Generalization Error of Linear Neural Networks in an Empirical Bayes Approach

Shinichi Nakajima [†‡] **and Sumio Watanabe** [†]
[†] Tokyo Institute of Technology
Mailbox R2-5, 4259 Nagatsuda, Midori-ku, Yokohama, Kanagawa, 226-8503 Japan
nakajima.s@cs.pi.titech.ac.jp, swatanab@pi.titech.ac.jp
[‡] Nikon Corporation
201-9 Oaza-Miizugahara, Kumagaya, Saitama, 360-8559 Japan

Abstract

It is well known that in unidentifiable models, the Bayes estimation has the advantage of generalization performance to the maximum likelihood estimation. However, accurate approximation of the posterior distribution requires huge computational costs. In this paper, we consider an empirical Bayes approach where a part of the parameters are regarded as hyperparameters, which we call a subspace Bayes approach, and theoretically analyze the generalization error of three-layer linear neural networks. We show that a subspace Bayes approach is asymptotically equivalent to a positive-part James-Stein type shrinkage estimation, and behaves similarly to the Bayes estimation in typical cases.

1 Introduction

Unidentifiable parametric models, such as neural networks, mixture models, and so on, have a wide range of applications. These models have singularities in the parameter space, hence the conventional learning theory of the regular statistical models does not hold. Recently, generalization performance of some unidentifiable models has been theoretically clarified. In the maximum likelihood (ML) estimation, which is asymptotically equivalent to the maximum a posterior (MAP) estimation, the generalization error of linear neural networks 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]. On the other hand, in the Bayes estimation, the generalization error of neural networks, linear neural networks, mixture models, and so on was proved to be less than that of the regular models [Watanabe, 2001; Aoyagi and Watanabe, 2004; Yamazaki and Watanabe, 2003].

However, the Bayes posterior distribution can seldom be exactly realized. Furthermore, Markov chain Monte Carlo (MCMC) methods, often used for approximation of the posterior distribution, require huge computational costs. As an alternative, the variational Bayes approach, where the correlation between parameters and the other parameters, or the correlation between the parameters and the hidden variables is neglected, was proposed [Hinton and van Camp, 1993; MacKay, 1995; Attias, 1999; Ghahramani and Beal, 2000].[1]

In this paper, we consider another alternative, which we call a subspace Bayes (SB) approach. An SB approach is an 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 an SB approach is thus much less than that of posterior distribution approximation by MCMC methods. At first in this paper, we prove that in three-layer linear neural networks, an SB approach is equivalent to a positive-part James-Stein (JS) type shrinkage estimation [James and Stein, 1961]. Then, we clarify its generalization error, also considering *delicate* situations, the most important situations in model selection problems and in statistical tests, when the Kullback-Leibler divergence of the true distribution from the singularities is comparable to the inverse of the number of training samples.[2] We conclude that an SB approach provides as good performance as the Bayes estimation in typical cases.

In Section 2, neural networks and linear neural networks are briefly introduced. The framework of the Bayes estimation, that of an EB approach, and that of an SB approach are described in Section 3. The significance of singularities for generalization performance and the importance of analysis of *delicate* situations are explained in Section 4. The SB solution and its generalization error are derived in Section 5. Discussions and conclusions follow in Section 6 and in Section 7, respectively.

2 Linear Neural Networks

Let $x \in \mathbb{R}^M$ be an input (column) 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

[1] We have just derived the variational Bayes solution of linear neural networks and clarified its generalization error and training error in [Nakajima and Watanabe, 2005b].

[2] We have also clarified the training error, which is put into [Nakajima and Watanabe, 2005a].

units is defined by

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

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. Assume that the output is observed with a noise subject to $\mathcal{N}_N(0, \sigma^2 I_N)$, where $\mathcal{N}_d(\mu, \Sigma)$ denotes the d-dimensional normal distribution with average vector μ and covariance matrix Σ, and I_d denotes the $d \times d$ identity matrix. Then, the conditional distribution is given by

$$p(y|x,w) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{\|y - f(x;w)\|^2}{2\sigma^2}\right). \tag{2}$$

In this paper, we focus on linear neural networks, whose activation function is linear, as the simplest multilayer models.[3] A linear neural network model (LNN) is defined by

$$f(x;A,B) = BAx, \tag{3}$$

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.(3) has trivial redundancy. Accordingly, the essential dimension of the parameter space is given by

$$K = H(M + N) - H^2. \tag{4}$$

We assume that $H \leq N \leq M$ throughout this paper.

3 Framework of Learning Methods

3.1 Bayes Estimation

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 a model $p(y|x, w)$ is given by

$$Z(Y^n|X^n) = \int \phi(w) \prod_{i=1}^{n} p(y_i|x_i, w) dw, \tag{5}$$

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

$$p(w|X^n, Y^n) = \frac{\phi(w) \prod_{i=1}^{n} p(y_i|x_i, w)}{Z(Y^n|X^n)}, \tag{6}$$

and the predictive distribution is defined as the average of the model over the posterior distribution as follows:

$$p(y|x, X^n, Y^n) = \int p(y|x, w) p(w|X^n, Y^n) dw. \tag{7}$$

The generalization error, a criterion of generalization performance, is defined by

$$G(n) = \langle G(X^n, Y^n) \rangle_{q(X^n, Y^n)}, \tag{8}$$

where

$$G(X^n, Y^n) = \int q(x)q(y|x) \log \frac{q(y|x)}{p(y|x, X^n, Y^n)} dxdy \tag{9}$$

is the Kullback-Leibler (KL) divergence of the predictive distribution from the true distribution, and $\langle \cdot \rangle_{q(X^n, Y^n)}$ denotes the expectation value over all sets of n training samples.

3.2 Empirical Bayes Approach and Subspace Bayes Approach

We often have little information about the prior distribution, with which an 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_1 such as

$$\phi(w) = \frac{1}{(2\pi\tau_1^2)^{K/2}} \exp\left(-\frac{\|w\|^2}{2\tau_1^2}\right), \tag{10}$$

the marginal likelihood, Eq.(5), also depends on τ_1. In an EB approach, τ_1 is estimated by maximizing the marginal likelihood or by a slightly different way [Efron and Morris, 1973; Akaike, 1980; Kass and Steffey, 1989]. Extending the idea above, we can introduce hyperparameters also in a model distribution. What we call an SB approach is an EB approach where a part of the parameters of a model are regarded as hyperparameters. In the following sections, we analyze two versions of SB approach: in the first one, we regard the output parameter matrix B of the map, Eq.(3), as a hyperparameter and then marginalize the likelihood in the input parameter space (MIP); and in the other one, we regard the input parameter matrix A, instead of B, as a hyperparameter and then marginalize in the output parameter space (MOP).

4 Unidentifiability and Singularities

We say that a parametric model is unidentifiable if the map from the parameter to the probability distribution is not one-to-one. A neural network model, Eq.(1), is unidentifiable because the model is independent of a_h when $b_h = 0$, or vice versa. The continuous points denoting the same distribution are called the singularities, because the Fisher information matrix on them degenerates. When the true model is not on the singularities, asymptotically they do not affect prediction, and therefore, the conventional learning theory of the regular models holds. On the other hand, when the true model is on the singularities, they significantly affect generalization performance as follows: in the ML estimation, the extent of the set of the points denoting the true distribution increases its neighborhoods and hence the flexibility of imitating noises, and therefore, accelerates overfitting; while in the Bayes estimation, the large entropy of the singularities increases the weights of the distributions near the true one, and therefore, suppresses overfitting. In LNNs, the former property appears as acceleration of overfitting by selection of the largest singular value components of a random matrix, and in the SB approaches of LNNs, the latter property appears as James-Stein type shrinkage, as shown in the following sections.

Suppression of overfitting accompanies insensitivity to the true components with small amplitude. There is a trade-off,

[3] A linear neural network model is not a toy but an useful model, known as a reduced-rank regression model, in many applications [Reinsel and Velu, 1998].

which would, however, be ignored in asymptotic analysis if we would consider only situations when the true model is *distinctly* on the singularities or not. Therefore, in this paper, we also consider *delicate* situations when the KL divergence of the true distribution from the singularities is comparable to the inverse of the number of training samples, n^{-1}, which are important situations in model selection problems and in statistical tests with finite number of 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.

5 Theoretical Analysis

5.1 Subspace Bayes Solution

By $\|$ we, hereafter, distinguish the hyperparameter τ from the parameter w, for example, $p(y|x, w\|\tau)$. Assume that the variance of a noise is known and equal to unity. Then, the conditional distribution of an LNN in the MIP version of SB approach is given by

$$p(y|x, A\|B) = \frac{1}{(2\pi)^{N/2}} \exp\left(-\frac{\|y - BAx\|^2}{2}\right). \quad (11)$$

We use the following prior distribution:

$$\phi(A) = \frac{1}{(2\pi)^{HM/2}} \exp\left(-\frac{\mathrm{tr}\,(A^t A)}{2}\right). \quad (12)$$

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 $H^* \leq H$. We denote by $*$ the true value of a parameter as above, and by a *hat* an estimator of a parameter, for example, $\hat{A}$, $\hat{b}_h$, etc.. For simplicity, we assume that the input vector is orthonormalized so that $\int xx^t q(x)dx = I_M$. Consequently, the central limit theorem leads to the following two equations:

$$Q(X^n) = n^{-1}\textstyle\sum_{i=1}^n xx^t = I_M + O(n^{-1/2}), \quad (13)$$

$$R(X^n, Y^n) = n^{-1}\textstyle\sum_{i=1}^n yx^t = B^*A^* + O(n^{-1/2}), \quad (14)$$

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. We find from Eq.(14) that γ_h for $H^* < h \leq H$ is of order $O(n^{-1/2})$. Hence, combining with Eq.(13), we get

$$\omega_{b_h} RQ^\rho = \omega_{b_h} R + O(n^{-1}) \quad \text{for } H^* < h \leq H, \quad (15)$$

where $-\infty < \rho \in \mathbb{R} < \infty$ 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 $L = M$ in the MIP version or $L = N$ in the MOP version, and $L'_h = \max(L, n\gamma_h^2)$. The SB estimator of the map of an LNN is given by*

$$\hat{B}\hat{A} = \textstyle\sum_{h=1}^H (1 - L_h'^{-1}L)\omega_{b_h}\omega_{b_h}^t RQ^{-1} + O(n^{-1}). \quad (16)$$

(The proof is given in Appendix A.)

Because the independence between A and B makes the posterior distribution localized, the following lemma holds.

Lemma 1 *The predictive distribution in the SB approaches can be written as follows:*

$$p(y|x, X^n, Y^n) = \left((2\pi)^N |\hat{V}|\right)^{-1/2} \cdot \exp\left(-(y - \hat{V}\hat{B}\hat{A}x)^t \frac{\hat{V}^{-1}}{2}(y - \hat{V}\hat{B}\hat{A}x)\right) + O(n^{-3/2}), \quad (17)$$

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

(Proof) The predictive distribution is written as follows:

$$p(y|x, X^n, Y^n) = \langle p(y|x, A\|B)\rangle_{p(A|X^n, Y^n\|B)} \propto q(y|x)\left\langle \exp\left(y^t(\hat{B}\hat{A} - B^*A^*)x\right)\right\rangle_{p(A|X^n, Y^n\|B)}, \quad (18)$$

where $\langle\cdot\rangle_p$ denotes the expectation value over a distribution p. Since $(\hat{B}\hat{A} - B^*A^*) = O(n^{-1/2})$ in the SB approaches, we can expand the predictive distribution as follows:

$$p(y|x, X^n, Y^n) \propto q(y|x)\left\langle 1 + y^t(\hat{B}\hat{A} - B^*A^*)x + \frac{y^t vv^t y}{2n}\right\rangle_{p(A|X^n, Y^n\|B)} + O(n^{-3/2}), \quad (19)$$

where $v = \sqrt{n}(\hat{B}\hat{A} - B^*A^*)x$ is an N-dimensional vector of order $O(1)$. Calculating the expectation value and expanding the logarithm of Eq.(19), we arrive at Lemma 1. (Q.E.D.)

Comparing Eq.(16) with the ML estimator

$$\hat{B}\hat{A}_{MLE} = \textstyle\sum_{h=1}^H \omega_{b_h}\omega_{b_h}^t RQ^{-1} \quad (20)$$

[Baldi and Hornik, 1995], we find that the SB estimator of each component is asymptotically equivalent to a positive-part JS type shrinkage estimator. Moreover, by virtue of Lemma 1, we can substitute the model at the SB estimator for the predictive distribution with asymptotically insignificant impact on generalization performance. Therefore, we conclude that the SB approach is asymptotically equivalent to the shrinkage estimation. Note that the variance of the prior distribution, Eq.(12), asymptotically has no effect upon prediction and hence upon generalization performance, as far as it is a positive, finite constant. We call L the degree of shrinkage. Remember that we can modify all the theorems in this paper for the ML estimation only by letting $L = 0$.

5.2 Generalization Error

Using the singular value decomposition of the true map B^*A^*, we can transform arbitrary A^* and B^* without change of the map into a matrix with its orthogonal row vectors and another matrix with its orthogonal column vectors, respectively. Accordingly, we assume the above orthogonalities without loss of generality. Then, Lemma 1 implies that the KL divergence, Eq.(9), with a set of n training samples is

given by

$$G(X^n, Y^n) = \left\langle \frac{\|(B^*A^* - \hat{B}\hat{A})x\|^2}{2} \right\rangle_{q(x)} + O(n^{-3/2})$$
$$= \sum_{h=1}^{H} G_h(X^n, Y^n) + O(n^{-3/2}), \quad (21)$$

where

$$G_h(X^n, Y^n) = \frac{1}{2}\mathrm{tr}\left((b_h^* a_h^{*t} - \hat{b}_h \hat{a}_h^t)^t (b_h^* a_h^{*t} - \hat{b}_h \hat{a}_h^t) \right) \quad (22)$$

is the contribution of the h-th component. Here tr$(\cdot)$ denotes the trace of a matrix. We denote by $\mathcal{W}_d(m, \Sigma, \Lambda)$ the d-dimensional Wishart distribution with m degrees of freedom, scale matrix Σ, and noncentrality matrix Λ, and abbreviate as $\mathcal{W}_d(m, \Sigma)$ the central Wishart distribution.

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

$$G(n) = \lambda n^{-1} + O(n^{-3/2}),$$

where the coefficient of the leading term, called the generalization coefficient in this paper, is given by

$$2\lambda = (H^*(M+N) - H^{*2}) + \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h'^2 > L) \left(1 - \frac{L}{\gamma_h'^2}\right)^2 \gamma_h'^2 \right\rangle_{q(\{\gamma_h'^2\})}. \quad (23)$$

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

(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 $O(n^{-1})$. 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 Eq.(4), we obtain the first term of Eq.(23) as the contribution of the first H^* components. On the other hand, we find from Eq.(15) and Theorem 1 that for a redundant component, identifying $RQ^{-1/2}$ with R affects the SB estimator only of order $O(n^{-1})$, which, 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 is singular value decomposed as $T = \Omega_b U \Omega_a$, 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.(22), we obtain the second term of Eq.(23) as the contribution of the last $(H-H^*)$ components. Thus, we complete the proof of Theorem 2. (Q.E.D.)

5.3 Large Scale Approximation

In a similar fashion to the analysis of the ML estimation in [Fukumizu, 1999], the second term of Eq.(23) 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: $\alpha = N'/M' = (N-H^*)/(M-H^*)$, $\beta = H'/N' = (H-H^*)/(N-H^*)$, and $\kappa = L/M' = L/(M-H^*)$. Let W be a random matrix subject to $\mathcal{W}_{N'}(M', I_{N'})$, and $\{u_1, \ldots, u_{N'}\}$ the eigenvalues of $M'^{-1}W$. The measure of the empirical distribution of the eigenvalues is defined by

$$p(u)du = N'^{-1} \left\{ \delta(u_1) + \delta(u_2) + \cdots + \delta(u_{N'}) \right\}, \quad (24)$$

where $\delta(u)$ denotes the Dirac measure at u. In the large scale limit, the measure, Eq.(24) , converges almost everywhere to

$$p(u)du = \frac{\sqrt{(u-u_m)(u_M-u)}}{2\pi\alpha u} \theta(u_m < u < u_M) du, \quad (25)$$

where $u_m = (\sqrt{\alpha}-1)^2$ and $u_M = (\sqrt{\alpha}+1)^2$ [Watcher, 1978]. Calculating moments of Eq.(25), we obtain the following theorem:

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

$$2\lambda = (H^*(M+N) - H^{*2}) + \frac{(M-H^*)(N-H^*)}{2\pi\alpha} \left\{ J(s_t; 1) - 2\kappa J(s_t; 0) + \kappa^2 J(s_t; -1) \right\}, \quad (26)$$

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 $s_t = \max\left((\kappa - (1+\alpha))/2\sqrt{\alpha}, J^{-1}(2\pi\alpha\beta; 0)\right)$. Here $J^{-1}(\cdot; k)$ denotes the inverse function of $J(s; k)$.

5.4 *Delicate* Situations

In ordinary asymptotic analysis, one considers only situations when the amplitude of each component of the true model is zero or *distinctly-positive*. Also Theorem 2 holds only in such situations. However, as mentioned in the last paragraph of Section 4, it is important to consider *delicate* situations when the true map B^*A^* has tiny but non-negligible singular values such that $0 < \sqrt{n}\gamma_h^* < \infty$. Theorem 1 still holds in such situations by replacing the second term of Eq.(16) with $o(n^{-1/2})$. We regard H^* as the number of *distinctly-positive* true singular values such that $\gamma_h^{*-1} = o(\sqrt{n})$. 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

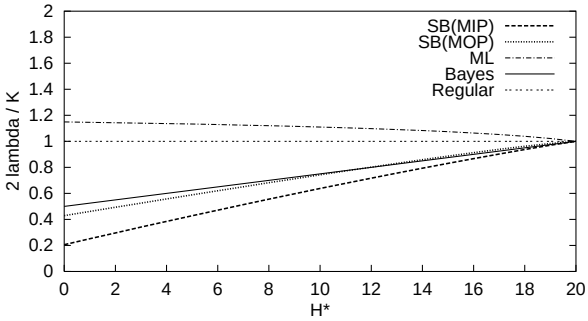

Figure 1: Generalization error.

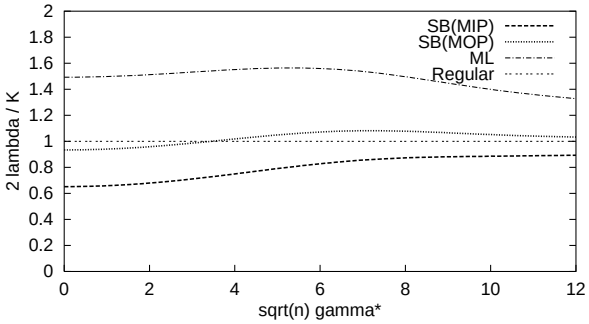

Figure 2: With *delicate* true components.

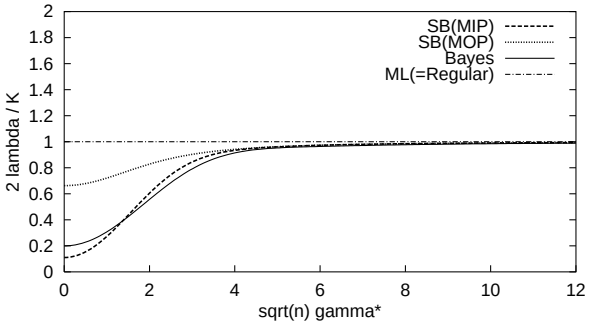

Figure 3: Single-output LNN.

first H^* columns and rows from B^*A^*. Then, D', defined in the proof of Theorem 2, is the general diagonalized matrix of $n^{-1/2}R''$, where R'' is a random matrix such that $R''R''^t$ is subject to $\mathcal{W}_{N-H^*}(M-H^*, I_{N-H^*}, nR''^*R''^*)$. Therefore, we obtain the following theorem:

Theorem 4 *The 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*

$$2\lambda = (H^*(M+N) - H^{*2}) + \sum_{h=H^*+1}^{H} n\gamma_h^{*2} + \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h''^2 > L) \left\{ \left(1 - \frac{L}{\gamma_h''^2}\right)^2 \gamma_h''^2 - 2\left(1 - \frac{L}{\gamma_h''^2}\right) \gamma_h'' \omega_{b_h}''^t \sqrt{n} R'^* \omega_{a_h}'' \right\} \right\rangle_{q(R'')}, \quad (27)$$

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.

6 Discussions

6.1 Comparison with the ML Estimation and with the Bayes Estimation

Figure 1 shows the generalization coefficients of an LNN with $M = 50$ input, $N = 30$ output, and $H = 20$ hidden units. The horizontal axis indicates the true rank H^*. The vertical axis indicates the coefficients normalized by the parameter dimension K, given by Eq.(4). The lines correspond to the generalization coefficients of the SB approaches, clarified in this paper, that of the ML estimation, clarified in [Fukumizu, 1999], that of the Bayes estimation, clarified in [Aoyagi and Watanabe, 2004], and that of the regular models, respectively.[4] The results in Fig. 1 have been calculated in the large scale approximation, i.e., by using Theorem 3. We have also numerically calculated them by creating samples subject to the Wishart distribution and then using Theorem 2, and thus found that the both results almost coincide with each other so that we can hardly distinguish. We see in Fig. 1 that the SB approaches provide as good performance as the Bayes estimation, and that the MIP, moreover, has no greater generalization coefficient than the Bayes estimation for arbitrary H^*, which might seem to be inconsistent with the proved superiority of the Bayes estimation to any other learning method when we use the true prior distribution. This suspicion is cleared by consideration of *delicate* situations in the following.

Using Theorem 4, we can numerically calculate the SB, as well as the ML, generalization error in *delicate* situations when the true distribution is near the singularities. Figure 2 shows the coefficients of an 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 $h = 2, \ldots, 4$. The Bayes generalization error in *delicate* situations was previously clarified [Watanabe and Amari, 2003], but unfortunately, only in single-output (SO) LNNs, i.e., $N = H = 1$.[5] Figure 3 shows the coefficients of an SOLNN with $M = 5$ input units on the assumption that $H^* = 0$ and the true singular value of the one component, indicated by the horizontal axis, is *delicate*. We see in Fig. 3 that the SB approaches have a property similar to the Bayes estimation, suppression of overfitting by the entropy of the singularities. We also see that in some *delicate* situations, the MIP is worse than the Bayes estimation, which shows consistency with the superiority of the Bayes estimation. We conclude that in typical cases, the suppression by the singularities in the MIP is comparable to, or sometimes stronger than, that

[4] In the regular models, the normalized generalization coefficient is always equal to one, which leads to the penalty term of Akaike's information criterion [Akaike, 1974].

[5] An SOLNN is regarded as a regular model at a view point of the ML estimation because the transform $b_1 a_1 \mapsto w \in \mathbb{R}^M$ makes the model linear and hence identifiable, and therefore, the ML generalization error is identical to that of the regular models. Nevertheless, an SOLNN has a property of unidentifiable models at a view point of the Bayesian learning methods, as shown in Fig. 3.

in the Bayes estimation.

It would be more fortunate if any of the SB approaches, which require much less computational costs than MCMC methods, would always provide comparable generalization performance to the Bayes estimation. However, the SB approaches have also a property similar to the ML estimation, acceleration of overfitting by selection of the largest singular values of a random matrix. Because of selection from a large number of random variables subject to non-compact support distribution, the $(H - H^*)$ largest eigenvalues of a random matrix subject to $\mathcal{W}_{N-H^*}(M - H^*, I_{N-H^*})$ are much greater than L when $(M - H^*) > (N - H^*) \gg (H - H^*)$. Therefore, the eigenvalues $\{\gamma_h'^2\}$ in Theorem 2 go out of the effective range of shrinkage, and consequently, the SB approaches approximate the ML estimation in such atypical cases. Actually, the generalization coefficient of an LNN in the case that $M = N = 80, H = 1, H^* = 0$ is $2\lambda/K \sim 1.04$, which is greater than that of the regular models, though in the Bayes estimation, the generalization coefficient never exceeds that of the regular models [Watanabe, 2001].

6.2 Relation to Shrinkage Estimation

The relation between an EB approach in a linear model and the JS estimator, into which the SB estimator, Eq.(16), in an SOLNN is changed by letting $L = (M - 2)$, was discussed in [Efron and Morris, 1973]. Based on the EB approach, the JS estimator can be derived as the solution of an equation with respect to an unbiased estimator of the hyperparameter τ_1^{-2}, introduced in Section 3.2. The similarity between the JS and the SB estimators is natural because in an SOLNN, the transform $b_1 a_1 \mapsto w \in \mathbb{R}^M$ makes not only the model linear but also the prior distribution as Eq.(10).

We focus on SOLNNs in this paragraph. In Fig. 3, the SB approaches and the Bayes estimation seem to be superior to the ML estimation regardless of the true distribution. The following asymptotic expansion of the generalization coefficient with respect to $\sqrt{n}\gamma_1^*$ provides a clue when it occurs:

$$2\lambda = M - \xi(\sqrt{n}\gamma_1^*)^{-2} + o\left((\sqrt{n}\gamma_1^*)^{-2}\right), \qquad (28)$$

where ξ is the coefficient of the leading term when γ_1^* 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 regular models. It was found that $\xi = (M - 1)(M - 3)$ in the Bayes estimation, which leads to the conjecture that the Bayes estimation would be superior to the ML estimation when $M \geq 4$ [Watanabe and Amari, 2003]. Similarly expanding Eq.(27), we get $\xi = M(M - 4)$ in the MIP, as well as $\xi = (2M - 5)$ in the MOP, which leads to the conjecture that the MIP when $M \geq 5$, as well as the MOP when $M \geq 3$, would be superior to the ML estimation. We have found that the numerical calculation by using Theorem 4 supports the conjecture above [Nakajima and Watanabe, 2005a]. We also find that $\xi = (M - 2)^2$ in the JS estimation, which is consistent with its proved superiority to the ML estimation when $M \geq 3$.

6.3 Relation to Variational Bayes Approach

The generalization error of the variational Bayes (VB) approach in LNNs has just been clarified [Nakajima and Watanabe, 2005b]. In the parameter subspace corresponding to the redundant components, the VB posterior distribution extends with its variance of order 1 in the larger dimension parameter subspace either input one or the output one; while the SB posterior distribution extends with its variance of order 1 in the parameter space, not in the hyperparameter space, as we find from Eqs. (30) and (37) in Appendix A. Consequently, the VB approach is asymptotically equivalent to the MIP version of SB approach.

6.4 Future Works

A future work can be consideration of the effect of non-linearity of the activation function, $\psi(\cdot)$ in Eq.(1). We expect that the non-linearity would extend the range of basis selection and hence increase the generalization error.

7 Conclusions

We have introduced a subspace Bayes (SB) approach, an empirical Bayes approach where a part of the parameters are regarded as hyperparameters, and derived the solution of two versions of SB approach in three-layer linear neural networks. We have also clarified its generalization error and concluded that the SB approaches have a property similar to the Bayes estimation and provide as good performance as the Bayes estimation in typical cases.

Acknowledgments

The authors would like to thank the reviewers who gave us meaningful advice, which motivates us to add Section 4 and some comments in other sections. They also would like to thank Kazuo Ushida, Masahiro Nei, and Nobutaka Magome of Nikon Corporation for encouragement to research.

A Proof of Theorem 1

First, we will prove in the MIP version. 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.(12), is maximized. Accordingly, we assume without loss of generality that the optimum value of B consists of its orthogonal column vectors. Consequently, the marginal (conditional) likelihood, as well as the posterior distribution, factorizes as

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

Then, substituting Eqs.(11) and (12) into Eq.(5), as well as Eq.(6), we can easily derive the marginal likelihood, as well as the posterior distribution, of each component as follows:

$$Z(Y^n|X^n\|b_h) \propto |S_h|^{-1/2} \exp\left(\frac{n b_h^t R S_h^{-1} R^t b_h}{2}\right), \qquad (29)$$

$$p(a_h|X^n, Y^n\|b_h) \propto \exp\left(-\left(a_h - S_h^{-1}R^t b_h\right)^t \frac{nS_h}{2}\left(a_h - S_h^{-1}R^t b_h\right)\right), \qquad (30)$$

where $S_h = (\|b_h\|^2 Q + n^{-1} I_M)$. Let $F'(Y^n|X^n\|b_h) = F(Y^n|X^n\|b_h) + \text{const.}$, where $F(Y^n|X^n\|b_h)$ is the stochastic complexity, i.e., the negative log marginal likelihood, of the h-th component. Then, we get

$$2F'(Y^n|X^n\|b_h) = -2\log Z(Y^n|X^n\|b_h) + \text{const.} \\ = \log|S_h| - n b_h^t R S_h^{-1} R^t b_h \quad (31)$$

Hereafter, separately considering the components imitating the positive true ones and the redundant components, we will minimize Eq.(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. Then, from Eq.(31), we get

$$2F'(b_h) = M\log\|b_h\|^2 - n\|b_h\|^{-2} b_h^t R Q^{-1} R^t b_h \\ + \|b_h\|^{-4} b_h^t R Q^{-2} R^t b_h + O(n^{-1}). \quad (32)$$

To minimize Eq.(32), the leading, second term dominates the determination of the direction cosine of b_h and leads to $b_h = \|b_h\|(\omega_{b_h} + O(n^{-1}))$. The first and the third terms determine the norm of b_h because the second term is independent of it. Thus, we get the optimal hyperparameter value as follows:

$$\hat{b}_h = \sqrt{\frac{\omega_{b_h}^t R Q^{-2} R^t \omega_{b_h}}{M}}\omega_{b_h} + O(n^{-1}). \quad (33)$$

Because the average of a_h over the posterior distribution, Eq.(30), is $\hat{a}_h = S_h^{-1} R^t b_h$, we obtain the SB estimator for the positive true component of the map BA as follows:

$$\hat{b}_h \hat{a}_h^t = \hat{\omega}_{b_h}\hat{\omega}_{b_h}^t R Q^{-1} + O(n^{-1}). \quad (34)$$

On the other hand, for a redundant component, $h > H^*$, Eq.(15) allows us to approximate Eq.(31) as follows:

$$2F'(b_h) = M\log\left(\|b_h\|^2 + n^{-1}\right) - \frac{n b_h^t R R^t b_h}{\|b_h\|^2 + n^{-1}} + O(n^{-1/2}). \quad (35)$$

Then, we find that the direction cosine of b_h, determined by the second term of Eq.(35), is approximated by ω_{b_h} with accuracy $O(n^{-1/2})$. After substituting $\gamma_h^2\|b_h\|^2(1+O(n^{-1/2}))$ for $b_h^t R R^t b_h$, we get the following extreme condition by partial differentiation of Eq.(35) with respect to the norm of b_h:

$$0 = 2\frac{\partial F'(b_h)}{\partial\|b_h\|^2} = \frac{M}{\left(\|b_h\|^2 + n^{-1}\right)^2}\left(\|b_h\|^2 - \frac{n\gamma_h^2 - M}{nM}\right) \\ + O(\|b_h\|^{-2} n^{-1/2}). \quad (36)$$

We find that Eq.(36) has no solution if γ_h is less than $\sqrt{M/n}$. Therefore, using the fact that Eq.(35) diverges to infinity with arbitrary n if $\|b_h\| \to \infty$, we get the optimum hyperparameter value as follows:

$$\hat{b}_h = \sqrt{\frac{L_h' - M}{nM}}\omega_{b_h} + O(n^{-1}). \quad (37)$$

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

$$\hat{b}_h \hat{a}_h^t = (1 - (L_h')^{-1} M)\omega_{b_h}\omega_{b_h}^t R + O(n^{-1}). \quad (38)$$

Selecting the largest singular value components minimizes Eq.(31). Hence, combining Eqs.(34) and (38), and then using Eq.(15), 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.)

References

[Akaike, 1974] H. Akaike. A new look at statistical model. *IEEE Trans. on Automatic Control*, 19:716–723, 1974.

[Akaike, 1980] H. Akaike. Likelihood and bayes procedure. In *J. M. Bernald, Bayesian statistics*, pages 143–166, Valencia, Italy, 1980. University Press.

[Aoyagi and Watanabe, 2004] M. Aoyagi and S. Watanabe. The generalization error of reduced rank regression in bayesian estimation. In *Proc. of ISITA*, pages 1068–1073, Parma, Italy, 2004.

[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*, 6:837–858, 1995.

[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, 1999] K. Fukumizu. Generalization error of linear neural networks in unidentifiable cases. In *Proc. of ALT*, pages 51–62. Springer, 1999.

[Ghahramani and Beal, 2000] Z. Ghahramani and M. J. Beal. Graphical models and variational methods. In *Advanced Mean Field Methods*. MIT Press, 2000.

[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*, 1993.

[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, Berkeley:University of California, 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.

[MacKay, 1995] 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.

[Nakajima and Watanabe, 2005a] S. Nakajima and S. Watanabe. Generalization performance of subspace bayes approach in linear neural networks. *Submitted to IEICE Trans.*, 2005.

[Nakajima and Watanabe, 2005b] S. Nakajima and S. Watanabe. Variational bayes solution of linear neural networks and its generalization error and training error. *Submitted to ICML*, 2005.

[Reinsel and Velu, 1998] G. C. Reinsel and R. P. Velu. *Multivariate Reduced-Rank Regression*. Springer, 1998.

[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, 2001] S. Watanabe. Algebraic information geometry for learning machines with singularities. *Advances in NIPS*, 13:329–336, 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, 2003] K. Yamazaki and S. Watanabe. Singularities in mixture models and upper bounds of stochastic complexity. *Neural Networks*, 16(7):1029–1038, 2003.