See Section 6.1.4.)

[6] Some improved MCMC methods have been proposed [Hukushima and Nemoto, 1996; Iba, 2001; 2005], but their computational costs are still large for some applications.

VB approach has been showing good generalization performance in many applications.

Some properties of the VB iterative algorithm, which are common to those of the EM algorithm, have been discussed [Wang and Titterington, 2004], and it has been shown that the VB algorithm is a type of the natural gradient descent with respect to the free energy [Sato, 2001]. However, its advantage of generalization performance over the EM algorithm when each algorithm converges to its optimum has not been theoretically shown yet. Although the VB free energies or their bounds of some singular models have been derived [Watanabe and Watanabe, 2005; Hosino *et al.*, 2005; Nakano and Watanabe, 2005], they, unfortunately, provide little information on their generalization performance, unlike in the Bayes estimation. It is because the simple relation, which corresponds to Eq.(1.23), between the free energy and the generalization error does not hold in the VB approach. Currently, in any singular model, the VB generalization error has never been theoretically clarified yet. One of the main purposes of this thesis is to clarify the VB generalization error in singular models.

In this thesis, we also consider another alternative, which we call a subspace Bayes (SB) approach. The SB approach is the extension of the empirical Bayes (EB) approach where a part of the parameters of a model are regarded as hyperparameters. If we regard the parameters of one layer as hyperparameters, we can analytically calculate the marginal likelihood in some three-layer models. Consequently, what we have to do is only to find the hyperparameter value maximizing the marginal likelihood. The computational costs of the SB approach is thus much less than those of posterior distribution approximation by MCMC methods. Unfortunately, the SB approach is not widely applicable, however, its analysis also provides an insight into the reason why the VB approach provides good generalization performance, as will be discussed in Chapter 6.

1.2 Purpose

In this thesis, we analyze the generalization properties of approximation methods of the Bayes estimation, the SB and the VB approaches, focusing on the simplest

Table 1.1: Current status in clarifying generalization properties. The symbol $+$ indicates that the coefficients have been clarified in some models, and $*$ indicates that the coefficient has not been clarified yet in any model.

	Regular		Singular			
	ML	Bayes	ML	Bayes	SB	VB
$2\lambda'$	—	$= K$	—	$+$	—	$+$
2λ	$= K$	$= K$	$+$	$= 2\lambda'$	$*$	$*$
-2ν	$= K$	$= K$	$+$	$*$	$*$	$*$

singular models, three-layer linear neural networks (LNNs). Since the generalization properties of LNNs in the standard learning methods, i.e., the ML estimation and the Bayes estimation, have been clarified, we can compare the properties of the approximation methods with those of the standard learning methods. As described in Section 1.1.3, the generalization error, the training error, and the free energy can generally be asymptotically expanded as follows:

$$G(n) = \lambda n^{-1} + o(n^{-1}),$$
$$T(n) = \nu n^{-1} + o(n^{-1}),$$
$$F(n) = \lambda' \log n + o(\log n),$$

and the CLT provides the coefficients of the regular models:

$$2\lambda'_{\text{Regular}} = 2\lambda_{\text{Regular}} = -2\nu_{\text{Regular}} = K,$$

which holds both in the ML estimation and the Bayes estimation. The coefficients of some singular models also have been clarified, as mentioned in Section 1.1.4 as well as in Section 1.1.5. Table 1.1 summarizes the current status in clarifying generalization properties. The symbol $+$ indicates that the coefficients (or their reasonably tight bounds) have been clarified in some singular models, and $*$ indicates that the coefficient has not been clarified yet in any singular model. On the other hand, Table 1.2 summarizes the current status in LNNs. The generalization coefficient in the ML estimation and that in the Bayes estimation have been clarified, which will be described in Section 2.3.2 and in Section 2.3.3, respectively.

Table 1.2: Current status in three-layer linear neural networks.

	Linear neural networks			
	ML	Bayes	SB	VB
$2\lambda'$	—	known	—	being clarified
2λ	known	$= 2\lambda'$	being clarified	being clarified
-2ν	known	unknown	being clarified	being clarified

The VB free energy coefficient, $\lambda'_{\rm VB}$, has been upper bounded in [Nakano and Watanabe, 2005], and we clarify its exact value in this thesis. The main purpose of this thesis is to clarify the generalization and the training coefficients of the SB and the VB approaches, which have never been clarified in any singular models.

Through the analysis, we will consider when and why the SB and the VB approaches provide good generalization performance. A key is the close relation with the James-Stein (JS) estimator [James and Stein, 1961], which was proved to dominate the ML estimator even in regular models. The JS estimator will be introduced in Section 2.4 with the definition of the verb *dominate*. The relation between Bayesian learning methods and the JS estimator was previously discussed: it was shown that the JS estimator can be derived as the solution of the EB approach where the prior distribution of a regular model has a hyperparameter as its variance [Efron and Morris, 1973];[7] it was pointed out that the generalization error of the Bayes estimation in a class of singular models behaves similarly to that of the JS estimator [Watanabe and Amari, 2003]. Discussion of the relation among the singularities, Bayesian learning methods, and the JS type *shrinkage* is another purpose of this thesis. We will also consider the *delicate* situations in Chapter 5, of which importance has been explained in the last paragraph of Section 1.1.4.

The purpose of this thesis is summarized in the following:

1. We will clarify the asymptotic behavior of the generalization properties of the SB and the VB approaches in LNNs, and compare them with those of the ML estimation and the Bayes estimation.

[7] The derivation is described in Appendix A.2. (See also Section 2.4.)

2. We will discuss their similarity to, and difference from, the Bayes estimation.

3. We will discuss their relation to the JS estimator, and consider when and why the SB and the VB approaches provide good generalization performance.

4. We will discuss the domination, considering the *delicate* situations when the true model is near the singularities.

1.3 Overview

This thesis consists of seven chapters and one appendix. Chapter 2 is devoted to the description of the previous works and the notation of this thesis. After that, Chapters 3–5 are devoted to the contributions of this thesis. Finally, Chapters 6 and 7 are devoted to discussion and conclusions, respectively. Some previous works, related with this thesis but a little complicated and not necessarily required to be known for understanding the main part of this thesis, are put into Appendix. They can be quoted only in discussion in Chapter 6.

In Chapter 2, we explain the previous works important and closely related with this thesis, and then summarize the notation for the later chapters. First, we describe the frameworks of the Bayes estimation and its subspecies, the EB approach and the VB approach, in Section 2.1. After that, we introduce the AGA, proposed to clarify the generalization performance of the Bayes estimation, in Section 2.2. In Section 2.3, we introduce linear neural networks, on which we focus in this thesis, and describe their generalization properties in the ML estimation and those in the Bayes estimation, which were previously clarified. In Section 2.4, the JS estimation, closely related with our results, is introduced. In Section 2.5, we explain the necessity of consideration of the *delicate* situations, and describe the clarified results in the Bayes estimation. At the end of Chapter 2, symbols, variables, and the definitions of words are summarized in Section 2.6. Many of them will have already appeared in Chapter 1 or in Chapter 2, and can appear without their definitions in the later chapters.

Chapter 3 is devoted to the analysis of the SB approach, which has been published in [Nakajima and Watanabe, 2005b] and will be published in [Nakajima and Watanabe, 2006b]. In Section 3.1, we first introduce the SB approach, extending the EB approach. Then, in Section 3.2, we derive the posterior distribution and the predictive distribution, and thus prove the asymptotic equivalence between the SB approach and a JS type shrinkage estimation. After that, in Section 3.3, we theoretically clarify its generalization error and training error, and then illustrate the numerical results. We discuss the similarity to, and the difference from, the Bayes estimation, comparing the results of the SB approach with those of the ML estimation and those of the Bayes estimation.

Chapter 4 is devoted to the analysis of the VB approach, which has been published in [Nakajima and Watanabe, 2005a]. The organization of this chapter is similar to that of Chapter 3. First, in Section 4.1, we apply the VB approach to LNNs. In Section 4.2, we deduce the variational condition, and then derive the posterior distribution and the predictive distribution, analytically solving the variational condition. Thus, we conclude that also the VB approach is asymptotically equivalent to a JS type shrinkage estimation, and that the SB and the VB approaches are similar to each other. After that, in Section 4.3, we clarify and illustrate the generalization properties of the VB approach, and discuss its performance.

Chapter 5 is devoted to consideration of the *delicate* situations, which will be published mainly in [Nakajima and Watanabe, 2006b]. Before analysis, we discuss the admissibility of the Bayes estimation, and its consistency with our results in Section 5.1. Then, we prove the theorems providing the generalization and the training errors of the SB and the VB approaches in Section 5.2. After that, we discuss the domination of the SB and the VB approaches over the ML estimation in Section 5.3.

Chapter 6 is devoted to discussion. In Section 6.1, we first discuss the reason of the asymptotic equivalence among the JS estimation, the SB approach, and the VB approach in LNNs. Then, we discuss the relation between the LNN and the automatic relevance determination model (ARD), which has recently been applied to a real application with the VB approach. We also discuss our results from the

viewpoint of statistical physics. In Section 6.2, we discuss the features of the SB and the VB approaches, and make some conjectures for other singular models. In Section 6.3, we make some suggestions on applying Bayesian learning methods to real problems.

Finally, we state conclusions and future work in Chapter 7. In Appendix, we describe the following three previous works: the proof of the superiority of the Bayes estimation in Appendix A.1; the derivation of the JS estimator as an EB estimator in Appendix A.2; and the introduction of the EM algorithm and the VB approach in the normal mixture models in Appendix A.3.

Chapter 2

Preliminaries

In this chapter, we describe the previous works important and closely related with this thesis, and then summarize the notation for the later chapters. In Section 2.1, the frameworks of the Bayes estimation, the empirical Bayes (EB) approach, and the variational Bayes (VB) approach are described. Then, the algebraic geometrical analysis (AGA), proposed to clarify the generalization properties of the Bayes estimation, is explained in Section 2.2. In this thesis, we focus on the simplest singular models, three-layer linear neural networks (LNNs). We introduce them in Section 2.3 with the theoretical results of the generalization properties in the ML estimation and in the Bayes estimation. After that, the James-Stein (JS) estimation and its subspecies, strongly related with the results of this thesis, are introduced in Section 2.4. In Section 2.5, we explain the necessity of consideration of *delicate* situations, and describe the known results in single-output (SO) LNNs in the Bayes estimation. Finally, in Section 2.6, we summarize the notation for the later chapters.

2.1 Bayesian Learning Methods

In this section, we first describe the Bayes estimation, and then the EB approach and the VB approach.

2.1.1 Bayes Estimation

Suppose we use a model $p(y|x,w)$, where x is an input vector, y is an output vector, and w is a parameter vector. Let $X^n = \{x_1, \ldots, x_n\}$ and $Y^n = \{y_1, \ldots, y_n\}$ be arbitrary n training samples independently and identically taken from the true distribution $q(x,y) = q(x)q(y|x)$. The marginal conditional likelihood of the model is given by

$$Z(Y^n|X^n) = \int \phi(w) \prod_{i=1}^{n} p(y_i|x_i,w) dw, \tag{2.1}$$

where $\phi(w)$ is the prior distribution of the parameter. The posterior distribution is given by

$$p(w|X^n,Y^n) = \frac{\phi(w)\prod_{i=1}^n p(y_i|x_i,w)}{Z(Y^n|X^n)}. \tag{2.2}$$

In the Bayes estimation, the predictive distribution is defined as the average of the model over the posterior distribution:

$$p_{\text{Bayes}}(y|x,X^n,Y^n) = \int p(y|x,w)p(w|X^n,Y^n)dw, \tag{2.3}$$

with which we predict the new output from a new input.

In the following, we define the variables characterizing the generalization properties of the Bayes estimation, whose significance has been explained in Section 1.1.3. The generalization error and the training error are defined by

$$G_{\text{Bayes}}(n) = \langle G_{\text{Bayes}}(X^n,Y^n)\rangle_{q(X^n,Y^n)}, \tag{2.4}$$

$$T_{\text{Bayes}}(n) = \langle T_{\text{Bayes}}(X^n,Y^n)\rangle_{q(X^n,Y^n)}, \tag{2.5}$$

respectively, where

$$G_{\text{Bayes}}(X^n,Y^n) = \int q(x)q(y|x)\log\frac{q(y|x)}{p_{\text{Bayes}}(y|x,X^n,Y^n)}dxdy, \tag{2.6}$$

$$T_{\text{Bayes}}(X^n,Y^n) = n^{-1}\sum_{i=1}^{n}\log\frac{q(y|x)}{p_{\text{Bayes}}(y|x,X^n,Y^n)}, \tag{2.7}$$

and $\langle\cdot\rangle_{q(X^n,Y^n)}$ denotes the expectation value over all sets of n training samples. Commonly, the generalization error and the training error can be asymptotically expanded as follows:

$$G_{\text{Bayes}}(n) = \frac{\lambda_{\text{Bayes}}}{n} + o(n^{-1}), \tag{2.8}$$

$$T_{\text{Bayes}}(n) = \frac{\nu_{\text{Bayes}}}{n} + o(n^{-1}). \tag{2.9}$$

The free energy, evidence, or stochastic complexity is defined by

$$F_{\text{Bayes}}(Y^n|X^n) = -\log Z(Y^n|X^n). \tag{2.10}$$

The average normalized free energy, defined by

$$F_{\text{Bayes}}(n) = \langle F_{\text{Bayes}}(Y^n|X^n) + \log q(Y^n|X^n)\rangle_{q(X^n,Y^n)}, \tag{2.11}$$

can be asymptotically expanded as follows:

$$F_{\text{Bayes}}(n) = \lambda'_{\text{Bayes}} \log n + o(\log n). \tag{2.12}$$

In the (rigorous) Bayes estimation, we can prove that the relation

$$\lambda'_{\text{Bayes}} = \lambda_{\text{Bayes}} \tag{2.13}$$

still holds even in singular models, by using the following well-known relation [Levin *et al.*, 1990]:[1]

$$G_{\text{Bayes}}(n) = F_{\text{Bayes}}(n+1) - F_{\text{Bayes}}(n). \tag{2.14}$$

Therefore, clarifying the Bayes free energy immediately informs us of the asymptotic behavior of the Bayes generalization error. By using the AGA, which will be introduced in Section 2.2, the generalization properties in some singular models have been clarified, and moreover, it has been proved that the following inequality holds for any singular model when we use a prior distribution having positive values on the singularities [Watanabe, 2001b]:

$$2\lambda'_{\text{Bayes}} = 2\lambda_{\text{Bayes}} \leq 2\lambda_{\text{Regular}} \ (= K), \tag{2.15}$$

which is one of the most preferable properties of the Bayes estimation.

[1] This relation can be easily derived. (See the footnote in Appendix A.1.)

Remark 1 The maximum a posterior (MAP) estimation essentially differs from the Bayes estimation. The MAP predictive distribution is denoted by *one* model as follows:

$$p_{\rm MAP}(y|x, X^n, Y^n) = p(y|x; \hat{w}_{\rm MAP}), \tag{2.16}$$

where

$$\hat{w}_{\rm MAP} = \mathop{\rm argmax}_{w} \left(\phi(w) \prod_{i=1}^{n} p(y_i|x_i; w) \right) \tag{2.17}$$

is the MAP estimator. Therefore, the MAP estimation is classified into the ML type. In addition, the prior distribution, $\phi(w)$, asymptotically does not affect the MAP estimator and hence the MAP predictive distribution, as far as we use an ordinary and constant prior distribution. Therefore, the MAP estimation is asymptotically equivalent to the ML estimation, defined by Eqs. (1.1) and (1.2).

Superiority of Bayes Estimation

The Bayes estimation is said to be superior to any other learning method in a certain meaning. Consider the case that we apply a learning method to many applications, where the true parameter w^* is different in each application and subject to $q(w^*)$. Note that the true distribution in each application is written as $q(y|x) = p(y|x, w^*)$. Although we have been abbreviating it, the generalization error, Eq.(1.5), naturally depends on the true distribution, so we here denote the dependence explicitly as follows: $G(n; w^*)$. Let

$$\bar{G}(n) = \langle G(n; w^*) \rangle_{q(w^*)} \tag{2.18}$$

be the average generalization error over $q(w^*)$.

Proposition 1 *When we know and use the true prior distribution, $q(w^*)$, the Bayes estimation minimizes the average generalization error over $q(w^*)$, i.e.,*

$$\bar{G}_{Bayes}(n) \leq \bar{G}_{Other}(n), \tag{2.19}$$

where $\bar{G}_{Other}(n)$ denotes the average generalization error of an arbitrary learning method.

(The proof is given in Appendix A.1.)

Remark 2 However, we consider the situations where we need model selection, i.e., we do not know even the true dimension of the parameter space. Accordingly, it can happen that an approximation method of the Bayes estimation provides better generalization performance than the Bayes estimation, as will be shown in Chapters 3 and 4.

2.1.2 Empirical Bayes Approach

We often have little information about the prior distribution, with which the empirical Bayes (EB) approach was originally proposed to cope. We can introduce hyperparameters in the prior distribution; for example, when we use a prior distribution that depends on a hyperparameter τ such as

$$\phi(w\|\tau) = \frac{1}{(2\pi\tau^2)^{K/2}} \exp\left(-\frac{\|w\|^2}{2\tau^2}\right), \tag{2.20}$$

where we distinguish the hyperparameter from the parameter by $\|$. Then, the marginal likelihood, Eq.(2.1), also depends on τ as follows:

$$Z(Y^n|X^n\|\tau) = \int \phi(w\|\tau) \prod_{i=1}^{n} p(y_i|x_i, w) dw. \tag{2.21}$$

In the EB approach, τ is estimated by maximizing the marginal likelihood, i.e.,

$$\hat{\tau} = \operatorname*{argmax}_{\tau} Z(Y^n|X^n\|\tau), \tag{2.22}$$

or by a slightly different way [Efron and Morris, 1973; Akaike, 1980; Kass and Steffey, 1989]. After the hyperparameter value is estimated, the posterior distribution and the predictive distribution are calculated in the same way as in the Bayes estimation.

As mentioned in Remark 1 in Section 2.1.1, the prior distribution does not affect asymptotic generalization performance, as far as we use a constant prior, i.e., which does not depend on the training data. However, the generalization performance of the EB approach significantly differs from that of the ML estimation.

The EB approach is also known as the hierarchical Bayesian learning. In fact, that kind of hierarchy is equivalent to the hierarchy of the parameters in singular models, and therefore, the EB approach in a regular model behaves as a singular model, which will be discussed in detail in Section 6.1.1.

It was found that the EB approach is closely related with the James-Stein (JS) estimator, which will be introduced in Section 2.4. In Appendix A.2, we show that the JS estimator can be derived as an EB estimator in a simple model [Efron and Morris, 1973]. In Chapter 3, we introduce the subspace Bayes (SB) approach, extending the EB approach, and clarify its generalization error and training error in LNNs. That analysis reveals the relation between the singularities of models and the JS type *shrinkage*.

2.1.3 Variational Bayes Approach

For an arbitrary trial posterior distribution $r(w|X^n, Y^n)$, we define the generalized free energy by

$$\bar{F}(Y^n|X^n) = \left\langle \log \frac{r(w|X^n, Y^n)}{p(Y^n|X^n, w)\phi(w)} \right\rangle_{r(w|X^n,Y^n)}, \tag{2.23}$$

where $\langle\cdot\rangle_p$ denotes the expectation value over a distribution p.[2] Jensen's inequality guarantees that the generalized free energy, Eq.(2.23), gives the upper bound of the Bayes free energy, Eq.(2.10), as follows:

$$\begin{aligned} F_{\text{Bayes}}(Y^n|X^n) &= -\log \int r(w|X^n, Y^n) \frac{p(Y^n|X^n, w)\phi(w)}{r(w|X^n, Y^n)} dw \\ &\leq -\int r(w|X^n, Y^n) \log \frac{p(Y^n|X^n, w)\phi(w)}{r(w|X^n, Y^n)} dw \\ &= \bar{F}(Y^n|X^n), \end{aligned} \tag{2.24}$$

and that the equality holds when and only when the posterior distribution is equal to the Bayes posterior distribution:

$$r_{\text{Bayes}}(w|X^n, Y^n) = \frac{p(Y^n|X^n, w)\phi(w)}{Z(Y^n|X^n)}.$$

[2]The generalized free energy, Eq.(2.23), is the functional of $r(w|X^n, Y^n)$ corresponding to the Helmholtz free energy in statistical physics. (See Section 6.1.4.)

In the VB approach, restricting the set of possible distributions for $r(w|X^n, Y^n)$, we minimize the generalized free energy, Eq.(2.23). The optimum of the trial distribution, which we denote by $\hat{r}(w|X^n, Y^n)$, is regarded as the posterior distribution in the VB approach. In addition, the minimum of the generalized free energy is called the VB free energy,[3] and denoted by $F_{\rm VB}(Y^n|X^n)$ in later chapters.

In the original VB approach, proposed for learning of neural networks, the posterior distribution is restricted to the normal distribution [Hinton and van Camp, 1993]. Recently, an iterative algorithm with good tractability has been proposed for models with hidden variables, such as mixture models, graphical models, etc., by using an appropriate class of prior distributions and restricting the set of possible posterior distributions such that the parameters and the hidden variables are independent of each other [Attias, 1999; Ghahramani and Beal, 2001]. It has a similar procedure to the expectation-maximization (EM) algorithm, which had been proposed to obtain the ML estimator [Dempster *et al.*, 1977].

In many applications, the VB approach experimentally shows significantly better generalization performance than the EM algorithm. It shows a model selecting effect, i.e., Occam's razor [MacKay, 1995b], without any intentional penalty with respect to the degree of freedom of the model. Since its proposal, the VB approach with good tractability and good generalization performance has been attracting people's attention. In Appendix A.3, we describe the derivation of the EM algorithm and the VB approach in the normal mixture models, the simplest models with hidden variables. The VB approach is also applied to the automatic relevance determination model (ARD) [MacKay, 1994; Neal, 1996], which will be introduced in Section 6.1.3, in a real application of brain current estimation [Sato *et al.*, 2004]. Actually, the ARD has a similar structure to the LNN, and therefore, the approach in [Sato *et al.*, 2004] is closely related with the results of this thesis. Further discussion will be in Section 6.1.3.

Some properties of the VB iterative algorithm, which are common to those of the EM algorithm, have been discussed [Wang and Titterington, 2004], and it has been shown that the VB algorithm is a type of the natural gradient descent

[3] Also $\bar{F}(Y^n|X^n)$ sometimes is called the VB free energy.

with respect to the generalized free energy, Eq.(2.23) [Sato, 2001]. However, its advantage of generalization performance over the EM algorithm when each algorithm converges to its optimum has not been theoretically shown yet. The bounds of the VB free energy of the normal mixture models have recently been derived [Watanabe and Watanabe, 2004]. Subsequently, the bounds of the VB free energies in other singular models, such as mixture models of exponential families, hidden Markov models, neural networks, etc., have been derived [Watanabe and Watanabe, 2005; Hosino *et al.*, 2005; Nakano and Watanabe, 2005]. However, the VB generalization errors of those models are still unknown, since the simple relation corresponding to Eq.(2.13) does not hold in the VB approach.

In Chapter 4, we apply the VB approach to LNNs, by restricting the posterior distribution such that the parameters of different layers, as well as those of different components, are independent of each other. Then, we derive the VB solution, analytically solving the variational condition. After that, we clarify the generalization error, the training error, and the free energy of the VB approach, and discuss the reason why the VB approach provides good generalization performance. Note that the restriction applied to the VB posterior distribution in models with hidden variables is as strong as, and hence results in, the independence between the parameters of different layers. Moreover, the independence between the parameters of different components is also guaranteed by introducing hidden variables. (See Remark 3 in Appendix A.3.) Therefore, the analysis of the VB approach in LNNs provides an insight into the properties of the VB approach in general singular models, which will be discussed in Section 6.2.2.

2.2 Algebraic Geometrical Analysis

The singularities had been preventing people from clarifying the generalization properties of singular models long time. Recently, the algebraic geometrical analysis (AGA) has been established [Watanabe, 2001a; 2001c], and subsequently, the generalization coefficients or its upper bounds have been clarified in some singular models [Yamazaki and Watanabe, 2003a; 2003b; 2003c; 2004; Rusakov and Geiger, 2002; Aoyagi and Watanabe, 2005]. In this section, we

describe the outline of the AGA, which was used for clarifying the Bayes generalization error of LNNs [Aoyagi and Watanabe, 2005], described in Section 2.3.3.

The goal is to obtain the asymptotic expansion of the average normalized free energy, Eq.(2.11), which informs us of the generalization coefficient through the relation Eq.(2.13). The free energy is bounded by

$$\tilde{F}(n) = -\log \int \exp(-nE(w))\phi(w)dw, \tag{2.25}$$

where

$$E(w) = \int q(x)q(y|x)\log\frac{q(y|x)}{p(y|x,w)}dxdy \tag{2.26}$$

is the KL divergence from the true distribution $q(x)q(y|x)$ to the model $q(x)p(y|x,w)$. Equation (2.25) involves an integral in multi-dimensional space, which is difficult to be evaluated because of the singularities.

As mentioned in Section 1.1.4. the *Bayesian effect* of the singularities comes from the large state density of the model denoted by the singularities. Actually, if we have the state density function of the energy value $s > 0$, i.e.,

$$v(s) = \int \delta(s - E(w))\phi(w)dw, \tag{2.27}$$

where $\delta(\cdot)$ denotes the Dirac delta function, the Laplace transform of $v(s)$ gives the free energy, Eq.(2.25), as follows:

$$\tilde{F}(n) = -\log \int \exp(-s)v\left(\frac{s}{n}\right)\frac{ds}{n}. \tag{2.28}$$

Consider the Mellin transform, an extension of the Laplace transform, of Eq.(2.27):

$$\zeta(z) = \int s^z v(s)ds = \int E(w)^z\phi(w)dw, \tag{2.29}$$

which is a function of a complex number z, and called the zeta function of $E(w)$ and $\phi(w)$. The zeta function, $\zeta(z)$, is a holomorphic function in the region where $\mathrm{Re}(z) > 0$, and can be analytically continued to the meromorphic function on the entire complex plane. It was proved that all the poles of $\zeta(z)$ are real, negative,

and rational numbers. So, let $0 > -\lambda_1 > -\lambda_2 > \cdots$ be the sequence of the poles in decreasing order, and $m_1, m_2, \ldots$ the corresponding orders of the poles.

The inverse Mellin transform of $\zeta(z)$ gives the state density, Eq.(2.27):

$$v(s) = \frac{1}{2\pi i} \int_{u-\infty i}^{u+\infty i} s^{-z-1} \zeta(z) dz, \tag{2.30}$$

where $u > 0$ is a real number, and i denotes the imaginary unit. By virtue of the Cauchy residue theorem, we can move the integral path such that $u \to -\infty$, counting the residues of the poles beyond which the path goes. Consequently, the asymptotic expansion of Eq.(2.30) for $s \to 0$ is given by the sum of the residues:

$$v(s) = \sum_{k=1}^{\infty} \sum_{m=1}^{m_k} c_{km} s^{\lambda_k - 1} (-\log s)^{m-1}. \tag{2.31}$$

Substituting Eq.(2.31) into Eq.(2.28), we obtain the asymptotic expansion of Eq.(2.25) as follows:

$$\tilde{F}(n) = \lambda_1 \log n - (m_1 - 1) \log \log n + O(1). \tag{2.32}$$

Since the first two leading terms of $F_{\text{Bayes}}(n)$ are equal to those of $\tilde{F}(n)$, we have

$$F_{\text{Bayes}}(n) = \lambda_1 \log n - (m_1 - 1) \log \log n + O(1). \tag{2.33}$$

By using the relation Eq.(2.14), we have the asymptotic expansion of the generalization error:

$$G_{\text{Bayes}}(n) = \frac{\lambda_1}{n} - \frac{m_1 - 1}{n \log n} + o\left(\frac{1}{n \log n}\right). \tag{2.34}$$

To sum up, finding the maximum pole of the zeta function $\zeta(z)$ gives the Bayes free energy coefficient, which is equal to the Bayes generalization coefficient. Finding poles needs a technique in algebraic geometry, called the resolution of singularities, and checking all the poles is not easy. However, any pole gives the upper bound of the coefficients. In some singular models, the upper bounds have been derived, and fortunately in LNNs, the maximum pole has been found [Aoyagi and Watanabe, 2005], which will be described in Section 2.3.3.

2.3 Linear Neural Networks

In this thesis, we focus on linear neural networks (LNNs), which are the simplest singular models. In the LNNs, the generalization coefficients both in the ML estimation and in the Bayes estimation have already been clarified. Therefore, we can compare our results with them, and discuss the properties of the Bayesian approximation methods. In Section 2.3.1, we first introduce the LNNs. Because an LNN can be embedded in a finite dimensional regular model, the *ML type effect* of the singularities, named in Section 1.1.4, is relatively soft, and its ML estimator is in a finite region, as shown below. However, the generalization error of the ML estimation was proved to be greater than that of the regular models whose dimension of the parameter space is the same when the model is redundant to learn the true distribution [Fukumizu, 1999], which is described in Section 2.3.2. On the other hand, the generalization error of the Bayes estimation was proved to be less than that of the regular models [Aoyagi and Watanabe, 2005], which is descried in Section 2.3.3.

2.3.1 Definition

Let $x \in \mathbb{R}^M$ be an input vector, $y \in \mathbb{R}^N$ an output vector, and w a parameter vector. A neural network model can be described as a parametric family of maps $\{f(\cdot; w) : \mathbb{R}^M \mapsto \mathbb{R}^N\}$. A three-layer neural network with H hidden units is defined by

$$f(x; w) = \sum_{h=1}^{H} b_h \psi\left(a_h^t x\right), \tag{2.35}$$

where $w = \{(a_h, b_h) \in \mathbb{R}^M \times \mathbb{R}^N; h = 1, \ldots, H\}$ summarizes all the parameters, $\psi(\cdot)$ is an activation function, which is usually a bounded, non-decreasing, antisymmetric, nonlinear function like $\tanh(\cdot)$, and ^{t} denotes the transpose of a matrix or vector. We denote by $\mathcal{N}_d(\mu, \Sigma)$ the d-dimensional normal distribution with average vector μ and covariance matrix Σ, and by $\mathcal{N}_d(\cdot; \mu, \Sigma)$ its density function. Assume that the output is observed with a noise subject to $\mathcal{N}_N(0, \Sigma)$.

Then, the conditional distribution is given by

$$p(y|x,w) = \mathcal{N}_N(y; f(x;w), \Sigma). \tag{2.36}$$

In this thesis, we focus on linear neural networks, whose activation functions are linear.

A linear neural network model (LNN) is defined by

$$f(x;A,B) = \sum_{h=1}^{H} b_h a_h^t x = BAx, \tag{2.37}$$

where

$$A = (a_1, \ldots, a_H)^t$$

is an $H \times M$ input parameter matrix, and

$$B = (b_1, \ldots, b_H)$$

is an $N \times H$ output parameter matrix. Because the transform $(A,B) \mapsto (TA, BT^{-1})$ does not change the map for any non-singular $H \times H$ matrix T, the parameterization in Eq.(2.37) has trivial redundancy. Accordingly, the *essential* dimension of the parameter space is given by

$$K = H(M+N) - H^2. \tag{2.38}$$

Let

$$L = \max(M,N), \tag{2.39}$$

$$l = \min(M,N). \tag{2.40}$$

We assume that $H \leq l$ throughout this thesis, since, at most, only l hidden units contribute to learning even when $H > l$. An LNN is also known as a reduced-rank regression model, and not a toy but a useful model in some applications [Reinsel and Velu, 1998]. It is used for multivariate factor analysis when the relevant dimension of the factors is considered to be less than l.

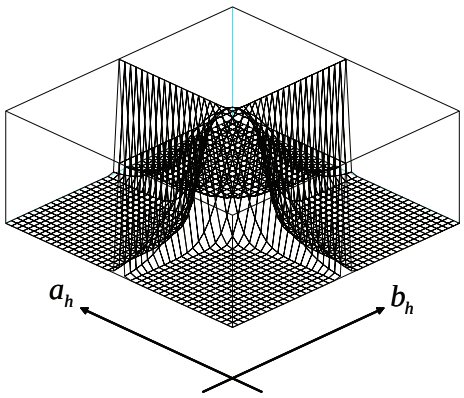

Figure 2.1: Likelihood function of an LNN.

When the map BA is full-rank, i.e., $H = l$, the region of the singularities coincides with that of the trivial degeneracy, and the transform $BA \to W$ makes the LNN identifiable. Therefore, when $H = l$, the LNN is regarded as a regular model from the viewpoint of the ML estimation. However, when $0 < H < l$, the LNN is classified as an *essentially* singular model, because no transform makes it identifiable. Although LNNs are simple, we expect that some phenomena caused by the singularities, and revealed in this thesis, would be observed also in other singular models, as will be discussed in 6.2.2. In addition, even when $H = l$, the LNN should be regarded as a singular model from the viewpoint of Bayesian learning methods, since also the singularities formed by the trivial degeneracy cause the *Bayesian effect*, although they do not cause the *ML type effect*. Consequently, in Bayesian learning methods, the generalization performance of the full-rank LNN is different from that of the regular models.

Figure 2.1 shows the likelihood function of an LNN when the true distribution is on the singularities, i.e., the h-th component is redundant, and the number of training samples, n, is sufficiently large. We can see in Fig. 2.1 that the likelihood function cannot be approximated by the normal distribution even in the asymptotic limit.

2.3.2 Maximum Likelihood Estimation

When $H = l$, the transform $W = BA$ makes the LNN identifiable, as mentioned in Section 2.3.1. In that case, we easily obtain the ML estimator as follows:

$$\hat{W}_{\text{MLE}} = R(X^n, Y^n)Q(X^n)^{-1}, \tag{2.41}$$

where

$$Q(X^n) = n^{-1}\sum_{i=1}^{n} x_i x_i^t \tag{2.42}$$

is an $M \times M$ symmetric matrix, which corresponds to the empirical Fisher information matrix when we regard W, instead of the set of A and B, as the parameter, and

$$R(X^n, Y^n) = n^{-1}\sum_{i=1}^{n} y_i x_i^t \tag{2.43}$$

is an $N \times M$ matrix. However, when $H < l$, the map should be restricted in lower rank, which makes the model *essentially* singular. Hereafter, we abbreviate $Q(X^n)$ as Q, and $R(X^n, Y^n)$ as R. Let γ_h be the h-th largest singular value of the matrix $RQ^{-1/2}$, ω_{a_h} the corresponding right singular vector, and ω_{b_h} the corresponding left singular vector, where $1 \leq h \leq H$. The ML estimator is given as follows:

Proposition 2 *[Baldi and Hornik, 1995] The ML estimator of an LNN is given by*

$$(\hat{B}\hat{A})_{MLE} = \sum_{h=1}^{H} \omega_{b_h}\omega_{b_h}^t RQ^{-1}. \tag{2.44}$$

Using Proposition 2, the generalization error in the ML estimation was previously clarified. Assume that the true distribution is given by $p(y|x, A^*, B^*)$, where B^*A^* is the true map with rank $H^* \leq H$. We denote by $\mathcal{W}_d(m, \Sigma)$ the d-dimensional central Wishart distribution with m degrees of freedom and scale matrix Σ.

Proposition 3 *[Fukumizu, 1999] The generalization error of an LNN in the ML estimation can be asymptotically expanded as*

$$G_{MLE}(n) = \frac{\lambda_{MLE}}{n} + O(n^{-3/2}),$$

where the generalization coefficient is given by

$$2\lambda_{MLE} = (H^*(L+l) - H^{*2}) + \left\langle \sum_{h=1}^{H-H^*} \gamma_h'^2 \right\rangle_{q(\{\gamma_h'^2\})}. \tag{2.45}$$

Here, $\gamma_h'^2$ is the h-th largest eigenvalue of a random matrix subject to $\mathcal{W}_{l-H^*}(L-H^*, I_{l-H^*})$, over which $\langle\cdot\rangle_{q(\{\gamma_h'^2\})}$ denotes the expectation value.

The second term of Eq.(2.45) can be analytically calculated in the large scale limit when M, N, H, and H^* go to infinity in the same order. We define the following scalars:

$$\alpha = (l - H^*)/(L - H^*), \tag{2.46}$$

$$\beta = (H - H^*)/(l - H^*). \tag{2.47}$$

Proposition 4 *[Fukumizu, 1999] The ML generalization coefficient of an LNN in the large scale limit is given by*

$$2\lambda_{MLE} \sim (H^*(L+l) - H^{*2}) + \frac{(L-H^*)(l-H^*)}{2\pi\alpha} J(s_t; 1), \tag{2.48}$$

where

$$J(s;1) = 2\alpha(-s\sqrt{1-s^2} + \cos^{-1} s),$$

and

$$s_t = \max\left(\frac{\kappa - (1+\alpha)}{2\sqrt{\alpha}}, J^{-1}(2\pi\alpha\beta; 0)\right).$$

Here $J^{-1}(\cdot;k)$ denotes the inverse function of $J(s;k)$.

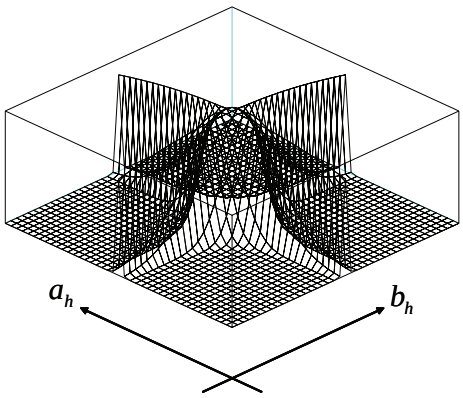

Figure 2.2: Bayes posterior distribution of an LNN.

2.3.3 Bayes Estimation

Figure 2.2 shows the Bayes posterior distribution of an LNN when the h-th component is redundant and n is sufficiently large. We can see in Fig. 2.2 that also the Bayes posterior distribution cannot be approximated by the normal distribution even in the asymptotic limit. In singular models, we cannot rigorously perform the Bayes estimation, nevertheless, its generalization coefficients or their upper bounds can be derived by using the AGA, as described in Section 2.2. Fortunately, the generalization coefficients in LNNs have been exactly clarified:

Proposition 5 *[Aoyagi and Watanabe, 2005] The generalization error of an LNN in the Bayes estimation can be asymptotically expanded as*

$$G_{Bayes}(n) = \frac{\lambda_{Bayes}}{n} - \frac{m-1}{n \log n} + o\left(\frac{1}{n \log n}\right),$$

where the generalization coefficient, as well as the coefficient of the second leading term, is given as follows:

1. *When* $N + H^* \leq M + H$, $M + H^* \leq N + H$, *and* $H^* + H \leq M + N$,

(a) If $M + H + N + H^*$ *is even, then* $m = 1$ *and*

$$\lambda_{Bayes} = \frac{-(H^* + H)^2 - M^2 - N^2}{8} + \frac{2(H^* + H)M + 2(H^* + H)N + 2MN}{8}. \quad (2.49)$$

(b) If $M + H + N + H^*$ *is odd, then* $m = 2$ *and*

$$\lambda_{Bayes} = \frac{-(H^* + H)^2 - M^2 - N^2}{8} + \frac{2(H^* + H)M + 2(H^* + H)N + 2MN + 1}{8}. \quad (2.50)$$

2. *When* $M + H < N + H^*$, *then* $m = 1$ *and*

$$\lambda_{Bayes} = \frac{HM - HH^* + NH^*}{2}. \quad (2.51)$$

3. *When* $N + H < M + H^*$, *then* $m = 1$ *and*

$$\lambda_{Bayes} = \frac{HN - HH^* + MH^*}{2}. \quad (2.52)$$

4. *When* $M + N < H + H^*$, *then* $m = 1$ *and*

$$\lambda_{Bayes} = \frac{MN}{2}. \quad (2.53)$$

Remember that the proposition above comes from the asymptotic expansion of the Bayes free energy:

$$F_{\text{Bayes}} = \lambda'_{\text{Bayes}} \log n - (m - 1) \log \log n + O(1),$$

where

$$\lambda'_{\text{Bayes}} = \lambda_{\text{Bayes}}. \quad (2.54)$$

2.4 James-Stein Estimation

In this section, considering the multi-dimensional mean estimation, a simple regular model, we introduce the James-Stein (JS) estimation, also known as the shrinkage estimation, and its subspecies [James and Stein, 1961; Efron and Morris, 1973; Kubokawa, 2004]. First of all, we notify that the predictive distribution of the JS estimation is given by the *one* model denoted by the JS estimator, and hence the JS estimation should be classified into the ML type; however, it behaves differently from the other ML type methods. So, we regard the JS estimation as an exception. The JS estimation is closely related with this thesis, because we will prove in Chapters 3 and 4 that the SB and the VB approaches are asymptotically equivalent to its subspecies. The relation between the EB approach and the JS estimation was discussed in [Efron and Morris, 1973], where the JS estimator was derived as an EB estimator. The derivation is put into Appendix A.2.

Suppose that $X^n = \{x_1, \ldots, x_n\}$ are arbitrary n training samples independently and identically taken from

$$p(x|\mu) = \mathcal{N}_M(x; \mu, I_M), \tag{2.55}$$

where μ is the mean parameter to be estimated, and I_d denotes the $d \times d$ identity matrix. We denote by $*$ the true value of a parameter and by a *hat* an estimator of a parameter.

The generalization error, defined in Eq.(1.5), of an estimator $\hat{\mu}$ is given as the mean squared error in this case:

$$G_{\hat{\mu}}(\mu^*) = \frac{1}{2}\langle \|\hat{\mu} - \mu^*\|^2 \rangle_{q(X^n)},$$

and also called the risk function. We define two words, *dominate* and *admissible*, in the following:

Definition 1 *We say that an estimator $\hat{\mu}_\alpha$ dominates another estimator $\hat{\mu}_\beta$ if*

$$G_{\hat{\mu}_\alpha}(\mu^*) \leq G_{\hat{\mu}_\beta}(\mu^*) \qquad \textit{for arbitrary } \mu^*,$$
$$\textit{and} \qquad G_{\hat{\mu}_\alpha}(\mu^*) < G_{\hat{\mu}_\beta}(\mu^*) \qquad \textit{for, at least, a certain } \mu^*.$$

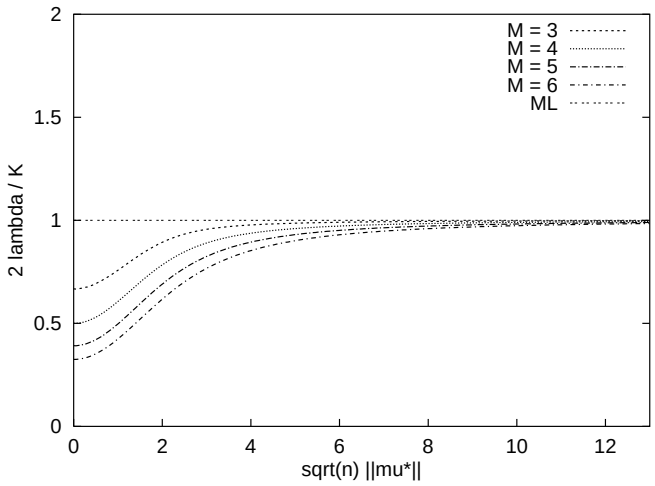

Figure 2.3: Generalization error of James-Stein estimator.

Definition 2 *We say that an estimator is admissible if no estimator dominates it.*

The usual ML estimator,

$$\hat{\mu}_{MLE} = n^{-1} \sum_{i=1}^{n} x_i, \tag{2.56}$$

is, naturally, an efficient estimator, i.e., which is never dominated by any unbiased estimator. However, the ML estimator has, surprisingly, been proved to be inadmissible when $M \geq 3$ [Stein, 1956]. The JS estimator, defined as follows, was, subsequently, introduced as an estimator dominating the ML estimator, Eq.(2.56) [James and Stein, 1961]:

$$\hat{\mu}_{JS} = \left(1 - \frac{\chi}{n\|\hat{\mu}_{MLE}\|^2}\right) \hat{\mu}_{MLE}, \tag{2.57}$$

where $\chi = (M - 2)$ is called the degree of shrinkage in this thesis.

Figure 2.3 shows the generalization error of the JS estimator. The horizontal axis indicates $\sqrt{n}\|\mu^*\|$. The scaling factor $\sqrt{n}$ is important to discuss the domination, as will be discussed in Section 2.5. The vertical axis indicates the normalized generalization error, $2\lambda(\mu^*)/K = 2nG(\mu^*)/K$, where $K = M$ is the parameter

dimension. The normalized generalization error of the ML estimator is always equal to one. We see in Fig. 2.3 that the JS estimator dominates the ML estimator when $K \geq 3$. Note that, even in the asymptotic limit, the shrinkage estimators do not converge to the ML estimator if $\mu^* = 0$, although they converge to the ML estimator otherwise.

The estimators expressed by Eq.(2.57) with arbitrary $\chi > 0$ are called the JS type shrinkage estimators. Furthermore, we can easily find that a positive-part JS type shrinkage estimator, defined by

$$\begin{aligned}\hat{\mu}_{PJS} &= \theta\left(n\|\hat{\mu}_{MLE}\|^2 > \chi\right)\left(1 - \frac{\chi}{n\|\hat{\mu}_{MLE}\|^2}\right)\hat{\mu}_{MLE} \\ &= \left(1 - \frac{\chi}{\max(\chi, n\|\hat{\mu}_{MLE}\|^2)}\right)\hat{\mu}_{MLE}, \end{aligned} \tag{2.58}$$

dominates the non-thresholding JS type shrinkage estimator, Eq.(2.57), with the same degree of shrinkage, χ. Here, $\theta(\cdot)$ is the indicator function of an event, defined by

$$\theta(\textit{event}) = \begin{cases} 1 & \text{if } \textit{event} \text{ is true} \\ 0 & \text{otherwise} \end{cases}. \tag{2.59}$$

2.5 *Delicate* Situations

Increasing the degree of shrinkage, $\chi < \infty$, of the positive-part JS type estimator, Eq.(2.58), we can have an estimator with arbitrary small generalization error when $\mu^* = 0$, and the same generalization error as that of the ML estimator when $\mu^* = O(1)$ in the asymptotic limit. The case that $\mu^* = 0$ and the case that $\mu^* = O(1)$ corresponds to $\sqrt{n}\|\mu^*\| = 0$ and $\sqrt{n}\|\mu^*\| \to \infty$ in Fig. 2.3. Naturally, such estimators will not perform well, because they can provide extremely worse generalization performance in the region where $0 < \sqrt{n}\|\mu^*\| < \infty$. We have to balance between the generalization performance in the region where $\|\mu^*\| = 0$ and that in the region where $0 < \sqrt{n}\|\mu^*\| < \infty$. This is the trade-off, mentioned in the last paragraph of Section 1.1.4, between suppression of overfitting and insensitivity to the true components with small amplitude. Therefore, consideration

of the cases that $\|\mu^*\| = O(n^{-1/2})$, called *delicate* situations in this thesis, is important.

However, in usual asymptotic analysis, including the discussion in Section 2.3, in Chapter 3, and in Chapter 4, we ignore the *delicate* situations. In [Watanabe and Amari, 2003], the importance of the *delicate* situations was pointed out and the Bayes generalization and training coefficients of the single-output (SO), i.e., $N = H = 1$, LNN were clarified. In this thesis, we will also consider the *delicate* situations in Chapter 5.

In the rest of this section, we describe the main result of [Watanabe and Amari, 2003]. An SOLNN is defined by

$$f(x; a, b) = ba^t x, \tag{2.60}$$

where $a \in \mathbb{R}^M$ and $b \in \mathbb{R}$ are the parameters.

Proposition 6 *The Bayes generalization error of an SOLNN with $M \geq 2$ input units in the general situations when the true map may be delicate such that $0 < \sqrt{n}b^*\|a^*\| < \infty$ can be asymptotically expanded as*

$$G_{Bayes}(n) = \frac{\lambda_{Bayes}}{n} + o(n^{-1}),$$

where the generalization coefficient is given by

$$2\lambda_{Bayes} = 1 + \left\langle \left((\sqrt{n}b^*\|a^*\|)^2 + \sqrt{n}b^* a^{*t} g\right) \frac{Y_M(g)}{Y_{M-2}(g)} \right\rangle_{q(g)}. \tag{2.61}$$

Here,

$$Y_M(g) = \int_0^{\pi/2} d\theta \sin^M \theta \exp\left(-\frac{1}{2}\|\sqrt{n}b^* a^{*t} + g\|^2 \sin^2 \theta\right),$$

and g is a random variable subject to $\mathcal{N}_1(0, 1)$, over which $\langle\cdot\rangle_{q(g)}$ denotes the distribution.

Figure 2.4 shows the Bayes generalization coefficients of the SOLNNs with $M = 2, \ldots, 6$ input units. The horizontal axis indicates $\sqrt{n}\gamma^*$, where $\gamma^* = b^*\|a^*\|$. The SOLNN has a full-rank map and hence is regular from the viewpoint

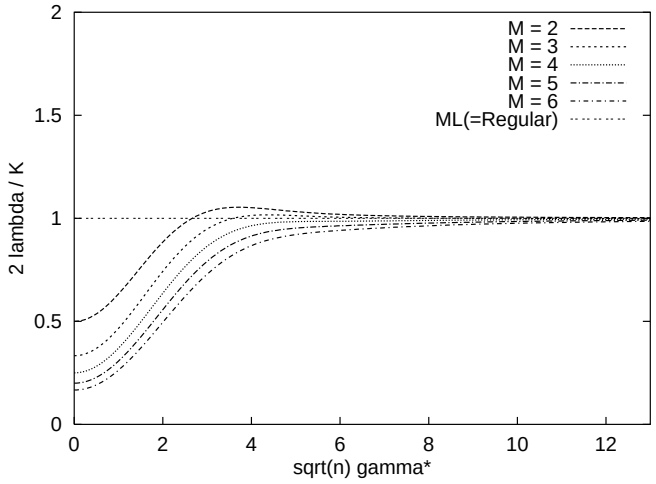

Figure 2.4: Bayes generalization error in *delicate* situations.

of the ML estimation, therefore, its ML generalization coefficient is identical to that of the regular models. Nevertheless, it has a property of singular models from the viewpoint of the Bayesian learning methods, as shown in Figure 2.4. We see in Fig. 2.4 its similarity to the JS estimation. The domination of the Bayes estimation over the ML estimation in the SOLNN was also discussed in [Watanabe and Amari, 2003], which will be described and compared with our results in Section 5.3.

2.6 Notation

In this section, we summarize symbols, variables, the definitions of words, etc.. Many of them have already appeared in the preceding, and can appear without their definitions in the later chapters.

Symbols

We use the following symbols:

$\mathbb{R}$: real number $\quad$ ^{t} : transpose

$\hat{}$: estimator $\quad$ * : true value

$\|\cdot\|$: norm $\quad$ $|\cdot|$: absolute value of scalar or determinant of matrix

$\mathrm{tr}(\cdot)$: trace of matrix $\quad$ I_d : $d \times d$ identity matrix

We use the indicator function:

$$\theta(\textit{event}) = \begin{cases} 1 & \text{if } \textit{event} \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

Distributions

We distinguish the *true* distribution from the others:

$p(\cdot)$: probabilistic distribution $\quad$ $q(\cdot)$: true distribution

We use the following symbol in the empirical and the subspace Bayes approaches:

$\|$: separator between the parameter and the hyperparameter,
for example, $p(x; w\|\tau)$, where x is a random variable, w is a parameter,
and τ is a hyperparameter

We denote the basic distributions as follows:

$\mathcal{N}_d(\mu, \Sigma)$: d-dimensional normal distribution with average μ and covariance Σ

$\mathcal{W}_d(m, \Sigma)$: d-dimensional central Wishart distribution
with m degrees of freedom and scale matrix Σ

$\mathcal{W}_d(m, \Sigma, \Lambda)$: d-dimensional Wishart distribution
with m degrees of freedom, scale matrix Σ, and noncentrality matrix Λ

Then, we denote the density function of the normal distribution as follows:

$\mathcal{N}_d(\cdot; \mu, \Sigma)$: the density function of the random variable subject to $\mathcal{N}_d(\mu, \Sigma)$

We denote the expectation value as follows:

$$\langle \cdot \rangle_p : \text{expectation value over distribution } p$$

We define the word *estimator* for the Bayesian learning methods as follows:

Definition 3 *We define the estimator of a parameter in a Bayesian learning method as the expectation value over the posterior distribution.*

Note that the Bayes predictive distribution is not necessarily equivalent to $p(y|x, \hat{w}_{\text{Bayes}})$, where $\hat{w}_{\text{Bayes}}$ denotes the Bayes estimator, because of the extent of the posterior distribution.

Variables

To describe learning models, we use the following variables:

M : input dimension (# of input units)	$x \in \mathbb{R}^M$: input vetor
N : output dimension (# of output units)	$y \in \mathbb{R}^N$: output vetor
K : *essential* parameter dimension	$w \in \mathbb{R}^K$: parameter vetor
$L = \max(M, N)$	$\varepsilon \in \mathbb{R}^N$: output noise
$l = \min(M, N)$	n : # of training samples

$$D^n = (X^n, Y^n) = (\{x_1, \ldots, x_n\}, \{y_1, \ldots, y_n\})$$
$$: n \text{ training samples independently and identically taken from the true distribution } q(x, y) = q(x)q(y|x).$$

Note that H will appear in the following.

Linear Neural Network Model

To describe linear neural networks (LNNs), we use the following variables:

$$\begin{aligned} BA &: \text{map of LNN } (y = BAx + \varepsilon) \\ H &: \text{rank of } BA \text{ (\# of hidden units)} \\ H^* &: \text{rank of } B^*A^* \text{ (true rank)} \\ A = (a_1, \ldots, a_H)^t &: H \times M \text{ input parameter matrix} \\ B = (b_1, \ldots, b_H) &: N \times H \text{ output parameter matrix} \end{aligned}$$

The efficient statistics of an LNN are given as follows:

$$Q = n^{-1} \sum_{i=1}^{n} x_i x_i^t. \qquad R = n^{-1} \sum_{i=1}^{n} y_i x_i^t,$$

The following variables related with the singular value decomposition of $RQ^{-1/2}$ are important:

$$\begin{aligned} h &: \text{index number of component} \\ \gamma_h &: h\text{-th largest singular value of } RQ^{-1/2} \\ \omega_{a_h} &: \text{right singular vector corresponding to } \gamma_h \\ \omega_{b_h} &: \text{left singular vector corresponding to } \gamma_h \end{aligned}$$

Hence, the following relations hold:

$$\begin{aligned} \gamma_h \omega_{a_h} &= Q^{-1/2} R^t \omega_{b_h} & \gamma_h \omega_{b_h} &= RQ^{-1/2} \omega_{a_h} \\ \gamma_h^2 \omega_{a_h} &= Q^{-1/2} R^t R Q^{-1/2} \omega_{a_h} & \gamma_h^2 \omega_{b_h} &= RQ^{-1} R^t \omega_{b_h} \end{aligned}$$

For single-output (SO) LNNs, i.e., $N = H = 1$, we use the following:

$$ba^t : \text{map of SOLNN} \qquad \gamma = b\|a\|$$

Generalization Properties

We generally use the following letters subscripted with 'Regular' for the regular models, or with an algorithm abbreviation for the singular models:

$G(X^n, Y^n)$: generalization error with a set of n training samples
$T(X^n, Y^n)$: training error with a set of n training samples
$F(Y^n|X^n)$: free energy with a set of n training samples

$G(n) = \langle G(X^n, Y^n) \rangle_{q(X^n, Y^n)}$: average generalization error over all training sets
$T(n) = \langle T(X^n, Y^n) \rangle_{q(X^n, Y^n)}$: average training error over all training sets
$F(n) = \langle F(Y^n|X^n) + \log q(Y^n|X^n) \rangle_{q(X^n, Y^n)}$: average normalized free energy over all training sets

λ : generalization coefficient ($G(n) = \lambda n^{-1} + o(n^{-1})$)
ν : training coefficient ($T(n) = \nu n^{-1} + o(n^{-1})$)
λ' : free energy coefficient ($F(n) = \lambda' \log n + o(\log n)$)

Note that the average normalized free energy, $F(n)$, is defined as the average of the free energy, $F(Y^n|X^n)$, subtracted by the entropy of the true distribution.

James-Stein Estimator

The following variable appears with the subscript indicating the algorithm related with the James-Stein type shrinkage estimation:

χ : degree of shrinkage

The following definitions are extended ones of Definitions 1 and 2, respectively, for Bayesian learning methods:

Definition 4 *We say that a learning method α dominates another learning method β if*

$$G_\alpha(n; w^*) \leq G_\beta(n; w^*) \qquad \textit{for arbitrary } w^*,$$
$$\textit{and} \qquad G_\alpha(n; w^*) < G_\beta(n; w^*) \qquad \textit{for, at least, a certain } w^*,$$

where $G_\kappa(n; w^*)$ denotes the average generalization error, which is a function of the true parameter w^*, of a learning method κ.

Definition 5 *We say that a learning method is admissible if no learning method dominates it.*

Chapter 3

Subspace Bayes Approach

In this chapter, we begin to describe our results. This chapter is devoted to the analysis of a subspace Bayes (SB) approach, which has been published in [Nakajima and Watanabe, 2005b] and will be published in [Nakajima and Watanabe, 2006b]. In Section 3.1, We introduce the SB approach. Then, in Section 3.2, we derive the SB posterior distribution and the SB predictive distribution in linear neural networks (LNNs), and thus show that the SB approach is asymptotically equivalent to a positive-part James-Stein (JS) type shrinkage estimation. After that, in Section 3.3, we prove the theorems of generalization properties, and illustrate the results.

3.1 Introduction

Extending the idea of the empirical Bayes (EB) approach, described in Section 2.1.2, we can introduce hyperparameters also in a model distribution. What we call a subspace Bayes (SB) approach is the EB approach where a part of the parameters of a model are regarded as hyperparameters. In the SB approach, we first separate the *whole* parameter w of an original model $p(y|x,w)$ into the parameter $\bar{w}$ and the hyperparameter τ, i.e., $w = \{\bar{w}, \tau\}$. Then, we have the model distribution $p(y|x,\bar{w}\|\tau)$. Using the prior distribution $\phi(\bar{w})$, we have the marginal

likelihood as follows:

$$Z(Y^n|X^n\|\tau) = \int \phi(\bar{w}) \prod_{i=1}^{n} p(y_i|x_i, \bar{w}\|\tau) d\bar{w}. \tag{3.1}$$

We estimate the hyperparameter value by maximizing Eq.(3.1):

$$\hat{\tau} = \operatorname*{argmax}_{\tau} Z(Y^n|X^n\|\tau). \tag{3.2}$$

Then, the SB posterior distribution is given by

$$p(\bar{w}|X^n, Y^n\|\hat{\tau}) = \frac{\phi(\bar{w}) \prod_{i=1}^{n} p(y_i|x_i, \bar{w}\|\hat{\tau})}{Z(Y^n|X^n\|\hat{\tau})}. \tag{3.3}$$

The SB predictive distribution, the SB generalization error, and the SB training error are given by:

$$p_{\mathrm{SB}}(y|x, X^n, Y^n) = \int p(y|x, \bar{w}\|\hat{\tau}) p(\bar{w}|X^n, Y^n\|\hat{\tau}) d\bar{w}, \tag{3.4}$$

$$G_{\mathrm{SB}}(n) = \langle G_{\mathrm{SB}}(X^n, Y^n) \rangle_{q(X^n, Y^n)}, \tag{3.5}$$

$$T_{\mathrm{SB}}(n) = \langle T_{\mathrm{SB}}(X^n, Y^n) \rangle_{q(X^n, Y^n)}, \tag{3.6}$$

respectively, where

$$G_{\mathrm{SB}}(X^n, Y^n) = \int q(x) q(y|x) \log \frac{q(y|x)}{p_{\mathrm{SB}}(y|x, X^n, Y^n)} dx dy, \tag{3.7}$$

$$T_{\mathrm{SB}}(X^n, Y^n) = n^{-1} \sum_{i=1}^{n} \log \frac{q(y|x)}{p_{\mathrm{SB}}(y|x, X^n, Y^n)}. \tag{3.8}$$

The generalization error and the training error can be asymptotically expanded as

$$G_{\mathrm{SB}}(n) = \frac{\lambda_{\mathrm{SB}}}{n} + o(n^{-1}), \tag{3.9}$$

$$T_{\mathrm{SB}}(n) = \frac{\nu_{\mathrm{SB}}}{n} + o(n^{-1}), \tag{3.10}$$

respectively.

In the following sections, we apply two versions of SB approach to linear neural network models (LNNs), defined by Eq.(2.37): in the first one, we regard the output parameter matrix B as a hyperparameter and then marginalize the likelihood in the input parameter space (MIP); and in the other one, we regard the input

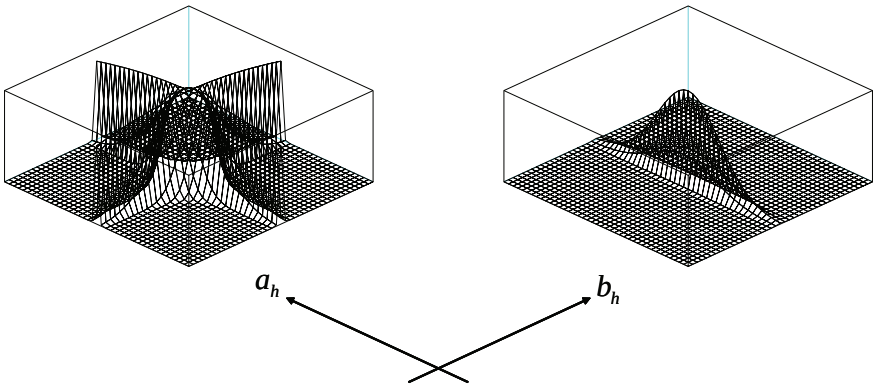

Figure 3.1: SB posterior distribution in the MIP version (right-hand side), which is substituted for the Bayes posterior distribution (left-hand side).

parameter matrix A, instead of B, as a hyperparameter and then marginalize in the output parameter space (MOP). Figure 3.1 virtually shows the Bayes posterior distribution (left-hand side) and the SB posterior distribution in the MIP version (right-hand side). Note that, in reality, the SB posterior distribution is the delta function with respect to b_h, so that it has infinite values.

3.2 Subspace Bayes Solution

Assume that the covariance matrix of a noise is known and equal to I_N. Then the conditional distribution of an LNN in the MIP version of SB approach is given by

$$p(y|x, A\|B) = \frac{1}{(2\pi)^{N/2}} \exp\left(-\frac{\|y - BAx\|^2}{2}\right). \tag{3.11}$$

We use the following prior distribution:

$$\phi(A) = \frac{1}{(2\pi)^{HM/2}} \exp\left(-\frac{\mathrm{tr}\,(A^t A)}{2}\right). \tag{3.12}$$

Note that we can similarly prepare $p(y|x, B\|A)$ and $\phi(B)$ for the MOP version. We assume that the true conditional distribution is $p(y|x, A^*\|B^*)$, where B^*A^* is the true map with rank $H^* \leq H$.

3.2.1 Subspace Bayes Estimator

For simplicity, we assume that the input vector is orthonormalized,[1] i.e,

$$\int xx^t q(x)dx = I_M. \tag{3.13}$$

Consequently, the central limit theorem leads to the following two equations:

$$Q(X^n) = n^{-1}\sum_{i=1}^{n} x_i x_i^t = I_M + O_p(n^{-1/2}), \tag{3.14}$$

$$R(X^n, Y^n) = n^{-1}\sum_{i=1}^{n} y_i x_i^t = B^* A^* + O_p(n^{-1/2}), \tag{3.15}$$

where $Q(X^n)$ is an $M \times M$ symmetric matrix and $R(X^n, Y^n)$ is an $N \times M$ matrix. Hereafter, we abbreviate $Q(X^n)$ as Q, and $R(X^n, Y^n)$ as R.

Let γ_h be the h-th largest singular value of the matrix $RQ^{-1/2}$, ω_{a_h} the corresponding right singular vector, and ω_{b_h} the corresponding left singular vector, where $1 \leq h \leq H$. We find from Eq.(3.15) that, in the asymptotic limit, the singular values corresponding to the necessary components to realize the true distribution converge to finite values; while the other ones corresponding to the redundant components converge to zero. Therefore, with probability 1, the largest H^* singular values correspond to the necessary components, and the others correspond to the redundant components. Combining Eqs.(3.14) and (3.15), we have

$$\omega_{b_h} RQ^{\rho} = \omega_{b_h} R + O_p(n^{-1}) \qquad \text{for } H^* < h \leq H, \tag{3.16}$$

where $-\infty < \rho < \infty$ is an arbitrary constant. The SB estimator, defined as the expectation value over the SB posterior distribution, is given by the following theorem:

Theorem 1 *Let*

$$\chi_{SB} = \begin{cases} M & (\textit{in the MIP version}) \\ N & (\textit{in the MOP version}) \end{cases}. \tag{3.17}$$

[1]This assumption is true if we apply the following procedure before learning: orthonormalize the input by using the given n training samples, which can be done with accuracy $O_p(n^{-1})$; then, use the prior distribution with diagonal covariance matrix, like Eq.(3.12), based on the new basis.

The SB estimator of the map of an LNN is given by

$$(\hat{B}\hat{A})_{SB} = \sum_{h=1}^{H} \left(1 - \frac{\chi_{SB}}{\max(\chi_{SB}, n\gamma_h^2)}\right) \omega_{b_h} \omega_{b_h}^t R Q^{-1} + O_p(n^{-1}). \tag{3.18}$$

(The proof is given in Section 3.2.3.)

Comparing Eq.(3.18) and Eq.(2.44), we find that the h-th component of the SB estimator can be written as follows:

$$(\hat{b}_h \hat{a}_h^t)_{\text{SB}} = \left(1 - \frac{\chi_{\text{SB}}}{\max(\chi_{\text{SB}}, n(\|\hat{b}_h\|^2 \|\hat{a}_h\|^2)_{\text{MLE}})}\right) (\hat{b}_h \hat{a}_h^t)_{\text{MLE}} + O_p(n^{-1}), \tag{3.19}$$

where $(\hat{b}_h \hat{a}_h^t)_{\text{MLE}}$ is the corresponding component of the ML estimator, and $(\|\hat{b}_h\|^2 \|\hat{a}_h\|^2)_{\text{MLE}} = \gamma_h^2$ is its squared amplitude. Therefore, we can say that the SB estimator of each component is asymptotically equivalent to a positive-part James-Stein (JS) type shrinkage estimator, defined by Eq.(2.58).

3.2.2 Predictive Distribution

Even if the SB estimator is asymptotically equivalent to the JS type estimator, the SB approach can differ from the JS type estimation because of the extent of the SB posterior distribution. The following lemma of the SB predictive distribution guarantees the asymptotic equivalence between the two learning methods.

Lemma 1 *The predictive distribution of an LNN in the SB approach can be written as follows:*

$$p_{SB}(y|x, X^n, Y^n) = \left((2\pi)^N |\hat{V}|\right)^{-1/2} \cdot \exp\left(-\frac{(y - \hat{V}(\hat{B}\hat{A})_{SB}\, x)^t \hat{V}^{-1} (y - \hat{V}(\hat{B}\hat{A})_{SB}\, x)}{2}\right) + O_p(n^{-3/2}), \tag{3.20}$$

where $\hat{V} = I_N + O_p(n^{-1})$, and $|\cdot|$ denotes the determinant of a matrix.

(Proof) We prove only in the MIP version, as we can do also in the MOP version in exactly the same way. The predictive distribution is written as follows:

$$\begin{aligned} p_{\rm SB}(y|x,X^n,Y^n) &= \left\langle p(y|x,A\|\hat{B}) \right\rangle_{p(A|X^n,Y^n\|\hat{B})} \\ &= q(y|x) \left\langle \frac{p(y|x,A\|\hat{B})}{q(y|x)} \right\rangle_{p(A|X^n,Y^n\|\hat{B})} \\ &\propto q(y|x) \left\langle \exp\left(y^t(\hat{B}A - B^*A^*)x \right) \right\rangle_{p(A|X^n,Y^n\|\hat{B})}. \end{aligned} \tag{3.21}$$

We find from Eqs.(3.28), (3.29), (3.35), and (3.39) in Section 3.2.3 that the random variable $(\hat{B}A - B^*A^*)$ is of order $O_p(n^{-1/2})$ when A is subject to $p(A|X^n,Y^n\|\hat{B})$. Hence, we can expand the predictive distribution, Eq.(3.21), as follows:

$$\begin{aligned} p_{\rm SB}(y|x,X^n,Y^n) \propto q(y|x) \left\langle 1 + y^t(\hat{B}A - B^*A^*)x + \frac{y^t v v^t y}{2n} \right\rangle_{p(A|X^n,Y^n\|\hat{B})} \\ + O_p(n^{-3/2}), \end{aligned} \tag{3.22}$$

where $v = \sqrt{n}(\hat{B}A - B^*A^*)x$ is an N-dimensional vector of order $O_p(1)$. Calculating the expectation value and expanding the logarithm of Eq.(3.22), we immediately arrive at Lemma 1. (Q.E.D.)

Lemma 1 states that the SB posterior distribution is sufficiently localized, so that we can substitute the model at the SB estimator for the SB predictive distribution with asymptotically insignificant impact upon generalization performance:

$$p_{\rm SB}(y|x,X^n,Y^n) \sim p_{\rm SB}(y|x;(\hat{B}\hat{A})_{\rm SB}). \tag{3.23}$$

Therefore, we conclude that the SB approach is asymptotically equivalent to the positive-part JS type shrinkage estimation. Note that the variance of the prior distribution, Eq.(3.12), asymptotically has no effect upon prediction and hence upon generalization performance, as far as it is a positive and finite constant.

3.2.3 Proof of Theorem 1

In this subsection, we prove Theorem 1. First, we consider the MIP version, where the conditional marginal likelihood is given by

$$Z(Y^n|X^n\|B) = \int \phi(A) \prod_{i=1}^{n} p(y_i|x_i, A\|B) dA$$
$$\propto \int \exp\left(-\frac{\sum_{i=1}^{n} \|y_i - BAx_i\|^2 + \mathrm{tr}(A^t A)}{2} \right) dA. \qquad (3.24)$$

Here $\int dA$ denotes the integral with respect to all the elements of the matrix A. We denote by $\otimes$ the Kronecker product and by $\mathrm{vec}(\cdot)$ the vector created from a matrix by stacking the column vectors below one another, for example, $\mathrm{vec}(V) = (v_1^t, \ldots, v_H^t)^t$ is the NH-dimensional column vector, where $V = (v_1, \ldots, v_H)$ is an $N \times H$ matrix. By using the Gaussian integral, we have the following form of the marginal likelihood:

$$Z(Y^n|X^n\|B) \propto \left(n|\tilde{S}|^{-1} \right)^{-1/2} \exp\left(\frac{n\tilde{b}^t \tilde{R} \tilde{S}^{-1} \tilde{R}^t \tilde{b}}{2} \right), \qquad (3.25)$$

where

$$\tilde{a} = \mathrm{vec}(A^t),$$
$$\tilde{b} = \mathrm{vec}(B),$$
$$\tilde{R} = I_M \otimes R,$$

and

$$\tilde{S} = \left(B^t B \otimes Q \right) + n^{-1} I_{HM}.$$

Similarly, we also have the following form of the posterior distribution:

$$p(A|X^n, Y^n\|B) = \frac{\phi(A) \prod_{i=1}^{n} p(y_i|x_i, A\|B)}{Z(Y^n|X^n\|B)}$$
$$\propto \exp\left(-\left(\tilde{a} - \tilde{S}^{-1} \tilde{R}^t \tilde{b} \right)^t \frac{n\tilde{S}}{2} \left(\tilde{a} - \tilde{S}^{-1} \tilde{R}^t \tilde{b} \right) \right). \qquad (3.26)$$

Given an arbitrary map BA, we can have A with its orthogonal row vectors and B with its orthogonal column vectors by using the singular value decomposition. Just in that case, the prior probability, Eq.(3.12), is maximized. Accordingly, we assume without loss of generality that the optimal value of B consists of its orthogonal column vectors. Consequently, the marginal likelihood, Eq.(3.25), and the posterior distribution, Eq.(3.26), factorize as

$$Z(Y^n|X^n\|B) = \prod_{h=1}^{H} Z(Y^n|X^n\|b_h),$$
$$p(A|X^n, Y^n\|B) = \prod_{h=1}^{H} p(a_h|X^n, Y^n\|b_h),$$

respectively, where

$$Z(Y^n|X^n\|b_h) \propto |S_h|^{-1/2} \exp\left(\frac{nb_h^t R S_h^{-1} R^t b_h}{2}\right), \tag{3.27}$$
$$p(a_h|X^n, Y^n\|b_h) \propto \exp\left(-\left(a_h - S_h^{-1}R^t b_h\right)^t \frac{nS_h}{2}\left(a_h - S_h^{-1}R^t b_h\right)\right). \tag{3.28}$$

Here

$$S_h = \|b_h\|^2 Q + n^{-1} I_M. \tag{3.29}$$

Therefore, the free energy can be written as the sum of the contributions of each component:

$$F(Y^n|X^n\|B) = \sum_{h=1}^{H} F(Y^n|X^n\|b_h), \tag{3.30}$$

where

$$\begin{aligned} 2F(Y^n|X^n\|b_h) &= -2\log Z(Y^n|X^n\|b_h) \\ &= \log|S_h| - nb_h^t R S_h^{-1} R^t b_h + \text{const.}. \end{aligned} \tag{3.31}$$

We also find that the posterior distribution of a_h, Eq.(3.28), is the normal distribution with the average

$$\hat{a}_h = S_h^{-1} R^t \hat{b}_h, \tag{3.32}$$

and the covariance matrix

$$\hat{\Sigma}_{a_h} = \frac{S_h^{-1}}{n}. \tag{3.33}$$

Hereafter, separately considering the components imitating the positive true ones, and the redundant ones, we will find the optimal hyperparameter value $\hat{b}_h$ that minimizes Eq.(3.31). We abbreviate $F(Y^n|X^n\|b_h)$ as $F(b_h)$. For a positive true component, $h \leq H^*$, the corresponding observed singular value γ_h of $RQ^{-1/2}$ is of order 1 with probability 1. Therefore, from Eqs.(3.31) and (3.29), we obtain

$$2F(b_h) = M \log \|b_h\|^2 - \frac{n b_h^t RQ^{-1}R^t b_h}{\|b_h\|^2} + \frac{b_h^t RQ^{-2}R^t b_h}{\|b_h\|^4} + O_p(n^{-1}) + \text{const.}. \tag{3.34}$$

In minimizing Eq.(3.34), the second term, which is the leading one, dominates the determination of the direction cosine of $\hat{b}_h$, which leads to

$$\hat{b}_h = \|\hat{b}_h\|(\omega_{b_h} + O_p(n^{-1})).$$

Then, the first and the third terms determine the norm of $\hat{b}_h$ because the second term is independent of it. Thus, we have the optimal hyperparameter value as follows:

$$\hat{b}_h = \sqrt{\frac{\omega_{b_h}^t RQ^{-2}R^t\omega_{b_h}}{M}}\omega_{b_h} + O_p(n^{-1}). \tag{3.35}$$

By using Eqs.(3.35) and (3.32), we obtain the SB estimator for the positive true component of the map as follows:

$$\begin{aligned}\hat{b}_h\hat{a}_h^t &= \frac{\hat{b}_h\hat{b}_h^t RQ^{-1}}{\|\hat{b}_h\|^2} + O_p(n^{-1}) \\ &= \omega_{b_h}\omega_{b_h}^t RQ^{-1} + O_p(n^{-1}).\end{aligned} \tag{3.36}$$

On the other hand, for a redundant component, $h > H^*$, Eq.(3.16) allows us to approximate Eq.(3.31) as follows:

$$2F(b_h) = M \log\left(\|b_h\|^2 + n^{-1}\right) - \frac{n b_h^t RR^t b_h}{\|b_h\|^2 + n^{-1}} + O_p(n^{-1/2}) + \text{const.}. \tag{3.37}$$

Then, we find that the direction cosine of $\hat{b}_h$, determined by the second term of Eq.(3.37), is approximated by ω_{b_h} with accuracy $O_p(n^{-1/2})$. After substituting $\gamma_h^2\|b_h\|^2(1+O_p(n^{-1/2}))$ for $b_h^t R R^t b_h$, we have the following condition by partial differentiation of Eq.(3.37) with respect to the norm of b_h:

$$\begin{aligned} 0 = 2\frac{\partial F(b_h)}{\partial \|b_h\|^2} &= \frac{M}{\|b_h\|^2+n^{-1}} - \frac{\gamma_h^2}{(\|b_h\|^2+n^{-1})^2} + O_p(\|b_h\|^{-2}n^{-1/2}) \\ &= \frac{M}{(\|b_h\|^2+n^{-1})^2}\left(\|b_h\|^2 - \frac{n\gamma_h^2 - M}{nM}\right) + O_p(\|b_h\|^{-2}n^{-1/2}). \end{aligned} \tag{3.38}$$

We find from Eq.(3.38) that Eq.(3.37) is an increasing function of $\|b_h\|$ if γ_h is less than $\sqrt{M/n}$. Therefore, we have the following optimal hyperparameter value:

$$\hat{b}_h = \sqrt{\frac{M_h' - M}{nM}}\omega_{b_h} + O_p(n^{-1}), \tag{3.39}$$

where

$$M_h' = \max(M, n\gamma_h^2).$$

Thus, we obtain the SB estimator of the redundant component as follows:

$$\begin{aligned} \hat{b}_h\hat{a}_h^t &= \frac{\hat{b}_h\hat{b}_h^t R}{\|\hat{b}_h\|^2+n^{-1}} + O_p(n^{-1}) \\ &= (1 - MM_h'^{-1})\omega_{b_h}\omega_{b_h}^t R + O_p(n^{-1}). \end{aligned} \tag{3.40}$$

Selecting the largest singular value components minimizes Eq.(3.31). Hence, combining Eq.(3.36) with the fact that $MM_h'^{-1} = O_p(n^{-1})$ for the positive true components, and Eq.(3.40) with Eq.(3.16), we obtain the SB estimator in Theorem 1. We can also derive the SB estimator in the MOP version in exactly the same way. (Q.E.D.)

3.3 Generalization Properties

3.3.1 Generalization Error

The existence of the singular value decomposition of any B^*A^* allows us to assume without loss of generality that A^* and B^* consist of its orthogonal row

vectors and of its orthogonal column vectors, respectively. Then, we find from Lemma 1 that the KL divergence, Eq.(3.7), with a set of n training samples is given by

$$\begin{aligned} G_{\rm SB}(X^n, Y^n) &= \int q(x)q(y|x) \log \frac{q(y|x)}{p_{\rm SB}(y|x, X^n, Y^n)} dxdy \\ &= \left\langle \frac{\|(B^*A^* - \hat{B}\hat{A})x\|^2}{2} \right\rangle_{q(x)} + O_p(n^{-3/2}) \\ &= \frac{1}{2} \sum_{h=1}^{H} {\rm tr} \left((b_h^* a_h^{*t} - \hat{b}_h \hat{a}_h^t)^t (b_h^* a_h^{*t} - \hat{b}_h \hat{a}_h^t) \right) + O_p(n^{-3/2}). \quad (3.41) \end{aligned}$$

The following theorem provides the SB generalization coefficient:

Theorem 2 *The generalization error of an LNN in the SB approach can be asymptotically expanded as*

$$G_{SB}(n) = \frac{\lambda_{SB}}{n} + O(n^{-3/2}),$$

where the generalization coefficient is given by

$$2\lambda_{SB} = (H^*(L+l) - H^{*2}) + \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h'^2 > \chi_{SB}) \left(1 - \frac{\chi_{SB}}{\gamma_h'^2}\right)^2 \gamma_h'^2 \right\rangle_{q(\{\gamma_h'^2\})}. \quad (3.42)$$

Here, $\theta(\cdot)$ is the indicator function and $\gamma_h'^2$ is the h-th largest eigenvalue of a random matrix subject to $\mathcal{W}_{l-H^*}(L-H^*, I_{l-H^*})$, over which $\langle\cdot\rangle_{q(\{\gamma_h'^2\})}$ denotes the expectation value.

(Proof) According to Theorem 1, the difference between the SB and the ML estimators of a true component with a positive singular value is of order $O_p(n^{-1})$. Furthermore, the generalization error of the ML estimator of the component is the same as that of the regular models because of its identifiability. Hence, from Eqs.(2.38) and (1.9), we obtain the first term of Eq.(3.42) as the contribution of the first H^* components. On the other hand, we find from Eq.(3.16) and Theorem 1 that, for a redundant component, identifying $RQ^{-1/2}$ with R affects the SB estimator only of order $O_p(n^{-1})$, and hence, does not affect the generalization

coefficient. We say that U is the general diagonalized matrix of an $N \times M$ matrix T if T has the following singular value decomposition: $T = \Omega_b U \Omega_a$, where Ω_a and Ω_b are an $M \times M$ and an $N \times N$ orthogonal matrices, respectively. Let D be the general diagonalized matrix of R, and D' the $(N - H^*) \times (M - H^*)$ matrix created by removing the first H^* columns and rows from D. Then, the first H^* diagonal elements of D correspond to the positive true singular value components, and D' consists only of noises. Therefore, D' is the general diagonalized matrix of $n^{-1/2}R'$, where R' is an $(N - H^*) \times (M - H^*)$ random matrix whose elements are independently subject to $\mathcal{N}_1(0, 1)$, so that $R'R'^t$ is subject to $\mathcal{W}_{N-H^*}(M - H^*, I_{N-H^*})$. The redundant components imitate $n^{-1/2}R'$. Hence, using Theorem 1 and Eq.(3.41) and noting that the distribution of the $(l - H^*)$ largest eigenvalues, which are not trivially equal to zero, of a random matrix subject to $\mathcal{W}_{L-H^*}(l - H^*, I_{L-H^*})$ are equal to that of the eigenvalues of another random matrix subject to $\mathcal{W}_{l-H^*}(L - H^*, I_{l-H^*})$, we obtain the second term of Eq.(3.42) as the contribution of the last $(H - H^*)$ components. Thus, we complete the proof of Theorem 2. (Q.E.D.)

Although the second term of Eq.(3.42) is not a simple function, it can, relatively easily, be numerically calculated by creating samples subject to the Wishart distribution. Furthermore, the more simple function approximating the term can be derived in the large scale limit when M, N, H, and H^* go to infinity in the same order, in a similar fashion to the analysis of the ML estimation [Fukumizu, 1999].

We define the following scalars:

$$\alpha = (l - H^*)/(L - H^*), \tag{3.43}$$

$$\beta = (H - H^*)/(l - H^*), \tag{3.44}$$

$$\kappa = \chi_{\mathrm{SB}}/(L - H^*). \tag{3.45}$$

Let W be a random matrix subject to $\mathcal{W}_{l-H^*}(L - H^*, I_{l-H^*})$, and $\{u_1, \ldots, u_{l-H^*}\}$ the eigenvalues of $(L - H^*)^{-1}W$. The measure of the empirical distribution of the eigenvalues is defined by

$$\delta P = (l - H^*)^{-1} \left\{ \delta(u_1) + \delta(u_2) + \cdots + \delta(u_{l-H^*}) \right\}, \tag{3.46}$$

where $\delta(u)$ denotes the Dirac measure at u. In the large scale limit, the measure, Eq.(3.46), converges to

$$p(u)du = \frac{\sqrt{(u-u_m)(u_M-u)}}{2\pi\alpha u}\,\theta(u_m < u < u_M)\,du, \tag{3.47}$$

where $u_m = (\sqrt{\alpha}-1)^2$ and $u_M = (\sqrt{\alpha}+1)^2$ [Watcher, 1978]. Here, the convergence is almost everywhere for training samples. Let

$$(2\pi\alpha)^{-1}J(u_t;k) = \int_{u_t}^{\infty} u^k p(u)du \tag{3.48}$$

be the k-th order moment of the distribution, Eq.(3.47), where u_t is the lower bound of the integration range. The second term of Eq.(3.42) consists of the terms proportional to the minus first, the zero, and the first order moments of the eigenvalues. Because only the eigenvalues greater than $\chi_{\rm SB}$ among the largest $(H-H^*)$ eigenvalues contribute to the generalization error, the moments with the lower bound $u_t = \max(\kappa, u_\beta)$ should be calculated, where u_β is the β-percentile point of $p(u)$, i.e.,

$$\beta = \int_{u_\beta}^{\infty} p(u)du = (2\pi\alpha)^{-1}J(u_\beta;0).$$

Using the transform

$$s = \frac{u-(u_m+u_M)/2}{2\sqrt{\alpha}},$$

we can calculate the moments and thus obtain the following theorem:

Theorem 3 *The SB generalization coefficient of an LNN in the large scale limit is given by*

$$\begin{aligned} 2\lambda_{SB} \sim\ & (H^*(L+l) - H^{*2}) \\ & + \frac{(L-H^*)(l-H^*)}{2\pi\alpha}\left\{J(s_t;1) - 2\kappa J(s_t;0) + \kappa^2 J(s_t;-1)\right\}, \end{aligned} \tag{3.49}$$

where

$$J(s;1) = 2\alpha(-s\sqrt{1-s^2} + \cos^{-1} s),$$
$$J(s;0) = -2\sqrt{\alpha}\sqrt{1-s^2} + (1+\alpha)\cos^{-1}s - (1-\alpha)\cos^{-1}\frac{\sqrt{\alpha}(1+\alpha)s + 2\alpha}{2\alpha s + \sqrt{\alpha}(1+\alpha)},$$
$$J(s;-1) = \begin{cases} 2\sqrt{\alpha}\frac{\sqrt{1-s^2}}{2\sqrt{\alpha}s+1+\alpha} - \cos^{-1} s + \frac{1+\alpha}{1-\alpha}\cos^{-1}\frac{\sqrt{\alpha}(1+\alpha)s+2\alpha}{2\alpha s+\sqrt{\alpha}(1+\alpha)} & (0<\alpha<1) \\ 2\sqrt{\frac{1-s}{1+s}} - \cos^{-1} s & (\alpha = 1) \end{cases},$$

and

$$s_t = \max\left(\frac{\kappa - (1+\alpha)}{2\sqrt{\alpha}}, J^{-1}(2\pi\alpha\beta; 0)\right).$$

Here $J^{-1}(\cdot;k)$ denotes the inverse function of $J(s;k)$.

3.3.2 Training Error

We find from Lemma 1 that the empirical KL divergence, Eq.(3.8), with a set of n training samples is given by

$$\begin{aligned} T_{\text{SB}}(X^n, Y^n) &= n^{-1}\sum_{i=1}^{n}\log\frac{q(y|x)}{p_{\text{SB}}(y|x, X^n, Y^n)} \\ &= -\frac{1}{2}\Big\{\text{tr}\Big((B^*A^* - \hat{B}\hat{A}_{\text{MLE}})^t(B^*A^* - \hat{B}\hat{A}_{\text{MLE}}) \\ &\qquad -(\hat{B}\hat{A} - \hat{B}\hat{A}_{\text{MLE}})^t(\hat{B}\hat{A} - \hat{B}\hat{A}_{\text{MLE}})\Big)\Big\} + O_p(n^{-3/2}). \quad (3.50) \end{aligned}$$

In a similar way to the analysis of the generalization error, we obtain the following theorems.

Theorem 4 *The training error of an LNN in the SB approach can be asymptotically expanded as*

$$T_{SB}(n) = \frac{\nu_{SB}}{n} + O(n^{-3/2}),$$

where the training coefficient is given by

$$\begin{aligned} \nu_{SB} &= -(H^*(L+l) - H^{*2}) \\ &\quad - \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h'^2 > \chi_{SB})\left(1 - \frac{\chi_{SB}}{\gamma_h'^2}\right)\left(1 + \frac{\chi_{SB}}{\gamma_h'^2}\right)\gamma_h'^2 \right\rangle_{q(\{\gamma_h'^2\})}. \quad (3.51) \end{aligned}$$

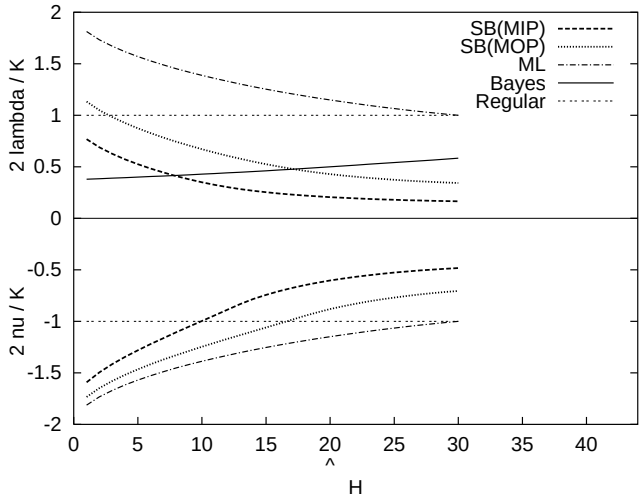

Figure 3.2: Generalization error (in the positive region) and training error (in the negative region) of the LNNs with $M = 50$, $N = 30$, and $H = 1, \ldots, 30$ units when $H^* = 0$.

Theorem 5 *The SB training coefficient of an LNN in the large scale limit is given by*

$$2\nu_{SB} \sim -(H^*(L+l) - H^{*2}) - \frac{(L-H^*)(l-H^*)}{2\pi\alpha} \left\{ J(s_t; 1) - \kappa^2 J(s_t; -1) \right\}. \tag{3.52}$$

3.3.3 Numerical Results

Figure 3.2 shows the theoretical results of the generalization and the training coefficients of the LNNs with $M = 50$ input, $N = 30$ output, and $H = 1, \ldots, 30$ hidden units, on the assumption that the true rank is equal to zero, $H^* = 0$. The horizontal axis indicates H, and the vertical axis indicates the coefficients normalized by the half of the *essential* parameter dimension K, given by Eq.(2.38). The lines in the positive region correspond to the generalization coefficients of the SB approaches given by Theorem 3, that of the ML estimation given by Proposition 4, that of the Bayes estimation given by Proposition 5, and that of the regular models, respectively; while the lines in the negative region correspond to the training

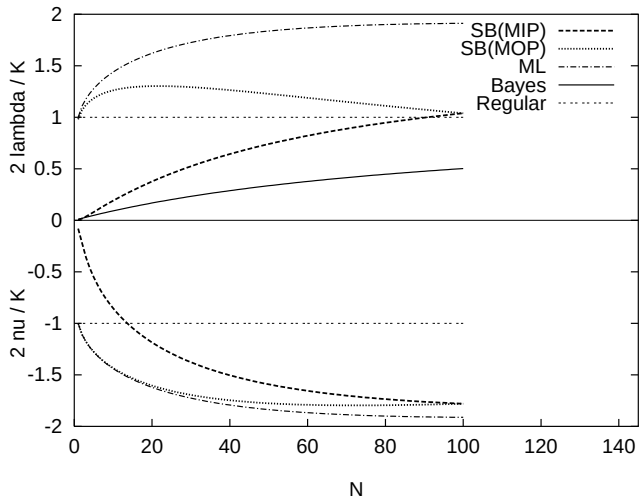

Figure 3.3: N dependence. ($M = 100$, $N = 1, \ldots, 100$, $H = 1$, and $H^* = 0$.)

coefficients of the SB approaches given by Theorem 5, that of the ML estimation given by Theorem 5 with the degree of shrinkage, $\chi_{\rm SB}$, set to zero,[2] and that of the regular models, respectively.[3] In singular models, it has not been known whether the Bayes generalization and training coefficients are symmetrical with each other, although we find from Theorems 2 and 4 that the ML generalization and training coefficients are symmetrical with each other. The Bayes training error, unfortunately, has not been clarified yet.

We see the following properties in Fig. 3.2:

1. Generally speaking, the SB approaches provide good generalization performance comparable to the Bayes estimation.

   It is not strange that there are some cases where the SB approaches provide better generalization performance than the Bayes estimation. (See Remark 2 in Section 2.1.1.)

[2]Because the SB generalization and training coefficients are equal to those of a positive-part JS type estimation, the theorems for the SB approach are changed for the ML estimation by setting the degree of shrinkage to zero, $\chi_{\rm SB} = 0$. Note that, then, $\kappa = 0$ in Theorem 5.

[3]We find from Eq.(1.9) that the normalized generalization and training coefficients are always equal to one and to minus one, respectively.

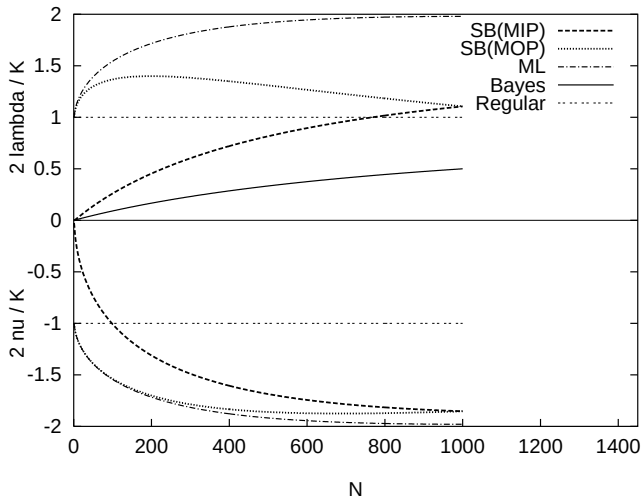

Figure 3.4: N dependence. ($M = 1000$, $N = 1, \ldots, 1000$, $H = 1$, and $H^* = 0$.)

2. The dependence of the SB generalization coefficient on the rank, H, is similar to that of the ML generalization coefficient.

3. The MIP is always better than the MOP.

   That is because $M > N$ in this case. We find from Theorem 2 that the version where we marginalize the larger dimension space provides better generalization performance than the other version, i.e., the MIP is better than MOP when $M > N$, and worse when $M < N$.

The second property is serious of the approximation method. We wish any of the SB approaches could inherit the one of the most preferable properties of the Bayes estimation, Eq.(2.15):

$$2\lambda_{\text{Bayes}} \leq 2\lambda_{\text{Regular}} \ (= K).$$

However, none of the SB approaches unfortunately does. According to the discussion that will be in Section 6.2.1, the worst cases are when $H = 1$, and M and N are large and equal to each other. Figure 3.3 shows one of those cases, i.e., the coefficients of the LNNs with $M = 100$ input units, $N = 1, \ldots, 100$ output units, indicated by the horizontal axis, and $H = 1$ hidden unit on the assumption

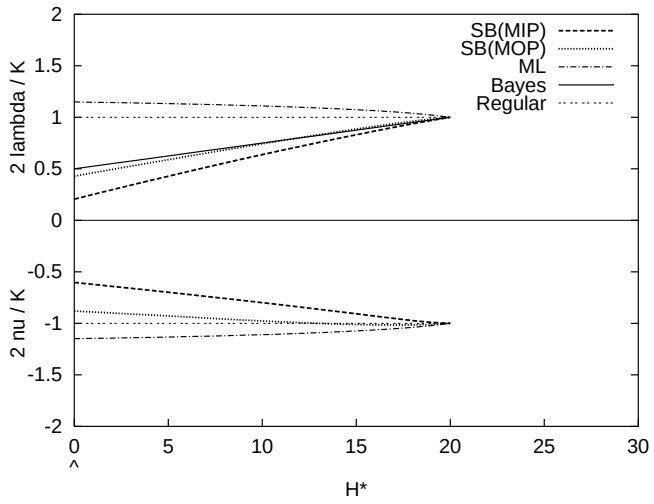

Figure 3.5: True rank dependence. ($M = 50$, $N = 30$, $H = 20$, and $H^* = 1, \ldots, 20$.)

that $H^* = 0$. We see in Fig. 3.3 that, when $N \sim 100$, the generalization error of the MIP, which is always no worse than the MOP in this case, is greater than that of the regular models. In LNNs, the generalization error, however, is not very bad even in the worst cases, because the *ML type effect* of the singularities of the LNNs are relatively soft, as mentioned in the first paragraph of Section 2.3. Figure 3.4 similarly shows the case that $M = 1000$. We see in Fig. 3.4 that, even when $M = N = 1000$ and $H = 1$, the generalization error of the MIP is only slightly greater than that of the regular models.

On the other hand, Fig. 3.5 shows the true rank, H^*, dependence of the LNN with $M = 50$ input, $N = 30$ output, and $H = 20$ hidden units.[4] The horizontal axis indicates the true rank H^*. We see in Fig. 3.5 that the SB approaches provide as good generalization performance as the Bayes estimation, and moreover, the MIP provides no worse generalization performance than the Bayes estimation regardless of H^*. That might seem to show the domination of the MIP over the Bayes estimation, which, however, is inconsistent with the superiority of the

[4]Note that the case that $H^* = 0$ in Fig.3.5 is the same as the case that $H = 20$ in Fig.3.2, both of which are marked with $\hat{}$.

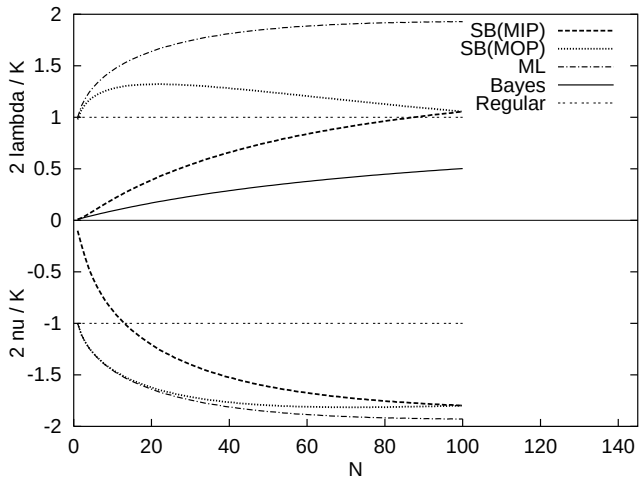

Figure 3.6: Results calculated by creating samples subject to the Wishart distribution and using Theorems 2 and 4 on the same condition as in Fig. 3.3.

Bayes estimation, Proposition 1. We will consider the *delicate* situations and deny the domination in Chapter 5.

Validity of Large Scale Approximation

The results shown in Figs. 3.2–3.5 have been calculated in the large scale approximation, i.e., by using Theorems 3 and 5. We have also numerically calculated them by creating samples subject to the Wishart distribution and then using Theorems 2 and 4. Thus we have found that the both results almost coincide with each other so that we can hardly distinguish. For example, Fig. 3.6 shows the coefficients calculated by using Theorems 2 and 4 on the same condition as in Fig. 3.3. We find only slight difference at $N \sim 100$, although it is a bad condition for the large scale approximation.

Chapter 4

Variational Bayes Approach

This chapter is devoted to the analysis of the variational Bayes (VB) approach in linear neural networks (LNNs), which have been published in [Nakajima and Watanabe, 2005a]. In Section 4.1, we apply the VB approach to the LNNs, restricting the space of the approximating posterior distribution such that the parameters of different layers, as well as those of different components, are independent of each other. In Section 4.2, we deduce the variational condition, and then derive the posterior distribution and the predictive distribution, analytically solving the variational condition. As a result, we find that the VB approach is asymptotically equivalent to a positive-part James-Stein (JS) type shrinkage estimation, like the subspace Bayes (SB) approach, analyzed in Chapter 3. Since we have already derived the generalization and the training coefficients of the positive-part JS type shrinkage estimation in Chapter 3, we obtain the theorems of those of the VB approach, which are described in Section 4.3. In the section, we also clarify the VB free energy and show that, in typical cases, the VB free energy well approximates the Bayes one; while the VB generalization error significantly differs from the Bayes one. This can throw a doubt on the model selection by minimizing the VB free energy to obtain better generalization performance.

4.1 Introduction

Given a trial posterior distribution $r(w|X^n, Y^n)$, we define the generalized free energy, a functional of $r(w|X^n, Y^n)$, by

$$\bar{F}(Y^n|X^n) = \left\langle \log \frac{r(w|X^n, Y^n)}{p(Y^n|X^n, w)\phi(w)} \right\rangle_{r(w|X^n,Y^n)}. \tag{4.1}$$

In the VB approach, restricting the set of possible distributions for $r(w|X^n, Y^n)$, we minimize the generalized free energy, Eq.(4.1). The optimum of the trial posterior distribution,

$$\hat{r}(w|X^n, Y^n) = \operatorname*{argmin}_r \bar{F}(Y^n|X^n), \tag{4.2}$$

is called the VB posterior distribution, and substituted for the Bayes posterior distribution. Thus, the VB predictive distribution is given by

$$p_{\rm VB}(y|x, X^n, Y^n) = \int p(y|x, w)\hat{r}(w|X^n, Y^n)dw. \tag{4.3}$$

The VB generalization error and the VB training error are given by

$$G_{\rm VB}(n) = \langle G_{\rm VB}(X^n, Y^n)\rangle_{q(X^n,Y^n)}, \tag{4.4}$$

$$T_{\rm VB}(n) = \langle T_{\rm VB}(X^n, Y^n)\rangle_{q(X^n,Y^n)}, \tag{4.5}$$

respectively, where

$$G_{\rm VB}(X^n, Y^n) = \int q(x)q(y|x) \log \frac{q(y|x)}{p_{\rm VB}(y|x, X^n, Y^n)} dxdy, \tag{4.6}$$

$$T_{\rm VB}(X^n, Y^n) = n^{-1} \sum_{i=1}^{n} \log \frac{q(y|x)}{p_{\rm VB}(y|x, X^n, Y^n)}. \tag{4.7}$$

The minimum of the generalized free energy is called the VB free energy:

$$F_{\rm VB}(Y^n|X^n) = \min_r \bar{F}(Y^n|X^n), \tag{4.8}$$

and the average normalized VB free energy is defined by

$$F_{\rm VB}(n) = \langle F_{\rm VB}(Y^n|X^n) + \log q(Y^n|X^n)\rangle_{q(X^n,Y^n)}. \tag{4.9}$$

The generalization error, the training error, and the normalized free energy, can be asymptotically expanded as follows:

$$G_{\rm VB}(n) = \frac{\lambda_{\rm VB}}{n} + o(n^{-1}), \tag{4.10}$$

$$T_{\rm VB}(n) = \frac{\nu_{\rm VB}}{n} + o(n^{-1}), \tag{4.11}$$

$$F_{\rm VB}(n) = \lambda'_{\rm VB} \log n + o(\log n). \tag{4.12}$$

Note that, in general, $\lambda'_{\rm VB} \neq \lambda_{\rm VB}$, because the relation corresponding to Eq.(2.14) does not hold in the VB approach. Accordingly, the VB free energy and the VB generalization error are less simply related with each other, as will be shown in the following.

4.2 Variational Bayes Solution

Assume that the covariance matrix of a noise is known and equal to I_N. Then, the conditional distribution of an LNN is given by

$$p(y|x, A, B) = \mathcal{N}_N(y; BAx, I_N). \tag{4.13}$$

We use the prior distribution that factorizes as $\phi(A, B) = \phi(A)\phi(B)$, where

$$\phi(A) = \mathcal{N}_{HM}(\mathrm{vec}(A^t); 0, c_a^2 I_{HM}), \tag{4.14}$$

$$\phi(B) = \mathcal{N}_{NH}(\mathrm{vec}(B); 0, c_b^2 I_{NH}). \tag{4.15}$$

Here, $0 < c_a^2, c_b^2 < \infty$ are the constants corresponding to the variances, and $\mathrm{vec}(\cdot)$ denotes the vector created from a matrix by stacking the column vectors below one another. Assume that the true conditional distribution is $p(y|x, A^*, B^*)$, where B^*A^* is the true map with rank $H^* \leq H$.

Hereafter, we abbreviate the trial posterior distribution $r(w|X^n, Y^n)$ as $r(w)$. In LNNs, we can obtain tractability by restricting the posterior distribution such that the input parameter matrix A and the output parameter matrix B are independent of each other, as shown below. Assume that the VB posterior distribution factorizes as

$$r(w) = r(A, B) = r(A)r(B). \tag{4.16}$$

Then, the generalized free energy, Eq.(4.1), is written as follows:

$$\bar{F}(Y^n|X^n) = \int r(A)r(B)\log\frac{r(A)r(B)}{p(Y^n|X^n,A,B)\phi(A,B)}dAdB, \tag{4.17}$$

where $\int dV$ denotes the integral with respect to all the elements of the matrix V. Using the variational method, we can easily show that the VB posterior distribution satisfies the following relations:

$$r(A) \propto \phi(A)\exp\langle\log p(Y^n|X^n,A,B)\rangle_{r(B)}, \tag{4.18}$$

$$r(B) \propto \phi(B)\exp\langle\log p(Y^n|X^n,A,B)\rangle_{r(A)}. \tag{4.19}$$

We find from Eqs.(4.18) and (4.19) that the VB posterior distributions are the normal when the prior distributions, $\phi(A)$ and $\phi(B)$, are the normal, because the log-likelihood of an LNN, $\log p(Y^n|X^n,A,B)$, is a biquadratic function of A and B.

For simplicity, we then apply another restriction, the independence of the parameters of different components, to the posterior distribution:

$$r(A,B) = \prod_{h=1}^{H} r(a_h)r(b_h). \tag{4.20}$$

This restriction makes the relations, Eqs.(4.18) and (4.19), decompose into those for each component:

$$r(a_h) \propto \phi(a_h)\exp\langle\log p(Y^n|X^n,A,B)\rangle_{r(A,B)/r(a_h)}, \tag{4.21}$$

$$r(b_h) \propto \phi(b_h)\exp\langle\log p(Y^n|X^n,A,B)\rangle_{r(A,B)/r(b_h)}, \tag{4.22}$$

where $\langle\cdot\rangle_{r(A,B)/r(a_h)}$, as well as $\langle\cdot\rangle_{r(A,B)/r(b_h)}$, denotes the expectation value over the trial posterior distribution of the parameters except a_h, as well as that except b_h.

Figure 4.1 shows the VB posterior distribution (right-hand side). The VB posterior distribution, which is the independent distribution with respect to different layer parameters minimizing the generalized free energy, is substituted for the Bayes posterior distribution (left-hand side).

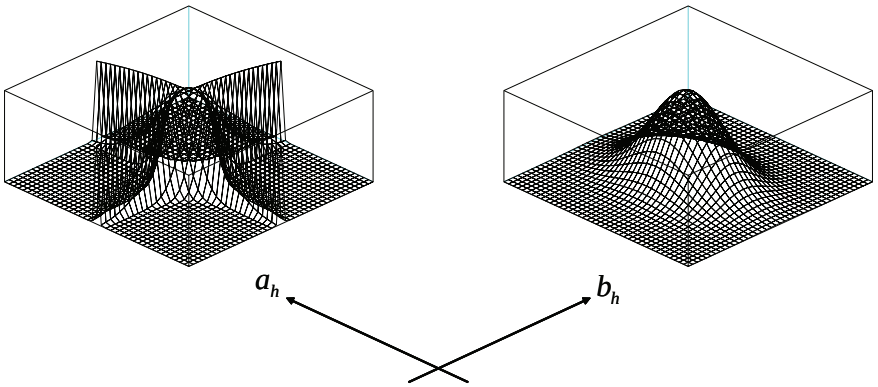

Figure 4.1: VB posterior distribution (right-hand side), which is the independent distribution with respect to different layer parameters to be substituted for the Bayes posterior distribution (left-hand side).

4.2.1 Variational Condition

Substituting Eqs.(4.13)–(4.15) into Eqs.(4.21) and (4.22), we find that the posterior distribution of the parameters of each component can be written as follows:

$$r(a_h) = \mathcal{N}_M(a_h; \mu_{a_h}, \Sigma_{a_h}), \tag{4.23}$$

$$r(b_h) = \mathcal{N}_N(b_h; \mu_{b_h}, \Sigma_{b_h}). \tag{4.24}$$

Given an arbitrary map BA, we can have A with its orthogonal row vectors and B with its orthogonal column vectors by using the singular value decomposition. Just in that case, both of the prior probabilities, Eqs.(4.14) and (4.15), are maximized. Accordingly, $\{\mu_{a_h}; h = 1, \ldots, H\}$, as well as $\{\mu_{b_h}; h = 1, \ldots, H\}$, of the optimum distribution is a set of vectors orthogonal to each other.

In exactly the same way as in Section 3.2.1, assuming the orthonormality of the input,

$$\int xx^t q(x)dx = I_M, \tag{4.25}$$

we have

$$Q = n^{-1}\sum_{i=1}^{n} x_i x_i^t = I_M + O_p(n^{-1/2}), \tag{4.26}$$

$$R = n^{-1}\sum_{i=1}^{n} y_i x_i^t = B^* A^* + O_p(n^{-1/2}), \tag{4.27}$$

$$\omega_{b_h} R Q^{\rho} = \omega_{b_h} R + O_p(n^{-1}) \qquad \text{for} \qquad H^* < h \leq H, \tag{4.28}$$

where $-\infty < \rho < \infty$ is an arbitrary constant. Here, Q and R are the sufficient statistics of an LNN, and γ_h, ω_{a_h}, and ω_{b_h} are the h-th largest singular value of $RQ^{-1/2}$, the corresponding right singular vector, and the corresponding left singular vector, respectively.

We easily obtain the following condition, called the variational condition, by substituting Eqs.(4.23) and (4.24) into Eqs.(4.21) and (4.22):

$$\mu_{a_h} = n\Sigma_{a_h} R^t \mu_{b_h}, \tag{4.29}$$

$$\Sigma_{a_h} = (n(\|\mu_{b_h}\|^2 + \mathrm{tr}(\Sigma_{b_h}))Q + c_a^{-2} I_M)^{-1}, \tag{4.30}$$

$$\mu_{b_h} = n\Sigma_{b_h} R \mu_{a_h}, \tag{4.31}$$

$$\Sigma_{b_h} = (n(\mu_{a_h}^t Q \mu_{a_h} + \mathrm{tr}(\Sigma_{a_h}^{1/2} Q \Sigma_{a_h}^{1/2})) + c_b^{-2})^{-1} I_N. \tag{4.32}$$

(For derivation, see the end of this subsection.) Then, we find from Eqs.(4.30), (4.32), and (4.26) that the covariance matrices can be written as follows:

$$\Sigma_{a_h} = \sigma_{a_h}^2 (I_M + O_p(n^{-1/2})), \tag{4.33}$$

$$\Sigma_{b_h} = \sigma_{b_h}^2 (1 + O_p(n^{-1/2})) I_N, \tag{4.34}$$

where $\sigma_{a_h}^2$ and $\sigma_{b_h}^2$ are the scalars corresponding to the variances. We also obtain the following form of the generalized free energy by substituting Eqs.(4.20), (4.23), and (4.24) into Eq.(4.17) and then using Eqs.(4.33) and (4.34):

$$\begin{aligned} 2\bar{F}(Y^n|X^n) = \sum_{h=1}^{H} \Big\{ &- \log \sigma_{a_h}^{2M} \sigma_{b_h}^{2N} + \frac{\|\mu_{a_h}\|^2 + M\sigma_{a_h}^2}{c_a^2} + \frac{\|\mu_{b_h}\|^2 + N\sigma_{b_h}^2}{c_b^2} \\ &- 2n\mu_{b_h}^t R \mu_{a_h} + n\big(\mu_{a_h}^t Q \mu_{a_h} + \mathrm{tr}(\Sigma_{a_h}^{1/2} Q \Sigma_{a_h}^{1/2})\big)\big(\|\mu_{b_h}\|^2 + \mathrm{tr}(\Sigma_{b_h})\big) \Big\} \\ &+ \sum_{i=1}^{n} \|y_i\|^2 + \text{const.}. \end{aligned} \tag{4.35}$$

(For derivation, see the following.)

The rest of this subsection is devoted to the derivation of the variational condition and the generalized free energy above.

Derivation of Eqs.(4.29)–(4.32)

By using Eqs.(4.23) and (4.24), we have

$$
\begin{aligned}
&\langle \log p(Y^n|X^n, A, B)\rangle_{r(A,B)/r(a_h)} + \text{const.} \\
&\quad = -\frac{1}{2}\sum_{i=1}^{n}\left\langle \left\| y_i - \textstyle\sum_{h=1}^{H} b_h a_h^t x_i \right\|^2 \right\rangle_{r(A,B)/r(a_h)} \\
&\quad = -\frac{1}{2}\sum_{i=1}^{n}\left\langle \left\| y_i - \textstyle\sum_{h=1}^{H} (b_h - \mu_{b_h})\, a_h^t x_i - \sum_{h=1}^{H} \mu_{b_h} a_h^t x_i \right\|^2 \right\rangle_{r(A,B)/r(a_h)} \\
&\quad = -\frac{1}{2}\sum_{i=1}^{n}\left\langle \left\| y_i - \textstyle\sum_{h=1}^{H} \mu_{b_h} a_h^t x_i \right\|^2 \right. \\
&\qquad\qquad \left. + \left(\textstyle\sum_{h=1}^{H}(b_h - \mu_{b_h}) a_h^t x_i\right)^t \left(\textstyle\sum_{h'=1}^{H}(b_{h'} - \mu_{b_{h'}}) a_{h'}^t x_i\right) \right\rangle_{r(A,B)/r(a_h)} \\
&\quad = -\frac{1}{2}\sum_{i=1}^{n}\left\langle \left\| y_i - \textstyle\sum_{h=1}^{H} \mu_{b_h} a_h^t x_i \right\|^2 + \textstyle\sum_{h=1}^{H} \|b_h - \mu_{b_h}\|^2 \left(a_h^t x_i\right)^2 \right\rangle_{r(A,B)/r(a_h)}.
\end{aligned}
\tag{4.36}
$$

Then, using Eqs.(4.26) and (4.27), and the orthogonality of $\{\mu_{b_h}\}$, we have

$$
\begin{aligned}
&\langle \log p(Y^n|X^n, A, B)\rangle_{r(A,B)/r(a_h)} + \text{const.} \\
&\qquad = -\frac{n}{2}\left(-2a_h^t R^t \mu_{b_h} + \left(\|\mu_{b_h}\|^2 + \text{tr}(\Sigma_{b_h})\right) a_h^t Q a_h\right). \qquad (4.37)
\end{aligned}
$$

Substituting Eq.(4.37) and the prior distribution, Eq.(4.14), into Eq.(4.21), we obtain the VB posterior distribution of a_h:

$$
\begin{aligned}
&\log r(a_h|X^n, Y^n) \\
&\quad = \log \phi(a) + \langle \log p(Y^n|X^n, A, B) \rangle_{r(A,B)/r(a_h)} + \text{const.} \\
&\quad = -\frac{n}{2} \left(-2a_h^t R^t \mu_{b_h} + a_h^t \left(\left(\|\mu_{b_h}\|^2 + \text{tr}(\Sigma_{b_h}) \right) Q + n^{-1} c_a^{-2} I_M \right) a_h \right) \\
&\qquad\qquad + \text{const.} \\
&\quad = -\frac{n}{2} \Big(a_h - \left(\left(\|\mu_{b_h}\|^2 + \text{tr}(\Sigma_{b_h}) \right) Q + n^{-1} c_a^{-2} I_M \right)^{-1} R\mu_{b_h} \Big)^t \\
&\qquad\qquad \left(\left(\|\mu_{b_h}\|^2 + \text{tr}(\Sigma_{b_h}) \right) Q + n^{-1} c_a^{-2} I_M \right) \\
&\qquad\qquad\quad \Big(a_h - \left(\left(\|\mu_{b_h}\|^2 + \text{tr}(\Sigma_{b_h}) \right) Q + n^{-1} c_a^{-2} I_M \right)^{-1} R\mu_{b_h} \Big) \\
&\qquad\qquad + \text{const.}. \qquad (4.38)
\end{aligned}
$$

In a similar fashion, we obtain the VB posterior distribution of b_h:

$$
\begin{aligned}
&\log r(b_h|X^n, Y^n) \\
&\quad = \log \phi(b) + \langle \log p(Y^n|X^n, A, B) \rangle_{r(A,B)/r(b_h)} + \text{const.} \\
&\quad = -\frac{n}{2} \Big(b_h - \left(\mu_{a_h}^t Q \mu_{a_h} + \text{tr} \left(\Sigma_{a_h}^{1/2} Q \Sigma_{a_h}^{1/2} \right) + n^{-1} c_b^{-2} \right)^{-1} R\mu_{a_h} \Big)^t \\
&\qquad\qquad \left(\mu_{a_h}^t Q \mu_{a_h} + \text{tr} \left(\Sigma_{a_h}^{1/2} Q \Sigma_{a_h}^{1/2} \right) + n^{-1} c_b^{-2} \right) I_N \\
&\qquad\qquad\quad \Big(b_h - \left(\mu_{a_h}^t Q \mu_{a_h} + \text{tr} \left(\Sigma_{a_h}^{1/2} Q \Sigma_{a_h}^{1/2} \right) + n^{-1} c_b^{-2} \right)^{-1} R\mu_{a_h} \Big) \\
&\qquad\qquad + \text{const.}. \qquad (4.39)
\end{aligned}
$$

Comparing Eqs.(4.38) and (4.39) with Eqs.(4.23) and (4.24), we obtain Eqs.(4.29)–(4.32).

Derivation of Eq.(4.35)

Substituting Eqs.(4.20), (4.23), and (4.24) into Eq.(4.17) and then using Eqs.(4.33) and (4.34), we have

$$
\begin{aligned}
&2\bar{F}(Y^n|X^n) \\
&\quad = 2\left\langle \log \frac{r(A)r(B)}{p(Y^n|X^n,A,B)\phi(A,B)} \right\rangle_{r(A)r(B)} \\
&\quad = 2\left\langle \textstyle\sum_{h=1}^H (\log r(a_h) + \log r(b_h)) - \log p(Y^n|X^n,A,B) \right\rangle_{r(A)r(B)} + \text{const.} \\
&\quad = -\sum_{h=1}^H \left(\log \sigma_{a_h}^{2M} + \log \sigma_{b_h}^{2N}\right) + \left\langle \sum_{h=1}^H \left(-\frac{\|a_h - \mu_{a_h}\|^2}{\sigma_{a_h}^2} - \frac{\|b_h - \mu_{b_h}\|^2}{\sigma_{b_h}^2} \right.\right. \\
&\qquad\qquad \left.\left. + \frac{\|a_h\|^2}{c_a^2} + \frac{\|b_h\|^2}{c_b^2} \right) + \sum_{i=1}^n \left\| y_i - \textstyle\sum_{h=1}^H b_h a_h^t x_i \right\|^2 \right\rangle_{r(A)r(B)} + \text{const.} \\
&\quad = \sum_{h=1}^H \left(-\log \sigma_{a_h}^{2M} - \log \sigma_{b_h}^{2N} + \frac{\|\mu_{a_h}\|^2 + M\sigma_{a_h}^2}{c_a^2} + \frac{\|\mu_{b_h}\|^2 + N\sigma_{b_h}^2}{c_b^2} \right) \\
&\qquad\qquad - H(M+N) + \sum_{i=1}^n \left\langle \left\| y_i - \textstyle\sum_{h=1}^H \mu_{b_h}\mu_{a_h}^t x_i \right\|^2 \right. \\
&\qquad\qquad\quad + \left\| \textstyle\sum_{h=1}^H \{(b_h - \mu_{b_h})(a_h - \mu_{a_h})^t x_i + \mu_{b_h}(a_h - \mu_{a_h})^t x_i \right. \\
&\qquad\qquad\qquad \left.\left. + (b_h - \mu_{b_h})\mu_{a_h}^t x_i \} \right\|^2 \right\rangle_{r(A)r(B)} + \text{const..}
\end{aligned}
\tag{4.40}
$$

Using Eqs.(4.26) and (4.27) and the orthogonality of $\{\mu_{b_h}\}$, we have

$$
\begin{aligned}
&2\bar{F}(Y^n|X^n) \\
&\quad = \sum_{h=1}^H \left(-\log \sigma_{a_h}^{2M} - \log \sigma_{b_h}^{2N} + \frac{\|\mu_{a_h}\|^2 + M\sigma_{a_h}^2}{c_a^2} + \frac{\|\mu_{b_h}\|^2 + N\sigma_{b_h}^2}{c_b^2} \right) \\
&\qquad + n\sum_{h=1}^H \left(-2\mu_{b_h}^t R \mu_{a_h} + \mu_{a_h}^t Q \mu_{a_h} \|\mu_{b_h}\|^2 + \text{tr}\left(\Sigma_{a_h}^{1/2} Q \Sigma_{a_h}^{1/2}\right) \text{tr}(\Sigma_{b_h}) \right. \\
&\qquad \left. + \text{tr}\left(\Sigma_{a_h}^{1/2} Q \Sigma_{a_h}^{1/2}\right) \|\mu_{b_h}\|^2 + \mu_{a_h}^t Q \mu_{a_h} \text{tr}\left(\Sigma_{b_h}\right) \right) + \sum_{i=1}^n \|y_i\|^2 + \text{const..}
\end{aligned}
\tag{4.41}
$$

Thus we obtain Eq.(4.35).

4.2.2 Variational Bayes Estimator

The variational condition, Eqs.(4.29)–(4.32), can be analytically solved, which leads to the following theorem:

Theorem 6 *Let*

$$\chi_{VB} = L \ (= \max(M, N)). \tag{4.42}$$

The VB estimator of the map of an LNN is given by

$$(\hat{B}\hat{A})_{VB} = \sum_{h=1}^{H} \left(1 - \frac{\chi_{VB}}{\max(\chi_{VB}, n\gamma_h^2)}\right) \omega_{b_h} \omega_{b_h}^t R Q^{-1} + O_p(n^{-1}). \tag{4.43}$$

(The proof is given in Section 4.2.4.)

4.2.3 Predictive Distribution

Similarly to in Section 3.2.2, expanding the predictive distribution divided by the true distribution, i.e., $p_{\rm VB}(y|x, X^n, Y^n)/q(y|x)$, we immediately obtain the following lemma:

Lemma 2 *The predictive distribution of an LNN in the VB approach can be written as follows:*

$$p_{VB}(y|x, X^n, Y^n) = \mathcal{N}_N(y; \hat{V}(\hat{B}\hat{A})_{VB}\, x, \hat{V}) + O_p(n^{-3/2}), \tag{4.44}$$

where $\hat{V} = I_N + O_p(n^{-1})$.

It can be said that the reason why Lemma 2 holds is because the independence between A and B makes the posterior distribution localized, as we can see in Fig. 4.1. By virtue of Lemma 2, we can substitute the model at the VB estimator for the VB predictive distribution with asymptotically insignificant impact upon generalization performance:

$$p_{\rm VB}(y|x, X^n, Y^n) \sim p_{\rm VB}(y|x; (\hat{B}\hat{A})_{\rm VB}). \tag{4.45}$$

Thus, we find that the VB approach is asymptotically equivalent to a positive-part JS type shrinkage estimation, like the SB approach. The VB approach, interestingly, results in a predictive distribution similar to that of the SB approach. The difference between the SB and the VB approaches is the degree of shrinkage. (Compare Eqs.(3.17) and (4.42).) The VB approach automatically selects the larger dimension, either the input one or the output one, as the degree of shrinkage, and therefore, it is asymptotically equivalent to the MIP version of SB approach when $M \geq N$, and to the MOP version when $M \leq N$. We will further discuss in Chapter 6 the coincidence between them, and the theoretical reason why the SB and the VB approaches provide good generalization performance.

Note that, in addition, the variances, c_a^2 and c_b^2, of the prior distributions, Eqs.(4.14) and (4.15), asymptotically have no effect upon prediction and hence upon generalization properties, as far as they are positive and finite constants. We say that the JS type *shrinkage* is caused by the singularities rather than by the prior distribution, and accordingly, it is considered to be the *Bayesian effect* of singularities, named in Section 1.1.4.[1]

4.2.4 Proof of Theorem 6

In this subsection, we prove Theorem 6. Both tr(Σ_{a_h}) and tr(Σ_{b_h}) are of no greater order than 1 because of the prior distributions, Eqs.(4.14) and (4.15). Both of them are also not equal to zero, otherwise the generalized free energy, Eq.(4.35), diverges to infinity with finite n. Consequently, the generalized free energy diverges to infinity when either $\|\mu_{a_h}\|$ or $\|\mu_{b_h}\|$ goes to infinity. Hence, the optimum values of μ_{a_h}, μ_{b_h}, Σ_{a_h}, and Σ_{b_h} necessarily satisfy the variational condition, Eqs.(4.29)–(4.32). Combining Eqs.(4.29) and (4.31), we have

$$\hat{\mu}_{a_h} = n^2 \hat{\Sigma}_{a_h} R^t \hat{\Sigma}_{b_h} R \hat{\mu}_{a_h}, \tag{4.46}$$

$$\hat{\mu}_{b_h} = n^2 \hat{\Sigma}_{b_h} R \hat{\Sigma}_{a_h} R^t \hat{\mu}_{b_h}, \tag{4.47}$$

and thus find that $\hat{\mu}_{a_h}$ and $\hat{\mu}_{b_h}$ are an eigenvector of $R^t R$ and an eigenvector of $R\hat{\Sigma}_{a_h} R^t$, respectively, or $\|\hat{\mu}_{a_h}\| = \|\hat{\mu}_{b_h}\| = 0$. Hereafter, separately considering

[1] However, we can interdict the *Bayesian effect* of singularities by using a prior distribution having zero value on the singularities. (See Section 6.3.2.)

the necessary components, which imitate the true ones with positive singular values, and the redundant ones, we will find the solution of the variational condition that minimizes the generalized free energy, Eq.(4.35).

For a necessary component, $h \leq H^*$, the (observed) singular value of R is of order $O_p(1)$. Hence, the free energy, Eq.(4.35), can be minimized only when both $\hat{\mu}_{a_h}$ and $\hat{\mu}_{b_h}$ are of order $O_p(1)$. Therefore, the variational condition, Eqs.(4.29)–(4.32), is approximated as follows:

$$\hat{\mu}_{a_h} = \|\hat{\mu}_{b_h}\|^{-2} Q^{-1} R^t \hat{\mu}_{b_h} + O_p(n^{-1}), \tag{4.48}$$

$$\hat{\Sigma}_{a_h} = n^{-1} \|\hat{\mu}_{b_h}\|^{-2} Q^{-1} + O_p(n^{-2}), \tag{4.49}$$

$$\hat{\mu}_{b_h} = (\hat{\mu}_{a_h}^t Q \hat{\mu}_{a_h})^{-1} R \hat{\mu}_{a_h} + O_p(n^{-1}), \tag{4.50}$$

$$\hat{\Sigma}_{b_h} = n^{-1} (\hat{\mu}_{a_h}^t Q \hat{\mu}_{a_h})^{-1} I_N + O_p(n^{-2}). \tag{4.51}$$

Thus, we obtain the VB estimator of the necessary component:

$$\hat{\mu}_{b_h} \hat{\mu}_{a_h}^t = \omega_{b_h} \omega_{b_h}^t R Q^{-1} + O_p(n^{-1}). \tag{4.52}$$

On the other hand, for a redundant component, $h > H^*$, Eqs.(4.28), (4.33), and (4.34) allow us to approximate the variational condition, Eqs.(4.29)–(4.32), as follows:

$$\hat{\mu}_{a_h} = n \hat{\sigma}_{a_h}^2 R^t \hat{\mu}_{b_h} (1 + O_p(n^{-1/2})), \tag{4.53}$$

$$\hat{\sigma}_{a_h}^2 = (n(\|\hat{\mu}_{b_h}\|^2 + N \hat{\sigma}_{b_h}^2) + c_a^{-2})^{-1}, \tag{4.54}$$

$$\hat{\mu}_{b_h} = n \hat{\sigma}_{b_h}^2 R \hat{\mu}_{a_h} (1 + O_p(n^{-1/2})), \tag{4.55}$$

$$\hat{\sigma}_{b_h}^2 = (n(\|\hat{\mu}_{a_h}\|^2 + M \hat{\sigma}_{a_h}^2) + c_b^{-2})^{-1}. \tag{4.56}$$

Combining Eqs.(4.54) and (4.56), we have

$$\hat{\sigma}_{a_h}^2 = \frac{-(n\hat{\eta}_h^2 - (M-N)) + \sqrt{(n\hat{\eta}_h^2 + M + N)^2 - 4MN}}{2nM(\|\hat{\mu}_{b_h}\|^2 + n^{-1} c_a^{-2})}, \tag{4.57}$$

$$\hat{\sigma}_{b_h}^2 = \frac{-(n\hat{\eta}_h^2 + (M-N)) + \sqrt{(n\hat{\eta}_h^2 + M + N)^2 - 4MN}}{2nN(\|\hat{\mu}_{a_h}\|^2 + n^{-1} c_b^{-2})}, \tag{4.58}$$

where

$$\hat{\eta}_h^2 = \left(\|\hat{\mu}_{a_h}\|^2 + \frac{c_b^{-2}}{n} \right) \left(\|\hat{\mu}_{b_h}\|^2 + \frac{c_a^{-2}}{n} \right). \tag{4.59}$$

We consider two cases of solutions:

1. Such that $\|\hat{\mu}_{a_h}\| = \|\hat{\mu}_{b_h}\| = 0$:

   Obviously, the following equations satisfy Eqs.(4.53) and (4.55):

   $$\hat{\mu}_{a_h} = 0, \tag{4.60}$$
   $$\hat{\mu}_{b_h} = 0. \tag{4.61}$$

   Substituting Eqs.(4.60) and (4.61) into Eqs.(4.57) and (4.58), we easily find the following solution:

   (a) When $M > N$:

   $$\hat{\sigma}^2_{a_h} = \frac{M-N}{Mc_a^{-2}} + O_p(n^{-1}), \tag{4.62}$$
   $$\hat{\sigma}^2_{b_h} = \frac{c_a^{-2}}{n(M-N)} + O_p(n^{-2}). \tag{4.63}$$

   (b) When $M = N$:

   $$\hat{\sigma}^2_{a_h} = \frac{c_a}{c_b\sqrt{nM}} + O_p(n^{-1}), \tag{4.64}$$
   $$\hat{\sigma}^2_{b_h} = \frac{c_b}{c_a\sqrt{nM}} + O_p(n^{-1}). \tag{4.65}$$

   (c) When $M < N$:

   $$\hat{\sigma}^2_{a_h} = \frac{c_b^{-2}}{n(N-M)} + O_p(n^{-2}), \tag{4.66}$$
   $$\hat{\sigma}^2_{b_h} = \frac{N-M}{Nc_b^{-2}} + O_p(n^{-1}). \tag{4.67}$$

2. Such that $\|\hat{\mu}_{a_h}\|, \|\hat{\mu}_{b_h}\| > 0$:

   Combining Eqs.(4.53) and (4.55), we find that $\hat{\mu}_{a_h}$ and $\hat{\mu}_{b_h}$ are the right and the corresponding left singular vectors of R, respectively. Let the h-th largest singular value component of R correspond to the h-th component of the estimator. Then we have

   $$\hat{\mu}_{a_h} = \|\hat{\mu}_{a_h}\|\omega_{a_h}, \tag{4.68}$$
   $$\hat{\mu}_{b_h} = \|\hat{\mu}_{b_h}\|\omega_{b_h}. \tag{4.69}$$

Substituting Eqs.(4.57), (4.58), (4.68), and (4.69) into Eqs.(4.53) and (4.55), we have the following equations:

$$\frac{2MN}{n\gamma_h^2} = n\hat{\eta}_h^2 + M + N - \sqrt{(n\hat{\eta}_h^2 + M + N)^2 - 4MN} + O_p(n^{-1/2}), \tag{4.70}$$

$$(Mc_a^{-2}\hat{\delta}_h - Nc_b^{-2}\hat{\delta}_h^{-1})(1 + O_p(n^{-1/2})) = n(M - N)(\gamma_h - \hat{\gamma}_h), \tag{4.71}$$

where

$$\hat{\gamma}_h = \|\hat{\mu}_{a_h}\|\|\hat{\mu}_{b_h}\|, \tag{4.72}$$

$$\hat{\delta}_h = \frac{\|\hat{\mu}_{a_h}\|}{\|\hat{\mu}_{b_h}\|}. \tag{4.73}$$

Equation (4.70) implies that $\xi = (n\gamma_h^2)^{-1}$ is a solution of the following quadratic equation:

$$MN\xi^2 - (n\hat{\eta}_h^2 + M + N)\xi + 1 = O_p(n^{-1/2}). \tag{4.74}$$

Therefore, we find that Eq.(4.70) has the solution below when and only when $n\gamma_h^2 \geq L$:

$$\hat{\eta}_h^2 = \left(1 - \frac{M}{n\gamma_h^2}\right)\left(1 - \frac{N}{n\gamma_h^2}\right)\gamma_h^2 + O_p(n^{-3/2}). \tag{4.75}$$

By using Eqs.(4.72) and (4.73), the definition of $\hat{\eta}_h$, Eq.(4.59), can be written as follows:

$$\hat{\eta}_h^2 = \left(1 + \frac{c_b^{-2}}{n\hat{\gamma}_h\hat{\delta}_h}\right)\left(1 + \frac{c_a^{-2}}{n\hat{\gamma}_h\hat{\delta}_h^{-1}}\right)\hat{\gamma}_h^2, \tag{4.76}$$

Hereafter, we will find the simultaneous solution of Eqs.(4.71), (4.75), and (4.76) in order to obtain the solution that exists only when $n\gamma_h^2 \geq L$. Equations (4.75) and (4.76) imply that $\gamma_h^2 > \hat{\eta}_h^2 > \hat{\gamma}_h^2$ and that $(\gamma_h^2 - \hat{\gamma}_h^2)$ is of order $O_p(n^{-1})$. We then find that $(\gamma_h - \hat{\gamma}_h)$ is of order $O_p(n^{-1/2})$ because γ_h is of order $O_p(n^{-1/2})$.

(a) When $M > N$:

In this case, Eq.(4.71) can be approximated as follows:

$$\hat{\delta}_h = \frac{n(M-N)(\gamma_h - \hat{\gamma}_h)}{Mc_a^{-2}} + O_p(n^{-1/2}). \tag{4.77}$$

Substituting Eqs.(4.75) and (4.77) into Eq.(4.76), we have

$$\hat{\gamma}_h^2 + \frac{(M-N)(\gamma_h - \hat{\gamma}_h)}{M}\hat{\gamma}_h - \left(1 - \frac{M}{n\gamma_h^2}\right)\left(1 - \frac{N}{n\gamma_h^2}\right)\gamma_h^2 = O_p(n^{-2}), \tag{4.78}$$

and hence

$$\left(\hat{\gamma}_h - \left(1 - \frac{M}{n\gamma_h^2}\right)\gamma_h\right)\left(\hat{\gamma}_h + \frac{M}{N}\left(1 - \frac{N}{n\gamma_h^2}\right)\gamma_h\right) = O_p(n^{-2}). \tag{4.79}$$

Therefore, we obtain the following solution with respect to $\hat{\gamma}_h$ and $\hat{\delta}_h$:

$$\hat{\gamma}_h = \left(1 - \frac{M}{n\gamma_h^2}\right)\gamma_h + O_p(n^{-3/2}), \tag{4.80}$$

$$\hat{\delta}_h = \frac{(M-N)}{c_a^{-2}\gamma_h} + O_p(n^{-1/2}). \tag{4.81}$$

We thus obtain the following solution:

$$\hat{\mu}_{a_h} = \left(\left(1 - \frac{M}{n\gamma_h^2}\right)\frac{M-N}{c_a^{-2}}\right)^{1/2}\omega_{a_h} + O_p(n^{-1}), \tag{4.82}$$

$$\hat{\sigma}_{a_h}^2 = \frac{M-N}{c_a^{-2}n\gamma_h^2} + O_p(n^{-1}), \tag{4.83}$$

$$\hat{\mu}_{b_h} = \left(\left(1 - \frac{M}{n\gamma_h^2}\right)\frac{c_a^{-2}}{M-N}\right)^{1/2}\gamma_h\omega_{b_h} + O_p(n^{-1}), \tag{4.84}$$

$$\hat{\sigma}_{b_h}^2 = \frac{c_a^{-2}}{n(M-N)} + O_p(n^{-2}). \tag{4.85}$$

(b) When $M = N$:

In this case, we find from Eq.(4.71) that

$$\hat{\delta}_h = \frac{c_a}{c_b}. \tag{4.86}$$

Substituting Eq.(4.86) and then Eq.(4.75) into Eq.(4.76), we have

$$\begin{aligned}\hat{\gamma}_h &= \hat{\eta}_h + O_p(n^{-1}) \\ &= \left(1 - \frac{M}{n\gamma_h^2}\right)\gamma_h + O_p(n^{-1}). \end{aligned} \tag{4.87}$$

We thus obtain the following solution:

$$\hat{\mu}_{a_h} = \left(\frac{c_a}{c_b}\left(1 - \frac{M}{n\gamma_h^2}\right)\gamma_h\right)^{1/2}\omega_{a_h} + O_p(n^{-1}), \tag{4.88}$$

$$\hat{\sigma}^2_{a_h} = \frac{c_a}{c_b n\gamma_h} + O_p(n^{-1}), \tag{4.89}$$

$$\hat{\mu}_{b_h} = \left(\frac{c_b}{c_a}\left(1 - \frac{M}{n\gamma_h^2}\right)\gamma_h\right)^{1/2}\omega_{b_h} + O_p(n^{-1}), \tag{4.90}$$

$$\hat{\sigma}^2_{b_h} = \frac{c_b}{c_a n\gamma_h} + O_p(n^{-1}). \tag{4.91}$$

(c) When $M < N$:

In exactly the same way as the case when $M > N$, we obtain the following solution:

$$\hat{\mu}_{a_h} = \left(\left(1 - \frac{N}{n\gamma_h^2}\right)\frac{c_b^{-2}}{N-M}\right)^{1/2}\gamma_h\omega_{a_h} + O_p(n^{-1}), \tag{4.92}$$

$$\hat{\sigma}^2_{a_h} = \frac{c_b^{-2}}{n(N-M)} + O_p(n^{-2}), \tag{4.93}$$

$$\hat{\mu}_{b_h} = \left(\left(1 - \frac{N}{n\gamma_h^2}\right)\frac{N-M}{c_b^{-2}}\right)^{1/2}\omega_{b_h} + O_p(n^{-1}), \tag{4.94}$$

$$\hat{\sigma}^2_{b_h} = \frac{N-M}{c_b^{-2}n\gamma_h^2} + O_p(n^{-1}). \tag{4.95}$$

We can find that, when the solution such that $\|\hat{\mu}_{a_h}\|, \|\hat{\mu}_{b_h}\| > 0$, Eqs.(4.82)–(4.85) and (4.88)–(4.95), exists, it makes the generalized free energy, Eq.(4.35), smaller than the other solution such that $\|\hat{\mu}_{a_h}\| = \|\hat{\mu}_{b_h}\| = 0$, Eqs.(4.60)–(4.67). Hence, letting

$$L'_h = \max(L, n\gamma_h^2),$$

we arrive at the following solution:

1. When $M > N$,

$$\hat{\mu}_{a_h} = \left(\left(1 - \frac{L}{L'_h} \right) \frac{L-l}{c_a^{-2}} \right)^{1/2} \omega_{a_h} + O_p(n^{-1}), \tag{4.96}$$

$$\hat{\sigma}^2_{a_h} = \frac{L-l}{c_a^{-2} L'_h} + O_p(n^{-1}), \tag{4.97}$$

$$\hat{\mu}_{b_h} = \left(\left(1 - \frac{L}{L'_h} \right) \frac{c_a^{-2}}{L-l} \right)^{1/2} \gamma_h \omega_{b_h} + O_p(n^{-1}), \tag{4.98}$$

$$\hat{\sigma}^2_{b_h} = \frac{c_a^{-2}}{n(L-l)} + O_p(n^{-2}). \tag{4.99}$$

2. When $M = N$,

$$\hat{\mu}_{a_h} = \left(\frac{c_a}{c_b} \left(1 - \frac{L}{L'_h} \right) \gamma_h \right)^{1/2} \omega_{a_h} + O_p(n^{-1}), \tag{4.100}$$

$$\hat{\sigma}^2_{a_h} = \frac{c_a}{c_b \sqrt{n L'_h}} + O_p(n^{-1}), \tag{4.101}$$

$$\hat{\mu}_{b_h} = \left(\frac{c_b}{c_a} \left(1 - \frac{L}{L'_h} \right) \gamma_h \right)^{1/2} \omega_{b_h} + O_p(n^{-1}), \tag{4.102}$$

$$\hat{\sigma}^2_{b_h} = \frac{c_b}{c_a \sqrt{n L'_h}} + O_p(n^{-1}). \tag{4.103}$$

3. When $M < N$,

$$\hat{\mu}_{a_h} = \left(\left(1 - \frac{L}{L'_h} \right) \frac{c_b^{-2}}{L-l} \right)^{1/2} \gamma_h \omega_{a_h} + O_p(n^{-1}), \tag{4.104}$$

$$\hat{\sigma}^2_{a_h} = \frac{c_b^{-2}}{n(L-l)} + O_p(n^{-2}), \tag{4.105}$$

$$\hat{\mu}_{b_h} = \left(\left(1 - \frac{L}{L'_h} \right) \frac{L-l}{c_b^{-2}} \right)^{1/2} \omega_{b_h} + O_p(n^{-1}), \tag{4.106}$$

$$\hat{\sigma}^2_{b_h} = \frac{L-l}{c_b^{-2} L'_h} + O_p(n^{-1}). \tag{4.107}$$

Selecting the largest singular value components minimizes the generalized free energy, Eq.(4.35). Hence, combining Eq.(4.52) with the fact that $L L_h'^{-1} = O_p(n^{-1})$ for the necessary components, and the solution above with Eq.(4.28), we obtain the VB estimator in Theorem 6. (Q.E.D.)

4.3 Generalization Properties

In Section 4.2, the VB approach has been proved to be asymptotically equivalent to the positive-part JS type shrinkage estimation whose generalization and training errors have already been clarified in Section 3.3.1 and in Section 3.3.2, respectively. Accordingly, we first describe the theorems of the VB generalization and training errors without proof in Section 4.3.1 and in Section 4.3.2, respectively. We also clarify the VB free energy, of which the theorem and its proof follow in Section 4.3.3. Finally, we illustrate numerical results with some figures in Section 4.3.4.

4.3.1 Generalization Error

Theorem 7 *The generalization error of an LNN in the VB approach can be asymptotically expanded as*

$$G_{VB}(n) = \frac{\lambda_{VB}}{n} + O(n^{-3/2}),$$

where the generalization coefficient is given by

$$2\lambda_{VB} = (H^*(L+l) - H^{*2}) + \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h'^2 > \chi_{VB}) \left(1 - \frac{\chi_{VB}}{\gamma_h'^2}\right)^2 \gamma_h'^2 \right\rangle_{q(\{\gamma_h'^2\})}. \tag{4.108}$$

Here, $\theta(\cdot)$ is the indicator function, and $\gamma_h'^2$ is the h-th largest eigenvalue of a random matrix subject to $\mathcal{W}_{l-H^*}(L-H^*, I_{l-H^*})$, over which $\langle\cdot\rangle_{q(\{\gamma_h'^2\})}$ denotes the expectation value.

Consider the large scale limit when M, N, H, and H^* go to infinity in the same order. Let

$$\alpha = (l - H^*)/(L - H^*), \tag{4.109}$$

$$\beta = (H - H^*)/(l - H^*), \tag{4.110}$$

$$\kappa = \chi_{\mathrm{VB}}/(L - H^*). \tag{4.111}$$

Theorem 8 *The VB generalization coefficient of an LNN in the large scale limit is given by*

$$2\lambda_{VB} \sim (H^*(L+l) - H^{*2}) + \frac{(L-H^*)(l-H^*)}{2\pi\alpha}\left\{J(s_t;1) - 2\kappa J(s_t;0) + \kappa^2 J(s_t;-1)\right\}, \tag{4.112}$$

where

$$J(s;1) = 2\alpha(-s\sqrt{1-s^2} + \cos^{-1} s),$$

$$J(s;0) = -2\sqrt{\alpha}\sqrt{1-s^2} + (1+\alpha)\cos^{-1}s - (1-\alpha)\cos^{-1}\frac{\sqrt{\alpha}(1+\alpha)s + 2\alpha}{2\alpha s + \sqrt{\alpha}(1+\alpha)},$$

$$J(s;-1) = \begin{cases} 2\sqrt{\alpha}\frac{\sqrt{1-s^2}}{2\sqrt{\alpha}s+1+\alpha} - \cos^{-1}s + \frac{1+\alpha}{1-\alpha}\cos^{-1}\frac{\sqrt{\alpha}(1+\alpha)s+2\alpha}{2\alpha s+\sqrt{\alpha}(1+\alpha)} & (0<\alpha<1) \\ 2\sqrt{\frac{1-s}{1+s}} - \cos^{-1} s & (\alpha = 1) \end{cases},$$

and

$$s_t = \max\left(\frac{\kappa - (1+\alpha)}{2\sqrt{\alpha}}, J^{-1}(2\pi\alpha\beta; 0)\right).$$

Here $J^{-1}(\cdot;k)$ denotes the inverse function of $J(s;k)$.

4.3.2 Training Error

Theorem 9 *The training error of an LNN in the VB approach can be asymptotically expanded as*

$$T_{VB}(n) = \frac{\nu_{VB}}{n} + O(n^{-3/2}),$$

where the training coefficient is given by

$$2\nu_{VB} = -(H^*(L+l) - H^{*2}) - \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h'^2 > \chi_{VB})\left(1 - \frac{\chi_{VB}}{\gamma_h'^2}\right)\left(1 + \frac{\chi_{VB}}{\gamma_h'^2}\right)\gamma_h'^2 \right\rangle_{q(\{\gamma_h'^2\})}. \tag{4.113}$$

Theorem 10 *The VB training coefficient of an LNN in the large scale limit is given by*

$$2\nu_{VB} \sim -(H^*(L+l) - H^{*2}) - \frac{(L-H^*)(l-H^*)}{2\pi\alpha}\left\{J(s_t;1) - \kappa^2 J(s_t;-1)\right\}. \tag{4.114}$$

4.3.3 Free Energy

We can obtain the normalized VB free energy by substituting the VB posterior distribution, derived in Section 4.2.4, into Eq.(4.35), where only the order of $\hat{\sigma}_{a_h}^2$ and that of $\hat{\sigma}_{b_h}^2$ signify because the leading term of the normalized free energy should be the first one in the curly braces of Eq.(4.35).

Theorem 11 *The normalized free energy of an LNN in the VB approach can be asymptotically expanded as*

$$F_{VB}(n) = \lambda'_{VB} \log n + O(1),$$

where the free energy coefficient is given by

$$2\lambda'_{VB} = H^*(L+l) + (H-H^*)l. \tag{4.115}$$

(Proof) We will separately consider the necessary components and the redundant components. For a necessary component, $h \leq H^*$, Eq.(4.52) implies that both $\hat{\mu}_{a_h}$ and $\hat{\mu}_{b_h}$ are of order $O_p(1)$. Then, we find from Eqs.(4.49) and (4.51) that both Σ_{a_h} and Σ_{b_h} are of order $O_p(n^{-1})$. Substituting the fact above into Eq.(4.35) leads to the first term of Eq.(4.115). On the other hand, for a redundant component, $h > H^*$, we find from the VB solution, Eqs.(4.96)–(4.107), that the orders of the estimators are as follows:

1. When $M > N$,

$$\hat{\mu}_{a_h} = O_p(1), \quad \hat{\sigma}_{a_h}^2 = O_p(1), \quad \hat{\mu}_{b_h} = O_p(n^{-1/2}), \quad \hat{\sigma}_{b_h}^2 = O_p(n^{-1}). \tag{4.116}$$

2. When $M = N$,

$$\hat{\mu}_{a_h} = O_p(n^{-1/4}), \ \ \hat{\sigma}_{a_h}^2 = O_p(n^{-1/2}), \ \ \hat{\mu}_{b_h} = O_p(n^{-1/4}), \ \ \hat{\sigma}_{b_h}^2 = O_p(n^{-1/2}). \tag{4.117}$$

3. When $M < N$,

$$\hat{\mu}_{a_h} = O_p(n^{-1/2}), \quad \hat{\sigma}_{a_h}^2 = O_p(n^{-1}), \quad \hat{\mu}_{b_h} = O_p(1), \quad \hat{\sigma}_{b_h}^2 = O_p(1). \tag{4.118}$$

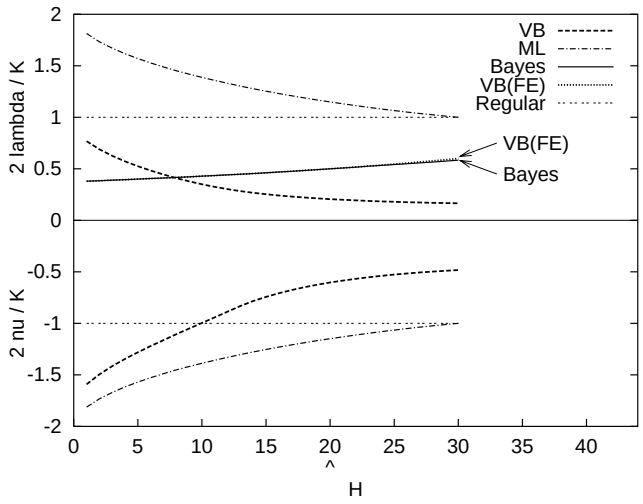

Figure 4.2: Generalization, training and free energy coefficients. ($L = 50$, $l = 30$, $H = 1, \ldots, 30$, and $H^* = 0$.)

Substituting Eqs.(4.116)–(4.118) into Eq.(4.35) leads to the second term of Eq.(4.115). (Q.E.D.)

Note that the contribution of the necessary components involves the trivial redundancy, because the independence between A and B prevents the VB posterior distribution from extending along the trivial degeneracy, which is a peculiarity of the LNNs. (Compare the first term of Eq.(4.115) with Eq.(2.38).)

4.3.4 Numerical Results

Now, we illustrate the generalization properties of the VB approach with some figures. Note that, if we assume that $M \geq N$, Figs 4.2, 4.3, and 4.5 are on the identical conditions to Figs 3.2, 3.3, and 3.5, respectively, and moreover, the generalization and the training coefficients of the VB approach coincide with those of the MIP version of SB approach. First, Fig. 4.2 shows the generalization, the training, and the free energy coefficients of the LNNs where $L = 50$ and $l = 30$ on the assumption that the true rank is equal to zero, $H^* = 0$. The horizontal axis indicates the rank of the learner, $H = 1, \ldots, 30$. The vertical axis indicates the

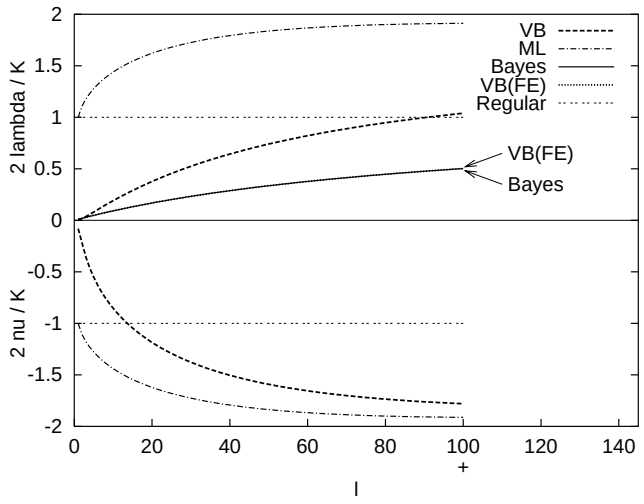

Figure 4.3: l dependence. ($L = 100$, $l = 1, \ldots, 100$, $H = 1$, and $H^* = 0$.)

coefficients normalized by the half of the *essential* parameter dimension K, given by Eq.(2.38).

The distinguishable lines in the positive region correspond to the generalization coefficient of the VB approach given by Theorem 8, that of the ML estimation given by Proposition 4, that of the Bayes estimation given by Proposition 5, and that of the regular models, respectively; while the lines in the negative region correspond to the training coefficient of the VB approach given by Theorem 10, that of the ML estimation given by Theorem 10 with the degree of shrinkage, $\chi_{\rm VB}$, set to zero, and that of the regular models, respectively. Actually, in Fig. 4.2, there is the line, labeled as VB(FE), corresponding to the VB free energy coefficient, given by Theorem 11. It almost coincides with the line corresponding to the Bayes generalization and the free energy coefficients, which are identical to each other. This shows that the VB free energy well approximates the Bayes free energy in this case; while the VB generalization error significantly differs from the Bayes generalization error.

Figure 4.3 similarly shows the coefficients of the LNNs where $L = 100$, $l = 1, \ldots, 100$, indicated by the horizontal axis, and $H = 1$ on the assumption that $H^* = 0$. Also in this case, the VB free energy well approximates the Bayes one, so

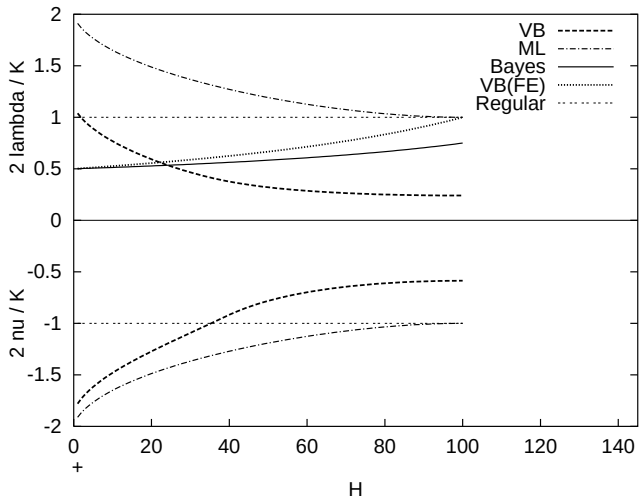

Figure 4.4: H dependence. ($L = l = 100$, $H = 1, \ldots, 100$, and $H^* = 0$.)

that we cannot distinguish it from the Bayes coefficient in Fig.4.3. Nevertheless, note that, when $l \sim 100$, the VB generalization coefficient exceeds that of the regular models, which never happens in the Bayes estimation.

On the other hand, the following two figures show the cases where the VB free energy does not well approximate: Fig. 4.4 shows the coefficients of the LNNs where $L = l = 100$, and $H = 1, \ldots, 100$, indicated by the horizontal axis, on the assumption that $H^* = 0$;[2] and Fig. 4.5 shows the true rank, $H^* = 1, \ldots, 20$, dependence of the coefficients of the LNN with $L = 50$, $l = 30$, and $H = 20$ units.[3] We conclude that the VB free energy behaves similarly to the Bayes one in general, and well approximates when $l/L \ll 1$ or $H \ll l$. Exceptionally, we see in Fig. 4.5 that the VB free energy strangely behaves and poorly approximates the Bayes one when H^* is large, which can, however, be regarded as a peculiarity of the LNNs, which have the trivial redundancy, rather than as a property of the VB approach in general singular models. (See the last paragraph of Section 4.3.3.)

We see in Figs. 4.2–4.5 that the VB generalization coefficient significantly

[2]Note that the case that $H = 1$ in Fig. 4.4 is the same as the case that $l = 100$ in Fig.4.3, both of which are marked with $+$.

[3]Note that the case that $H^* = 0$ in Fig. 4.5 is the same as the case that $H = 20$ in Fig.4.2, both of which are marked with $\hat{}$.

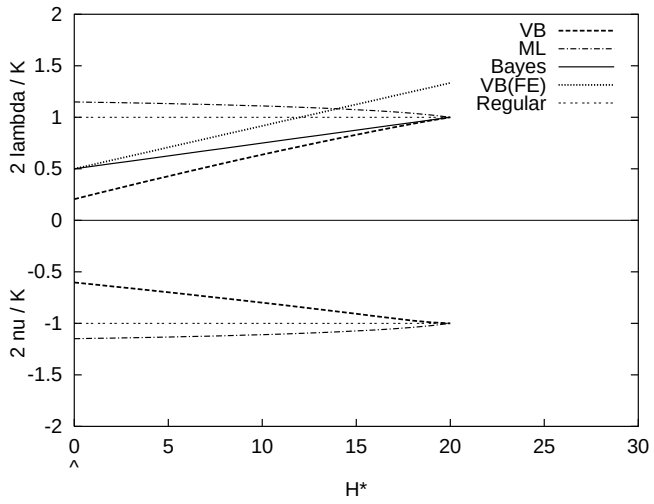

Figure 4.5: True rank dependence. ($L = 50$, $l = 30$, $H = 20$, and $H^* = 1, \ldots, 20$.)

differs from the Bayes one even in the case when the VB free energy well approximates the Bayes one. Furthermore, Fig. 4.4 shows the following facts: that, when $H \ll l$, the VB free energy well approximates the Bayes one, while the VB approach provides much worse generalization performance than the Bayes estimation; and that, when $H \sim l$, the VB free energy is significantly larger than the Bayes one, while the VB approach provides much better generalization performance than the Bayes estimation. These can throw a doubt on the model selection by minimizing the VB free energy to obtain better generalization performance.

The following summarize the properties of the VB approach that we have found in this chapter, the first two of which are common to the SB approach:

1. Generally speaking, the VB approach provides good generalization performance comparable to the Bayes estimation. In addition, the VB approach is better than the Bayes estimation in some cases.

2. The dependence of the VB generalization coefficient on the rank, H, is similar to that of the ML generalization coefficient.

3. The VB free energy well approximates the Bayes one when $l/L \ll 1$ or

$H \ll l$. Exceptionally, the VB free energy behaves strangely when H^* is large, because of the trivial redundancy, a peculiarity of the LNNs.

4. The gap in the free energies of the VB approach and the Bayes estimation behaves differently from the gap in the generalization errors of them.

Further discussion of the features of the VB approach will be in Section 6.2.

Chapter 5

Delicate Situations

In ordinary asymptotic analysis, one considers only the situations when the amplitude of each component of the true model is zero or *distinctly-positive*. The propositions for the maximum likelihood (ML) estimation described in Section 2.3.2, that for the Bayes estimation described in Section 2.3.3, and the theorems for the SB and the VB approaches proved in Chapters 3 and 4, respectively, also hold only in such situations. As a result, we have seen some results that seem to be inconsistent with the superiority of the Bayes estimation, Proposition 1. This chapter is devoted to consideration of the *delicate* situations when the true map B^*A^* has tiny but non-negligible singular values such that $0 < \sqrt{n}\gamma_h^* < \infty$. As mentioned in the last paragraph of Section 1.1.4 and in Section 2.5, the *delicate* situations are important in model selection problems and statistical tests, and moreover, needed to be considered when we discuss the domination. First, in Section 5.1, we discuss the admissibility of the Bayes estimation, derived from its superiority, and discuss its consistency with our results, derived in Chapters 3 and 4. Then, in Section 5.2, we derive the theorems of the generalization error and the training error of the SB and the VB approaches that apply to the *delicate* situations. At the end of this chapter, focusing on single-output (SO) linear neural networks (LNNs), we discuss the domination of the Bayesian learning methods over the ML estimation in Section 5.3.

5.1 Admissibility of Bayes Estimation

As mentioned in Remark 2 in Section 2.1.1, it is not inconsistent with the superiority of the Bayes estimation that there is any learning method providing good generalization performance than the Bayes estimation in some situations, because we consider the case when we do not know the *true* prior distribution. However, if there would be any learning method dominating the Bayes estimation, it would obviously be inconsistent with the superiority of the Bayes estimation, Proposition 1. Hence, we can say that:

Proposition 7 *The Bayes estimation is admissible.*

Note that the words, *domination* and *admissibility*, are defined by Definitions 4 and 5, respectively, at the very end of Chapter 2.6. Nevertheless, Figs. 3.5 and 4.5 show that the MIP version of SB approach and the VB approach provide no worse generalization performance than the Bayes estimation regardless of the true rank, H^*. We show in the following that consideration of the *delicate* situations consistently denies the domination.

5.2 Generalization Error and Training Error

Theorem 1, as well as Theorem 6, still holds in the *delicate* situations when the true map B^*A^* has singular values such that $0 < \sqrt{n}\gamma_h^* < \infty$ if we replace the second term of Eq.(3.18), as well as that of Eq.(4.43), with $o_p(n^{-1/2})$. We regard H^* as the number of *distinctly-positive* true singular values such that $\gamma_h^{*-1} = o(\sqrt{n})$.

5.2.1 Theorems

We denote by $\mathcal{W}_d(m, \Sigma, \Lambda)$ the d-dimensional (non-central) Wishart distribution with m degrees of freedom, scale matrix Σ, and noncentrality matrix Λ. Without loss of generality, we assume that B^*A^* is a non-negative, general diagonal matrix with its diagonal elements arranged in non-increasing order. Let R''^* be the true submatrix created by removing the first H^* columns and rows from B^*A^*. We

say that U is the general diagonalized matrix of an $N \times M$ matrix T if T has the following singular value decomposition: $T = \Omega_b U \Omega_a$, where Ω_a and Ω_b are an $M \times M$ and an $N \times N$ orthogonal matrices, respectively. Let D be the general diagonalized matrix of R, and D' the $(N - H^*) \times (M - H^*)$ matrix created by removing the first H^* columns and rows from D. Then, the first H^* diagonal elements of D correspond to the *distinctly* positive true singular value components, and D' consists of the *delicate* true components and noises. Therefore, D' is the general diagonalized matrix of $n^{-1/2}R''$, where R'' is a random matrix such that $R''R''^t$ is subject to $\mathcal{W}_{N-H^*}(M - H^*, I_{N-H^*}, nR''^*R''^{*t})$. Hence, we obtain the following theorem:

Theorem 12 *Let*

$$\chi = \begin{cases} \chi_{SB} & (\textit{in the SB approach}) \\ \chi_{VB} & (\textit{in the VB approach}) \end{cases}. \tag{5.1}$$

The SB, as well as the VB, generalization coefficient of an LNN in the general situations when the true map B^*A^* may have delicate singular values such that $0 < \sqrt{n}\gamma_h^* < \infty$ is given by

$$2\lambda = (H^*(L+l) - H^{*2}) + \sum_{h=H^*+1}^{H} n\gamma_h^{*2} + \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h''^2 > L) \left\{ \left(1 - \frac{\chi}{\gamma_h''^2}\right)^2 \gamma_h''^2 - 2\left(1 - \frac{\chi}{\gamma_h''^2}\right) \gamma_h'' \omega_{b_h}''^t \sqrt{n} R''^* \omega_{a_h}'' \right\} \right\rangle_{q(R'')}, \tag{5.2}$$

where γ_h'', ω_{a_h}'', and ω_{b_h}'' are the h-th largest singular value of R'', the corresponding right singular vector, and the corresponding left singular vector, respectively, of which $\langle \cdot \rangle_{q(R'')}$ denotes the expectation value over the distribution.

Similarly, we obtain the following theorem of the training error.

Theorem 13 *The SB, as well as the VB, training coefficient of an LNN in the general situations when the true map B^*A^* may have delicate singular values*

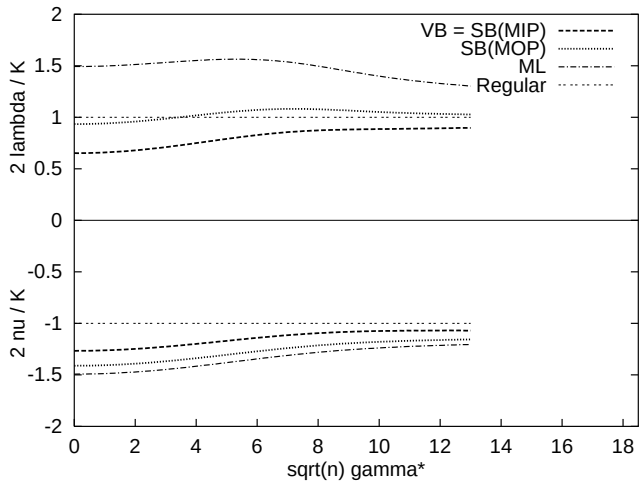

Figure 5.1: *Delicate* situations.

such that $0 < \sqrt{n}\gamma_h^* < \infty$ *is given by*

$$2\nu = -(H^*(L+l) - H^{*2}) + \sum_{h=H^*+1}^{H} n\gamma_h^{*2} - \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h''^2 > L)\left(1 - \frac{\chi}{\gamma_h''^2}\right)\left(1 + \frac{\chi}{\gamma_h''^2}\right)\gamma_h''^2 \right\rangle_{q(R'')}. \quad (5.3)$$

5.2.2 Numerical Results

Using Theorems 12 and 13, we can numerically calculate the SB, as well as the VB, generalization and training coefficients in the *delicate* situations. Also those of the ML estimation are given by setting $\chi = 0$ in those theorems. Figure 5.1 shows the coefficients of the LNN with $M = 50$ input, $N = 30$ output, and $H = 5$ hidden units, on the assumption that the true map consists of $H^* = 1$ *distinctly-positive* component, three *delicate* components whose singular values are identical to each other, and the other one null component. The horizontal axis indicates $\sqrt{n}\gamma^*$, where $\gamma_h^* = \gamma^*$ for $h = 2, \ldots, 4$. Note that, even in the ML estimation, the generalization error and the training error are asymmetrical with each other in the *delicate* situations.

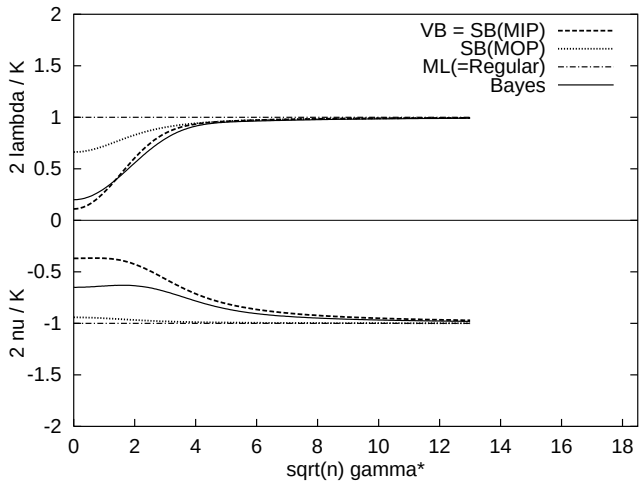

Figure 5.2: *Delicate* situations in SOLNN.

The Bayes generalization error in the *delicate* situations is given by Proposition 6, but unfortunately, only in single-output (SO) LNNs, i.e., $N = H = 1$. Figure 5.2 shows the coefficients of the SOLNN with $M = 5$ input units, on the assumption that $H^* = 0$ and the one true singular value, indicated by the horizontal axis, is *delicate*.[1] We see that, in some *delicate* situations, the MIP of SB approach and the VB approach provide worse generalization performance than the Bayes estimation, although this is the case that they can seem to dominate the Bayes estimation without consideration of the *delicate* situations. Since the maps of the SOLNNs are full-rank, the SOLNNs are considered to be regular from the viewpoint of the ML estimation, as mentioned in Section 2.3.1. Accordingly, the ML generalization coefficient is identical to that of the regular models, as we see in Fig. 5.2. Nevertheless, Fig. 5.2 shows that the SOLNNs have a property of singular models, i.e., James-Stein (JS) type *shrinkage*, from the viewpoint of the Bayesian learning methods.

The domination of the SB, as well as the VB, approach over the Bayes estimation has been denied in the SOLNNs, in the discussion above. Also in the case that

[1]The training error of the Bayes estimation in the *delicate* situations was also derived in [Watanabe and Amari, 2003], although we omitted it from Section 2.5.

$N > 1$, we expect similar results denying the domination. We conclude that, in typical cases, suppression of overfitting, i.e., the *Bayesian effect* of singularities, in the SB, as well as the VB, approach is comparable to, and sometimes stronger than, that in the Bayes estimation.

5.3 Asymptotic Domination over ML Estimation

In Fig. 5.2, all the Bayesian learning methods, i.e., the SB approaches, the VB approach, and the Bayes estimation seem to dominate the ML estimation. In this section, focusing on the SOLNNs, defined by Eq.(2.60), we discuss the domination of the SB and the VB approaches over the ML estimation. Let $\gamma = b\|a\|$. It was suggested that the following asymptotic expansion of the generalization coefficient with respect to $\sqrt{n}\gamma^*$ provides a clue, not a proof though, to find when the domination occurs [Watanabe and Amari, 2003]:

$$2\lambda = M - \frac{\xi}{(\sqrt{n}\gamma^*)^2} + o\left(\frac{1}{(\sqrt{n}\gamma^*)^2}\right), \tag{5.4}$$

where ξ is the coefficient of the leading term when γ^* increases to be *distinctly-positive*. The sign of ξ indicates the direction of approach to the line $2\lambda = M$, which corresponds to the generalization coefficient of the ML estimation. In the Bayes estimation, it was shown that

$$\xi_{\text{Bayes}} = (M-1)(M-3), \tag{5.5}$$

which leads to the conjecture that the Bayes estimation would dominate the ML estimation when $M \geq 4$ [Watanabe and Amari, 2003]. Figure 2.4 seems to support the conjecture.

Now we consider the SB and the VB approaches. Let $a'^* = \sqrt{n}b^*a^*$ be an M-dimensional vector, so that $\|a'^*\|^2 = n\gamma^{*2}$. Then, Eq.(5.2) can be asymptotically

expanded when $\sqrt{n}\gamma^*$ goes to infinity as follows:

$$
\begin{aligned}
2\lambda &= \|a'^*\|^2 + \left\langle \left(1 - \frac{\chi}{\|a'^* + g\|^2}\right)^2 \|a'^* + g\|^2 \right. \\
&\qquad \left. -2\left(1 - \frac{\chi}{\|a'^* + g\|^2}\right) a'^{*t}(a'^* + g) \right\rangle_{q(g)} + o\left(\frac{1}{\|a'^*\|^2}\right) \\
&= \left\langle \|g\|^2 + \frac{1}{\|a'^*\|^2}\left\{\chi^2 - 2\chi\|g\|^2 + 8\chi\frac{(a'^{*t}g)^2}{\|a'^*\|^2}\right.\right. \\
&\qquad \left.\left. -4\chi\frac{(a'^{*t}g)^2}{\|a'^*\|^2}\right\} \right\rangle_{q(g)} + o\left(\frac{1}{\|a'^*\|^2}\right), \qquad (5.6)
\end{aligned}
$$

where g is a random variable subject to $\mathcal{N}_M(0, I_M)$, over which $\langle\cdot\rangle_{q(g)}$ denotes the expectation value. Because $\langle\|g\|^2\rangle_{q(g)} = M$ and $\langle(a'^{*t}g)^2\rangle_{q(g)} = \|a'\|^2$, we have

$$
2\lambda = M - \frac{\chi(2M - \chi - 4)}{\|a'^*\|^2} + o\left(\frac{1}{\|a'^*\|^2}\right). \qquad (5.7)
$$

Comparing Eqs.(5.4) and (5.7), we find that

$$
\xi = \chi(2M - \chi - 4). \qquad (5.8)
$$

Thus, we have

$$
\xi_{\text{MIP}} = \xi_{\text{VB}} = M(M - 4), \qquad (5.9)
$$

which leads to the conjecture that the MIP version of SB approach and the VB approach, which are asymptotically equivalent to each other in the SOLNNs, would dominate the ML estimation when $M \geq 5$; while we have

$$
\xi_{\text{MOP}} = (2M - 5), \qquad (5.10)
$$

which leads to the conjecture that the MOP version of SB approach would dominate the ML estimation when $M \geq 3$.

In the same format as Fig. 5.2, Figs. 5.3 and 5.4 show the input dimension, M, dependence of the generalization and the training coefficients in SOLNNs, numerically calculated by using Theorems 12 and 13. Figure 5.3 shows the coefficients

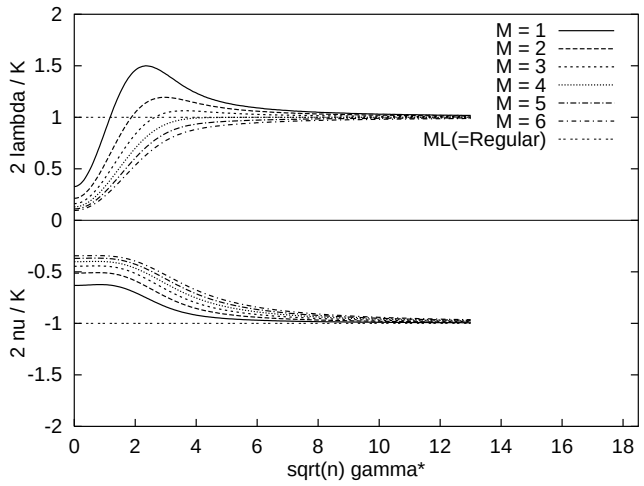

Figure 5.3: Domination of MIP and VB over ML.

of the MIP and the VB approach; while Fig. 5.4 shows those of the MOP. Those figures seem to support the conjectures above. We also find from Eq.(5.8) that

$$\xi_{\rm JS} = (M-2)^2 \tag{5.11}$$

in the JS estimation,[2] i.e., the case that $\chi = (M-2)$, which is consistent with its proved domination over the ML estimation when $M \geq 3$ [James and Stein, 1961].

The training coefficient can be also asymptotically expanded as follows:

$$2\nu = -M + \frac{\iota}{(\sqrt{n}\gamma^*)^2} + o\left(\frac{1}{(\sqrt{n}\gamma^*)^2}\right), \tag{5.12}$$

where ι is the leading coefficient. It was found that

$$\iota_{\rm Bayes} = (M-1)^2, \tag{5.13}$$

in the Bayes estimation [Watanabe and Amari, 2003]. Expanding Eq.(5.3), we have

$$\iota = \chi^2 \tag{5.14}$$

in the SB and the VB approaches, as well as in the JS type estimation.

[2]The indicator function of the positive-part JS type estimation does not affect ξ.

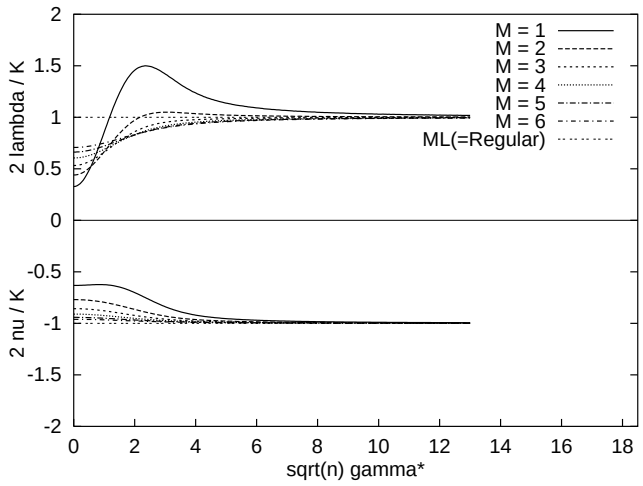

Figure 5.4: Domination of MOP over ML.

In addition, we find that ξ and ι in the Bayes estimation are given by letting $\chi = (M - 3)$ in Eq.(5.8) and by letting $\chi = (M - 1)$ in Eq.(5.14), respectively. It does not seem to be trivial that ξ and ι can, even in the Bayes estimation, be expressed by the forms of Eqs.(5.8) and (5.14), respectively, of which consideration is future work.

Chapter 6

Discussion

In this chapter, we discuss the results that we have shown in the previous chapters. In Section 6.1, we first consider the relation among the James-Stein (JS) shrinkage estimation, the subspace Bayes (SB) approach, and the variational Bayes (VB) approach. Then, we mention the relation between linear neural networks (LNNs) and the automatic relevance determination models (ARDs), which are also useful in real applications. After that, we discuss the results of this thesis from the viewpoint of statistical physics. In Section 6.2, we discuss the features of the SB and the VB approaches, and consider when and why the approximation methods perform well or not. We also consider what the results of this thesis imply for general singular models. At the end of this chapter, based on the discussion in Section 6.1 and Section 6.2, we make some suggestions on using the Bayesian learning methods.

6.1 Relations

6.1.1 JS Estimation and SB Approach

It was shown that the JS estimator is derived as an empirical Bayes (EB) estimator in multi-dimensional mean estimation [Efron and Morris, 1973]. (See Appendix A.2.) Now, we consider the following linear regression model, which is a

simple regular model:

$$p(y|x,w) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{\|y - w^t x\|^2}{2}\right), \tag{6.1}$$

where $x \in \mathbb{R}^M$ is an input, $y \in \mathbb{R}$ is an output, and $w \in \mathbb{R}^M$ is a parameter. Suppose that we perform the EB approach, using the following prior distribution:

$$\phi(w\|b) = \frac{1}{(2\pi b^2)^{M/2}} \exp\left(-\frac{\|w\|^2}{2b^2}\right), \tag{6.2}$$

where $b \in \mathbb{R}$ is a hyperparameter. Since the mean estimation can be considered to be the special case of regression where the input can be controlled, the EB approach above results in the JS type estimation. On the other hand, the transform $w \mapsto ba$, where $a \in \mathbb{R}^M$, makes the model and the prior distribution, Eqs. (6.1) and (6.2), equal to those of a single-output (SO) LNN, which has been analyzed in this thesis:

$$p(y|x,a\|b) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{\|y - ba^t x\|^2}{2}\right), \tag{6.3}$$

$$\phi(a) = \frac{1}{(2\pi)^{M/2}} \exp\left(-\frac{\|a\|^2}{2}\right). \tag{6.4}$$

Therefore, the asymptotic equivalence, proved in Chapter 3, between the JS estimation and the SB approach is natural. We can say that the parameters of one layer of a singular model regarded as the hyperparameters in the SB approach play the same role as the deviation hyperparameters of the prior distribution in the EB approach.

6.1.2 SB Approach and VB Approach

In Chapter 4, we have proved that the VB approach is asymptotically equivalent to the MIP version of SB approach when $M \geq N$, and to the MOP version when $M \leq N$. To consider the reason, we focus on the extent of the posterior distribution.

In the MIP version, we find from Eqs.(3.29), (3.33), (3.35), and (3.39) that the

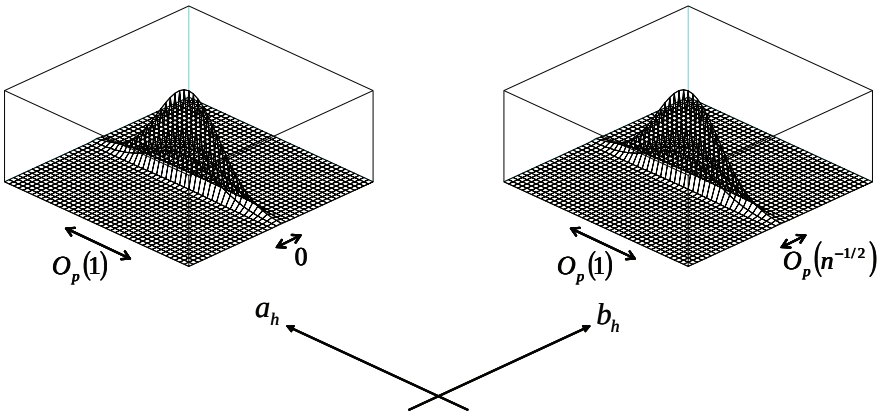

Figure 6.1: SB (MIP) posterior (left-hand side) and VB posterior (right-hand side) for $H^* < h \leq H$ when $M > N$.

SB posterior distribution of a_h extends with its variance of the following order:

$$\begin{aligned}\left(\hat{\sigma}^2_{a_h}\right)_{\text{MIP}} &= O_p(n^{-1}\|b_h\|^{-2}) \\ &= \begin{cases} O_p(n^{-1}) & \text{for } 1 \leq h \leq H^* \\ O_p(1) & \text{for } H^* < h \leq H \end{cases} .\end{aligned} \tag{6.5}$$

We can regard the SB approach as an approximation method with the trial distribution restricted such that the distribution of the parameters of one layer is the delta function. So, we can formally write

$$\left(\hat{\sigma}^2_{b_h}\right)_{\text{MIP}} = 0. \tag{6.6}$$

On the other hand, we can find from Eqs. (4.49), (4.51), (4.52), and (4.116) that the variances of the VB posterior distributions of a_h and b_h are of the following order when $M > N$:

$$\left(\hat{\sigma}^2_{a_h}\right)_{\text{VB}} = \begin{cases} O_p(n^{-1}) & \text{for } 1 \leq h \leq H^* \\ O_p(1) & \text{for } H^* < h \leq H \end{cases} , \tag{6.7}$$

$$\left(\hat{\sigma}^2_{b_h}\right)_{\text{VB}} = O_p(n^{-1}). \tag{6.8}$$

Figure 6.1 shows the (virtual) SB posterior distribution (left-hand side), and the VB posterior distribution (right-hand side). We can expect that they converge to

similar distributions in the asymptotic limit.[1] Therefore, the asymptotic equivalence between the MIP and the VB approach is also natural.

The SB and the VB approaches have a difference. In the SB approach, the marginalized space is intentionally selected by choosing one of the two versions. On the other hand, we find from Eqs. (4.116)–(4.118) and Eq.(4.42) that the VB approach automatically selects the larger dimension space, either the input one or the output one, as the marginalized space, and consequently, the larger dimension, L, as the degree of shrinkage, χ. The dimension of the space in which the singularities allow the posterior distribution to extend is essentially related to the degree of shrinkage, and hence, to the magnitude of suppression of overfitting. So, we can say that the JS type *shrinkage* is the *Bayesian effect*, i.e., the attraction of the singularities caused by their large state density, explained in Section 1.1.4. (See also Section 6.1.4.)

We have, finally, revealed a link from the JS type shrinkage estimation, a classic in statistics, to the VB approach, a rising method in the field of machine learning, via the EB approach and the SB approach.

6.1.3 Automatic Relevance Determination and LNN

The automatic relevance determination model (ARD) [MacKay, 1994; Neal, 1996] is closely related with this thesis, because the ARD based on a linear model is similar to an LNN. Recently, the VB approach has been applied to that model, and showed good performance in a real application of brain current estimation [Sato *et al.*, 2004]. Hereafter, we shortly introduce the ARD in a linear model, and then discuss its relation to the LNN.

The ARD was proposed to eliminate irrelevant connections from networks [MacKay, 1994; Neal, 1996]. Although the ARD was proposed originally in learning of neural networks, we here consider its application to a linear model where one estimates the electric current distribution in a brain from the magnetic fields that are induced by the current and observed on the head. [Sato *et al.*, 2004]. For simplicity, we assume that the current and the field are scalar. Let $a' \in \mathbb{R}^M$ be

[1] Note that the objective function of the SB approach and that of the VB approach, however, are not exactly the same as each other.

the current vector of which each element corresponds to the current value at each site in a brain, and $y \in \mathbb{R}^N$ the field vector of which each element corresponds to the field value at each site on the head. Then, the linear regression model is given by

$$y = Va' + \varepsilon, \tag{6.9}$$

where V is the $N \times M$ matrix, called the lead field matrix, that represents the field induced by the current, and ε is a noise. Assume that the noise is independently subject to $\mathcal{N}_N(0, I_N)$. Then, the probability density of the field is given by

$$p(y|a') \propto \exp\left(-\frac{\|y - Va'\|^2}{2}\right). \tag{6.10}$$

We use the following prior distribution of a':

$$\phi(a'\|B^2) \propto \exp\left(-\frac{a'^t B^{-2} a'}{2}\right), \tag{6.11}$$

where the $M \times M$ matrix B^2 is the hyperparameter corresponding to the covariance matrix. In the simplest ARD, we assume that B^2 is diagonal. Then, reducing the (m, m)-th element of B^2 eliminates the m-th element of the current, a', as an irrelevant one. If we consider B^2 to be constant, the model above is just a regular one and its asymptotic generalization performance will not improve. However, if we estimate B^2 by maximizing the free energy, the generalization performance will significantly improve. We can estimate the value of B^2 based on the EB approach, and also can estimate the distribution of B^2 by introducing the prior distribution of the hyperparameter, called the hyperprior distribution. To the latter method, we can apply the VB approach, assuming the posterior distribution with the independence between a' and B^2. In a real application, the VB approach in that model showed good performance [Sato *et al.*, 2004].

Let $x' \in \mathbb{R}^M$ be the formal input vector of which all the elements are equal to one. Then, similarly to in Section 6.1.1, the transform

$$a' \to Ba, \tag{6.12}$$

where $a \in \mathbb{R}^M$, makes the model distribution and the prior distribution like those of an LNN, as follows:

$$p(y|x', A\|B) \propto \exp\left(-\frac{\|y - VBAx'\|^2}{2}\right), \tag{6.13}$$

$$\phi(A) \propto \exp\left(-\frac{\mathrm{tr}(A^2)}{2}\right), \tag{6.14}$$

where A is the $M \times M$ diagonal matrix whose (m, m)-th element is equal to the m-th element of a. So, we expect that some results of this thesis apply also to that application. However, the existence of the lead field matrix, V in Eq. (6.13), makes difference, although the ARD is equivalent to an LNN when V is general diagonal. We do not like to transform the basis of the current vector, a', so that V is general diagonal, since the purpose of that application is to find the few sites where synapses fire. Actually, we have just found a certain relation between a simple case of the ARD and the JS type estimation [Nakajima and Watanabe, 2006a], and further consideration is future work.

6.1.4 From Viewpoint of Statistical Physics

The relation between the Bayesian learning and statistical physics is well-known. The Bayes posterior distribution, Eq.(2.2), can be written in the form of the Boltzmann distribution as follows:

$$p(w|X^n, Y^n) = \frac{1}{Z(Y^n|X^n)} \exp\left(-nE(w; X^n, Y^n)\right), \tag{6.15}$$

where the factor

$$E(w; X^n, Y^n) = -\frac{1}{n} \log\left(\phi(w) \prod_{i=1}^{n} p(y_i|x_i, w)\right) \tag{6.16}$$

corresponds to the Hamiltonian or energy, the marginal likelihood,

$$Z(Y^n|X^n) = \int \phi(w) \prod_{i=1}^{n} p(y_i|x_i, w) dw, \tag{6.17}$$

corresponds to the partition function, and the number of training samples, n, corresponds to the inverse temperature. We have mentioned in the first paragraph of

Section 1.1.5 that the Markov chain Monte Carlo (MCMC) methods can, therefore, be utilized for approximation of the Bayes posterior distribution.

Now, we discuss the VB approach. The generalized free energy, Eq.(2.23), for an arbitrary distribution $r(w)$ can be written as follows:

$$\bar{F}(r) = -S(r) + n\bar{E}(r), \tag{6.18}$$

where

$$S(r) = -\langle \log r(w) \rangle_{r(w)} \tag{6.19}$$

is the entropy of $r(w)$, and

$$\bar{E}(r) = \langle E(w; X^n, Y^n) \rangle_{r(w)} \tag{6.20}$$

is the average energy over $r(w)$. Therefore, $n^{-1}\bar{F}(r)$ corresponds to the Helmholtz free energy:

$$F_{\text{Helmholtz}}(r) = n^{-1}\bar{F}(r) = -TS(r) + \bar{E}(r), \tag{6.21}$$

where $T = n^{-1}$ denotes the temperature.

In statistical physics, it is known that the equilibrium distribution given the temperature, T, minimizes the Helmholtz free energy. Therefore, the Bayes posterior distribution, which minimizes the generalized free energy, Eq.(6.18), corresponds to the equilibrium distribution given the number of training samples, n, as well as their values, which determine the (random) energy, Eq.(6.16). Physicists sometimes face the case that minimizing the free energy is mathematically difficult. In such a case, they utilize the mean field approximation, where they assume that the equilibrium distribution factorizes, in exactly the same way as in the VB approach. That is the reason why the VB approach is also called the mean field approximation. We can say that, if God forbade the particles, the *ensemble* of parameter values, to have the same correlations that we forbid the trial distribution to have in the VB approach, the VB posterior distribution could be realized in equilibrium.

Since the free energy, Eq.(6.21), is the sum of the negative entropy multiplied by $T = n^{-1}$ and the energy, the entropy dominates the determination of the optimum distribution and makes it extend when the number of training samples is small; while the energy dominates the determination of the optimum distribution and makes it shrink at the minimum energy point, i.e., the maximum a posterior (MAP) estimator, Eq.(2.17), when the number of training samples is large. Consequently, in singular models, the large entropy caused by the large state density of the singularities leashes the optimum distribution near the singularities according to the number of training samples. This is the reason why the Bayesian learning methods provide Occam's razor. (See also Section 6.3.1.)

As the solution of the VB approach in LNNs, we have obtained the following positive-part JS type estimator in Section 4.2.2:

$$(\hat{B}\hat{A})_{\rm VB} = \sum_{h=1}^{H} \theta(n\gamma_h^2 > \chi_{\rm VB}) \left(1 - \frac{\chi_{\rm VB}}{n\gamma_h^2}\right) \omega_{b_h} \omega_{b_h}^t R Q^{-1} + O_p(n^{-1}), \qquad (6.22)$$

where $\chi_{\rm VB} = L$. The thresholding by the indicator function $\theta(n\gamma_h^2 > L)$ can be considered to be a kind of phase transition. When $n\gamma_h^2 \leq L$, the free energy has the minimum at the origin, i.e., the singularities, since the entropy dominates; while, when $n\gamma_h^2 > L$, the free energy has the minimum between the origin and the MAP estimator since the energy becomes influential. Actually, in SOLNNs, the thresholding by the factor $\theta(n\gamma_h^2 > L)$ is equivalent to the model selection minimizing Akaike's information criterion (AIC).[2] It is interesting that this phase transition provides the model selection equivalent to AIC but not to Bayesian information criterion (BIC), although we minimize not the generalization error but the free energy.

6.2 Features

We summarize the features of the VB approach in LNNs in the following, the first and the fourth ones of which hold also in the SB approach:

[2]We assumed in Eq. (4.13) that the variance of the noise was equal to unity. When the variance, which is assumed to be known, is not equal to unity, the degree of shrinkage, $\chi_{\rm VB}$, should be multiplied by the variance.

1. The VB approach is asymptotically equivalent to a positive-part JS type shrinkage estimation.

   Hence, the VB estimator of a necessary component to realize the true distribution is asymptotically equivalent to the ML estimator; while the VB estimator of a redundant component is not equivalent to the ML estimator even in the asymptotic limit.

2. In the VB approach, the asymptotic behavior of the free energy and that of the generalization error are less simply related to each other, unlike in the Bayes estimation.

   In typical cases, the VB free energy well approximates the Bayes one; while the VB generalization error significantly differs from the Bayes one. Moreover, their dependences on the redundant parameter dimension are different from each other, which can throw a doubt on the model selection by minimizing the VB free energy to obtain better generalization performance.

3. Although the VB free energy is, by definition, never less than the Bayes one, the VB generalization error can be much less than the Bayes one.

   That does not mean the domination of the VB approach over the Bayes estimation, and is consistent with the proved superiority of the Bayes estimation.

4. The VB approach has not only a property similar to the Bayes estimation but also another property similar to the ML estimation.

5. The more different the parameter space dimensions of the different layers are, the smaller the generalization error is.

Further discussion on 4 and on 5 follow in Section 6.2.1 and in Section 6.2.2, respectively.

6.2.1 Two Aspects — ML like and Bayes like

We see in Figs. 4.2 and 4.4 that the VB generalization coefficient depends on H similarly to the ML one. Moreover, we see in Fig. 4.3 that, when $l \sim 100$, the

VB generalization coefficient exceeds that of the regular models, which the Bayes one never exceeds [Watanabe, 2001b]. That is for the following reason: because of the asymptotic equivalence to the shrinkage estimation, the VB approach would approximate the ML estimation and hence provide poor generalization performance in models where the ML estimator of a redundant component would go out of the effective range of shrinkage, i.e., be much greater than $\sqrt{L/n}$. When $(l - H^*) \gg (H - H^*)$, the $(H - H^*)$ largest eigenvalues of a random matrix subject to $\mathcal{W}_{l-H^*}(L - H^*, I_{l-H^*})$ are much greater than L, because of selection from a large number of random variables subject to the non-compact support distribution. Therefore, the eigenvalues $\{\gamma_h'^2\}$ in Theorem 7 go out of the effective range of shrinkage.

To summarize, we can say that, in the VB approach in LNNs, the *ML type effect*, the first effect of the singularities itemized in Section 1.1.4, is observed as the acceleration of overfitting by selection of the largest singular values of a random matrix; while the *Bayesian effect*, the second effect, is observed as the JS type *shrinkage*.

6.2.2 Conjecture for General Singular Models

We conjecture that the two aspects of the VB approach, discussed in Section 6.2.1, are essentially caused by the localization of the posterior distribution, which we can see in Fig. 4.1. Because of the independence between the parameters of different layers, the posterior distribution cannot extend so as to fill up the chink of singularities, which changes the strength of the *Bayesian effect*, sometimes increases, and sometimes decreases.

In the VB approach in models with hidden variables, such as mixture models, hidden Markov models, etc., the VB posterior distribution is restricted such that the parameters and the hidden variables are independent of each other [Attias, 1999; Ghahramani and Beal, 2001]. That restriction results in the independence between the parameters of different layers, and consequently, leads to the localization of the posterior distribution.[3] (See Appendix A.3.) In fact, the order of the

[3] In models with hidden variables, the other restriction that we applied in LNNs, i.e., the inde-

extent of the VB posterior distribution in mixture models when we use the prior distribution having positive values on the singularities has recently been derived to be similar to that in LNNs, [Watanabe and Watanabe, 2005]. So, we expect that the two aspects of the VB approach would hold also in more general singular models.

According to the fifth feature itemized at the beginning of Section 6.2, the VB generalization error is very small when the difference between the parameter space dimensions of different layers is very large, i.e., $l/L \ll 1$. That is because the *Bayesian effect*, i.e., the JS type *shrinkage*, is enhanced relatively to the *ML type effect*, i.e., the number of the random variables among which the redundant components of the VB estimator choose. We conjecture that the advantage of the VB approach over the EM algorithm would be enhanced in mixture models, where the dimension of the upper layer parameters, i.e., the mixing coefficients, is one per component and that of the lower layer parameters is usually more.

However, in general, non-linearity increases the *ML type effect* of singularities. In (general) neural networks, for example, we expect that the non-linearity of the activation function, $\psi(\cdot)$ in Eq.(2.35), would extend the range of basis selection and hence increase the generalization error. An important question to be solved is how the VB approach behaves in the models where the ML estimator diverges because of the *ML type effect*, that is, whether the *Bayesian shrinkage effect* leashes the VB estimator in a finite region or not.

6.3 Suggestions

6.3.1 Occam's Razor caused by Singularities

The Bayesian learning methods are said to unintentionally posses Occam's Razor [MacKay, 1995b]. As discussed in the preceding, it is the *Bayesian effect* of singularities and, in our case, observed as the JS type *shrinkage*. The strength of the effect depends on model and algorithm without philosophical justification. The author thinks that the singularities *unfairly* distribute huge weight to the model

pendence between the parameters of different components, is also guaranteed by introducing the hidden variables. (See Remark 3 in Appendix A.3.)

denoted by them, and this *unfairness* provides the very useful property, i.e., Occam's razor. Our first suggestion is that we should choose a set of a model and an algorithm, i.e., a learning machine, not clinging onto philosophical justification but considering its performance if it has been clarified. In addition, we should not forget to pay attention also to the trade-off, discussed in the first paragraph of Section 2.5, between suppression of overfitting and insensitivity to the true components with small amplitude.

6.3.2 Prior Selection

Letting the hyperparameters, c_a^2 and c_b^2, in the prior distributions, Eqs.(4.14) and (4.15), go to infinity, we have a kind of non-informative prior distribution. It seems to be valid that we use non-informative prior when we do not have any information on the parameter value. However, if we use the prior distribution above, the VB estimator diverges because the free energy can go to minus infinity when $M \neq N$, and the learning machine will not work well. As we have shown in Chapter 4, the hyperparameters, c_a^2 and c_b^2, do not affect the leading term of the generalization error, but only the lower order terms, as far as they are positive and finite constants. On the other hand, in singular models, whether we use an informative or non-informative prior affects the leading term, and hence, is much more important than what the hyperparameter values are. In addition, the Jeffreys prior, which *fairly* distributes equal weight to each model and hence zero weight to each point on the singularities, does not provide good generalization performance in the Bayesian learning methods [Watanabe, 2001b], because it interdicts the *unfair Bayesian effect* of singularities. The second suggestion is that we should use an informative prior even when we have no information on the parameter, if we need the advantage of generalization performance in the Bayesian learning methods.

Note the difference from the regular models. In the regular models, whether we use an informative or non-informative prior does not affect the leading term, and hence, the lower terms are important. In addition, we would say that the Bayesian procedure is philosophically valid in the regular models.

6.3.3 Other Learning Methods

Some approximation algorithms of the Bayes estimation have been developing to enhance and control the *Bayesian effect* of singularities. In [Nagata and Watanabe, 2005], pursuing the most preferable property, i.e., $2\lambda_{\text{Bayes}} \leq K$, the authors have used mixture models, which will better fill up the energy chink of the singularities, for approximation of the Bayes posterior distribution. We expect that this method would provide a different generalization coefficient from that of the VB approach if we use the proportional number of components to n^{α}, where $0 < \alpha \leq 1$ is a constant. Another method, where we can balance between suppression of overfitting and insensitivity to the true components with small amplitude, has been proposed [Takamatsu *et al.*, 2005]. The authors of [Takamatsu *et al.*, 2005] substitute a trial posterior distribution for the Bayes one, like in the VB approach, and control the trade-off by changing the restriction applied to the posterior distribution. Note that the trade-off is determined only by the geometry of the singularities in the Bayes estimation.

Chapter 7

Conclusions and Future Work

We have analyzed a subspace Bayes (SB) approach and the variational Bayes (VB) approach in three-layer linear neural networks (LNNs). The contributions of this thesis are as follows:

1. We have found that, in LNNs, the SB and the VB approaches are asymptotically equivalent to a positive-part James-Stein (JS) type shrinkage estimation and another one, respectively, which are similar to each other.

2. We have clarified the asymptotic behavior of the generalization error and the training error of the SB and the VB approaches, as well as the VB free energy, and compared them with those of the maximum likelihood (ML) estimation and the Bayes estimation.

3. We have found that, in the VB approach, the free energy and the generalization error are less simply related to each other, unlike in the Bayes estimation.

4. We have shown that the SB and the VB approaches have not only a property similar to the Bayes estimation but also another property similar to the ML estimation.

5. We have discussed when and why the SB and the VB approaches provide good generalization performance, through consideration of the rela-

tion among the JS estimation, the empirical Bayes (EB) approach, the SB approach, and the VB approach.

6. We have considered the *delicate* situations when the Kullback-Leibler divergence of the true distribution from the singularities is comparable to the inverse of the number of training samples, and discussed the domination of the SB and the VB approaches. As a result, we have found that the SB and the VB approaches dominate the ML estimation in many cases, although they never dominate the Bayes estimation.

Analysis of the SB and the VB approaches in other singular models is future work. We expect that our method in this thesis can be applied to analysis of the automatic relevance determination model (ARD), because the ARD is similar to an LNN, as discussed in Section 6.1.3. Clarifying the generalization performance of the VB approach in models with hidden variables, such as mixture models, hidden Markov models, etc., is also important, although it seems to be difficult. We would also like to consider the relation between the SB and the VB approaches in general singular models. We expect that some kind of similarity between them could be observed not only in LNNs.

Appendix A

Supplements about Previous Works

The appendix is devoted to supplements about previous works. We describe the proof of the superiority of the Bayes estimation, the derivation of the James-Stein (JS) estimator as an empirical Bayes (EB) estimator, and the introduction of the expectation-maximization (EM) algorithm and the variational Bayes (VB) approach in the normal mixture models, in Sections A.1–A.3, respectively.

A.1 Proof of Bayes Superiority (Proposition 1)

In this section, we prove Proposition 1, described in Section 2.1.1. Regarding the new test sample as the $(n+1)$-th sample, we can write the Bayes predictive distribution as follows:

$$\begin{aligned} p_{\text{Bayes}}(y_{n+1}|X^{n+1}, Y^n) &= \frac{\int \phi(w) \prod_{i=1}^{n+1} p(y_i|x_i, w) dw}{\int \phi(w') \prod_{i=1}^{n} p(y_i|x_i, w') dw'} \\ &= \frac{Z(Y^{n+1}|X^{n+1})}{Z(Y^n|X^n)}, \end{aligned} \tag{A.1}$$

where $Z(Y^n|X^n)$ is the marginal likelihood, defined by Eq.(2.1).[1] On the other hand, the average generalization error, defined by Eq.(2.18), of a learning method

[1] We can easily derive also Eq. (2.14) by using the expression Eq.(A.1).

with its predictive distribution $r(y|x)$ is given by

$$\begin{aligned}
\bar{G}(n) &= -\int q(w^*) \int q(X^{n+1}) p(Y^{n+1}|X^{n+1}, w^*) \\
&\qquad\qquad \log r(y|x)\, dX^{n+1} dY^{n+1} dw^* - S \\
&= -\int q(X^{n+1}) \left(\int q(w^*) p(Y^{n+1}|X^{n+1}, w^*) dw^* \right) \\
&\qquad\qquad \log r(y|x)\, dX^{n+1} dY^{n+1} - S \\
&= -\int q(X^{n+1}) Z^*(Y^{n+1}|X^{n+1}) \log r(y|x)\, dX^{n+1} dY^{n+1} - S \\
&= -\int q(X^{n+1}) Z^*(Y^n|X^n) \frac{Z^*(Y^{n+1}|X^{n+1})}{Z^*(Y^n|X^n)} \log r(y|x)\, dX^{n+1} dY^{n+1} - S,
\end{aligned} \tag{A.2}$$

where $Z^*(Y^n|X^n)$ is the marginal likelihood of the Bayes estimation with the *true* prior distribution, i.e., $\phi(w) = q(w)$, and

$$S = -\left\langle \int q(x) p(y|x, w^*) \log p(y|x, w^*) dx dy \right\rangle_{q(w^*)}$$

is the average entropy of the true model, which is independent of the predictive distribution, $r(y|x)$. We find from Eq.(A.2) that the average generalization error is minimized when

$$r(y|x) = \frac{Z^*(Y^{n+1}|X^{n+1})}{Z^*(Y^n|X^n)}. \tag{A.3}$$

Remembering Eq.(A.1), we find that Eq.(A.3) is equal to the Bayes predictive distribution with the *true* prior distribution, which completes the proof of the superiority of the Bayes estimation. (Q.E.D.)

A.2 Derivation of JS Estimator as an EB Estimator

The James-Stein (JS) estimator can be derived based on the empirical Bayes (EB) approach, as shown below [Efron and Morris, 1973; Kubokawa, 2004]. Suppose that we use the following set of the model distribution and the prior distribution

with a hyperparameter $0 \leq \tau \in \mathbb{R}$:

$$p(x|\mu) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{\|x-\mu\|^2}{2}\right), \tag{A.4}$$

$$\phi(\mu\|\tau) = \frac{1}{(2\pi\tau^2)^{M/2}} \exp\left(-\frac{\|\mu\|^2}{2\tau^2}\right), \tag{A.5}$$

where $x \in \mathbb{R}^M$ and $\mu \in \mathbb{R}^M$ are a random variable and a parameter, respectively. Then, the marginal likelihood can be analytically calculated as follows:

$$\begin{aligned}
Z(X^n|\tau) &= \int \prod_{i=1}^n p(x_i|\mu)\phi(\mu|\tau)d\mu \\
&= \frac{1}{(2\pi)^{nM/2}(2\pi\tau^2)^{M/2}} \int \exp\left(-\frac{1}{2}\sum_{i=1}^n \|x_i-\mu\|^2 - \frac{\|\mu\|^2}{2\tau^2}\right) d\mu \\
&= \frac{1}{(2\pi)^{nM/2}(2\pi\tau^2)^{M/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^n \|x_i\|^2 + \frac{n^2\|\hat{\mu}_{\text{MLE}}\|^2}{2(n+\tau^{-2})}\right) \\
&\qquad \cdot \int \exp\left(-\frac{1}{2}(n+\tau^{-2})\left\|\mu - \frac{n\hat{\mu}_{\text{MLE}}}{n+\tau^{-2}}\right\|^2\right) d\mu \\
&= \frac{1}{(2\pi)^{nM/2}} \frac{1}{(n\tau^2+1)^{M/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^n \|x_i\|^2 + \frac{n^2\|\hat{\mu}_{\text{MLE}}\|^2}{2(n+\tau^{-2})}\right) \\
&= \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^n \|x_i - \hat{\mu}_{\text{MLE}}\|^2\right)}{(2\pi)^{(n-1)M/2}} \frac{\exp\left(n\|\hat{\mu}_{\text{MLE}}\|^2/2(n\tau^2+1)\right)}{(2\pi)^{(n-1)M/2}(n\tau^2+1)^{M/2}}.
\end{aligned} \tag{A.6}$$

The posterior distribution and the predictive distribution are given by

$$\begin{aligned}
p(\mu|X^n) &= \frac{\prod_{i=1}^n p(x_i|\mu)\phi(\mu|\tau)}{\int \prod_{i=1}^n p(x_i|\mu')\phi(\mu'|\tau)d\mu'} \\
&= \mathcal{N}_M\left(\mu; \hat{\mu}_{\text{EB}}, \frac{I_M}{n(1+n^{-1}\tau^{-2})}\right),
\end{aligned} \tag{A.7}$$

$$\begin{aligned}
p(x|X^n) &= \int p(x|\mu)p(\mu|X^n)d\mu \\
&= \mathcal{N}_M\left(x; \hat{\mu}_{\text{EB}}, \left(1 + \frac{1}{n+\tau^{-2}}\right) I_M\right),
\end{aligned} \tag{A.8}$$

respectively, where

$$\hat{\mu}_{\text{EB}} = \hat{\mu}_{\text{MLE}} - \frac{\hat{\mu}_{\text{MLE}}}{n\tau^2+1} \tag{A.9}$$

is the EB estimator.

The marginal likelihood, Eq.(A.6), implies that $\hat{\mu}_{\mathrm{MLE}}\sqrt{n/(n\tau^2+1)}$ is a random variable subject to $\mathcal{N}_M(0, I_M)$, of which the minus second order moment is equal to $(M-2)^{-1}$. Therefore, we have

$$\left\langle \frac{n\tau^2+1}{n\|\hat{\mu}_{\mathrm{MLE}}\|^2} \right\rangle_{q(X^n)} = \frac{1}{M-2},$$

and hence,

$$\left\langle \frac{M-2}{n\|\hat{\mu}_{\mathrm{MLE}}\|^2} \right\rangle_{q(X^n)} = \frac{1}{n\tau^2+1}. \tag{A.10}$$

Accordingly, $(M-2)/n\|\hat{\mu}_{\mathrm{MLE}}\|^2$ is an unbiased estimator of the factor $(n\tau^2+1)^{-1}$. Replacing the factor $(n\tau^2+1)^{-1}$ in Eq.(A.9) with its unbiased estimator $(M-2)/n\|\hat{\mu}_{\mathrm{MLE}}\|^2$, we obtain the following JS estimator:

$$\hat{\mu}_{\mathrm{EB}} = \left(1 - \frac{M-2}{n\|\hat{\mu}_{\mathrm{MLE}}\|^2}\right)\hat{\mu}_{\mathrm{MLE}}\,. \tag{A.11}$$

In addition, if we estimate τ by maximizing the marginal likelihood, Eq.(A.6), we obtain the positive-part JS type estimator, Eq.(2.58), with the degree of shrinkage $\chi = M$.

A.3 EM Algorithm and VB Approach in Mixture Models

Hereafter, we deduce the expectation-maximization (EM) algorithm [Dempster *et al.*, 1977] and the iterative algorithm based on the variational Bayes (VB) approach [Attias, 1999] in simple mixture models, and then, describe some remarks on the similarity to and the difference from the case analyzed in the main part of this thesis. Let $x \in \mathbb{R}^M$ be an observed random variable, and $w = \{(a_h, \mu_h); a_h \in \mathbb{R}, \mu_h \in \mathbb{R}^M, h = 1, \ldots, H, a_h \geq 0, \sum_{h=1}^H a_h = 1\}$ summarizes the parameters. The probability density of a normal mixture model with H components is given by

$$p(x|w) = \sum_{h=1}^{H} a_h \mathcal{N}_M(x; \mu_h, I_M), \tag{A.12}$$

where the covariance matrices of all the components are assumed to be equal to I_M, for simplicity. Let $X^n = (x_1, \ldots, x_n)$ be n training samples. Then, the likelihood is given by

$$p(X^n|w) = \prod_{i=1}^{n} \left(\sum_{h=1}^{H} a_h \mathcal{N}_M(x_i; \mu_h, I_M) \right), \tag{A.13}$$

which is difficult to be calculated because of the summation in the multiplication. However, if we know the class label denoting the component from which each sample comes, the parameters of different components are conditionally independent of each other. In the framework of the EM algorithm, we regard the following class labels as the hidden variables for $1 \leq i \leq n$ and $1 \leq h \leq H$:

$$y_{ik} = \begin{cases} 1 & \text{if the } i\text{-th sample comes from the } h\text{-th component} \\ 0 & \text{otherwise} \end{cases}, \tag{A.14}$$

and substitute the following function, called the likelihood of the *complete* data, of the parameter w and the hidden variables $Y^n = \{y_{ih}\}$ for the *original* likelihood, Eq. (A.13):

$$p(X^n, Y^n|w) = \prod_{i=1}^{n} \prod_{h=1}^{H} \left(a_h \mathcal{N}_M(x_i; \mu_h, I_M) \right)^{y_{ih}}. \tag{A.15}$$

Expectation-Maximization Algorithm

We can easily derive the ML estimators of the parameters as follows, regarding the estimators of the hidden variables, $\{\hat{y}_{ih}\}$, to be constant and maximizing the *complete* likelihood, Eq.(A.15):

$$\hat{a}_h = \frac{\bar{N}_h}{n}, \tag{A.16}$$

$$\hat{\mu}_h = \bar{x}_h, \tag{A.17}$$

where

$$\bar{N}_h = \sum_{i=1}^{n} \hat{y}_{ih}, \tag{A.18}$$

$$\bar{x}_h = \frac{1}{\bar{N}_h} \sum_{i=1}^{n} \hat{y}_{ih} x_i. \tag{A.19}$$

We can also derive the ML estimators of the hidden variables as follows, regarding the estimators of the parameters, $\{\hat{a}_h, \hat{\mu}_h\}$, to be constant:

$$\hat{y}_{ih} \propto \exp\left(-\frac{1}{2}\|x_i - \hat{\mu}_h\|^2\right). \tag{A.20}$$

Thus, we have the algorithm, called the expectation-maximization (EM) algorithm, where we iteratively estimate the parameters, $\{a_h, \mu_h\}$, by using Eqs.(A.16)–(A.19), and the hidden variables, $\{y_{ih}\}$, by using Eq.(A.20) [Dempster *et al.*, 1977].

Variational Bayes Approach

The EM algorithm is to provide the ML estimator, so it tends to seriously overfit to noise. Then, the variational Bayes (VB) approach has been applied to models with hidden variables [Attias, 1999]. We use the following conjugate prior distributions:

$$p(\{a_h\}) = \mathcal{D}(\{a_h\}; \{\alpha_0, \ldots, \alpha_0\}), \tag{A.21}$$

$$p(\mu_h) = \mathcal{N}(\mu_h; \beta_0, c^2), \tag{A.22}$$

where

$$\mathcal{D}(\{a_h\}; \{\alpha_h\}) = \frac{\Gamma(\sum_{h=1}^{H} \alpha_h)}{\prod_{h=1}^{H} \Gamma(\alpha_h)} \prod_{h=1}^{H} a_h^{\alpha_h - 1} \tag{A.23}$$

is the Dirichlet distribution. Here, $\Gamma(\cdot)$ denotes the Gamma function.

The generalized free energy of a trial posterior distribution $r(Y^n, w|X^n)$ based on the *complete* data is given by

$$\bar{F}(X^n) = \sum_{Y^n} \int r(Y^n, w|X^n) \log \frac{r(Y^n, w|X^n)}{p(Y^n, w|X^n)} dw. \tag{A.24}$$

Hereafter, we abbreviate $r(Y^n, w|X^n)$ as $r(Y^n, w)$. In the framework of the VB approach, we restrict the posterior distribution such that the parameters and the hidden variables are independent of each other, i.e.,

$$r(Y^n, w) = r(Y^n) r(w). \tag{A.25}$$

The conditional independence, caused by introducing the hidden variables, between the parameters of different components makes the posterior distribution factorize as

$$r(Y^n, w) = \left(\prod_{i=1}^{n} r(y_i)\right)\left(\prod_{h=1}^{H} r(a_h) r(\mu_h)\right), \tag{A.26}$$

where $r(y_i) = r(\{y_{ih}; h = 1, \ldots, H)\})$ denotes the trial distribution of the label of the i-th sample.[2]

By using the variational method, we have the following relations:

$$r(a_h) \propto \phi(a_h) \exp\langle \log p(X^n, Y^n|w)\rangle_{r(Y^n)r(w)/r(a_h)}, \tag{A.27}$$

$$r(\mu_h) \propto \phi(\mu_h) \exp\langle \log p(X^n, Y^n|w)\rangle_{r(Y^n)r(w)/r(\mu_h)}, \tag{A.28}$$

$$r(y_i) \propto \phi(a_h) \exp\langle \log p(X^n, Y^n|w)\rangle_{r(Y^n)r(w)/r(y_i)}. \tag{A.29}$$

Substituting the *complete* data likelihood, Eq.(A.15), into Eqs.(A.27) and (A.28), we have

$$r(a_h) \propto a_h^{\bar{N}_h + \alpha_0 - 1}, \tag{A.30}$$

$$r(\mu_h) \propto \exp\left(-\frac{\|\mu_h - \beta_0\|^2}{2c^2} + \bar{N}_h \bar{x}_h^t \mu_h - \bar{N}_h \frac{\|\mu_h\|^2}{2}\right), \tag{A.31}$$

and thus the following VB posterior distribution of the parameters:

$$\hat{r}(\{a_h\}) = \mathcal{D}(\{a_h\}; \{\hat{\alpha}_h\}), \tag{A.32}$$

$$\hat{r}(\mu_h) = \mathcal{N}_M(\mu_h; \hat{\mu}_h, \hat{\sigma}_h^2), \tag{A.33}$$

where

$$\hat{\alpha}_h = \bar{N}_h + \alpha_0, \tag{A.34}$$

$$\hat{\mu}_h = \bar{\sigma}_{\mu_h}^2 \left(\bar{N}_h \bar{x}_h + \beta_0 c^{-2}\right), \tag{A.35}$$

$$\hat{\sigma}_h^2 = \frac{1}{\bar{N}_h + c^{-2}}. \tag{A.36}$$

[2]Only formally, $r(\{a_h\})$ factorizes as $r(\{a_h\}) = \prod_{h=1}^{H} r(a_h)$. In reality, $r(\{a_h\})/r(a_h)$ still depends on a_h because of the condition $\sum_{h=1}^{H} a_h = 1$. Similarly, Eqs.(A.27) and (A.30) are formal expressions as well.

Here,

$$\bar{N}_h = \sum_{i=1}^{n} \bar{y}_{ih}, \tag{A.37}$$

$$\bar{x}_h = \frac{1}{\bar{N}_h} \sum_{i=1}^{n} \bar{y}_{ih} x_i, \tag{A.38}$$

where

$$\bar{y}_{ih} = \langle y_{ih} \rangle_{\hat{r}(y_i)} . \tag{A.39}$$

Substituting Eq.(A.15) into Eq.(A.29), we also have the VB posterior distribution of the hidden variables as follows:

$$\hat{r}(y_i) \propto \exp\left(\sum_{h=1}^{H} y_{ih} \left\{ \Psi(\hat{\alpha}_h) - \Psi\left(\sum_{h'=1}^{H} \hat{\alpha}_{h'} \right) - \frac{1}{2} \left(\|x_i - \hat{\mu}_h\|^2 + M\hat{\sigma}^2_{\mu_h} \right) \right\} \right), \tag{A.40}$$

where

$$\Psi(x) = \frac{\partial \log \Gamma(x)}{\partial x} \tag{A.41}$$

is the Digamma function. Therefore, Eq.(A.39) is given by

$$\bar{y}_{ih} = \hat{r}(y_{ih} = 1) \propto \exp\left(\Psi(\hat{\alpha}_h) - \frac{1}{2} \left(\|x_i - \hat{\mu}_h\|^2 + M\hat{\sigma}^2_{\mu_h} \right) \right). \tag{A.42}$$

Thus, we have the algorithm, similar to the EM algorithm, where we iteratively estimate the VB posterior distribution of the parameters by using Eqs.(A.34)–(A.38), and that of the hidden variables by using Eq.(A.42) [Attias, 1999]. The framework above can be applied to other models with hidden variables such as hidden Markov models, Bayesian networks, etc..

Remark 3 Note that the VB algorithm in models with hidden variables results in the following two independences of the posterior distribution: the independence between the parameters of different layers, i.e., the mixing coefficients, $\{a_h\}$, and the mean parameters, $\{\mu_h\}$, in the case of the normal mixture model Eq.(A.12);

and the independence between the parameters of different components. The former is caused by the assumed independence between the parameters and the hidden variables, and the latter is caused by the substitution of the *complete* likelihood, Eq.(A.15), for the *original* likelihood, Eq.(A.13), with introducing the hidden variables, as well as the independence above. In Chapter 4, we apply the VB approach to linear neural networks, directly assuming the independences above.

Remark 4 In mixture models, as well as in hidden Markov models, we can select whether the VB posterior distribution extends in the lower parameter space for the redundant components, i.e., $\{\mu_h; H^* < h \leq H\}$ in the case of the normal mixture model Eq.(A.12), by controlling the hyperparameter of the Dirichlet prior distribution, i.e., α_0 in Eq.(A.21) [Watanabe and Watanabe, 2005; Hosino *et al.*, 2005].[3] According to the results of the main part of this thesis, we expect that the VB posterior distribution extends either in the upper parameter space, i.e., $\{a_h; H^* < h \leq H\}$, or in the lower one, so as to enlarge the entropy. However, it has been shown that, even when α_0 is so large that the VB posterior distribution shrinks in the lower parameter space, the VB posterior distribution does not extend in the upper parameter space, but shrinks both in the upper and the lower parameter spaces. Consequently, the VB free energy coefficient is equal to that of the regular models [Watanabe and Watanabe, 2005; Hosino *et al.*, 2005]. That is because of the substitution of the *complete* likelihood for the *original* likelihood with introducing the hidden variables. In fact, the *complete* model, Eq.(A.15), does not have a part of the singularities that the *original* model, Eq.(A.13), has when we estimate the distribution of the hidden variables independent of the parameters, as explained below. Consider the case that $H = 2$ for simplicity. Then, the *original* likelihood is given by

$$p(X^n|w) = \prod_{i=1}^{n} \left(a_1 \mathcal{N}_M(x_i; \mu_1, I_M) + (1 - a_1)\mathcal{N}_M(x_i; \mu_2, I_M)\right), \qquad \text{(A.43)}$$

where $(1-a_1)$ corresponds to a_2. Equation (A.43) is invariant for μ_h when $a_h = 0$, and also invariant for a_1 when $\mu_1 = \mu_2$. Nevertheless, although the *complete*

[3] Note that the prior distribution in the region where $a_h = 0$ and in the region where $a_h = 1$ for any h, which is a part of the singularities, is zero when $\alpha_0 > 1$. That can make a big difference from our results. (See Section 6.3.2.)

likelihood

$$
\begin{aligned}
p(X^n, Y^n|w) &= \prod_{i=1}^{n} \left((a_1 \mathcal{N}_M(x; \mu_1, I_M))^{y_{i1}} \left((1-a_1) \mathcal{N}_M(x_i; \mu_2, I_M) \right)^{1-y_{i1}} \right) \\
&= a_1^{\sum_{i=1}^n y_{i1}} (1-a_1)^{n-\sum_{i=1}^n y_{i1}} \\
&\qquad \cdot \prod_{i=1}^{n} \left((\mathcal{N}_M(x_i; \mu_1, I_M))^{y_{i1}} (\mathcal{N}_M(x_i; \mu_2, I_M))^{1-y_{i1}} \right) \quad \text{(A.44)}
\end{aligned}
$$

is invariant for μ_h when $y_{ih} = 0$ for any $1 \leq i \leq n$, it depends on a_1, as well as $N_1 = \sum_{i=1}^n y_{i1}$, even when $\mu_1 = \mu_2$ because of the factor $a_1^{N_1}(1-a_1)^{n-N_1}$. Accordingly, the *complete* model, Eq.(A.15), in general, does not have the singularities along the a_h axis, and consequently, the VB posterior distribution can not extend in the a_h space.

Bibliography

[Akaike, 1974] H. Akaike. A New Look at Statistical Model. *IEEE Trans. on Automatic Control*, 19(6):716–723, 1974.

[Akaike, 1980] H. Akaike. Likelihood and Bayes Procedure. In J. M. Bernald, editor, *Bayesian Statistics*, pages 143–166. University Press, 1980.

[Amari *et al.*, 2002] S. Amari, H. Park, and T. Ozeki. Geometrical Singularities in the Neuromanifold of Multilayer Perceptrons. In *Advances in NIPS*, volume 14, pages 343–350, Cambridge, MA, 2002. MIT Press.

[Aoyagi and Watanabe, 2005] M. Aoyagi and S. Watanabe. Stochastic Complexities of Reduced Rank Regression in Bayesian Estimation. *Neural Networks*, 18(7):924–933, 2005.

[Attias, 1999] H. Attias. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In *Proc. of UAI*, 1999.

[Baldi and Hornik, 1995] P. F. Baldi and K. Hornik. Learning in Linear Neural Networks: a Survey. *IEEE Trans. on Neural Networks*, 6(4):837–858, 1995.

[Bickel and Chernoff, 1993] P. Bickel and H. Chernoff. Asymptotic Distribution of the Likelihood Ratio Statistic in a Prototypical Non Regular Problem. pages 83–96. Wiley Eastern Limited, 1993.

[Cramer, 1949] H. Cramer. *Mathematical Methods of Statistics*. University Press, Princeton, 1949.

[Dacunha-Castelle and Gassiat, 1997] D. Dacunha-Castelle and E. Gassiat. Testing in Locally Conic Models, and Application to Mixture Models. *Probability and Statistics*, 1:285–317, 1997.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood for Incomplete Data Via the EM Algorithm. *J. R. Statistical Society*, 39-B:1–38, 1977.

[Efron and Morris, 1973] B. Efron and C. Morris. Stein's Estimation Rule and its Competitors—an Empirical Bayes Approach. *J. of Am. Stat. Assoc.*, 68:117–130, 1973.

[Fukumizu *et al.*, 2004] K. Fukumizu, S. Kuriki, K. Takeuchi, and M. Akahira. *Statitical Theory of Singular Models (In Japanese)*. Iwanami, 2004.

[Fukumizu, 1999] K. Fukumizu. Generalization Error of Linear Neural Networks in Unidentifiable Cases. In *Proc. of ALT*, pages 51–62. Springer, 1999.

[Fukumizu, 2003] K. Fukumizu. Likelihood Ratio of Unidentifiable Models and Multilayer Neural Networks. *Annals of Statistics*, 31(3):833–851, 2003.

[Ghahramani and Beal, 2001] Z. Ghahramani and M. J. Beal. Graphical Models and Variational Methods. In *Advanced Mean Field Methods*, pages 161–177. MIT Press, 2001.

[Hagiwara, 2002] K. Hagiwara. On the Problem in Model Selection of Neural Network Regression in Overrealizable Scenario. *Neural Computation*, 14:1979–2002, 2002.

[Hartigan, 1985] J. A. Hartigan. A Failure of Likelihood Ratio Asymptotics for Normal Mixtures. In *Proc. of the Berkeley Conference in Honor of J. Neyman and J. Kiefer*, pages 807–810, 1985.

[Hinton and van Camp, 1993] G. E. Hinton and D. van Camp. Keeping Neural Networks Simple by Minimizing the Description Length of the Weights. In *Proc. of COLT*, pages 5–13, 1993.

[Hosino *et al.*, 2005] T. Hosino, K. Watanabe, and S. Watanabe. Stochastic Complexity of Variational Bayesian Hidden Markov Models. In *Proc. of IJCNN*, 2005.

[Hukushima and Nemoto, 1996] K. Hukushima and K. Nemoto. Exchange Monte Carlo Method and Application to Spin Glass Simulation. *J. of Physical Society of Japan*, 65(6):1604–1608, 1996.

[Iba, 2001] Y. Iba. Extended Ensemble Monte Carlo. *Int. J. of Modern Physics*, C12:623–656, 2001.

[Iba, 2005] Y. Iba. Part I. In *Computational Statistics II (in Japanese)*. Iwanami, 2005.

[Jaakkola and Jordan, 2000] T. S. Jaakkola and M. I. Jordan. Bayesian Parameter Estimation via Variational Methods. *Statistics and Computing*, 10:25–37, 2000.

[James and Stein, 1961] W. James and C. Stein. Estimation with Quadratic Loss. In *Proc. of the 4th Berkeley Symp. on Math. Stat. and Prob.*, pages 361–379, 1961.

[Kass and Steffey, 1989] R. E. Kass and D. Steffey. Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models). *J. of the Am. Stat. Assoc.*, 84:717–726, 1989.

[Kubokawa, 2004] T. Kubokawa. Stein's Paradox and Shrinkage Estimation (In Japanese). In *Model Selection*. Iwanami, 2004.

[Kuriki and Takemura, 2001] S. Kuriki and A. Takemura. Tail Probabilities of the Maxima of Multilinear Forms and Their Applications. *Annals of Statistics*, 29(2):328–371, 2001.

[Levin *et al.*, 1990] E. Levin, N. Tishby, and S. A. Solla. A Statistical Approaches to Learning and Generalization in Layered Neural Networks. In *Proc. of IEEE*, volume 78, pages 1568–1674, 1990.

[MacKay, 1992] D. J. C. MacKay. Bayesian Interpolation. *Neural Computation*, 4(2):415–447, 1992.

[MacKay, 1994] D. J. C. MacKay. Bayesian Non-linear Modeling for the Energy Prediction Competition. *ASHRAE Transactions*, 100(2):1053–1062, 1994.

[MacKay, 1995a] D. J. C. MacKay. Developments in Probabilistic Modeling with Neural Networks—Ensemble Learning. In *Proc. of the 3rd Ann. Symp. on Neural Networks*, pages 191–198, 1995.

[MacKay, 1995b] D. J. C. MacKay. Probable Networks and Plausible Predictions—A Review of Practical Bayesian Methods for Supervised Neural Networks. *Network*, 6(3):469–505, 1995.

[Nagata and Watanabe, 2005] K. Nagata and S. Watanabe. A Method to Approximate the Bayesian Posterior Distribution in Singular Learning Machines. In *Proc. of ICONIP*, pages 31–35, Taipei, Taiwan, 2005.

[Nakajima and Watanabe, 2005a] S. Nakajima and S. Watanabe. Generalization Error and Free Energy of Variational Bayes Approach of Linear Neural Networks. In *Proc. of ICONIP*, pages 55–60, Taipei, Taiwan, 2005.

[Nakajima and Watanabe, 2005b] S. Nakajima and S. Watanabe. Generalization Error of Linear Neural Networks in an Empirical Bayes Approach. In *Proc. of IJCAI*, pages 804–810, Edinburgh, U.K., 2005.

[Nakajima and Watanabe, 2006a] S. Nakajima and S. Watanabe. Analysis of Hierarchical Variational Bayes Approach in Linear Inverse Problem. *To Appear in Technical Report of IEICE*, March 2006.

[Nakajima and Watanabe, 2006b] S. Nakajima and S. Watanabe. Generalization Performance of Subspace Bayes Approach in Linear Neural Networks. *To Appear in IEICE Trans.*, E89-D(3):1128–1138, 2006.

[Nakano and Watanabe, 2005] N. Nakano and S. Watanabe. Stochastic Complexity of Layered Neural Networks in Mean Field Approximation. In *Proc. of ICONIP*, Taipei, Taiwan, 2005.

[Nakano *et al.*, 2005] N. Nakano, K. Takahashi, and S. Watanabe. On the Evaluation Criterion of the MCMC Method in Singular Learning Machines (In Japanese). *IEICE Trans.*, J88-D2(10):2011–2020, 2005.

[Neal, 1996] R. M. Neal. *Bayesian Learning for Neural Networks*. Springer, 1996.

[Reinsel and Velu, 1998] G. C. Reinsel and R. P. Velu. *Multivariate Reduced-Rank Regression*. Springer, 1998.

[Rissanen, 1986] J. Rissanen. Stochastic Complexity and Modeling. *Annals of Statistics*, 14(3):1080–1100, 1986.

[Rusakov and Geiger, 2002] D. Rusakov and D. Geiger. Asymptotic Model Selection for Naive Bayesian Networks. In *Proc. of UAI*, pages 438–445, Alberta, Canada, 2002.

[Sakamoto *et al.*, 1983] T. Sakamoto, M. Ishiguro, and G. Kitagawa. *Information Criterion Statistics (In Japanese)*. Kyoritsu, 1983.

[Sakamoto *et al.*, 1986] T. Sakamoto, M. Ishiguro, and G. Kitagawa. *Akaike Information Criterion Statistics*. D.Reidel Publishing Company, 1986.

[Sato *et al.*, 2004] M. Sato, T. Yoshioka, S. Kajihara, K. Toyama, N. Goda, K. Doya, and M. Kawato. Hierarchical Bayesian Estimation for MEG inverse problem. *Neuro Image*, 23:806–826, 2004.

[Sato, 2001] M. Sato. Online Model Selection Based on the Variational Bayes. *Neural Computation*, 13:1649–1681, 2001.

[Schwarz, 1978] G. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464, 1978.

[Stein, 1956] C. Stein. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. In *Proc. of the 3rd Berkeley Symp. on Math. Stat. and Prob.*, pages 197–206, 1956.

[Takamatsu *et al.*, 2005] S. Takamatsu, S. Nakajima, and S. Watanabe. Generalization Error of Localized Bayes Estimation in Reduced Rank Regression (In Japanese). In *Proc. of IBIS*, pages 81–86, Tokyo, Japan, 2005.

[Takemura and Kuriki, 1997] A. Takemura and S. Kuriki. Weights of Chi-bar-square Distribution for Smooth or Piecewise Smooth Cone Alternatives. *Annals of Statistics*, 25(6):2368–2387, 1997.

[Wang and Titterington, 2004] B. Wang and D. M. Titterington. Convergence and Asymptotic Normality of Variational Bayesian Approximations for Exponential Family Models with Missing Values. In *Proc. of UAI*, pages 577–584, Banff, Canada, 2004.

[Watanabe and Amari, 2003] S. Watanabe and S. Amari. Learning Coefficients of Layered Models When the True Distribution Mismatches the Singularities. *Neural Computation*, 15:1013–1033, 2003.

[Watanabe and Watanabe, 2004] K. Watanabe and S. Watanabe. Lower Bounds of Stochastic Complexities in Variational Bayes Learning of Gaussian Mixture Models. In *Proc. of IEEE on CIS*, pages 99–104, Singapore, 2004.

[Watanabe and Watanabe, 2005] K. Watanabe and S. Watanabe. Stochastic Complexity for Mixture of Exponential Families in Variational Bayes. In *Proc. of ALT*, pages 107–121, Singapore, 2005.

[Watanabe, 1995] S. Watanabe. A Generalized Bayesian Framework for Neural Networks with Singular Fisher Information Matrices. In *Proc. of NOLTA*, volume 2, pages 207–210, 1995.

[Watanabe, 2001a] S. Watanabe. Algebraic Analysis for Nonidentifiable Learning Machines. *Neural Computation*, 13(4):899–933, 2001.

[Watanabe, 2001b] S. Watanabe. Algebraic Information Geometry for Learning Machines with Singularities. In *Advances in NIPS*, volume 13, pages 329–336, 2001.

[Watanabe, 2001c] S. Watanabe. *Learning Machines and Algorithms (In Japanese)*. Kyoritsu, 2001.

[Watcher, 1978] K. W. Watcher. The Strong Limits of Random Matrix Spectra for Sample Matrices of Independent Elements. *Ann. Prob.*, 6:1–18, 1978.

[Yamazaki and Watanabe, 2003a] K. Yamazaki and S. Watanabe. Singularities in Mixture Models and Upper Bounds of Stochastic Complexity. *Neural Networks*, 16(7):1029–1038, 2003.

[Yamazaki and Watanabe, 2003b] K. Yamazaki and S. Watanabe. Stochastic Complexities of Hidden Markov Models. In *Proc. of Neural Networks for Signal Processing XIII (NNSP)*, pages 179–188, Toulouse, France, 2003.

[Yamazaki and Watanabe, 2003c] K. Yamazaki and S. Watanabe. Stochastic Complexity of Bayesian Networks. In *Proc. of UAI*, pages 592–599, Acapulco, Mexico, 2003.

[Yamazaki and Watanabe, 2004] K. Yamazaki and S. Watanabe. Newton Diagram and Stochastic Complexity in Mixture of Binomial Dstributions. In *Proc. of ALT*, Padova, Italy, October 2004. Springer.

[Yamazaki *et al.*, 2005] K. Yamazaki, Kenji Nagata, and S. Watanabe. A New Method of Model Selection Based on Learning Coefficient. In *Proc. of NOLTA*, pages 389–392, Bruge, Belgium, 2005.