ARTICLE Communicated by Noboru Murata

Singularities Affect Dynamics of Learning in Neuromanifolds

Shun-ichi Amari
amari@brain.riken.jp
RIKEN Brain Science Institute, Saitama, 351-0198, Japan

Hyeyoung Park
hypark@knu.ac.kr
Kyungpook National University, Korea

Tomoko Ozeki
tomoko@brain.riken.jp
RIKEN Brain Science Institute, Saitama, 351-0198, Japan

The parameter spaces of hierarchical systems such as multilayer perceptrons include singularities due to the symmetry and degeneration of hidden units. A parameter space forms a geometrical manifold, called the neuromanifold in the case of neural networks. Such a model is identified with a statistical model, and a Riemannian metric is given by the Fisher information matrix. However, the matrix degenerates at singularities. Such a singular structure is ubiquitous not only in multilayer perceptrons but also in the gaussian mixture probability densities, ARMA time-series model, and many other cases. The standard statistical paradigm of the Cramér-Rao theorem does not hold, and the singularity gives rise to strange behaviors in parameter estimation, hypothesis testing, Bayesian inference, model selection, and in particular, the dynamics of learning from examples. Prevailing theories so far have not paid much attention to the problem caused by singularity, relying only on ordinary statistical theories developed for regular (nonsingular) models. Only recently have researchers remarked on the effects of singularity, and theories are now being developed.

This article gives an overview of the phenomena caused by the singularities of statistical manifolds related to multilayer perceptrons and gaussian mixtures. We demonstrate our recent results on these problems. Simple toy models are also used to show explicit solutions. We explain that the maximum likelihood estimator is no longer subject to the gaussian distribution even asymptotically, because the Fisher information matrix degenerates, that the model selection criteria such as AIC, BIC, and MDL fail to hold in these models, that a smooth Bayesian prior becomes singular in such models, and that the trajectories of dynamics of learning are strongly affected by the singularity, causing plateaus or slow

manifolds in the parameter space. The natural gradient method is shown to perform well because it takes the singular geometrical structure into account. The generalization error and the training error are studied in some examples.

1 Introduction

The multilayer perceptron is an adaptive nonlinear system that receives input signals and transforms them into adequate output signals. Learning takes place by modifying the connection weights and the thresholds of neurons. Since perceptrons are specified by a set of these parameters, we may regard the whole set of perceptrons as a high-dimensional space or manifold whose coordinate system is given by these modifiable parameters. We call this a neuromanifold.

Let us assume that the behavior of a perceptron is disturbed by noise, so that by receiving input signal $\boldsymbol{x}$, a perceptron emits output y stochastically. This stochastic behavior is determined by the parameters. Let us also assume that a pair $(\boldsymbol{x}, y)$, of input signal $\boldsymbol{x}$ and corresponding answer y is given from the outside by a teacher. A number of examples are generated by an unknown probability distribution, $p_0(\boldsymbol{x}, y)$ or the conditional probability $p_0(y|\boldsymbol{x})$, of the teacher. Let us denote the set of examples as $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)$. A perceptron learns from these examples to imitate the stochastic behavior of the teacher.

The behavior of a perceptron under noise is described by a conditional probability distribution $p(y|\boldsymbol{x})$, which is the probability of output y given input $\boldsymbol{x}$, so it can be regarded as a statistical model that includes a number of unknown parameters. From the statistical point of view, estimation of the parameters is carried out from examples generated by an unknown probability of the teacher network. Learning, especially online learning, is a type of estimation where the parameters are modified sequentially, using examples one by one. The parameters change by learning, forming a trajectory in the neuromanifold. Therefore, we need to study the geometrical features of the neuromanifold to elucidate the behavior of learning.

The neuromanifold of multilayer perceptrons is a special statistical model because it includes singular points, where the Fisher information matrix degenerates. This is due to the symmetry of hidden units, and the number of hidden units substantially decreases when two hidden neurons are identical. The identifiability of parameters is lost at such singular positions. Such a structure was described in the pioneering work of Brockett (1976) in the case of linear systems, and in the case of multilayer perceptrons by Chen, Lu, and Hecht-Nielsen (1993), Sussmann (1992), Kůrková & Kainen (1994), and Rüger & Ossen (1997). This type of structure is ubiquitous in many hierarchical models such as the model of probability densities given by gaussian mixtures, the ARMA time-series model, and the model of linear systems whose transfer functions are given by rational functions. The

Riemannian metric degenerates at such singular points, which are not isolated but form a continuum. In all of these models, when we summarize the parameters that give the same behavior (probability distribution), the set of behaviors is known to have a generic cone-type singularity embedded in a finite-dimensional, sometimes infinite-dimensional, regular manifold (Dacunha-Castelle & Gassiat, 1997).

Many interesting problems arise in such singular models. Since the Fisher information matrix degenerates at singular points, its inverse does not exist. Therefore, the Cramér-Rao paradigm of classic statistical theory cannot be applied. The maximum likelihood estimator is no longer subject to the gaussian distribution even asymptotically, although the consistency of behavior holds. The criteria used for model selection (such as AIC, BIC, and MDL) are derived from the gaussianity of the maximum likelihood estimator, with the covariance given by the inverse of the Fisher information, so that their validity is lost in such hierarchical models. The generalization error has so far been evaluated based on the Cramér-Rao paradigm, so we need a new theoretical method to attack the problem. This is related to the strange behavior of the log-likelihood ratio statistic in a singular model. These problems have not been fully explored by conventional statistics.

When the true distribution lies at a regular point, the classical Cramér-Rao paradigm is still valid, provided it is sufficiently separated from a singular point. However, the dynamics of learning is a global problem that takes place throughout the entire neuromanifold. It has been shown that once parameters are attracted to singular points, at the parameters are very slow to move away from them. This is the plateau phenomenon ubiquitously observed in backpropagation learning. Therefore, even when the true point lies at a regular point, the singular structure strongly affects the dynamics of learning. Although there is no unified theory, problems caused by the singular structure in hierarchical models have been remarked by many researchers, and various new approaches have been proposed. This is a new area of research that is attracting much attention. Hagiwara, Toda, and Usui (1993) noticed this problem first. They used AIC (the Akaike information criterion) to determine the size of perceptrons to be used for learning and found that AIC did not work well. AIC is a criterion to determine the model that minimizes the generalization error. However, it has been reported for a long time that AIC does not give good model selection performance in the case of multilayer perceptrons. Hagiwara et al. found that this is because of the singular structure of such a hierarchical model and investigated ways to overcome this difficulty (Hagiwara, 2002a, 2002b; Hagiwara, Hayasaka, Toda, Usui, and Kuno, 2001; Hagiwara et al., 1993; Kitahara, Hayasaka, Toda, & Usui, 2000).

To accelerate the dynamics of learning, Amari (1998) proposed the natural or Riemannian gradient method of learning, which takes into account the geometrical structure of the neuromanifold. Through this method, one can avoid the plateau phenomenon in learning (Amari, 1998; Amari &

Ozeki, 2001; Amari, Park, & Fukumizu, 2000; Park, Amari, & Fukumizu, 2000). However, the Fisher information matrix, or the Riemannian metric, degenerates at some places because of singularity, so we need to develop a new theory of dynamics of learning (see Amari, 1967, for dynamics of learning for regular multilayer perceptrons) to understand its behavior, in particular the effects of singularity in the ordinary backpropagation method and the natural gradient method. Work done regarding this aspect includes that of Fukumizu and Amari (2000) as well as the statistical-mechanical approaches taken by Saad and Solla (1995), Rattray, Saad, and Amari (1998), and Rattray and Saad (1999).

Fukumizu used a simple linear model and indicated that the generalization error of multilayer perceptrons with singularities is different from that of the regular statistical model (Fukumizu, 1999). This problem is related to the analysis of the log-likelihood-ratio statistic of the perceptron model at a singularity (Fukumizu, 2003). The strange behavior of the log-likelihood statistic in the gaussian mixture has been remarked since the time of Hotelling (1939) and Weyl (1939), but only recently has it become possible to derive its asymptotic behavior (Hartigan, 1985; Liu & Shao, 2003). Fukumizu (2003) extended the idea of the gaussian random field (Hartigan, 1985; Dacunha-Castelle & Gassiat, 1997) to make it applicable to multilayer perceptrons and formulated the asymptotics of the generalization error of multilayer perceptrons. Watanabe (2001a, 2001b, 2001c) was the first who studied the effect of singularity in Bayesian inference. He and his colleagues introduced algebraic geometry and algebraic analysis by using Hironaka's theorem of singularity resolution and Sato's formula in algebraic analysis to evaluate the asymptotic performance of the Bayesian predictive distribution in various hierarchical singular models; remarkable results have been derived (Watanabe, 2001a, 2001b, 2001c; Yamazaki & Watanabe, 2002, 2003; Watanabe & Amari, 2003).

In this article, we give an overview concerning the strange behaviors of singular models so far studied. They include estimation, testing, Bayesian inference, model selection, generalization, and training errors. Special attention is paid to the dynamics of learning from the information-geometric point of view by summarizing our previous results (Amari & Ozeki, 2001; Amari, Park, & Ozeki, 2001, 2002; Amari, Ozeki, & Park, 2003). In particular, we show new results concerning the fast and slow submanifolds in learning of gaussian mixtures.

The article is organized as follows. We give various examples of singular models in section 2. They include models of simple multilayer perceptrons with one hidden unit and two hidden units, the gaussian mixture model of probability distributions, and a toy model of the cone that is used to give an exact analysis. The analysis of the gaussian mixture is newly presented here. In section 3, we explain the theory developed by Dacunha-Castelle and Gassiat (1997), which shows the generic cone structure of a singular model. This elucidates why strange behaviors emerge in a singular model.

We then show that such models with singularity have strange behaviors, differing from those of ordinary regular statistical models, in parameter estimation, Bayesian predictive distribution, the dynamics of learning, and model selection in section 4. Section 5 is devoted to a detailed analysis of the dynamics of learning. We use simple models to show that there appear slow and fast manifolds in the neighborhood of a singularity, which explains the plateau phenomenon in the ordinary gradient learning method. The natural gradient method is shown to resolve this difficulty. Section 6 deals with the generalization error and training error in simple singular models, where the gaussian random field (Hartigan, 1985; Dacunha-Castelle & Gassiat, 1997; Fukumizu, 2003) is used and the special potential functions are introduced (Amari & Ozeki, 2001). Explicit formulas will be given in the cases of the maximum likelihood estimator and the Bayesian estimator.

2 Singular Statistical Models and Their Geometrical Structure

The manifolds, or the parameter spaces, of many hierarchical models such as multilayer perceptrons inevitably include singularities. Typical examples are presented in this section.

2.1 Single Hidden Unit Perceptron. We begin with a trivial model of a single hidden unit perceptron, which receives input vector $\boldsymbol{x} = (x_1, \cdots, x_m)$ and emits scalar output y. The hidden unit calculates the weighted sum $\boldsymbol{w} \cdot \boldsymbol{x} = \sum w_i x_i$ of the input, where $\boldsymbol{w} = (w_1, \cdots, w_m)$ is the weight vector, and emits its nonlinear function $\varphi(\boldsymbol{w} \cdot \boldsymbol{x})$ as the output, where $\varphi(u)$ is the activation function $\varphi(u) = \tanh(u)$. The output unit is linear, so its output is $v\varphi(\boldsymbol{w} \cdot \boldsymbol{x})$, which is eventually disturbed by gaussian noise ϵ. Hence, the final output is

$$y = v\varphi(\boldsymbol{w} \cdot \boldsymbol{x}) + \epsilon. \tag{2.1}$$

The parameters to specify a perceptron are summarized into a single vector $\boldsymbol{\theta} = (\boldsymbol{w}, v)$. The average of y, given $\boldsymbol{x}$, is

$$E[y] = f(\boldsymbol{x}, \boldsymbol{\theta}) = v\varphi(\boldsymbol{w} \cdot \boldsymbol{x}), \tag{2.2}$$

where E denotes expectation, given $\boldsymbol{x}$.

Any point $\boldsymbol{\theta}$ in the $(m+1)$-dimensional parameter space $M = \{\boldsymbol{\theta}\}$ specifies a perceptron and its average output function $f(\boldsymbol{x}, \boldsymbol{\theta})$. However, the parameters are redundant or unidentifiable in some cases. Since φ is an odd function, $(\boldsymbol{w}, v)$ and $(-\boldsymbol{w}, -v)$ give the same function,

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = f(\boldsymbol{x}, -\boldsymbol{\theta}). \tag{2.3}$$

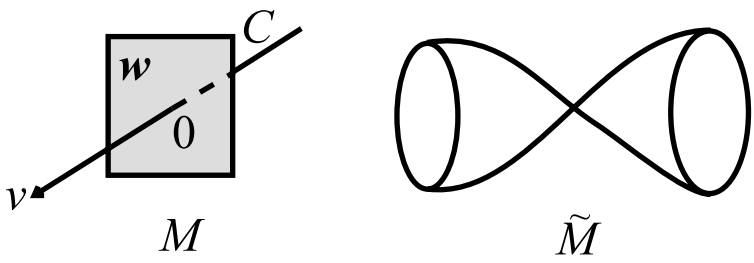

Figure 1: (Left) Parameter space M of a single hidden unit perceptron. (Right) Space $\tilde{M}$ of the average output functions of perceptrons with a singular structure.

Moreover, when $v = 0$ (or $\boldsymbol{w} = 0$), $f(\boldsymbol{x}, \boldsymbol{\theta}) = 0$, whatever value $\boldsymbol{w}$ takes (v takes). Hence, the function $f(\boldsymbol{x}, \boldsymbol{\theta})$ is the same in the set $C \subset M$,

$$C = \{(v, \boldsymbol{w}) | \; v\|\boldsymbol{w}\| = 0\} \tag{2.4}$$

(see Figure 1, left), which we call the critical set. Therefore, if we summarize the points that have the same average output function $f(\boldsymbol{x}, \boldsymbol{\theta})$ into one, the parameter space shrinks such that the critical set C is reduced to a single point. The parameter space reduces to $\tilde{M}$, which consists of all the different average output functions of perceptrons. It consists of two components connected by a single point (see Figure 1, right). In other words, $\tilde{M}$ has singularity. Note that M is the parameter space, each point of which corresponds to a perceptron. However, the behaviors of some perceptrons are the same even if their parameters are different. The reduced $\tilde{M}$ is the set of perceptron behaviors that correspond to the average output functions or the probability distributions specified thereby.

The probability density function of the input-output pair $(\boldsymbol{x}, y)$ is given by

$$p(y, \boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}} q(\boldsymbol{x}) \exp\left\{-\frac{1}{2}(y - f(\boldsymbol{x}, \boldsymbol{\theta}))^2\right\}, \tag{2.5}$$

where ϵ in equation 2.1 is subject to the standard gaussian distribution $N(0, 1)$ and $q(\boldsymbol{x})$ is the probability density function of input $\boldsymbol{x}$. The Fisher information matrix is defined by

$$G(\boldsymbol{\theta}) = E\left[\frac{\partial l(y, \boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial l(y, \boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^T\right], \tag{2.6}$$

where E denotes the expectation with respect to $p(y, \boldsymbol{x}, \boldsymbol{\theta})$ and $l(y, \boldsymbol{x}, \boldsymbol{\theta}) = \log p(y, \boldsymbol{x}, \boldsymbol{\theta}) = -\frac{1}{2}(y - f(\boldsymbol{x}, \boldsymbol{\theta}))^2 + \log q(\boldsymbol{x})$, is a fundamental quantity in statistics. It is positive definite in a regular statistical model and plays the role of the Riemannian metric of the parameter space, as is shown by the information geometry (Amari & Nagaoka, 2000).

The Fisher information gives the average amount of information included in one pair $(y, \boldsymbol{x})$ of data, which is used to estimate the parameter $\boldsymbol{\theta}$.

Cramér-Rao Theorem. *Let $\hat{\boldsymbol{\theta}}$ be an unbiased estimator from n examples in a regular statistical model. Then the error covariance of $\hat{\boldsymbol{\theta}}$ satisfies*

$$E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T] \geq \frac{1}{n} G^{-1}(\boldsymbol{\theta}). \tag{2.7}$$

Moreover, the equality holds asymptotically (i.e., for large n) for the MLE (maximum likelihood estimator) $\hat{\boldsymbol{\theta}}$, and it is asymptotically subject to the gaussian distribution with mean $\boldsymbol{\theta}$ and covariance matrix $(1/n)G^{-1}(\boldsymbol{\theta})\}$.

However, this does not hold when $v = 0$ or $\boldsymbol{w} = 0$, because

$$\frac{\partial l(y, \boldsymbol{x}, \boldsymbol{\theta})}{\partial v} = 0 \tag{2.8}$$

or

$$\frac{\partial l(y, \boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{w}} = 0. \tag{2.9}$$

When $v = 0$, $p(y, \boldsymbol{x}, \boldsymbol{\theta})$ is kept constant even when $\boldsymbol{w}$ changes, and the same situation holds when $\boldsymbol{w} = 0$. Hence, the Fisher information matrix, equation 2.6, degenerates, and its inverse $G^{-1}(\boldsymbol{\theta})$ does not exist in the set $C : v\|\boldsymbol{w}\| = 0$. The Cramér-Rao theorem is no longer valid at the critical set C. The model is singular on C. This makes it difficult to analyze the performance of estimation and learning when the true distribution is in C or in the neighborhood of C.

2.2 Gaussian Mixture. The Gaussian mixture is a statistical model of probability distributions that has long been known to include singularities (Hotelling, 1939; Weyl, 1939). Let us assume that the probability density of the real variable x is given by a mixture of k gaussian distributions as

$$p(x, \boldsymbol{\theta}) = \sum_{i=1}^{k} v_i \psi(x - \mu_i), \tag{2.10}$$

where $\psi(x) = (1/\sqrt{2\pi})\exp\{-x^2/2\}$. The parameters are $\boldsymbol{\theta} = (v_1, \cdots, v_k; \mu_1, \cdots, \mu_k)$ and $0 \le v_i \le 1$, $\sum v_i = 1$. When we know the number k of components, all we have to do is to estimate the unknown parameters $\boldsymbol{\theta}$ from observed data $x_1, x_2, \cdots, x_n$. However, in the usual situation, k is unknown.

To make the story simple, let us consider the case of $k = 2$. The model is given by

$$p(x, \boldsymbol{\theta}) = v\psi(x - \mu_1) + (1 - v)\psi(x - \mu_2). \tag{2.11}$$

The parameter space M is three-dimensional, $\boldsymbol{\theta} = (v, \mu_1, \mu_2)$. If $\mu_1 = \mu_2$ holds, $p(x, \boldsymbol{\theta})$ actually consists of one component, as is the case with $k = 1$. In this case, the distribution is the same whatever value v takes, so we cannot identify v. Moreover if $v = 0$ or $v = 1$, the distribution is the same whatever value μ_1 or μ_2 takes, so the parameters are unidentifiable. Hence, in the parameter space $M = \{\boldsymbol{\theta}\}$, some parameters are unidentifiable in the region C:

$$C : v(1 - v)(\mu_1 - \mu_2) = 0. \tag{2.12}$$

This is depicted in the shaded area in Figure 2. We call this the critical set. In the critical set, the determinant of the Fisher information matrix becomes 0, and there is no inverse matrix. Let us look at the critical set $C \subset M$ carefully. In C, any point on the three lines—$\mu_1 = \mu_2 = \mu_0$, $v = 0$, $\mu_2 = \mu_0$, and $v = 1$, $\mu_1 = \mu_0$—(see Figure 2 left), represents the same gaussian distribution, $\psi(x - \mu_0)$. If we regard the parameter points representing the same distribution as one and the same, these three lines shrink to one point.

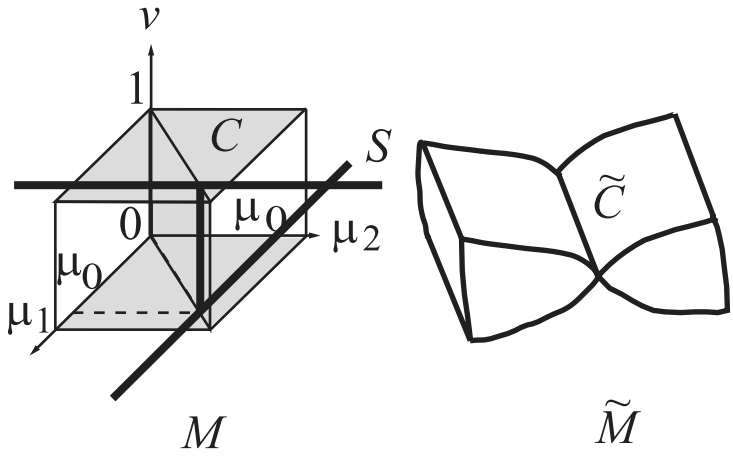

Figure 2: Parameter space of gaussian mixture M and singular structure $\tilde{M}$.

Mathematically speaking, we obtain the residue class $\tilde{M} = M/\approx$ through the equivalence relation $\approx$. Then we get the space $\tilde{M}$ depicted in Figure 2 (right), which is the set of probability distributions (not the set M of parameters). This is the space where singularities are accumulated on a line $\tilde{C}$ and the dimensionality is reduced on the line. The line corresponds to the critical set C.

To analyze the nature of singularity, let us introduce new variables w and u : w is the center of gravity of the two peaks, or the mean value of the distribution, and u is the difference between the locations of the two peaks,

$$w = v\mu_1 + (1 - v)\mu_2 \tag{2.13}$$

$$u = \mu_2 - \mu_1. \tag{2.14}$$

Estimation of the mean parameter w of the distribution is easy, because the Fisher information matrix is nondegenerate in this direction. The problem is estimation of u and v when $uv(1 - v)$ is small, because the critical region C is given by $uv(1 - v)$. To make discussions simpler, consider the case where $w = 0$ is known, and only u and v are unknown. The distribution is then written as

$$p(x, u, v) = v\psi\left\{x - (1 - v)u\right\} + (1 - v)\psi(x + vu). \tag{2.15}$$

Let us consider only the region where $u \approx 0$ and $v(1 - v) > c$ holds for some constant c. By Taylor expansion of the above equation around $u = 0$, we get

$$p(x, u, v) \approx \psi(x)\left\{1 + \frac{1}{2}c_2(v)H_2(x)u^2 + \frac{1}{6}c_3(v)H_3(x)u^3 + \frac{1}{24}c_4(v)H_4(x)u^4 + \frac{1}{120}c_5(v)H_5(x)u^5 + \cdots\right\}, \tag{2.16}$$

where

$$c_i(v) = v(1 - v)^i + (1 - v)(-v)^i,$$

meaning that

$$c_2(v) = v(1 - v),$$
$$c_3(v) = v(1 - v)(1 - 2v),$$
$$c_4(v) = v(1 - v)(1 - 3v + 3v^2),$$

and

$$H_2(x) = x^2 - 1,$$
$$H_3(x) = x^3 - 3x,$$
$$H_4(x) = x^4 - 6x^2 + 3$$

are the Hermite polynomials. We can embed this singular model locally in a regular model S as its singular subset. This is our new strategy of studying the singular structure by locally embedding it in an exponential family of distributions whose structure is well known and regular statistical analysis is possible.

Let us consider a regular exponential family S specified by the regular parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$,

$$p(x, \boldsymbol{\theta}) = \psi(x) \exp\left\{\theta_1 H_2(x) + \theta_2 H_3(x) + \theta_3 H_4(x)\right\}, \tag{2.17}$$

where $\psi(x)$ is a dominating measure. Let us denote by S the parameter space of $\boldsymbol{\theta}$. Now we calculate $l(x, u, v) = \log p(x, u, v)$, by taking the logarithm of equation 2.16 and performing Taylor expansion again,

$$l(x, u, v) = \log \psi + \frac{u^2}{2} c_2(v) H_2(x) + \frac{u^3}{6} c_3(v) H_3(x) + \frac{u^4}{24}\{c_4(v) H_4(x) - 3c_2(v)^2 H_2(x)^2\}. \tag{2.18}$$

Hence, the parameter space M of gaussian mixtures is approximately embedded in S by

$$\theta_1 = \frac{1}{2} c_2(v) u^2, \tag{2.19}$$

$$\theta_2 = \frac{1}{6} c_3(v) u^3, \tag{2.20}$$

$$\theta_3 = \frac{1}{24}\{c_4(v) - 3c_2(v)^2\} u^4. \tag{2.21}$$

This embedding from (u, v) to $\boldsymbol{\theta}$ is singular. We first consider how M is embedded in the two-dimensional space (θ_1, θ_2) by equations 2.19 and 2.20 in Figure 3a, where θ_3 is ignored. The shape near the singularity $u \approx 0$ becomes clear by using (u, v)-coordinates of M and their map in S. Let us consider a line in M where v is constant. This line $v = \text{const}$ is

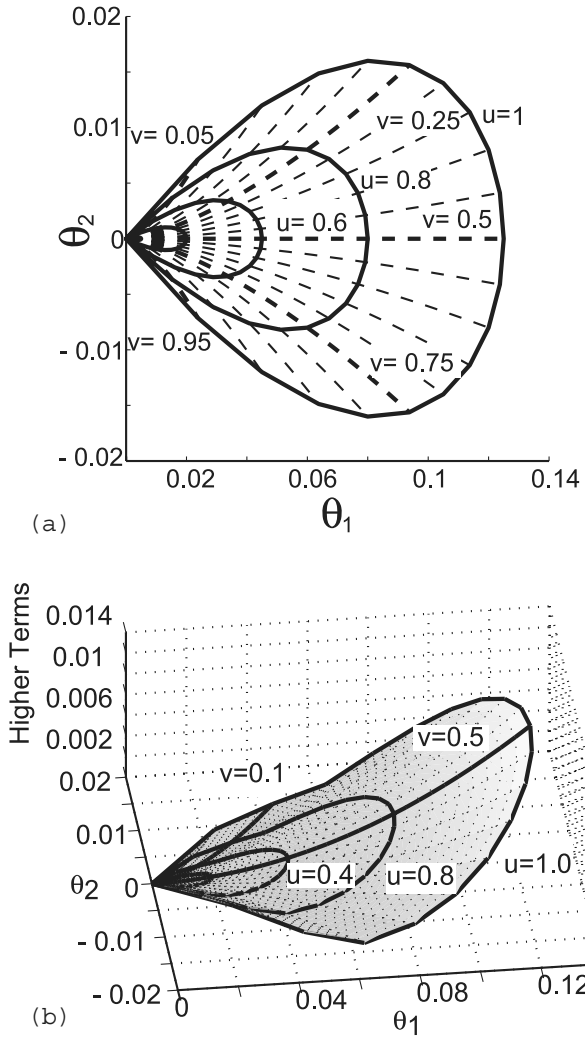

Figure 3: Gaussian mixture distribution M embedded in S. (a) M embedded in the parameter space S, (θ_1, θ_2). (b) The picture where M cannot be embedded in S and sticks out into the higher dimension.

embedded in S as

$$\theta_1 = a_1 u^2$$
$$\theta_2 = a_2 u^3,$$

where a_1 and a_2 are constants depending on v. This curve is not smooth but

is a cusp in S. The transformation from (u, v) to (θ_1, θ_2) is singular at $u = 0$, and the Jacobian determinant,

$$|J| = \det \begin{vmatrix} \frac{\partial \theta_1}{\partial u} & \frac{\partial \theta_2}{\partial u} \\ \frac{\partial \theta_1}{\partial v} & \frac{\partial \theta_2}{\partial v} \end{vmatrix}, \tag{2.22}$$

vanishes at $u = 0$. To elucidate the dynamics of learning near the singularity $u = 0$, θ_3 is necessary, as we will show later.

Note that the above approximation, which uses Taylor expansion of u, is not applicable in the case where u is not small, but v is close to 0 or 1. Equation 2.16 is valid only in the case of small u. Taylor expansion is not suitable when $v(1 - v)$ becomes small and u is large. Another consideration in terms of a random gaussian field is necessary, because an infinite-dimensional regular space is required to include the singular M. We have drawn the picture of M embedded in the three dimensions of $S = \{\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)\}$ by calculating the higher-order term of u^4 in Figure 3b. If u becomes large or v approaches 0 or 1, the surface of the embedded M includes higher terms that cannot be represented in equation 2.17 and embedded $\tilde{M}$ sticks out and is wrapped in higher dimensions. As $v(1 - v)$ approaches 0, the points in $\tilde{M}$ shrink toward the origin as $u \to 0$ but expand into infinite dimensions. This situation arises in the non-Donsker case (Dacunha-Castelle & Gassiat, 1997).

Our primary interest is to know the influence of the singularity on the dynamics of learning. Because the natural gradient learning method takes the geometrical structure into account, it will not be greatly affected by the singularity. However, the effect of the singularity is not negligible when the true distribution is at, or near, the singularity C. In gradient descent learning algorithms, the critical set C works as a pseudo-attractor (the Milnor attractor) even when the true distribution is at a regular point far from it. The singular point resembles a black hole in the dynamics of learning, and it is necessary to investigate its influence on the dynamics, as we discuss in a later section. The Fisher information matrix degenerates in the critical set. We calculate it explicitly in a later section.

2.3 Multilayer Perceptron. We consider here the multilayer perceptron with hidden units (Rosenblatt, 1961; Amari, 1967; Minsky & Papert, 1969; Rumelhart, Hinton, & Williams, 1986). It is a singular model where the output y is written as

$$y = \sum_{i=1}^{k} v_i \varphi(\boldsymbol{w}_i \cdot \boldsymbol{x}) + \varepsilon. \tag{2.23}$$

Here, $\boldsymbol{x}$ is an input vector, $\varphi(\boldsymbol{w}_i \cdot \boldsymbol{x})$ is the output of the ith neuron of the hidden layer, and $\boldsymbol{w}_i$ is its weight vector. The neuron of the output layer summarizes all the outputs of the hidden layer by taking the weighted sum with weights v_i. Gaussian noise $\varepsilon \sim N(0, 1)$ is added at the end, so the output y is a random variable.

Let us summarize all the modifiable parameters in one vector, $\boldsymbol{\theta} = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_k;\ v_1, \cdots, v_k)$, and then the average output is given by

$$E[y] = f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{k} v_i \varphi\left(\boldsymbol{w}_i \cdot \boldsymbol{x}\right). \tag{2.24}$$

This model has a structure similar to the gaussian mixture distribution.

When the parameter $\boldsymbol{w}_i$ of the ith neuron is 0, this neuron is useless because $\varphi(0) = 0$ and v_i may take any value. Moreover, when $v_i = 0$, whatever value $\boldsymbol{w}_i$ takes, this term is 0. In the meantime, if $\boldsymbol{w}_i = \boldsymbol{w}_j$ (or $\boldsymbol{w}_i = -\boldsymbol{w}_j$), these two neurons emit the same output (or the negative output). Then $v_i\varphi(\boldsymbol{w}_i \cdot \boldsymbol{x}) + v_j\varphi(\boldsymbol{w}_j \cdot \boldsymbol{x})$ is the same, not depending on particular values of v_i and v_j, provided $v_i + v_j$ (or $v_i - v_j$) takes a fixed value. Therefore, we can identify their sum (or difference), but each of v_i and v_j remains unidentifiable.

The neuromanifold M of perceptrons is a space whose admissible coordinate system is given by $\boldsymbol{\theta}$. The critical set C is defined by the join of the two subsets,

$$v_i \left\|\boldsymbol{w}_i\right\| = 0, \quad \boldsymbol{w}_i = \pm \boldsymbol{w}_j, \tag{2.25}$$

in which the parameters are unidentifiable. In other words, there is a direction such that the behavior (the input-output relation) is the same even when the parameters change in this direction. Since the Fisher information is 0 along this direction, the Fisher information matrix degenerates, and its inverse diverges to infinity.

In statistical study, the Cramér-Rao theorem guarantees that the error of estimation from a large number of data is given by the inverse of the Fisher information matrix in the regular case. However, because the inverse of the Fisher information matrix diverges, the classical theory is not applicable here. In geometrical terms, the Riemannian metric, which is determined by the Fisher information matrix, degenerates. Therefore, the distance becomes 0 along a certain direction. This is what the singular structure gives rise to.

Remark. We have absorbed the bias term in the weighted sum as

$$\boldsymbol{w}_i \cdot \boldsymbol{x} = \sum w_i x_i + w_{i0} x_0 = \tilde{\boldsymbol{w}}_i \cdot \tilde{\boldsymbol{x}} + w_{i0}, \tag{2.26}$$

where $x_0 = 1$. Because x_0 is constant, this causes another nonidentifiability. When

$$\tilde{\boldsymbol{w}}_i = \tilde{\boldsymbol{w}}_j = 0, \tag{2.27}$$

if the other parameters satisfy

$$v_i\varphi(w_{i0}) + v_j\varphi(v_{j0}) = \text{const.}, \tag{2.28}$$

they give the same behavior. In this article, we ignore such a case for simplicity. There are no other types of nonidentifiability (see Sussmann, 1992; Kůrková & Kainen, 1994).

We regard two perceptrons as being equivalent when their behaviors are the same even if their parameters are different. Let us shrink the manifold M by reducing the equivalent points of M to one point, giving the reduced $\tilde{M}$. In the mathematical sense, the reduced $\tilde{M}$ corresponds to the residue class of M because of the equivalence. In this case, degeneration of the dimensionality occurs in the critical set of the neuromanifold. We showed this in the trivial case of one hidden neuron. We now show this through another simple example. Let us consider a perceptron with two hidden units with $\|\boldsymbol{w}\| = 1$, $\boldsymbol{w} = (\cos\theta, \sin\theta)$, which is represented by

$$f(\boldsymbol{x}, v, \theta) = v\varphi(x_1\cos\theta + x_2\sin\theta + b), \tag{2.29}$$

where b is a fixed bias term. In this case, the parameter space is two-dimensional with coordinates $\boldsymbol{\theta} = (v, \theta)$. Consider the space S consisting of functions of the form $f(\boldsymbol{x}, \boldsymbol{\theta}) = \theta_3\varphi(\theta_1 x_1 + \theta_2 x_2 + b)$. Then equation 2.29 is embedded in S by $\theta_1 = \cos\theta$, $\theta_2 = \sin\theta$, and $\theta_3 = v$. The embedded $\tilde{M}$ in S,

$$\tilde{M} = \{\boldsymbol{\theta} |\ \theta_1 = \cos\theta, \theta_2 = \sin\theta, \theta_3 = v\},$$

is a cone depicted in Figure 4, and the apex is the singular point.

2.4 Cone Model. Amari and Ozeki (2001) analyzed a toy model, the cone model, to examine the exact behavior of the estimation error and the dynamics of learning. We introduce this model here for later use. Let us consider the statistical model described by a random variable $\boldsymbol{x} \in R^{d+2}$, which is subject to a gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix I, where I is the identity matrix,

$$p(\boldsymbol{x}; \boldsymbol{\mu}) = \frac{1}{(\sqrt{2\pi})^{d+2}} \exp\left\{-\frac{1}{2}||\boldsymbol{x} - \boldsymbol{\mu}||^2\right\}. \tag{2.30}$$

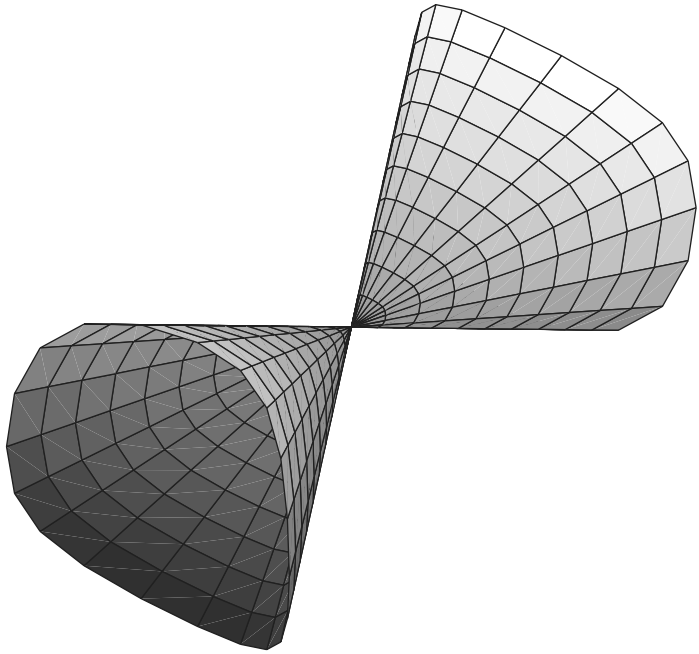

Figure 4: Singular structure of the neuromanifold.

The parameter space $S = \{\boldsymbol{\mu}\}$ is a $(d+2)$-dimensional Euclidean space with a coordinate system $\boldsymbol{\mu}$. When the mean parameter $\boldsymbol{\mu}$ is restricted on the surface of a cone in S, the family of the gaussian distributions is called the cone model M. We first consider the case with $d = 1$, that is, $d + 2 = 3$, in which the cone is given by

$$\mu_1 = \xi, \mu_2 = \xi \cos\theta, \mu_3 = \xi \sin\theta \tag{2.31}$$

(see Figure 5), where (ξ, θ) are the parameters used to specify M.

In the general case, the cone is parameterized by $(\xi, \boldsymbol{\omega})$,

$$M : \boldsymbol{\mu} = \frac{\xi}{\sqrt{1+c^2}} \begin{pmatrix} 1 \\ c\boldsymbol{\omega} \end{pmatrix} = \xi\, \boldsymbol{a}(\boldsymbol{\omega}), \tag{2.32}$$

where c is a constant that specifies the shape of the cone, $\boldsymbol{\omega}$ is a vector on the d-dimensional unit sphere that specifies the directions of the cone, and ξ is the distance from the origin. In the case of $d = 1$, the sphere is a circle (see Figure 5), and the parameter $\boldsymbol{\omega}$ is replaced by θ. The model M, which

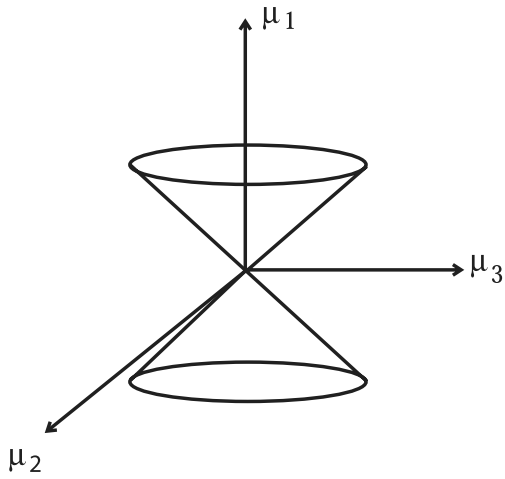

Figure 5: Cone Model.

is given by the parameters $(\xi, \boldsymbol{\omega})$, is embedded in the $d+2$–dimensional space S, and consists of two cones, one for $\xi \geq 0$ and the other for $\xi \leq 0$, of $d+1$ dimensions, $R \times S^d$, which are connected at the apex $\xi = 0$. The apex $\xi = 0$ is the singularity. Amari et al. analyzed the behavior of the maximum likelihood estimator in the case where the true parameters are on the singularity (Amari et al., 2001, 2002, 2003). In this case, the true distribution $p_0(\boldsymbol{x})$ is given by $\boldsymbol{x} \sim N(0, I_{d+2})$.

The simple multilayer perceptron with one hidden unit (Amari et al., 2001, 2002, 2003) has a similar cone structure. Through a transformation of parameters—$\beta = \|\boldsymbol{w}\|$, $\xi = v\beta$, and $\boldsymbol{\omega} = \frac{\boldsymbol{w}}{\beta} \in S^{d-1}$—we get

$$y = \xi \varphi_\beta(\boldsymbol{\omega} \cdot \boldsymbol{x}) + \varepsilon, \tag{2.33}$$

where we put $\varphi_\beta(u) = \frac{1}{\beta}\varphi(\beta u)$. This is a cone in the space of $(\xi, \boldsymbol{\omega})$. In previous articles, we have analyzed the behavior of learning when the true parameter is on the singularity ($\xi = 0$). In this article, we analyze the behavior of learning when the true parameter is not necessarily at the singularity and show that the singularity strongly affects the dynamics.

2.5 Other Models. There are many other statistical models with singular structures. Hierarchical models include such singular structures in many cases. The space of the ARMA time-series model and that of linear rational systems are good examples (Amari, 1987; Brockett, 1976), but little is known about the effects of singularity. The estimation of points of change,

which is called the Nile River problem, is also a well-known example of singularity.

Let us consider another model, the model of population coding with multiple stimuli in a neural field (Amari, 1977), on which neurons are lined up continuously along a one-dimensional positional axis z. When a stimulus from the outside is applied at a specific place corresponding to $z = \mu$, the neurons located around $z = \mu$ are excited. Excitation of the neural field—that is, the firing rate of a neuron at z—is written in this case as

$$r(z) = v\psi(z - \mu) + \varepsilon(z), \tag{2.34}$$

where v is the strength of the stimulus, μ is its center, and $\varepsilon(z)$ is a noise term dependent on z. Function ψ is unimodal and is called the tuning function. This model has been applied in population coding where the problem is to estimate μ from the neural response $r(z)$. Its statistical analysis has been given in terms of the Fisher information in many reports (for example, Wu, Nakahara, & Amari, 2001; Wu, Amari, & Nakahara, 2002).

When two stimuli are simultaneously given at locations μ_1 and μ_2 with intensities v and $1 - v$, the response of the field is written as

$$r(z; v, \mu_1, \mu_2) = v\psi(z - \mu_1) + (1 - v)\psi(z - \mu_2) + \varepsilon(z). \tag{2.35}$$

In this case, the same singular structure as that of the gaussian mixture appears in the parameter space $\boldsymbol{\theta} = (v, \mu_1, \mu_2)$. The strange behavior of the maximum likelihood estimator is analyzed in Amari and Burnashev (2003) and Amari and Nakahara (2005).

3 Locally Conic Structure and Gaussian Random Field

The local structure of singular statistical model is studied by Dacunha-Castelle and Gassiat (1997) in a unified manner. The local structure of a regular statistical model is represented by the tangent space of the manifold of the statistical model, where the first-order asymptotic theory is well formulated. The concepts of affine connections and related e- and m-curvature are necessary for the higher-order asymptotic theory, as is shown by information geometry (Amari & Nagaoka, 2000). A singular statistical model does not have the tangent space at singularity, and instead the tangent cone is useful for analyzing its local structure.

We summarize the results of Dacunha-Castelle and Gassiat (1997) in this section without mathematical rigor but intuitively. The locally conic structure and the related random gaussian field play a fundamental role in analyzing the behaviors of the likelihood ratio statistics (Hartigan, 1985; Fukumizu, 2003) and also of the MLE and its generalization ability (Amari

& Ozeki, 2001). Another general framework for singular models is given from algebraic geometry, which we do not summarize here. (See Watanabe, 2000a, 2000b, and 2000c, and related papers.)

3.1 Locally Conic Structure. All the examples in section 2 have a locally conic structure. For a singular statistical model $M = \{p(\boldsymbol{x}, \boldsymbol{\theta}), \boldsymbol{\theta} \in R^k\}$ in which the critical set C exists, Dacunha-Castelle and Gassiat (1997) introduced the following parameters in the neighborhood of C. Let us denote by $p_0(x)$ a probability density in C where the identifiability is lost. Given $p(\boldsymbol{x}, \boldsymbol{\theta})$, let ξ be the Hellinger distance between $p(x, \boldsymbol{\theta})$ and C,

$$\xi = \inf_{p_0(x) \in C} \int (\sqrt{p_0(x)} - \sqrt{p(x, \boldsymbol{\theta})})^2 dx. \tag{3.1}$$

It is further assumed that $p(x, \boldsymbol{\theta})$ can be parameterized in a neighborhood of $p_0(\boldsymbol{x}) \in C$ by $(\xi, \boldsymbol{\omega})$, where $\boldsymbol{\omega} \in \Omega$ is $(k-1)$-dimensional.

We thus have the new parameterization $p(x, \xi, \boldsymbol{\omega})$, where

$$\lim_{\xi \to 0} p(x, \xi, \boldsymbol{\omega}) = p(x, 0, \boldsymbol{\omega}) = p_0(x). \tag{3.2}$$

The critical set C is given by $\xi = 0$, where $p(x, 0, \boldsymbol{\omega})$ represents $p_0(\boldsymbol{x}) \in C$ so that $\boldsymbol{\omega}$ is not identifiable. The score function with respect to ξ is the directional derivative of log likelihood and is denoted by

$$v(x, \boldsymbol{\omega}) = \frac{d}{d\xi} \log p(x, 0, \boldsymbol{\omega}) \tag{3.3}$$

at $\xi = 0$. It depends on $\boldsymbol{\omega} \in \Omega$. Since ξ is the Hellinger distance, we have,

$$E[\{v(x, \boldsymbol{\omega})\}^2] = 1, \tag{3.4}$$

that is, the Fisher information in the ξ-direction is normalized to 1 at any $\boldsymbol{\omega}$.

Now consider

$$l(x, \xi, \boldsymbol{\omega}) = \log p(x, \xi, \boldsymbol{\omega}) \tag{3.5}$$

in the function space of random variable x. The model M is embedded in it, where the points in C are reduced to equivalent points, so that the dimension reduction takes place. Its image is the reduced set $\tilde{M}$. Let us fix a point $p_0(\boldsymbol{x})$ in C. In its neighborhood, the Taylor expansion gives

$$l(x, \xi, \boldsymbol{\omega}) = \log p_0(x) + \xi v(x, \boldsymbol{\omega}), \tag{3.6}$$

so that when ξ is very small, the image of M forms a cone in the space of functions of x, whose apex is $\log p_0(x)$ and whose edges are spanned by $v(x, \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega$. This is called the tangent cone and is different from the tangent space of a regular statistical model.

3.2 MLE and Gaussian Random Field. Given n examples $D = \{x_1, \cdots, x_n\}$ from a singular model M, the log likelihood is written as

$$L(D, \xi, \boldsymbol{\omega}) = \sum_{i=1}^{n} \log p(x_i, \xi, \boldsymbol{\omega}). \tag{3.7}$$

The MLE is the maximizer of L, but it is difficult to calculate because the derivatives of L with respect to $\boldsymbol{\omega}$ are 0 at $\xi = 0$ in some directions,

$$\frac{\partial L(D, 0, \boldsymbol{\omega})}{\partial \boldsymbol{\omega}} = \frac{\partial^2 L(D, 0, \boldsymbol{\omega})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}} = \cdots = 0. \tag{3.8}$$

We now fix $\boldsymbol{\omega}$ and calculate the maximizer $\hat{\xi}(D, \boldsymbol{\omega})$ of L. By expansion with respect to ξ, we have

$$L(D, \xi, \boldsymbol{\omega}) = L(D, 0, \boldsymbol{\omega}) + \frac{\partial L}{\partial \xi}\xi + \frac{1}{2}\frac{\partial^2 L}{\partial \xi^2}\xi^2 + \cdots. \tag{3.9}$$

Since $\hat{\xi}$ is small when the true distribution is $p_0(x)$, that is, $\xi = 0$, the maximizer is given from

$$-\frac{\partial^2 L}{\partial \xi^2}\hat{\xi} = \frac{\partial L}{\partial \xi}. \tag{3.10}$$

The term of the second derivative,

$$-\frac{1}{n}\frac{\partial^2 L}{\partial \xi^2} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log p(x_i, \xi, \boldsymbol{\omega})}{\partial \xi^2}, \tag{3.11}$$

converges to the Fisher information in the direction $\boldsymbol{\omega}$, because of the law of large numbers. The second term of the first derivative is

$$Y_n(\boldsymbol{\omega}) = \frac{1}{\sqrt{n}}\frac{\partial L}{\partial \xi} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\partial l(x_i, 0, \boldsymbol{\omega})}{\partial \xi} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} v(x_i, \boldsymbol{\omega}), \tag{3.12}$$

which converges to the gaussian random variable $Y(\boldsymbol{\omega})$ in law because of the central limit theorem. For $\boldsymbol{\omega} \neq \boldsymbol{\omega}'$, $Y(\boldsymbol{\omega})$ and $Y(\boldsymbol{\omega}')$ are correlated in general.

A family of gaussian distributions $\{Y(\boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$ forms a random gaussian field over Ω. The maximizer $\hat{\xi}$ is given by

$$\hat{\xi}(D, \boldsymbol{\omega}) = \frac{1}{\sqrt{n}} Y_n(\boldsymbol{\omega}). \tag{3.13}$$

3.3 MLE. Let us substitute $\hat{\xi}$ in equation 3.7, obtaining the partially maximized likelihood,

$$\begin{aligned} \hat{L}(D, \boldsymbol{\omega}) &= L(D, \hat{\xi}(D, \boldsymbol{\omega}), \boldsymbol{\omega}) \\ &= \sum \log p_0(x_i) + \frac{1}{2} Y_n(\boldsymbol{\omega})^2. \end{aligned}$$

Hence, the MLE is given by the maximizer of the random field,

$$\hat{\boldsymbol{\omega}} = \operatorname{argmax} Y_n(\boldsymbol{\omega})^2. \tag{3.14}$$

It is difficult to calculate this and study the properties of the MLE in general.

3.4 Likelihood Ratio Statistics. The log likelihood ratio statistics,

$$\lambda = 2 \sum \log \frac{p(x_i, \hat{\boldsymbol{\theta}})}{p(x_i, \boldsymbol{\theta}_0)}, \tag{3.15}$$

is used for testing the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against the alternative $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, where $\hat{\boldsymbol{\theta}}$ is the mle. The statistic λ is asymptotically subject to the χ^2 distribution with k degrees of freedom in a regular model, and hence

$$E[\lambda] = k \tag{3.16}$$

asymptotically. However, this does not hold in a singular model.

The log likelihood ratio statistics λ in a singular model is

$$\lambda = 2 \sup_{\xi, \boldsymbol{\omega}} \sum \log \frac{p(x_i, \xi, \boldsymbol{\omega})}{p_0(x_i)} \tag{3.17}$$

$$= 2 \sup_{\boldsymbol{\omega}} \{L(D, \hat{\xi}, \boldsymbol{\omega}) - L(D, 0, \boldsymbol{\omega})\} \tag{3.18}$$

$$= 2 \sup_{\boldsymbol{\omega}} Y_n(\boldsymbol{\omega})^2. \tag{3.19}$$

Hence, it is given by the supremum of the gaussian random field.

Hartigan (1985) suggested that $\lambda \sim \log \log n$ in the gaussian mixture model by extracting $m = \log n$ almost independent $Y(\boldsymbol{\omega}_1), \cdots, Y(\boldsymbol{\omega}_m)$.

Fukumizu (2003) followed the idea to evaluate λ in the case of multilayer perceptrons and derived $\lambda \sim \log n$.

4 Singularity Causes Strange Behaviors in Estimation and Learning

In the framework of singular statistical models, we give a glimpse of strange behaviors of estimation, testing, model selection, and online learning. We study three cases. In the first case, the true distribution, or the distribution that best approximates the true distribution, is exactly at the singularity. In this case, the parameters are not identifiable and the model is redundant (a smaller model suffices), but we can estimate its behavior (or the equivalent class) consistently. The gaussian random field plays a key role. In the second case, the true distribution is near the singularity. In the last case, the true distribution is at a regular point. In the last case, the classical theory can be applied locally. However, when studying the dynamics of learning, we need to take the influence of the singularity into account. The trajectories of learning cover the entire space, so it is a global problem in the entire space.

The log likelihood ratio test and MLE are known to be asymptotically optimal in the regular model. The likelihood principle is the belief that statistical inference should be based on likelihood. In singular models, however, this is not always true, and their optimality is not guaranteed (Amari & Burnashev, 2003). The behavior of Bayesian estimation and estimation with a regularization term also shows a different aspect from the regular case. There are many interesting problems to be studied, such as learning and its dynamics.

4.1 Statistical Testing in the Neighborhood of Critical Set. Statistical testing is a general method to judge from data whether the true distribution lies at the singularity. In the case of gaussian mixtures, we judge whether $k = 1$ or $k = 2$ through a statistical test. We take equation 2.12 as the null hypothesis and perform testing against the alternative that this equation is not true. In a general regular case, the log likelihood ratio statistic λ obeys the χ^2 distribution with the degrees of freedom equal to the number of parameters when the number of data is large enough. However, when the model is singular, the log likelihood ratio statistic may not be subject to the χ^2 distribution and may diverge to infinity in proportion to the number n of observations. This was shown in the classical works of Weyl (1939) and Hotelling (1939). Only recently was a precise asymptotic evaluation of the log likelihood ratio statistic given (Fukumizu, 2003; Hartigan, 1985; Liu & Shao, 2003) for some singular models. It is unfortunate that such tangled problems have usually been excluded as pathological cases and have not been well studied. Such cases are not pathological; they are ubiquitous in hierarchical engineering models.

Let us consider the statistical test $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. When the true point $\boldsymbol{\theta}_0$ is a regular point—that is, it is not in the critical set—the

MLE is asymptotically subject to a gaussian distribution with mean $\boldsymbol{\theta}_0$ and a variance-covariance matrix $G^{-1}(\boldsymbol{\theta}_0)/n$, where $G(\boldsymbol{\theta})$ is the Fisher information matrix. In such a case, the log likelihood ratio statistic is expanded in the Taylor series, giving

$$\lambda = n\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^T G^{-1}(\boldsymbol{\theta}_0)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right). \tag{4.1}$$

Hence, this is subject to the χ^2 distribution of the degrees of freedom equal to the number k of parameters. Its expectation is

$$E[\lambda] = k. \tag{4.2}$$

However, when the true distribution $\boldsymbol{\theta}_0$ lies on the critical set, the situation changes. The Fisher information matrix degenerates, and G^{-1} diverges, so the expansion is no longer valid. The expectation of the log likelihood estimator is asymptotically written as

$$E[\lambda] = c(n)k, \tag{4.3}$$

where the term $c(n)$ takes various forms depending on the nature of singularities. By evaluating the property of the gaussian random field $Y(\boldsymbol{\omega})$, Fukumizu (2003) showed that

$$c(n) = \log n \tag{4.4}$$

in the case of multilayer perceptrons under a certain condition. In the case of the gaussian mixture,

$$c(n) = \sqrt{\log \log n} \tag{4.5}$$

holds (Hartigan, 1985; Liu & Shao, 2003).

4.2 Estimation, Training Error, and Generalization Error. When the true parameter is at the singularity (or close to it), the MLE is no longer subject to the gaussian distribution, even asymptotically. This causes strange behaviors of training and generalization errors. The standard theory (Amari & Murata, 1993; Murata, Yoshizawa, & Amari, 1994) does not hold. This will be discussed in more detail in a later section.

4.3 Bayesian Estimator. The Bayesian estimator is used in many cases where an adequate prior distribution $\pi(\boldsymbol{\theta})$ is assumed. When the prior distribution penalizes complex models, it plays a role equivalent to the regularization term. When a set of independently and identically (i.i.d) data

$D = \{x_1, \cdots, x_n\}$ generated by $p(x; \boldsymbol{\theta}_0)$ is given, the posterior distribution of the parameters is written as

$$p(\boldsymbol{\theta}|D) = \frac{\pi(\boldsymbol{\theta}) p(D|\boldsymbol{\theta})}{p(D)}, \tag{4.6}$$

where

$$p(D) = \int p(D|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{4.7}$$

is the distribution of data D. The maximum a posterior (MAP) estimator is given by the parameter $\hat{\boldsymbol{\theta}}$ that maximizes the posterior distribution. Also, the Bayesian predictive distribution is used as the distribution of a new data x based on D. It is given by averaging the distribution $p(x|\boldsymbol{\theta})$ over the posterior distribution $p(\boldsymbol{\theta}|D)$,

$$p(x|D) = \int p(x, \boldsymbol{\theta}) p(\boldsymbol{\theta}|D) d\boldsymbol{\theta}. \tag{4.8}$$

It is empirically known that the Bayesian predictive distribution or MAP behaves well in the case of large-scale neural networks. In such a case, one uses a smooth prior $\pi(\boldsymbol{\theta}) > 0$ on the neuromanifold. Obviously, if $\pi(\boldsymbol{\theta}_0) = \infty$ at a specific point, the MAP estimator is attracted to that specific point. This is not fair. When $\pi(\boldsymbol{\theta}) > 0$ is smooth, its influence decreases as n approaches ∞, and it approaches the MLE, which is regarded as the MAP under the uniform prior.

However, a smooth prior on M is singular in the equivalence class $\tilde{M}$ of the neuromanifold, because a singular point in this class includes infinitely many equivalent parameters of M. Hence, the prior density is infinitely large on the singular points compared with that at regular points. This implies that the Bayesian smooth prior is in favor of singular points (perceptrons with a smaller number of hidden units) with an infinitely large factor. Hence, the Bayesian method works well in such a case to avoid overfitting. One may use a very large perceptron with a smooth Bayesian prior, and an adequate smaller model will be selected, although no theory exists that explains how to choose the prior.

The Bayesian estimator of singular models was studied by Watanabe (2001a, 2001b) and Yamazaki and Watanabe (2002, 2003) by using algebraic geometry, in particular, Hironaka's theory of singularity resolution and Sato's formula in the theory of algebraic analysis.

4.4 Model Selection. To obtain an adequate model, one should select a model from many alternatives based on the data. In the case of hierarchical models, one should determine the preferred model size, that is, the number

of hidden units. This is the problem of model selection. AIC, BIC, and MDL have been widely used as model selection criteria. NIC is a version of AIC applicable to the general cost function of a neural network (Murata et al., 1994).

AIC (Akaike, 1974) is a criterion to minimize the generalization error. The model that minimizes

$$\text{AIC} = 2 \times \text{training error} + \frac{2k}{n} \tag{4.9}$$

is selected according to this criterion. This is derived from asymptotic statistical analysis, where the MLE estimator $\hat{\boldsymbol{\theta}}$ is assumed to be asymptotically subject to the gaussian distribution with the covariance matrix equal to the inverse of the Fisher information matrix divided by n.

MDL (Rissanen, 1986) is a criterion to minimize the length of encoding for the observed data by using a family of parametric models. It is given asymptotically by the minimizer of

$$\text{MDL} = \text{training error} + \frac{\log n}{2n} k. \tag{4.10}$$

The Bayesian BIC (Schwarz, 1978) gives the same criterion as MDL. These criteria are derived also through the same assumption regarding the Gaussianity of the MLE.

In the case of multilayer perceptrons, the neuromanifold with a smaller number of hidden units is included in that with a larger number, but the smaller one forms a critical set within the larger neuromanifold. Therefore, the MLE (or any other efficient estimator) is no longer subject to the gaussian distribution, even asymptotically, provided the true distribution belongs to a smaller model. Model selection is required when the estimator is close to the critical set, but the validity of AIC and MDL fails to hold. Akaho and Kappen (2000) noted this in the gaussian mixture model. One should evaluate the log likelihood ratio statistic more carefully in such a case (Amari, 2003). The situation is the same in other hierarchical models with singularity.

There have been reported many comparisons of AIC and MDL by computer simulations. Sometimes AIC works better, while MDL does better in other cases. Such confusing reports seem to be the result of the difference between regular and singular models and also the difference in the nature of singularities.

4.5 Dynamics of Learning. Let us consider online learning of multilayer perceptrons through the gradient descent method. Let us define the error by the square of the difference between the network output and the teacher's signal. When the noise term is gaussian, the square of the error is

equal to the negative of the log likelihood. The minimization of the error is then equivalent to the maximization of the likelihood, and the result of learning locally converges to the maximum likelihood estimator. The stochastic gradient descent method was proposed by Amari (1967) and was named the backpropagation method (Rumelhart et al., 1986), while the natural gradient method (Amari, 1998; Amari et al., 2000; Park et al., 2000) takes the Riemannian structure of the space into account.

For an input-output example $(\boldsymbol{x}, y)$, the loss function or the negative log likelihood is given by

$$l(\boldsymbol{x}, y; \boldsymbol{\theta}) = \frac{1}{2}\left\{y - f(\boldsymbol{x}, \boldsymbol{\theta})\right\}^2. \tag{4.11}$$

Its expectation is given by averaging it with respect to the true distribution $p_0(\boldsymbol{x}, y)$,

$$L(\boldsymbol{\theta}) = E_{p_0}\left[l(\boldsymbol{x}, y; \boldsymbol{\theta})\right]. \tag{4.12}$$

The backpropagation and natural gradient learning algorithm (Amari, 1998) are written, respectively, as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla l\left(\boldsymbol{x}_t, y_t; \boldsymbol{\theta}_t\right) \tag{4.13}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta G^{-1}(\boldsymbol{\theta}_t)\nabla l(\boldsymbol{x}_t, y_t; \boldsymbol{\theta}_t), \tag{4.14}$$

when example $(\boldsymbol{x}_t, y_t)$ is given at time $t = 1, 2, \cdots$. Here, η is a learning constant, ∇ is the gradient $\partial/\partial\boldsymbol{\theta}$, and G is the Fisher information matrix. It is generally difficult to calculate $G(\boldsymbol{\theta})$, because the distribution $q(\boldsymbol{x})$ of inputs is unknown. Moreover, its inversion is costly. The adaptive natural gradient method estimates $G^{-1}(\boldsymbol{\theta}_t)$ adaptively from data (Amari et al., 2000; Park et al., 2000). It has been shown that the natural gradient method is locally equivalent to the Newton method, giving a Fisher efficient estimator, while the backpropagation is not Fisher efficient. The natural gradient method is capable of near-optimal performances (Rattray et al., 1998; Rattray & Saad, 1999). In the present formulation, the natural gradient is equivalent to the adaptive version of the Gauss-Newton method, but it is different and more powerful in other cost functions (Park et al., 2000).

The adaptive update of $G_t^{-1} = G^{-1}(\boldsymbol{\theta}_t)$ is calculated online by

$$G_{t+1}^{-1} = (1+\tau)G_t^{-1} - \tau G_t^{-1}\nabla l(\boldsymbol{x}_t, y_t, \hat{\boldsymbol{\theta}}_t)[G_t^{-1}\nabla l(\boldsymbol{x}_t, y_t, \hat{\boldsymbol{\theta}}_t)]^T. \tag{4.15}$$

The learning constant τ should not be large in order to guarantee the stability of the estimation of G^{-1}, but should not be too small to guarantee that the estimator $G_t = G(\hat{\boldsymbol{\theta}}_t)$ traces the change of $\hat{\boldsymbol{\theta}}_t$ well. Inoue, Park and

Okada (2003) show that the ratio η/τ should be kept within an adequate range.

Since examples are generated stochastically, the dynamics of learning, equations 4.13 and 4.14, are represented by the stochastic difference equations. However, when η is small, stochastic fluctuation is averaged out. Hence, we investigate the behavior of the averaged learning equation where ∇l is replaced by its expectation,

$$\nabla L(\boldsymbol{\theta}) = E[\nabla l(\boldsymbol{x}, y; \boldsymbol{\theta})]. \tag{4.16}$$

If continuous time is used, these become differential equations:

$$\begin{aligned} \frac{d\boldsymbol{\theta}}{dt} &= -\eta \nabla L(\boldsymbol{\theta}), \\ \frac{d\boldsymbol{\theta}}{dt} &= -\eta G^{-1}(\boldsymbol{\theta}) \nabla L(\boldsymbol{\theta}). \end{aligned} \tag{4.17}$$

The solution draws a trajectory in the neuromanifold. The problem is how the trajectory is influenced by the singular structure.

Kang et al. used a three-layer perceptron with binary weights and found that the parameters are attracted to the critical set that forms a singularity and are very slow to move away from it (Kang, Oh, Kwon, & Park, 1993). Saad and Solla (1995) and Riegler and Biehl (1995) analyzed the dynamics in a more general case and showed that such a phenomenon is universal. They argued that the slowness in backpropagation learning, or the plateau phenomenon, is caused by this singularity.

The natural gradient learning method takes the geometrical structure into account. It enables the influence of the singular structure to be reduced, and the trajectory is not trapped in the plateaus. Rattray et al. analyzed the dynamics of natural gradient learning by means of statistical physics and showed that it is almost ideal (Rattray & Saad, 1999; Rattray et al., 1998).

To examine the dynamics of learning in more detail, let us consider perceptrons consisting of two hidden units. The parameter space is $M = \{\boldsymbol{\theta}\}$, $\boldsymbol{\theta} = (\boldsymbol{w}_1, \boldsymbol{w}_2, v_1, v_2)$, and let us consider the subset $Q(\boldsymbol{w}, v)$ specified by $\boldsymbol{w}$ and v,

$$Q(\boldsymbol{w}, v) = \{\boldsymbol{w}_1 = \boldsymbol{w}_2 = \boldsymbol{w}, v_1 + v_2 = v\}, \tag{4.18}$$

which is included in the critical set C. The behavior of each perceptron in Q is the same and corresponds to that of a perceptron having only one hidden unit, where the weight vector is $\boldsymbol{w}$ and the output weight is v, and the behavior is $y = v\varphi(\boldsymbol{w} \cdot \boldsymbol{x}) + \epsilon$. Let the true parameters be $\boldsymbol{\theta}_0 = \{\boldsymbol{w}_1, \boldsymbol{w}_2, v_1, v_2\}$, where $\boldsymbol{w}_1 \neq \boldsymbol{w}_2$, so two different hidden units are used.

Let $\bar{\boldsymbol{\theta}} = (\bar{\boldsymbol{w}}, \bar{v})$ be the perceptron with only one hidden unit that best approximates the input-output function $f(\boldsymbol{x}, \boldsymbol{\theta}_0)$ of the true perceptron. Then all the perceptrons of two hidden units on the line $Q(\bar{\boldsymbol{w}}, \bar{v})$,

$$\boldsymbol{w}_1 = \boldsymbol{w}_2 = \bar{\boldsymbol{w}}, \quad v_1 + v_2 = \bar{v} \tag{4.19}$$

correspond to the best approximation. Let us transform the two weights as

$$\boldsymbol{w} = \frac{1}{2}(\boldsymbol{w}_1 + \boldsymbol{w}_2), \quad \boldsymbol{u} = \frac{1}{2}(\boldsymbol{w}_1 - \boldsymbol{w}_2). \tag{4.20}$$

The derivative of $L(\boldsymbol{\theta})$ along the line Q is then 0 because all the perceptrons are equivalent along the line. The derivatives in the direction of changing $\bar{\boldsymbol{w}}$ and $\bar{v}$ are also zero, because they are the best approximators. The derivative in the direction of $\boldsymbol{u}$ is again 0, because the perceptron having $\boldsymbol{u}$ is equivalent to that having $-\boldsymbol{u}$, which is derived by interchanging the two hidden units. Hence, the line Q forms critical points of the cost function. This implies that it is very difficult to get rid of it once the parameters are attracted to $Q(\bar{\boldsymbol{w}}, \bar{v})$.

Fukumizu and Amari (2000) calculated the Hessian of L on the line. When it is positive definite, the line is attractive. When it includes negative eigenvalues, the state eventually escapes in these directions. They showed that in some cases, part of the line can be truly attractive, although it is not a usual asymptotically stable equilibrium but has directions of escape (even though the derivative is 0) in other parts. This is not a usual saddle point and belongs to the special type called the Milner attractor. In such a case, the perceptron is truly attracted to the line and stays inside the line $Q(\bar{\boldsymbol{w}}, \bar{v})$, fluctuating around it because of random noise, until it finds a place from which it can escape. This explains the plateau phenomenon. The problem of plateau cannot be resolved by simply increasing η, because even when the state goes outside Q because of a large η, it may again return to it.

To show why the natural gradient method works well, we need to evaluate its behavior in the neighborhood of the critical points. We can then prove that the natural gradient has the effect of strong repulsion to escape from the neighborhood of the critical set, and the plateau phenomenon will disappear. Computer simulations confirm this observation.

Inoue et al. (2003) investigated the trajectory of learning by using the committee machine perceptron with two hidden units. They observed the following behavior of the trajectory. It approaches the singularity and stays near the singularity for a while before escaping from it. More precisely, they observed that $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ first came close to each other, both approaching $\bar{\boldsymbol{w}}$, which is the optimal in C, and then they moved away in different directions. Inoue et al. (2003) also studied the effectiveness of the adaptive natural gradient method, and showed the importance of controlling the two learning constants η and τ.

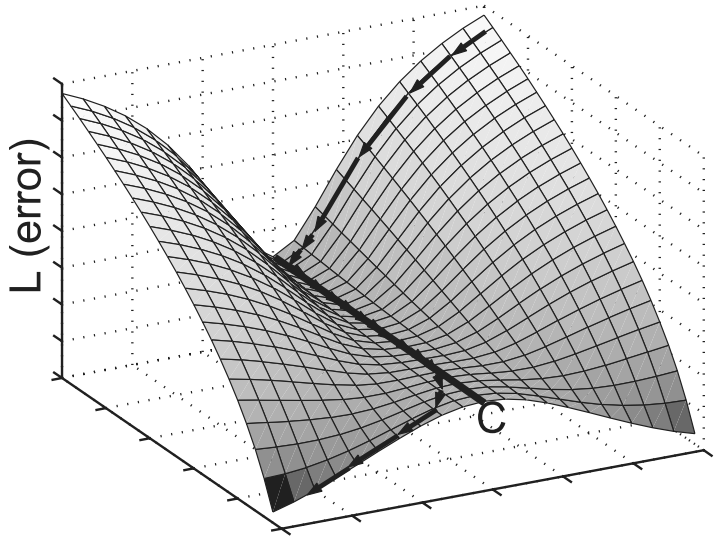

Figure 6: Learning trajectory near the singularities.

What is the trajectory of learning when the true parameters are on the singularity? Park, Inoue, and Okada (2003) investigated this problem by using three-layer perceptrons with two hidden units. Once the trajectory reaches C in Figure 6, all points are equivalent and suboptimal. Where does the trajectory enter C? To answer this question, we need to examine the dynamics near C. Because the component of the flow entering C is extremely slow near C, the flow component parallel to C is relatively strong. Analysis of the dynamics makes it clear that the trajectory does not stop at any point in the line where $\boldsymbol{w}_1 = \boldsymbol{w}_2 = \boldsymbol{w}$ and $v_1 + v_2 = v$ is constant, $v_1 \neq 0, v_2 \neq 0$, and that it approaches the point where either v_1 or v_2 is zero. This is an interesting observation.

What is the trajectory of learning in the natural gradient method? Since the metric degenerates in C, its inverse diverges to infinity. However, $\nabla L = 0$ even when the true distribution is outside C. If we consider $G^{-1}\nabla L$, it becomes a multiplication of 0 and ∞. However, if we evaluate $G^{-1}\nabla L$ near C, we can see that the infinitely strong repulsive force works in the direction of escape from C (Fukumizu & Amari, 2000). That is, the force going out from the plateau is strong, and the trajectory moves away without being attracted in the natural gradient.

5 Dynamics of Learning: Slow and Fast Manifolds

In this section, we show in detail the effect of singularities in the dynamics of learning for three simple models: the one-dimensional cone, the simple

MLP, and the gaussian mixture. Note that the structure of the gaussian mixture is very similar to that of the multilayer perceptron. We calculate the average trajectories of the standard and natural gradient learning methods to show how the trajectories approach the optimal point. We show that a slow manifold emerges around the critical set to which the state is quickly attracted by fast dynamics, and then the state escapes toward the optimal point slowly in the slow manifold. This is a universal feature of the plateau phenomenon.

5.1 Cone Model. Here, we investigate the dynamics of learning in the cone model introduced in section 2. The parameter space M is two-dimensional with coordinates (ξ, θ). The cost function is defined as the negative log likelihood

$$l(\boldsymbol{x}; \xi, \theta) = \frac{1}{2}\left\{(x_1 - \xi)^2 + (x_2 - c\xi\cos\theta)^2 + (x_3 - c\xi\sin\theta)^2\right\}. \tag{5.1}$$

For the average learning dynamics under the standard gradient learning method, we can easily obtain

$$\begin{pmatrix}\dot{\xi}(t)\\ \dot{\theta}(t)\end{pmatrix}_{\mathrm{SGD}} = -\eta_t \begin{pmatrix}\bar{x}_1 + c(\bar{x}_2\cos\theta + \bar{x}_3\sin\theta) - (1+c^2)\xi\\ -c\xi(\bar{x}_2\sin\theta - \bar{x}_3\cos\theta)\end{pmatrix}, \tag{5.2}$$

where $\bar{x}_i = E[x_i]$. Similarly, the average dynamics of natural gradient learning can also be obtained as

$$\begin{pmatrix}\dot{\xi}(t)\\ \dot{\theta}(t)\end{pmatrix}_{\mathrm{NGD}} = -\eta_t \begin{pmatrix}\frac{1}{1+c^2}\left(\bar{x}_1 + c(\bar{x}_2\cos\theta + \bar{x}_3\sin\theta) - (1+c^2)\xi\right)\\ -\frac{1}{c\xi}\left(\bar{x}_2\sin\theta - \bar{x}_3\cos\theta\right)\end{pmatrix}, \tag{5.3}$$

where we calculate the Fisher information matrix as

$$G(\xi, \theta) = \begin{bmatrix}1+c^2 & 0\\ 0 & c^2\xi^2\end{bmatrix}. \tag{5.4}$$

To consider the effect of singularity (i.e., $\xi = 0$) on the dynamics of learning, we define two submanifolds satisfying $\dot{\xi} = 0$ and $\dot{\theta} = 0$, respectively. These are

$$M_\xi = \{(\theta, \xi) : \bar{x}_1 + c(\bar{x}_2\cos\theta + \bar{x}_3\sin\theta) - (1+c^2)\xi = 0\} \tag{5.5}$$

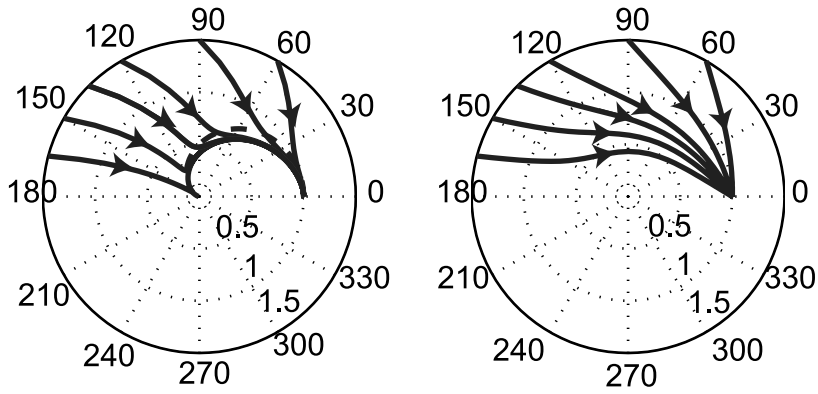

Figure 7: Learning trajectories in the cone model. (Left) Standard gradient (the dotted line is the slow manifold). (Right) Natural gradient.

and

$$M_\theta = \{(\theta, \xi) : \bar{x}_2 \sin\theta - \bar{x}_3 \cos\theta = 0\}. \tag{5.6}$$

The intersection of M_ξ and M_θ is the equilibrium of the dynamics. From the standard gradient learning equation, we see that $\dot{\xi}$ is of order $O(1)$, whereas $\dot{\theta}$ is of order $O(\xi)$. Therefore, in the neighborhood of the singularity where ξ is small, the speed of change in ξ is much faster than that of θ. Therefore, the state is attracted toward M_ξ (the dashed line in Figure 7 (left)) by the fast dynamics. Then the state moves along the line $\partial l/\partial \xi = 0$ or M_ξ, which is the slow manifold. The dynamics becomes especially slow when ξ is small (slow dynamics). This explains the plateau phenomenon in leaning curves. On the other hand, with the natural gradient learning equation 5.3, one can see that $\dot{\xi}$ is of order 1 and $\dot{\theta}$ is of order ξ^{-1}, so no slow manifolds appear. Moreover, the update term around the singularity is large, so that a strong repulsive force acts from the singularity. This explains why the plateau disappears in the natural gradient.

In computer simulations, we set $c = 1$. For the true parameters, we took $\xi^* = 1$ and $\theta^* = 0$, so we had $(\bar{x}_1, \bar{x}_2, \bar{x}_3) = (1, 1, 0)$. For the standard gradient and the natural gradient, we traced a number of trajectories with different initial values of ξ and θ. The trajectories in the parameter space are shown in Figure 7 using polar coordinates. Note that the center of the polar coordinates, $(0, 0)$, is the singular point corresponding to the apex of the cone.

In Figure 7 (left) for the standard gradient, we can see that the trajectories were attracted to the singular point or the slow manifold and then finally

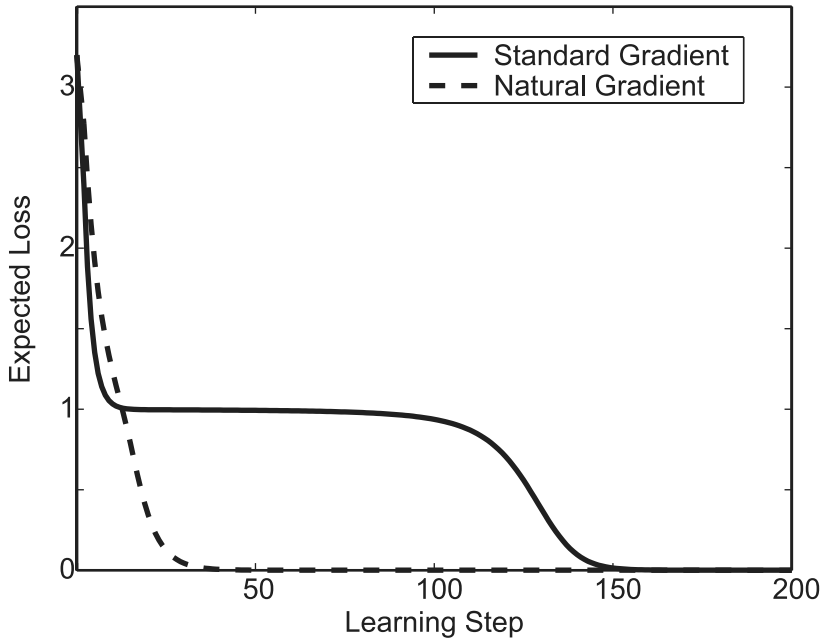

Figure 8: Time evolution of expected loss in the cone model. (Solid line: standard gradient, dashed line: natural gradient).

were attracted to the optimal point. In contrast, for the natural gradient, Figure 7 (right), such attraction and retardation did not appear.

Figure 8 shows the time evolution of the expected loss for the initial condition $(\xi, \theta) = (1.5, 5\pi/6)$. We can see a clear plateau in the standard gradient learning curve, whereas there was no plateau in the natural gradient learning curve.

5.2 Simple MLP. We also investigated the dynamics of learning of the simple MLP defined by equation 2.29, which is also discussed in section 2. The loss function, which is the squared error or the negative log likelihood, is given by

$$l(y, \boldsymbol{x}; \xi, \theta) = \frac{1}{2} \left\{ y - \xi \varphi(\cos\theta x_1 + \sin\theta x_2 + b) \right\}^2 . \tag{5.7}$$

The mathematical analysis is similar, and the fast and slow manifolds are obtained from $\dot{\theta} = 0$ and $\dot{\xi} = 0$, respectively.

For our computer simulations, we set $b = 0.5$. For the true parameters, we took $\xi^* = 1$ and $\theta^* = 0$. As for the cone model, we traced a number of

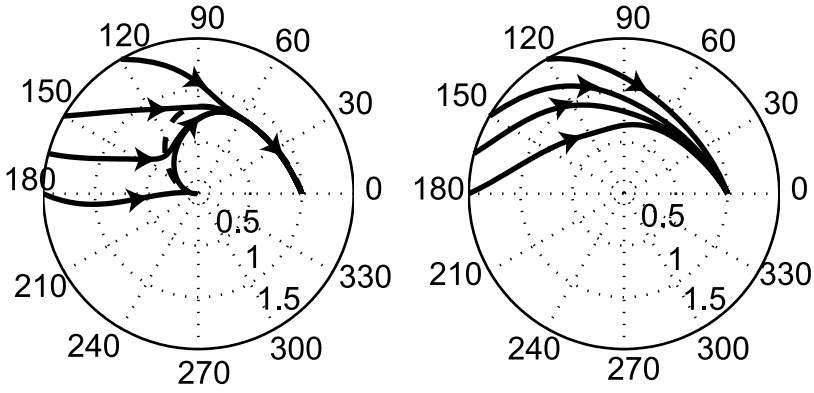

Figure 9: Learning trajectories in the simple MLP model. (Left) Standard gradient. (Right) Natural gradient.

trajectories with different initial values of ξ and θ. The trajectories in the parameter space are shown in Figure 9 using polar coordinates. Note that for this MLP model, the center of the polar coordinates is the singular point, which corresponds to the shrink point of Figure 4.

In Figure 9 (left) for the standard gradient, we can see phenomena similar to those for the cone model. That is, the trajectory was attracted to the plateau near the singular point and then slowly reached the optimal point. For the natural gradient, Figure 7 (right), we did not see that kind of attraction and retardation.

Figure 10 shows the time evolutions of the expected loss for the initial condition $(\xi, \theta) = (1.5, 5\pi/6)$. We can see a clear plateau in the curve of standard gradient learning, but none in that of natural gradient learning.

5.3 Gaussian Mixture. Next, we consider a more realistic model, the gaussian mixture. This is an original study in this article. To investigate the dynamics of learning of the gaussian mixture model, we begin with the Taylor expansion, equation 2.16, where $v(1-v) > c$ should be kept in mind. In this model, the singularity exists at $u = 0$, so the Taylor expansion for small values of u is useful. The cost function, the negative of the log likelihood, is further expanded as

$$l(x; u, v) = -\left\{\log \psi(x) + \frac{1}{2}c_2(v)H_2(x)u^2 + \frac{1}{6}c_3(v)H_3(x)u^3 + \frac{1}{24}c_4(v)H_4(x)u^4 - \frac{1}{8}c_2^2(x)H_2^2(x)u^4 + O(u^5)\right\}. \tag{5.8}$$

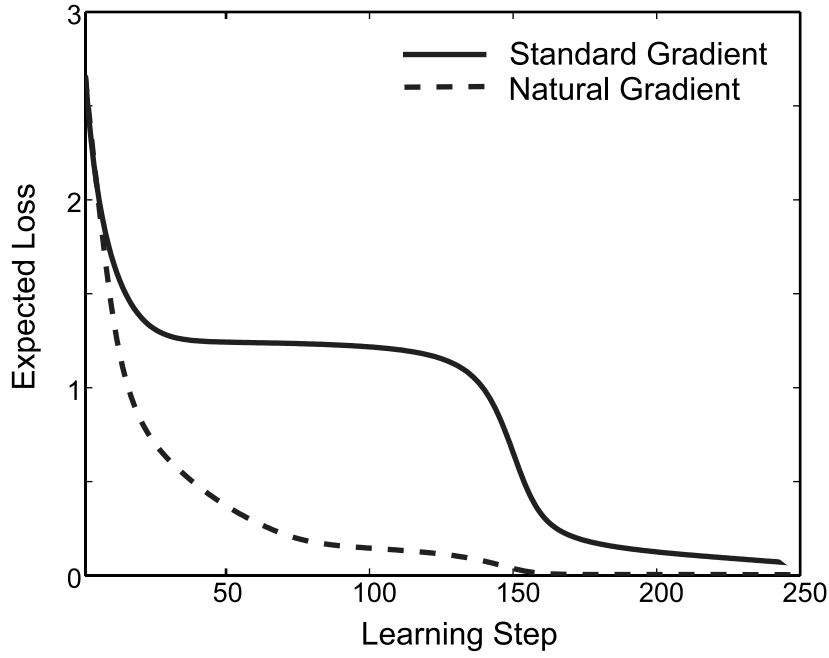

Figure 10: Time evolutions of expected loss in simple MLP. (Solid line: standard gradient, dashed line: natural gradient).

Let (u^*, v^*) be the true parameter from which the learning data are generated. We can calculate the average learning dynamics around small u.

Lemma. *When u is small, the gradient of l evaluated at the true parameter (u^*, v^*), is given by*

$$E_*[\partial_u l] = -c_2(v)\{u^{*2}c_2(v^*) - u^2c_2(v)\}u - \frac{1}{2}c_3(v)\{u^{*3}c_3(v^*) - u^3c_3(v))\}u^2 + O(u^3) \tag{5.9}$$

$$E_*[\partial_v l] = -\frac{1}{2}c_2'(v)\{u^{*2}c_2(v^*) - u^2c_2(v)\}u^2 - \frac{1}{6}c_3'(v)\{u^{*3}c_3(v^*) - u^3c_3(v))\}u^3 + O(u^4), \tag{5.10}$$

where E_* denotes the expectation with respect to $p(x, u^*, v^*)$ and $c_i'(v) = dc_i(v)/dv$. The proof is given in the appendix.

The averaged equation for the standard gradient is given by

$$\begin{pmatrix} \dot{u}(t) \\ \dot{v}(t) \end{pmatrix}_{\mathrm{SGD}} = -\eta \begin{pmatrix} E_*[\partial_u l] \\ E_*[\partial_v l] \end{pmatrix}. \tag{5.11}$$

We put

$$f_1(u, v) = u^{*2} c_2(v^*) - u^2 c_2(v),$$
$$f_2(u, v) = u^{*3} c_3(v^*) - u^3 c_3(v).$$

The averaged learning equations are then given by

$$\begin{pmatrix} \dot{u}(t) \\ \dot{v}(t) \end{pmatrix}_{\mathrm{SGD}} = -\eta \begin{pmatrix} c_2(v) f_1(u, v) u + \frac{1}{2} c_3(v) f_2(u, v) u^2 \\ \frac{1}{2} c_2'(v) f_1(u, v) u^2 + \frac{1}{6} c_3'(v) f_2(u, v) u^3 \end{pmatrix} + \begin{pmatrix} O(u^3) \\ O(u^4) \end{pmatrix}. \tag{5.12}$$

We now consider the trajectories of learning in two cases: $u^* = 0$ (singularity) and $u^* \neq 0$ (regular).

Case I: $u^* = 0$. By putting $u^* = 0$ in equation 5.12 and ignoring higher-order terms, we have

$$\dot{u} = -\eta c_2(v)^2 u^3,$$
$$\dot{v} = -\eta \frac{c_2(v) c_2'(v)}{2} u^4. \tag{5.13}$$

From this, the trajectory of dynamics is given by

$$\frac{dv}{du} = \frac{\dot{v}}{\dot{u}} = \frac{1 - 2v}{2v(1 - v)} u. \tag{5.14}$$

The equation can be integrated to give

$$u^2 = \frac{1}{4}(1 - 2v)^2 - \frac{1}{2} \log(1 - 2v) + c, \tag{5.15}$$

where c is constant. When u is small, $\dot{v}$ is of $O(u^4)$ and is much smaller than $\dot{u}$, which is of order u^3. Hence, $dv/du \approx 0$, and the trajectories are almost parallel to the u-axis, as shown in Figure 11. In other words, u converges to 0 without significantly changing v.

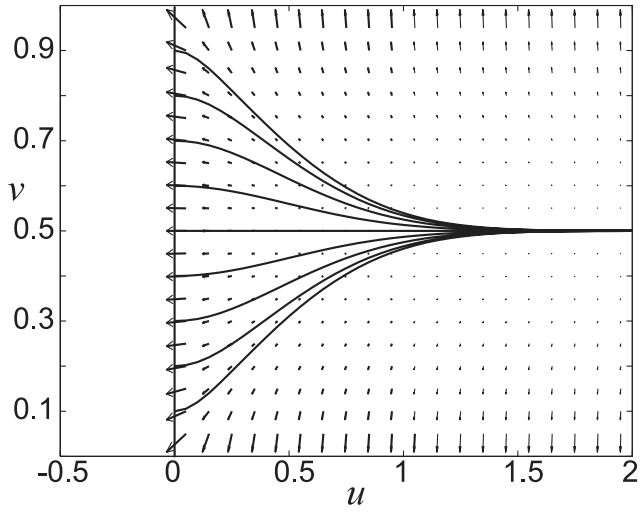

Figure 11: Trajectories of learning in the gaussian mixture model.

Case II: $u^* \neq 0$. We also evaluate the dynamics when u is small. The equation is

$$\dot{u} = \eta c_2(v) c_2(v^*) u^{*2} u \tag{5.16}$$

$$\dot{v} = \frac{\eta}{2} c_2'(v) c_2(v^*) u^{*2} u^2 \tag{5.17}$$

and

$$\frac{dv}{du} = \frac{u}{2} \frac{c_2'(v)}{c_2(v)} \tag{5.18}$$

irrespective of (u^*, v^*). Incidentally, the equation is the same as that of equation 5.14; hence, the trajectories are the same (see Figure 11), but the directions are opposite, and the state is escaping from $u = 0$ toward u^* in this case; that is, the directions in Figure 11 are reversed.

The equilibrium is given by the intersection of the two manifolds,

$$M_F : f_2(u, v) = 0,$$
$$M_S : f_1(u, v) = 0.$$

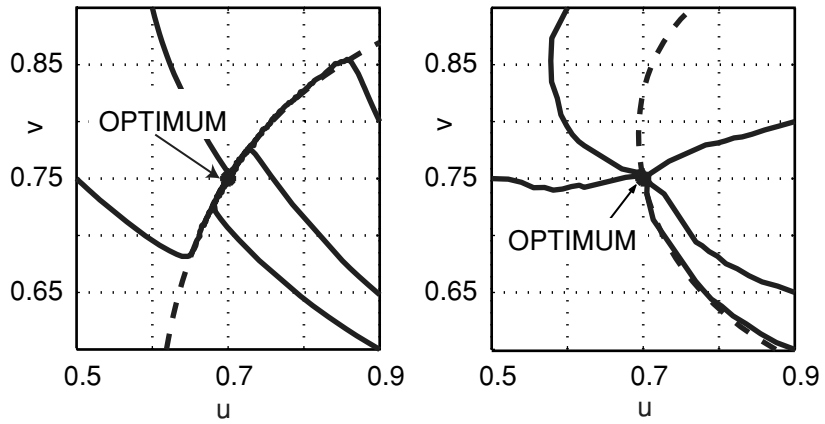

Figure 12: Learning trajectories in the simple gaussian mixture model. (Left) Standard gradient. (Right) Natural gradient.

When u is small, the first term of $f_1(u, v)$ dominates, and the state is quickly attracted to M_S. Then it moves in M_S slowly to the intersection of M_S and M_F. Computer simulations confirm this observation (see Figure 12 left).

We next studied the natural gradient method. Using an approximation, equation 5.8, we can also obtain an explicit form of the Fisher information matrix. This is given by

$$G(u, v) = \begin{bmatrix} 2c_2^2(v)u^2 + \frac{3}{2}c_2^2(v)u^4 & c_3(v)u^3 + \frac{1}{2}c_3(v)(2c_2(v)+1)u^5 \\ c_3(v)u^3 + \frac{1}{2}c_3(v)(2c_2(v)+1)u^5 & \frac{1}{2}(c_2'(v))^2u^4 + (\frac{1}{6} - c_2(v))u^6 \end{bmatrix}. \tag{5.19}$$

For the natural gradient method, the dynamics of learning is

$$\begin{pmatrix} \dot{u}(t) \\ \dot{v}(t) \end{pmatrix}_{\mathrm{NGD}} = \frac{\eta}{u^3} \begin{pmatrix} \tilde{c_1} f_2(u, v)u + \tilde{c_2} f_1(u, v)u^2 \\ \tilde{c_3} f_2(u, v) + \tilde{c_4} f_1(u, v)u \end{pmatrix}, \tag{5.20}$$

where $\tilde{c_i}$ are functions of v. The equilibrium is again the intersection of M_S and M_F, but the roles of M_S and M_F are reversed. Repulsion is strong when u is small, and no plateau appears.

For our computer simulations, we set the true parameters as $u^* = 0.75$ and $v^* = 0.7$. Since the analytic expression of the average dynamics given in equations 5.12 and 5.20 are an approximation around a small u, we cannot apply this for the whole trajectory. Therefore, we use the Monte Carlo

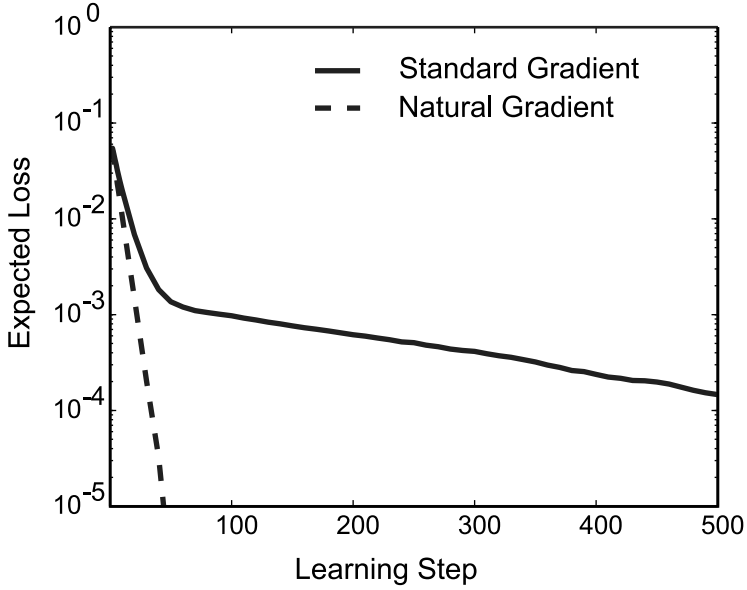

Figure 13: Time evolutions of the expected loss in the simple gaussian mixture model. Solid line: standard gradient, dashed line: natural gradient.

method to get the expectation of $\partial l/\partial u$ and $\partial l/\partial v$. At each learning step, we generate 10^6 samples according to the true input distribution and take the sample means. We traced a number of trajectories with different initial values for u and v. The trajectories in the parameter space are shown in Figure 12.

In Figure 12 (left), we can see the line (slow manifold) on which the parameters first converge. Therefore, the state proceeds by learning toward the line satisfying

$$u^{*2}c_2(v^*) = u^2c_2(v),$$

which is shown as a dashed line in Figure 12 (left). However, for the natural gradient, Figure 12 (right), the update terms of $\dot{u}$ and $\dot{v}$ have terms of order $O(u^{-2})$ and $O(u^{-3})$, respectively, which lead to much faster dynamics around the singularity.

Figure 13 shows the time evolutions of the expected loss for the initial condition $(u, v) = (0.9, 0.6)$. We can see the slow convergence in the standard gradient learning curve, whereas this sort of retardation is not apparent in natural gradient learning.

6 Generalization and Training Errors When the True Distribution Is at a Singular Point

It is important for model selection to evaluate the generalization error relative to the training error. Since AIC and MDL have been derived from the asymptotic gaussianity of the estimator, they cannot be applied to the singular case. In particular, for hierarchical models such as multilayer perceptrons and gaussian mixtures, smaller models are embedded in the larger models as critical sets. Therefore, we need to find new model selection criteria for the singular case. In the regular case, the MLE and Bayes predictive estimators give asymptotically the same estimation performance. However, these are not guaranteed in the singular case. As a preliminary study, we analyzed the asymptotic behavior of the MLE and Bayes predictive distribution by using simple toy models when the true distribution lay at a singular point, that is, in a smaller model.

The behavior of an estimator is evaluated by the relation between the expected generalization error and the expected training error. Let $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be observed data, or $D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ in the case of perceptrons. When we have an estimated probability density function $\hat{p}(\boldsymbol{x}; D)$, the generalization error can be defined by the Kullback-Leibler divergence from the true probability density p_o to the estimated density function,

$$KL[p_o(\boldsymbol{x}) : \hat{p}(\boldsymbol{x}; D)] = E_{p_o}\left[\log \frac{p_o(\boldsymbol{x})}{\hat{p}(\boldsymbol{x}; D)}\right]. \tag{6.1}$$

In the case of perceptrons, the estimated density function is the conditional probability density $\hat{p}(y|\boldsymbol{x}; D)$, and a similar formulation follows. For the evaluation, we take an expectation of the generalization error with respect to the data, and we call it the expected generalization error,

$$E_{gen} = E_D E_{p_o}\left[\log \frac{p_o(\boldsymbol{x})}{\hat{p}(\boldsymbol{x}; D)}\right], \tag{6.2}$$

where E_D denotes expectation with respect to the observed data D. Similarly, the training error of the estimated density function $\hat{p}(\boldsymbol{x}; D)$ is defined by the sample average,

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{p_o(\boldsymbol{x}_i)}{\hat{p}(\boldsymbol{x}_i; D)}, \tag{6.3}$$

which is the expectation of $\log(p_0/\hat{p})$ with respect to the empirical distribution of data D,

$$p_{emp}(\boldsymbol{x}) = \frac{1}{n}\sum \delta(\boldsymbol{x} - \boldsymbol{x}_i). \tag{6.4}$$

The expected training error is defined as

$$E_{train} = E_D\left[\frac{1}{n}\sum_{i=1}^{n}\log\frac{p_o(\boldsymbol{x}_i)}{\hat{p}(\boldsymbol{x}_i; D)}\right]. \tag{6.5}$$

For the MLE, the estimated density function is given by $p(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}$ is the MLE. One can also see the relation between the expected training error and the likelihood ratio statistics λ in equation 3.17,

$$E_{train} = -\frac{1}{2n}E_D[\lambda].$$

For the Bayes estimation, the estimation density function is given by the Bayes predictive distribution of the form

$$\hat{p}_{\text{Bayes}}(\boldsymbol{x}|D) = \int p(\boldsymbol{x}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}|D) d\boldsymbol{\theta}. \tag{6.6}$$

6.1 Maximum Likelihood Estimator

6.1.1 Cone Model. For the cone model defined in equation 2.30, the log likelihood of data $D = \{\boldsymbol{x}_i\}_{i=1,\ldots,n}$ is written as

$$L(D, \xi, \boldsymbol{\omega}) = -\frac{1}{2}\sum_{i=1}^{n}||\boldsymbol{x}_i - \xi\boldsymbol{a}(\boldsymbol{\omega})||^2. \tag{6.7}$$

The MLE is the one that maximizes $L(D, \xi, \boldsymbol{\omega})$. However, $\partial^k L/\partial\boldsymbol{\omega}^k = 0$ at $\xi = 0$ for any k, so we cannot analyze the behaviors of the MLE by Taylor expansion at $\xi = 0$. Therefore, we first fix $\boldsymbol{\omega}$ and search for the ξ that maximizes L. The maximum $\hat{\xi}$ is given by

$$\hat{\xi}(\boldsymbol{\omega}) = \text{argmax}_{\xi} L(D, \xi, \boldsymbol{\omega}) = \frac{1}{\sqrt{n}} Y_n(\boldsymbol{\omega}), \tag{6.8}$$

where

$$Y_n(\boldsymbol{\omega}) = \boldsymbol{a}(\boldsymbol{\omega}) \cdot \tilde{\boldsymbol{x}}, \quad \tilde{\boldsymbol{x}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{x}_i. \tag{6.9}$$

Its limit $Y(\boldsymbol{\omega})$ is a zero-mean gaussian random field with covariance $\boldsymbol{a}(\boldsymbol{\omega}) \cdot \boldsymbol{a}(\boldsymbol{\omega}')$, because $\tilde{\boldsymbol{x}} \sim N(0, I)$ when the true distribution is at $\xi = 0$. The MLE $\hat{\boldsymbol{\omega}}$ is given by the maximizer of

$$\hat{\boldsymbol{\omega}} = \text{argmax}_{\boldsymbol{\omega}} Y_n^2(\boldsymbol{\omega}). \tag{6.10}$$

Using the MLE, we obtain the expected generalization and training errors in the following theorem.

Theorem 1. *For the cone model, when the true distribution is at $\xi = 0$, the MLE satisfies*

$$E_{gen} = E_D E_{p_o}\left[log \frac{p_o(\boldsymbol{x})}{p(\boldsymbol{x}|\hat{\xi}, \hat{\boldsymbol{\omega}})}\right] = \frac{1}{2n} E_D\left[max_{\boldsymbol{\omega}} Y^2(\boldsymbol{\omega})\right], \tag{6.11}$$

$$E_{train} = E_D\left[\frac{1}{n}\sum_{i=1}^{n} log \frac{p_o(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i|\hat{\xi}, \hat{\boldsymbol{\omega}})}\right] = -\frac{1}{2n} E_D\left[max_{\boldsymbol{\omega}} Y^2(\boldsymbol{\omega})\right]. \tag{6.12}$$

A more detailed derivation is given in the appendix. In addition, we can obtain the explicit value of $E_D\left[\max_{\boldsymbol{\omega}} Y^2(\boldsymbol{\omega})\right]$ within the limit of large d.

Corollary 1. *When d is large, the MLE satisfies*

$$E_{gen} \approx \frac{1 + 2c\sqrt{\frac{2}{\pi}}\sqrt{d} + c^2(d+1)}{2n(1+c^2)} \approx \frac{c^2 d}{2n(1+c^2)}, \tag{6.13}$$

$$E_{train} \approx -\frac{1 + 2c\sqrt{\frac{2}{\pi}}\sqrt{d} + c^2(d+1)}{2n(1+c^2)} \approx -\frac{c^2 d}{2n(1+c^2)}. \tag{6.14}$$

The proof is also given in the appendix. Among the results is an interesting one concerning the antisymmetry between E_{gen} and E_{train}; that is, $E_{gen} = -E_{train}$, which is proved in the regular case (Amari & Murata, 1993). Note also that the generalization and training errors depend on the shape parameter c as well as the dimension number d. In the regular case, they depend only on d. As one can easily see, when c is small, the cone looks like a needle, and its behavior resembles a one-dimensional model. When c is large, the cone resembles two $(d+1)$-dimensional hypersurfaces, so its behavior is like a $(d+1)$–dimensional regular model. Such observations are confirmed by equations 6.13 and 6.14.

6.1.2 Simple MLP with One Hidden Unit. For the simple MLP defined in equation 2.33, we can also apply the same approach. The log likelihood of data set $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1,\ldots,n}$ is written as

$$L(D; \xi, \boldsymbol{\omega}) = -\frac{1}{2}\sum_{i=1}^{n}\left\{y_i - \xi\varphi_\beta\left(\boldsymbol{\omega}\cdot\boldsymbol{x}_i\right)\right\}^2. \tag{6.15}$$

Let us define two random variables depending on D and $\boldsymbol{\omega}$,

$$Y_n(\boldsymbol{\omega}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} y_i\varphi_\beta\left(\boldsymbol{\omega}\cdot\boldsymbol{x}_i\right), \tag{6.16}$$

$$A_n(\boldsymbol{\omega}) = \frac{1}{n}\sum_{i=1}^{n}\varphi_\beta^2(\boldsymbol{\omega}\cdot\boldsymbol{x}_i). \tag{6.17}$$

Note that $A_n(\boldsymbol{\omega})$ converges to $A(\boldsymbol{\omega}) = E_{\boldsymbol{x}}[\varphi_\beta^2(\boldsymbol{\omega}\cdot\boldsymbol{x})]$ as n goes to infinity, but is not normalized to 1 in the present case. $Y(\boldsymbol{\omega})$ defines a gaussian random field on Ω, with mean 0 and covariance $A(\boldsymbol{\omega}, \boldsymbol{\omega}') = E_{\boldsymbol{x}}[\varphi_\beta(\boldsymbol{\omega}\cdot\boldsymbol{x})\varphi_\beta(\boldsymbol{\omega}'\cdot\boldsymbol{x})]$. One should be careful that $A_n(\boldsymbol{\omega}, \beta)$ of equation 6.17 approaches 0 as $\beta \to 0$. This belongs to the non-Donsker class, and our theory does not hold in such a case. Using the MLE, we get the following theorem.

Theorem 2. *For the simple MLP model, when the teacher perceptron is $\xi = 0$, the MLE satisfies*

$$E_{gen} = E_D E_{p_o,q}\left[log\frac{p_o(y|\boldsymbol{x})}{p(y|\boldsymbol{x}, \hat{\xi}, \hat{\boldsymbol{\omega}})}\right] = \frac{1}{2n}E_D\left[sup_{\boldsymbol{\omega}}\frac{Y^2(\boldsymbol{\omega})}{A_n(\boldsymbol{\omega})}\right], \tag{6.18}$$

$$E_{train} = E_D\left[\frac{1}{n}\sum_{i=1}^{n} log\frac{p_o(y|\boldsymbol{x}_i)}{p(y|\boldsymbol{x}_i, \hat{\xi}, \hat{\boldsymbol{\omega}})}\right] = -\frac{1}{2n}E_D\left[sup_{\boldsymbol{\omega}}\frac{Y^2(\boldsymbol{\omega})}{A_n(\boldsymbol{\omega})}\right]. \tag{6.19}$$

The details of this derivation are given in the appendix. From the results, we can see a nice correspondence between the cone model and MLP. However, note that there is no sufficient statistic in the MLP case, while all the data are summarized in the sufficient statistic $\tilde{\boldsymbol{x}}$ in the cone model. In addition, due to the nonlinearity of the hidden unit, we cannot easily determine the explicit relation through which the training and generalization errors depend on the dimension number of the parameters, which we have for the cone model in corollary 1.

6.2 Bayes Predictive Distribution

6.2.1 Cone Model. Different from the regular case, the asymptotic behavior of the Bayesian predictive distribution depends on the prior. Let us define the prior as $\pi(\xi, \boldsymbol{\omega})$. The probability density of the observed sample is then given by

$$Z_n = p(D) = \int \pi(\xi, \boldsymbol{\omega}) \prod_{i=1}^{n} p(\boldsymbol{x}_n|\xi, \boldsymbol{\omega}) d\xi d\boldsymbol{\omega}. \tag{6.20}$$

When new data $\boldsymbol{x}_{n+1}$ are given, we can similarly obtain the joint probability density $p(\boldsymbol{x}_{n+1}, D)$ as

$$Z_{n+1} = p(\boldsymbol{x}_{n+1}, D) = \int \pi(\xi, \boldsymbol{\omega}) \prod_{i=1}^{n+1} p(\boldsymbol{x}_i|\xi, \boldsymbol{\omega}) d\xi d\boldsymbol{\omega}. \tag{6.21}$$

From the Bayes theorem, we can easily see that the Bayes predictive distribution is given by

$$\hat{p}_{\text{Bayes}}(\boldsymbol{x}|D) = \frac{Z_{n+1}}{Z_n}, \tag{6.22}$$

where $\boldsymbol{x} = \boldsymbol{x}_{n+1}$.

When we assume a specific prior for the parameter ξ and $\boldsymbol{\omega}$, we can calculate Z_n explicitly. When $\pi(\xi) = 1$ and $\boldsymbol{\omega}$ is uniform on Ω, we can obtain the Bayes predictive distribution and the generalization error explicitly, as in the following theorems.

Theorem 3. *Under the uniform prior on ξ, the Bayes predictive distribution of the cone model is given by*

$$\hat{p}_{BAYES}(\boldsymbol{x}|D) = \frac{1}{(\sqrt{2\pi})^{d+2}} exp\left\{-\frac{1}{2}\|\boldsymbol{x}\|^2\right\} \times \left\{1 + \frac{1}{\sqrt{n}} \nabla log S_d^U(\tilde{\boldsymbol{x}}) \cdot \boldsymbol{x} + \frac{1}{2n} tr\left(\frac{\nabla\nabla S_d^U}{S_d^U} H_2(\boldsymbol{x})\right)\right\}, \tag{6.23}$$

where $H_2(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{x}^T - I$ and

$$S_d^U(\tilde{\boldsymbol{x}}) = \int exp\left\{\frac{1}{2} Y_n(\boldsymbol{\omega})^2\right\} d\boldsymbol{\omega}, \quad Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} a(\boldsymbol{\omega}) \cdot \boldsymbol{x}_i = \boldsymbol{a}(\boldsymbol{\omega}) \cdot \tilde{\boldsymbol{x}}.$$

Theorem 4. *Under the uniform prior on ξ, the generalization and training errors of the Bayes predictive distribution of the cone model are given by*

$$E_{gen} = E_D E_{p_o}\left[log\frac{p_o(\boldsymbol{x})}{p(\boldsymbol{x}|D)}\right] = \frac{1}{2n}E_D\left[\left\|\nabla log\, S_d^U(\tilde{\boldsymbol{x}})\right\|^2\right] = \frac{1}{2n}, \tag{6.24}$$

$$E_{train} = \frac{1}{n}\sum_{i=1}^{n} E_D\left[log\frac{p_o(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i|D)}\right] = E_{gen} - \frac{1}{n}E_D\left[\nabla log\, S_d^U(\tilde{\boldsymbol{x}})\cdot\tilde{\boldsymbol{x}}\right]. \tag{6.25}$$

The details of this derivation are given in the appendix.

Remark. For any prior $\pi(\xi,\boldsymbol{\omega})$ that is positive and smooth, theorems 3 and 4 also hold asymptotically without any change.

The Jeffreys' prior is given by the square root of the determinant of the Fisher information matrix,

$$\pi(\xi,\boldsymbol{\omega}) \propto \sqrt{|G(\xi,\boldsymbol{\omega})|}. \tag{6.26}$$

In this case, $\pi(\xi)\propto|\xi|^d$ and $\boldsymbol{\omega}$ are uniformly distributed on S^d. The Jeffreys prior is not smooth and is 0 at $\xi=0$. This is completely different from the regular case. We can conduct a similar analysis and obtain the following theorems.

Theorem 5. *Under the Jeffreys' prior, the Bayes predictive distribution of the cone model is given asymptotically by*

$$\hat{p}_{BAYES}(\boldsymbol{x}|D) = \frac{1}{(\sqrt{2\pi})^{d+2}}exp\left\{-\frac{1}{2}\|\boldsymbol{x}\|^2\right\} \times\left\{1+\frac{1}{\sqrt{n}}\nabla log\, S_d^J(\tilde{\boldsymbol{x}})\cdot\boldsymbol{x}+\frac{1}{2n}tr\left(\frac{\nabla\nabla S_d^J}{S_d^J}H_2(\boldsymbol{x})\right)\right\}, \tag{6.27}$$

where

$$S_d^J(\tilde{\boldsymbol{x}}) = \int I_d\left(Y_n(\boldsymbol{\omega})\right)exp\left\{\frac{1}{2}Y_n(\boldsymbol{\omega})^2\right\}d\boldsymbol{\omega}, \tag{6.28}$$

$$I_d(u) = \frac{1}{\sqrt{2\pi}}\int|z+u|^d exp\left\{-\frac{1}{2}z^2\right\}dz. \tag{6.29}$$

Theorem 6. *Under the Jeffreys' prior, the generalization and training errors of the Bayes predictive distribution of the cone model are given asymptotically by*

$$E_{gen} = \frac{1}{2n}E_D\left[\left\|\nabla log\, S_d^J(\tilde{\boldsymbol{x}})\right\|^2\right] = \frac{d+1}{2n} \tag{6.30}$$

$$E_{train} = E_{gen} - \frac{1}{n} E_D \left[\nabla log S_d^l(\tilde{\boldsymbol{x}}) \cdot \tilde{\boldsymbol{x}} \right]. \tag{6.31}$$

For the proof, the same derivation process as that of the uniform case can be applied, although the process is fairly complicated.

These results are rather surprising. Under the uniform prior, the generalization error is constant and does not depend on d, which is the complexity of the model. Hence, no overfitting occurs whatever complex models we use. This is completely different from the regular case. However, this striking result arises from the uniform prior on ξ. The uniform prior puts a strong emphasis on the singularity because there are infinitely many equivalent points at $\xi = 0$, so the prior density is infinitely large if we consider the space $\tilde{M}$ of the probability distributions or behaviors. Hence, one should be very careful in choosing a prior when the model includes singularities. In the case of Jeffreys' prior, the generalization error increases in proportion to d, which is similar to the regular case. In addition, the antisymmetric duality between E_{gen} and E_{train} does not hold for both the uniform prior and Jeffreys' prior.

6.2.2 Simple MLP with One Hidden Unit. For the simple MLP model, we conducted a similar analysis for the uniform prior and Jeffreys' prior, and obtained the following theorems.

Theorem 7. *Under the uniform prior on ξ, the Bayes predictive distribution of the simple MLP model is given by*

$$\hat{p}_{BAYES}(y|\boldsymbol{x}, D) = \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{12}{y^2}\right\} \left\{1 + \frac{y}{\sqrt{n}} \frac{\int \nabla Q_d^U(Y_n, \boldsymbol{\omega}) \varphi_\beta(\boldsymbol{\omega} \cdot \boldsymbol{x}) d\boldsymbol{\omega}}{P_d^U(Y_n)} + \frac{1}{2n} H_2(y) \frac{\int \nabla\nabla Q_d^U(Y_n, \boldsymbol{\omega}) A_n(\boldsymbol{\omega}) d\boldsymbol{\omega}}{P_d^U(Y_n)} + O\left(\frac{1}{n^2}\right)\right\}, \tag{6.32}$$

where

$$Q_d^U(Y_n, \boldsymbol{\omega}) = \frac{1}{\sqrt{A(\boldsymbol{\omega})}} exp\left\{\frac{1}{2} \frac{Y_n(\boldsymbol{\omega})^2}{A(\boldsymbol{\omega})}\right\}, \tag{6.33}$$

$$P_d^U(Y_n) = \int Q_d^U(Y_n, \boldsymbol{\omega}) d\boldsymbol{\omega}, \tag{6.34}$$

$$Y_n(\boldsymbol{\omega}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} y_i \varphi_\beta(\boldsymbol{\omega} \cdot \boldsymbol{x}_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} y_i \varphi_i, \tag{6.35}$$

$$A(\boldsymbol{\omega}) = E_x \left[\varphi_\beta^2(\boldsymbol{\omega} \cdot \boldsymbol{x})\right]. \tag{6.36}$$

Theorem 8. *Under the uniform prior on ξ, the generalization and training errors of the Bayes predictive distribution of the simple MLP model are given by*

$$\begin{aligned} E_{gen} &= E_D E_{p_o q}\left[log \frac{p_o(y|\boldsymbol{x})}{p(y|\boldsymbol{x}, D)} \right] \\ &= \frac{1}{2n} E_D \left[\frac{\int \nabla Q_d^U(Y_n, \boldsymbol{\omega}) \nabla Q_d^U(Y_n, \boldsymbol{\omega}') A(\boldsymbol{\omega}\boldsymbol{\omega}') d\boldsymbol{\omega}\boldsymbol{\omega}'}{\left(P_d^U(Y_n)\right)^2} \right] \\ &= \frac{1}{2n} \end{aligned} \tag{6.37}$$

$$\begin{aligned} E_{train} &= \frac{1}{n} \sum_{i=1}^{n} E_D \left[log \frac{p_o(y_i|\boldsymbol{x}_i)}{p(y_i|\boldsymbol{x}_i, D)} \right] \\ &= E_{gen} - \frac{1}{n} E_D \left[\frac{\int \nabla Q_d^U(Y_n, \boldsymbol{\omega}) Y_n(\boldsymbol{\omega}) d\boldsymbol{\omega}}{P_d^U(Y_n)} \right]. \end{aligned} \tag{6.38}$$

Theorem 9. *Under Jeffreys' prior on ξ, the Bayes predictive distribution of the simple MLP model is given by*

$$\begin{aligned} \hat{p}_{BAYES}(y|\boldsymbol{x}, D) = \frac{1}{\sqrt{2\pi}} exp\left\{ -\frac{1}{2} y^2 \right\} \left\{ 1 + \frac{y}{\sqrt{n}} \frac{\int \nabla Q_d^J(Y_n, \boldsymbol{\omega}) \varphi_\beta(\boldsymbol{\omega} \cdot \boldsymbol{x}) d\boldsymbol{\omega}}{P_d^J(Y_n)} \right. \\ \left. + \frac{1}{2n} H_2(y) \frac{\int \nabla \nabla Q_d^J(Y_n, \boldsymbol{\omega}) A_n(\boldsymbol{\omega}) d\boldsymbol{\omega}}{P_d^J(Y_n)} + O\left(\frac{1}{n^2} \right) \right\}, \end{aligned} \tag{6.39}$$

where

$$Q_d^J(Y_n, \boldsymbol{\omega}) = \frac{1}{\sqrt{A(\boldsymbol{\omega})}^{d+1}} I_d\left(\frac{Y_n(\boldsymbol{\omega})}{\sqrt{A(\boldsymbol{\omega})}} \right) exp\left\{ \frac{1}{2} \frac{Y_n(\boldsymbol{\omega})^2}{A(\boldsymbol{\omega})} \right\} \tag{6.40}$$

$$P_d^J(D) = \int Q_d^J(Y_n, \boldsymbol{\omega}) d\boldsymbol{\omega}. \tag{6.41}$$

$$I_d(u) = \frac{1}{\sqrt{2\pi}} \int |z + u|^d exp\left\{ -\frac{1}{2} z^2 \right\} dz. \tag{6.42}$$

Theorem 10. *Under Jeffreys' prior on ξ, the generalization and training errors of the Bayes predictive distribution of the simple MLP model are given by*

$$\begin{aligned} E_{gen} &= \frac{1}{2n} E_D \left[\frac{\int \nabla Q_d^J(Y_n, \boldsymbol{\omega}) \nabla Q_d^J(Y_n, \boldsymbol{\omega}') A(\boldsymbol{\omega}\boldsymbol{\omega}') d\boldsymbol{\omega}\boldsymbol{\omega}'}{\left(P_d^J(D)\right)^2} \right] \\ &= \frac{d+1}{2n} \end{aligned} \tag{6.43}$$

$$E_{train} = E_{gen} - \frac{1}{n} E_D \left[\frac{\int \nabla Q_d^J(Y_n, \boldsymbol{\omega}) Y_n(\boldsymbol{\omega}) d\boldsymbol{\omega}}{P_d^J(D)} \right]. \tag{6.44}$$

All of these results for the simple MLP correspond well with those for the cone model. For both the cone and MLP models, we can see that the generalization error is strongly dependent on the prior distribution of the parameters. This differs from the classic theory for the regular models.

7 Conclusion

It has long been known that some kinds of statistical model violate ordinary conditions such as the existence of a regular Fisher information matrix. Unfortunately, in classical statistical theories, singular models of this type have been regarded as pathological and have received little attention. However, to understand the behavior of hierarchical models such as multilayer perceptrons, singularity problems cannot be ignored. The singularity is closely related to basic problems regarding these models—such as the slow dynamics of learning in plateaus and a strange relation between generalization and training errors—and also the criteria of model selection.

It is premature to give a general theory of estimation, testing, Bayesian inference, and learning dynamics for singular models. In this article, we have summarized our recent results regarding these problems using simple toy models. Although the results are preliminary, we believe that we have succeeded in elucidating various aspects of the singularity and have found some interesting paths to follow in future studies.

Appendix: Proofs of Theorems

Proof of Theorem 1. The log likelihood of D is given by

$$L(D, \xi, \boldsymbol{\omega}) = -\frac{1}{2} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \xi \boldsymbol{a}(\boldsymbol{\omega})\|^2. \tag{A.1}$$

From the definition of the generalization and training errors, we obtain

$$\begin{aligned} E_{gen} &= E_D E_{p_o} \left[-\hat{\xi}(\hat{\boldsymbol{\omega}}) \boldsymbol{a}(\hat{\boldsymbol{\omega}}) \cdot \boldsymbol{x} + \frac{1}{2} \hat{\xi}^2(\hat{\boldsymbol{\omega}}) \left(\boldsymbol{a}(\hat{\boldsymbol{\omega}}) \cdot \boldsymbol{a}(\hat{\boldsymbol{\omega}}) \right) \right] \\ &= E_D \left[\frac{1}{2} \hat{\xi}^2(\hat{\boldsymbol{\omega}}) \right] \\ &= \frac{1}{2n} E_D \left[\max_{\boldsymbol{\omega}} Y_n^2(\boldsymbol{\omega}) \right]. \end{aligned} \tag{A.2}$$

$$
\begin{aligned}
E_{train} &= E_D\left[\frac{1}{n}\sum_{i=1}^{n}\left\{-\hat{\xi}(\hat{\boldsymbol{\omega}})\boldsymbol{a}(\hat{\boldsymbol{\omega}})\cdot\boldsymbol{x}_i+\frac{1}{2}\hat{\xi}^2(\hat{\boldsymbol{\omega}})\left(\boldsymbol{a}(\hat{\boldsymbol{\omega}})\cdot\boldsymbol{a}(\hat{\boldsymbol{\omega}})\right)\right\}\right]\\
&= E_D\left[-\hat{\xi}(\hat{\boldsymbol{\omega}})\left(\frac{1}{\sqrt{n}}\boldsymbol{a}(\hat{\boldsymbol{\omega}})\cdot\tilde{\boldsymbol{x}}\right)+\frac{1}{2}\hat{\xi}^2(\hat{\boldsymbol{\omega}})\right]\\
&= E_D\left[-\frac{1}{2}\hat{\xi}^2(\hat{\boldsymbol{\omega}})\right]\\
&= -\frac{1}{2n}E_D\left[\max_{\boldsymbol{\omega}}Y_n^2(\boldsymbol{\omega})\right]. \qquad \text{(A.3)}
\end{aligned}
$$

Proof of Corollary 7. In order to get the final results, we need to calculate $\max_{\boldsymbol{\omega}}Y_n^2(\boldsymbol{\omega})$. Let

$$
\boldsymbol{a}(\boldsymbol{\omega})\cdot\tilde{\boldsymbol{x}}=\frac{1}{\sqrt{1+c^2}}(\tilde{x}_1+c\boldsymbol{\omega}\cdot\tilde{\boldsymbol{x}}')
$$

where $\tilde{\boldsymbol{x}}'=(\tilde{x}_2,\ldots,\tilde{x}_{d+2})^T$. Then

$$
\begin{aligned}
Y_n^2(\boldsymbol{\omega})=\|\boldsymbol{a}(\boldsymbol{\omega})\cdot\tilde{\boldsymbol{x}}\|^2&=\frac{1}{1+c^2}\left\{\tilde{x}_1^2+2c\tilde{x}_1\boldsymbol{\omega}\cdot\tilde{\boldsymbol{x}}'+c^2(\boldsymbol{\omega}\cdot\tilde{\boldsymbol{x}}')^2\right\}\\
&=\frac{1}{1+c^2}\left\{\tilde{x}_1^2+2c\tilde{x}_1A\boldsymbol{e}\cdot\boldsymbol{\omega}+c^2A^2(\boldsymbol{e}\cdot\boldsymbol{\omega})^2\right\}, \qquad \text{(A.4)}
\end{aligned}
$$

where $\tilde{\boldsymbol{x}}'=\|\tilde{\boldsymbol{x}}'\|\boldsymbol{e}=A\boldsymbol{e}$, $\|\boldsymbol{e}\|=1$. Then we obtain

$$
\hat{\boldsymbol{\omega}}=\operatorname{argmax}_{\boldsymbol{\omega}}Y_n^2(\boldsymbol{\omega})=\operatorname{sgn}(\tilde{x}_1)\boldsymbol{e},
$$

and

$$
\max_{\boldsymbol{\omega}}Y_n^2(\boldsymbol{\omega})=\frac{1}{1+c^2}\left\{\tilde{x}_1^2+2c|\tilde{x}_1|A+c^2A^2\right\}.
$$

From the fact that

$$
\begin{aligned}
&E_D[A^2]=d+1,\\
&E_D[A]=E\left[\sqrt{\tilde{x}_2^2+\cdots+\tilde{x}_{d+2}^2}\right]=\frac{d!!}{(d-1)!!}\sqrt{\frac{2}{\pi}}^{(-1)^d}\approx\sqrt{d},
\end{aligned}
$$

where $d!!=d(d-2)(d-4)\cdots$, we finally get

$$E_D\left[\max_{\boldsymbol{\omega}} Y_n^2(\boldsymbol{\omega})\right] = \frac{1}{1+c^2}\left\{1 + 2cE_D[|\tilde{x}_1|]E_D[A] + c^2(d+1)\right\}$$
$$\approx \frac{1 + 2c\sqrt{\frac{2}{\pi}}\sqrt{d} + c^2(d+1)}{(1+c^2)} \approx \frac{c^2 d}{1+c^2}. \tag{A.5}$$

Proof of Theorem 2. The log likelihood of D is given by

$$L(D, \xi, \boldsymbol{\omega}) = -\frac{1}{2}\sum_{i=1}^{n}\{y_i - \xi\varphi_\beta(\boldsymbol{\omega}\cdot\boldsymbol{x}_i)\}^2$$
$$= -\frac{1}{2}\sum_{i=1}^{n} y_i^2 + \xi\sum_{i=1}^{n} y_i\varphi_\beta(\boldsymbol{\omega}\cdot\boldsymbol{x}_i) + -\frac{1}{2}\xi^2\sum_{i=1}^{n}\varphi_\beta^2(\boldsymbol{\omega}\cdot\boldsymbol{x}_i). \tag{A.6}$$

By using

$$L(\hat{\xi}(\boldsymbol{\omega}), \boldsymbol{\omega}) = -\frac{1}{2}\sum_{i=1}^{n} y_i^2 + \frac{1}{2}\frac{Y_n^2(\boldsymbol{\omega})}{A_n(\boldsymbol{\omega})}, \tag{A.7}$$

$$\hat{\boldsymbol{\omega}} = \operatorname{argmax}_{\boldsymbol{\omega}} \frac{Y_n^2(\boldsymbol{\omega})}{A_n(\boldsymbol{\omega})}, \tag{A.8}$$

$$E_{gen} = E_D E_{y,\boldsymbol{x}}\left[-\hat{\xi}(\hat{\boldsymbol{\omega}})y\varphi_\beta(\hat{\boldsymbol{\omega}}\cdot\boldsymbol{x}) + \frac{1}{2}\hat{\xi}^2(\hat{\boldsymbol{\omega}})\varphi_\beta^2(\hat{\boldsymbol{\omega}}\cdot\boldsymbol{x})\right]$$
$$= E_D E_{\boldsymbol{x}}\left[\frac{1}{2}\hat{\xi}^2(\hat{\boldsymbol{\omega}})A_n(\hat{\boldsymbol{\omega}})\right] = \frac{1}{2n}E_D\left[\sup_{\boldsymbol{\omega}}\frac{Y^2(\boldsymbol{\omega})}{A_n(\boldsymbol{\omega})}\right], \tag{A.9}$$

$$E_{train} = E_D\left[\frac{1}{n}\sum_{i=1}^{n}\left\{-\hat{\xi}(\hat{\boldsymbol{\omega}})y_i\varphi_\beta(\boldsymbol{\omega}\cdot\boldsymbol{x}_i) + \frac{1}{2}\hat{\xi}^2(\hat{\boldsymbol{\omega}})\varphi_\beta^2(\hat{\boldsymbol{\omega}}\cdot\boldsymbol{x}_i)\right\}\right]$$
$$= E_D\left[-\hat{\xi}(\hat{\boldsymbol{\omega}})\left(\frac{1}{\sqrt{n}}Y(\hat{\boldsymbol{\omega}})\right) + \frac{1}{2}\hat{\xi}^2(\hat{\boldsymbol{\omega}})A_n(\hat{\boldsymbol{\omega}})\right]$$
$$= E_D\left[-\frac{1}{2}\hat{\xi}^2(\hat{\boldsymbol{\omega}})A_n(\hat{\boldsymbol{\omega}})\right] = -\frac{1}{2n}E_D\left[\sup_{\boldsymbol{\omega}}\frac{Y^2(\boldsymbol{\omega})}{A_n(\boldsymbol{\omega})}\right]. \tag{A.10}$$

Proof of Theorem 3. Let us define

$$Z_n = p(D) = \int \pi(\xi, \boldsymbol{\omega})\prod_{i=1}^{n} p(\boldsymbol{x}_i|\xi, \boldsymbol{\omega})d\xi d\boldsymbol{\omega}, \tag{A.11}$$

$$Z_{n+1} = p(\boldsymbol{x}, D) = \int \pi(\xi, \boldsymbol{\omega}) \prod_{i=1}^{n+1} p(\boldsymbol{x}_i|\xi, \boldsymbol{\omega}) d\xi d\boldsymbol{\omega}. \tag{A.12}$$

Then the Bayesian predictive distribution can be written by

$$\hat{p}_{\mathrm{BAYES}}(\boldsymbol{x}|D) = \frac{Z_{n+1}}{Z_n}. \tag{A.13}$$

Under the uniform prior, we can easily get

$$\begin{aligned} Z_n &= \frac{1}{(\sqrt{2\pi})^{n(d+2)}} \int \exp\left\{-\frac{1}{2}\sum \|\boldsymbol{x}_i\|^2 + \xi\, \boldsymbol{a}(\boldsymbol{\omega}) \cdot \sum \boldsymbol{x}_i - \frac{n}{2}\xi^2\right\} d\xi d\boldsymbol{\omega} \\ &= \frac{\sqrt{2\pi}}{(\sqrt{2\pi})^{n(d+2)}\sqrt{n}} \exp\left\{-\frac{1}{2}\sum \|\boldsymbol{x}_i\|^2\right\} \int \exp\left\{\frac{1}{2} Y_n(\boldsymbol{\omega})^2\right\} d\boldsymbol{\omega}. \end{aligned} \tag{A.14}$$

From equations A.14 and A.13, we obtain

$$\hat{p}_{\mathrm{BAYES}}(\boldsymbol{x}|D) = \frac{1}{(\sqrt{2\pi})^{d+2}} \sqrt{\frac{n}{n+1}} \exp\left\{-\frac{\|\boldsymbol{x}\|^2}{2}\right\} \frac{S_d^U(\tilde{\boldsymbol{x}}_{n+1})}{S_d^U(\tilde{\boldsymbol{x}})}, \tag{A.15}$$

where

$$\tilde{\boldsymbol{x}}_{n+1} = \frac{1}{\sqrt{n+1}} \left(\boldsymbol{x} + \sum \boldsymbol{x}_i\right), \; S_d^U(\tilde{\boldsymbol{x}}) = \int \exp\left\{\frac{1}{2} Y_n(\boldsymbol{\omega})^2\right\} d\boldsymbol{\omega}.$$

Using the approximation of the form,

$$\tilde{\boldsymbol{x}}_{n+1} \approx \tilde{\boldsymbol{x}} + \left(\frac{\boldsymbol{x}}{\sqrt{n}} - \frac{1}{2n}\tilde{\boldsymbol{x}}\right) = \tilde{\boldsymbol{x}} + \delta\boldsymbol{x},$$

and from equation A.15, we obtain

$$\begin{aligned}\hat{p}_{\mathrm{BAYES}}(\boldsymbol{x}|D) &= \frac{1}{(\sqrt{2\pi})^{d+2}} \exp\left\{-\frac{1}{2}\|\boldsymbol{x}\|^2\right\}\left\{1-\frac{1}{2n}\right\} \\ &\quad \times \left\{1+\frac{\nabla S_d^U(\tilde{\boldsymbol{x}})}{S_d^U(\tilde{\boldsymbol{x}})}\cdot\delta\boldsymbol{x}+\frac{1}{2}\mathrm{tr}\left(\frac{\nabla\nabla S_d^U(\tilde{\boldsymbol{x}})}{S_d^U(\tilde{\boldsymbol{x}})}\delta\boldsymbol{x}\delta\boldsymbol{x}^T\right)\right\} \\ &= \frac{1}{(\sqrt{2\pi})^{d+2}} \exp\left\{-\frac{1}{2}\|\boldsymbol{x}\|^2\right\}\left\{1+\frac{1}{\sqrt{n}}\frac{\nabla S_d^U(\tilde{\boldsymbol{x}})}{S_d^U(\tilde{\boldsymbol{x}})}\cdot\boldsymbol{x}\right. \\ &\quad \left.+\frac{1}{2n}\left(\mathrm{tr}\left(\frac{\nabla\nabla S_d^U}{S_d^U}\boldsymbol{x}\boldsymbol{x}^T\right)-\frac{\nabla S_d^U}{S_d^U}\cdot\tilde{\boldsymbol{x}}-1\right)\right\}. \end{aligned} \tag{A.16}$$

Using the fact that $\nabla S_d^U(\tilde{\boldsymbol{x}})\cdot\tilde{\boldsymbol{x}}+S_d^U(\tilde{\boldsymbol{x}})=\mathrm{tr}\{\nabla\nabla S_d^U(\tilde{\boldsymbol{x}})\}$, we can obtain the final result.

Proof of Theorem 4. From equation A.16,

$$\begin{aligned}E_{gen} &= E_D E_{p_o}\left[-\log\left\{1+\frac{1}{\sqrt{n}}\nabla\log S_d^U(\tilde{\boldsymbol{x}})\cdot\boldsymbol{x}+\frac{1}{2n}\mathrm{tr}\left(\frac{\nabla\nabla S_d^U}{S_d^U}H_2(\boldsymbol{x})\right)\right\}\right] \\ &= E_D E_{p_o}\left[-\frac{1}{\sqrt{n}}\nabla\log S_d^U(\tilde{\boldsymbol{x}})\cdot\boldsymbol{x}-\frac{1}{2n}\mathrm{tr}\left(\frac{\nabla\nabla S_d^U}{S_d^U}H_2(\boldsymbol{x})\right)\right. \\ &\quad \left.+\frac{1}{2n}\left(\nabla\log S_d^U(\tilde{\boldsymbol{x}})\cdot\boldsymbol{x}\right)^2\right] \\ &= \frac{1}{2n}E_D\left[\left\|\nabla\log S_d^U(\tilde{\boldsymbol{x}})\right\|^2\right] \end{aligned} \tag{A.17}$$

Similarly, for the training error, we get

$$\begin{aligned}E_{train} &= \frac{1}{n}\sum_{i=1}^n E_D\left[-\log\left\{1+\frac{1}{\sqrt{n}}\nabla\log S_d^U(\tilde{\boldsymbol{x}})\cdot\boldsymbol{x}_i+\frac{1}{2n}\left(\frac{\nabla\nabla S_d^U}{S_d^U}H_2(\boldsymbol{x}_i)\right)\right\}\right] \\ &= \frac{1}{n}\sum_{i=1}^n E_D\left[-\frac{1}{\sqrt{n}}\nabla\log S_d^U(\tilde{\boldsymbol{x}})\cdot\boldsymbol{x}_i-\frac{1}{2n}\mathrm{tr}\left(\frac{\nabla\nabla S_d^U}{S_d^U}H_2(\boldsymbol{x}_i)\right)\right. \\ &\quad \left.+\frac{1}{2n}\left(\nabla\log S_d^U(\tilde{\boldsymbol{x}})\cdot\boldsymbol{x}_i\right)^2\right] \\ &= -\frac{1}{n}E_D\left[\nabla\log S_d^U(\tilde{\boldsymbol{x}})\cdot\tilde{\boldsymbol{x}}\right] \\ &\quad +\frac{1}{n}\sum_{i=1}^n E_D\left[-\frac{1}{2n}\mathrm{tr}\left(\frac{\nabla\nabla S_d^U}{S_d^U}H_2(\boldsymbol{x}_i)\right)+\frac{1}{2n}\left(\nabla\log S_d^U(\tilde{\boldsymbol{x}})\cdot\boldsymbol{x}_i\right)^2\right]. \end{aligned}$$

Here, we use the expansion

$$\tilde{\boldsymbol{x}} \approx \tilde{\boldsymbol{x}}_i + \frac{1}{\sqrt{n}} \boldsymbol{x}_i, \tag{A.18}$$

$$f(\tilde{\boldsymbol{x}}) \approx f(\tilde{\boldsymbol{x}}_i) - \frac{1}{\sqrt{n}} \nabla f(\tilde{\boldsymbol{x}}_i) \cdot \boldsymbol{x}_i, \tag{A.19}$$

where $\tilde{\boldsymbol{x}}_i = \frac{1}{\sqrt{n}} \sum_{j \neq i} \boldsymbol{x}_j$, and we finally get

$$\begin{aligned} E_{train} &= -\frac{1}{n} E_D \left[\nabla \log S_d^U(\tilde{\boldsymbol{x}}) \cdot \tilde{\boldsymbol{x}} \right] + \frac{1}{n} \sum_{i=1}^{n} E_D \left[\frac{1}{2n} \left(\nabla \log S_d^U(\tilde{\boldsymbol{x}}_i) \cdot \boldsymbol{x} \right)^2 \right] \\ &= E_{gen} - \frac{1}{n} E_D \left[\nabla \log S_d^U(\tilde{\boldsymbol{x}}) \cdot \tilde{\boldsymbol{x}} \right]. \end{aligned} \tag{A.20}$$

On the other hand, since $Y_n(\boldsymbol{\omega})$ and $Y_{n+1}(\boldsymbol{\omega})$ have the same distributions, we easily get

$$E_{gen} = H_0 + \frac{1}{n},$$

where H_0 is the entropy of the distribution $p_0(\boldsymbol{x})$.

Proof of Theorem 5. Let us define

$$Z_n^J = p(D) = \int |\xi|^d \prod_{i=1}^{n} p(\boldsymbol{x}_i | \xi, \boldsymbol{\omega}) d\xi d\boldsymbol{\omega}, \tag{A.21}$$

$$Z_{n+1}^J = p(\boldsymbol{x}, D) = \int |\xi|^d \prod_{i=1}^{n+1} p(\boldsymbol{x}_i | \xi, \boldsymbol{\omega}) d\xi d\boldsymbol{\omega}. \tag{A.22}$$

Then the Bayesian predictive distribution can be written as

$$\hat{p}_{\text{BAYES}}(\boldsymbol{x} | D) = \frac{Z_{n+1}^J}{Z_n^J}. \tag{A.23}$$

Under Jeffreys' prior, we get

$$\begin{aligned} Z_n^J &= \frac{1}{(\sqrt{2\pi})^{n(d+2)}} \int |\xi|^d \exp\left\{ -\frac{1}{2} \sum \|\boldsymbol{x}_i\|^2 + \xi \boldsymbol{a}(\boldsymbol{\omega}) \cdot \sum \boldsymbol{x}_i - \frac{n}{2} \xi^2 \right\} d\xi d\boldsymbol{\omega} \\ &= \frac{1}{(\sqrt{2\pi})^{n(d+2)}} \exp\left\{ -\frac{1}{2} \sum \|\boldsymbol{x}_i\|^2 \right\} \\ &\quad \times \int |\xi|^d \exp\left\{ -\frac{n}{2} \xi^2 + \sqrt{n} (\boldsymbol{a}(\boldsymbol{\omega}) \cdot \tilde{\boldsymbol{x}}) \xi \right\} d\xi d\boldsymbol{\omega} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{(\sqrt{2\pi})^{n(d+2)}} \exp\left\{-\frac{1}{2}\sum \|\boldsymbol{x}_i\|^2\right\} \\
&\quad \times \int |\xi|^d \exp\left\{-\frac{n}{2}\left(\xi - \frac{\boldsymbol{a}(\boldsymbol{\omega})\cdot\tilde{\boldsymbol{x}}}{\sqrt{n}}\right)^2\right\} \exp\left\{\frac{(\boldsymbol{a}(\boldsymbol{\omega})\cdot\tilde{\boldsymbol{x}})^2}{2}\right\} d\xi d\boldsymbol{\omega} \\
&= \frac{1}{(\sqrt{2\pi})^{n(d+2)}\sqrt{n}^{d+1}} \exp\left\{-\frac{1}{2}\sum \|\boldsymbol{x}_i\|^2\right\} \\
&\quad \times \int \left|z + \boldsymbol{a}(\boldsymbol{\omega})\cdot\tilde{\boldsymbol{x}}\right|^d \exp\left\{-\frac{z^2}{2}\right\} \exp\left\{\frac{(\boldsymbol{a}(\boldsymbol{\omega})\cdot\tilde{\boldsymbol{x}})^2}{2}\right\} dz d\boldsymbol{\omega} \\
&= \frac{\sqrt{2\pi}}{(\sqrt{2\pi})^{n(d+2)}\sqrt{n}^{d+1}} \exp\left\{-\frac{1}{2}\sum \|\boldsymbol{x}_i\|^2\right\} \\
&\quad \times \int I_d(\boldsymbol{a}(\boldsymbol{\omega})\cdot\tilde{\boldsymbol{x}}) \exp\left\{\frac{(\boldsymbol{a}(\boldsymbol{\omega})\cdot\tilde{\boldsymbol{x}})^2}{2}\right\} d\boldsymbol{\omega}.
\end{aligned}
$$

Therefore, the predictive distribution is given from equation A.14 as

$$
\hat{p}_{\mathrm{BAYES}}(\boldsymbol{x}|D) = \frac{1}{(\sqrt{2\pi})^{d+2}}\sqrt{\frac{n}{n+1}}^{d+1} \exp\left\{-\frac{\|\boldsymbol{x}\|^2}{2}\right\} \frac{S_d^J(\tilde{\boldsymbol{x}}_{n+1})}{S_d^J(\tilde{\boldsymbol{x}})}. \tag{A.24}
$$

By again using the same expansion as for the uniform case, we obtain

$$
\begin{aligned}
\hat{p}_{\mathrm{BAYES}}(\boldsymbol{x}|D) &= \frac{1}{(\sqrt{2\pi})^{d+2}} \exp\left\{-\frac{1}{2}\|\boldsymbol{x}\|^2\right\}\left\{1 + \frac{d}{2n}\right\} \\
&\quad \times \left\{1 + \frac{\nabla S_d^J(\tilde{\boldsymbol{x}})}{S_d^J(\tilde{\boldsymbol{x}})}\cdot\delta\boldsymbol{x} + \frac{1}{2}\mathrm{tr}\left(\frac{\nabla\nabla S_d^J(\tilde{\boldsymbol{x}})}{S_d^J(\tilde{\boldsymbol{x}})}\delta\boldsymbol{x}\delta\boldsymbol{x}^T\right)\right\} \\
&= \frac{1}{(\sqrt{2\pi})^{d+2}} \exp\left\{-\frac{1}{2}\|\boldsymbol{x}\|^2\right\} \\
&\quad \times \left\{1 + \frac{1}{\sqrt{n}}\nabla \log S_d^J(\tilde{\boldsymbol{x}})\cdot\boldsymbol{x} + \frac{1}{2n}\mathrm{tr}\left(\frac{\nabla\nabla S_d^J}{S_d^J}H_2(\boldsymbol{x})\right)\right\}.
\end{aligned} \tag{A.25}
$$

Proof of Theorem 7. Let us define

$$
Z_n = \int \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i, \xi, \boldsymbol{\omega}) d\xi d\boldsymbol{\omega}, \tag{A.26}
$$

$$
Z_{n+1} = \int \prod_{i=1}^{n+1} p(y_i|\boldsymbol{x}_i, \xi, \boldsymbol{\omega}) d\xi d\boldsymbol{\omega}. \tag{A.27}
$$

Then the Bayesian predictive distribution can be written as

$$\hat{p}_{\mathrm{BAYES}}(y|\boldsymbol{x}, D) = \frac{Z_{n+1}}{Z_n}. \tag{A.28}$$

Under the uniform prior, we can easily get

$$\begin{aligned} Z_n &= \frac{1}{\sqrt{2\pi}^n} \int \exp\left\{-\frac{1}{2}\sum y_i^2 + \xi \sum y_i \varphi_i - \frac{1}{2}\sum \varphi_i^2\right\} d\xi d\boldsymbol{\omega} \\ &= \frac{\sqrt{2\pi}}{\sqrt{2\pi}^n \sqrt{n}} \exp\left\{-\frac{1}{2}\sum y_i^2\right\} \int \frac{1}{A_n} \exp\left\{\frac{1}{2}\frac{Y_n(\boldsymbol{\omega})^2}{A_n(\boldsymbol{\omega})}\right\} d\boldsymbol{\omega}. \end{aligned} \tag{A.29}$$

Therefore, the predictive distribution can be written as

$$\hat{p}_{\mathrm{BAYES}}(y|\boldsymbol{x}, D) = \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{n+1}} \exp\left\{-\frac{1}{2}y^2\right\} \frac{P_d^U(D_{n+1})}{P_d^U(D_n)}, \tag{A.30}$$

where

$$\begin{aligned} P_d^U(D_n) &= \int \frac{1}{A_n} \exp\left\{\frac{1}{2}\frac{Y_n(\boldsymbol{\omega})^2}{A_n(\boldsymbol{\omega})}\right\} d\boldsymbol{\omega} \\ A_n(\boldsymbol{\omega}) &= \frac{1}{n}\sum_{i=1}^{n} \varphi_\beta^2(\boldsymbol{\omega}\cdot\boldsymbol{x}_i). \end{aligned}$$

Noting that $A_n(\boldsymbol{\omega})$ converges to $A(\boldsymbol{\omega})$ within the limit of large n, we substitute $P_d^U(D_n) = P_d^U(Y_n)$. Using the approximation of the form,

$$Y_{n+1} \approx Y_n + \left(\frac{y\varphi}{\sqrt{n}} - \frac{1}{2n}Y_n\right), \tag{A.31}$$

$$\begin{aligned} Q(Y_{n+1}, \boldsymbol{\omega}) \approx\ & Q(Y_n, \boldsymbol{\omega}) + \frac{1}{\sqrt{n}}\nabla Q(Y_n, \boldsymbol{\omega}) y\varphi \\ & + \frac{1}{2n}\left(\nabla\nabla Q(Y_n, \boldsymbol{\omega}) y^2\varphi^2 - \nabla Q(Y_n, \boldsymbol{\omega}) Y_n\right), \end{aligned} \tag{A.32}$$

$$\begin{aligned} P(Y_{n+1}) \approx\ & P(Y_n) + \frac{y}{\sqrt{n}} \int \nabla Q(Y_n, \boldsymbol{\omega})\varphi d\boldsymbol{\omega} \\ & + \frac{1}{2n}\left(y^2 \int \nabla\nabla Q(Y_n, \boldsymbol{\omega})\varphi^2 d\boldsymbol{\omega} - \int \nabla Q(Y_n, \boldsymbol{\omega}) Y_n d\boldsymbol{\omega}\right), \end{aligned} \tag{A.33}$$

we obtain

$$\begin{aligned}\hat{p}_{\mathrm{BAYES}}(y|\boldsymbol{x}, D) &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}y^2\right\}\left(1-\frac{1}{2n}\right)\frac{P_d^U(Y_{n+1})}{P_d^U(Y_n)} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}y^2\right\}\left\{1+\frac{y}{\sqrt{n}}\frac{\int \nabla Q(Y_n, \boldsymbol{\omega})\varphi d\boldsymbol{\omega}}{P(Y_n)}\right. \\ &\quad + \frac{1}{2n}\left(y^2\frac{\int \nabla\nabla Q(Y_n, \boldsymbol{\omega})\varphi^2 d\boldsymbol{\omega}}{P(Y_n)}\right. \\ &\quad \left.\left. - \frac{\int \nabla Q(Y_n, \boldsymbol{\omega})Y_n d\boldsymbol{\omega}}{P(Y_n)} - \frac{P(Y_n)}{P(Y_n)}\right)\right\}. \end{aligned} \tag{A.34}$$

By using the fact that

$$\frac{\partial^2}{\partial Y^2}Q(Y)A(\boldsymbol{\omega}) = \frac{\partial}{\partial Y}Q(Y)A(\boldsymbol{\omega}) + Q(Y),$$

we can finally obtain the result.

Proof of Theorem 8. From equation 6.39 and the definition of the generalization error, we get

$$\begin{aligned} E_{gen} &= -E_D E_{y,\boldsymbol{x}}\left[\log\left\{1+\frac{y}{\sqrt{n}}\frac{\int \nabla Q_d^U(Y_n, \boldsymbol{\omega})\varphi_\beta(\boldsymbol{\omega}\cdot\boldsymbol{x})d\boldsymbol{\omega}}{P_d^U(Y_n)}\right.\right. \\ &\quad \left.\left. + \frac{1}{2n}H_2(y)\frac{\int \nabla\nabla Q_d^U(Y_n, \boldsymbol{\omega})A_n(\boldsymbol{\omega})d\boldsymbol{\omega}}{P_d^U(Y_n)}\right\}\right] \\ &\approx -E_D E_{y,\boldsymbol{x}}\left[\frac{y}{\sqrt{n}}\frac{\int \nabla Q_d^U(Y_n, \boldsymbol{\omega})\varphi_\beta(\boldsymbol{\omega}\cdot\boldsymbol{x})d\boldsymbol{\omega}}{P_d^U(Y_n)}\right. \\ &\quad + \frac{1}{2n}H_2(y)\frac{\int \nabla\nabla Q_d^U(Y_n, \boldsymbol{\omega})A_n(\boldsymbol{\omega})d\boldsymbol{\omega}}{P_d^U(Y_n)} \\ &\quad \left. - \frac{y^2}{2n}\left(\frac{\int \nabla Q_d^U(Y_n, \boldsymbol{\omega})\varphi_\beta(\boldsymbol{\omega}\cdot\boldsymbol{x})d\boldsymbol{\omega}}{P_d^U(Y_n)}\right)^2\right] \\ &= \frac{1}{2n}E_D\left[\frac{\int \nabla Q_d^U(Y_n, \boldsymbol{\omega})\nabla Q_d^U(Y_n, \boldsymbol{\omega}')A(\boldsymbol{\omega}\boldsymbol{\omega}')d\boldsymbol{\omega}\boldsymbol{\omega}'}{\left(P_d^U(Y_n)\right)^2}\right]. \end{aligned} \tag{A.35}$$

Similarly, for the training error, we get

$$E_{train} = -\frac{1}{n}\sum_{i=1}^{n} E_D\left[\log\left\{1+\frac{y_i}{\sqrt{n}}\frac{\int \nabla Q_d^U(Y_n, \boldsymbol{\omega})\varphi_\beta(\boldsymbol{\omega}\cdot\boldsymbol{x}_i)d\boldsymbol{\omega}}{P_d^U(Y_n)}\right.\right.$$

$$
\begin{aligned}
&+\frac{1}{2n}H_2(y_i)\frac{\int \nabla\nabla Q_d^U(Y_n,\boldsymbol{\omega})A_n(\boldsymbol{\omega})d\boldsymbol{\omega}}{P_d^U(Y_n)}\Bigg\}\Bigg]\\
&\approx -\frac{1}{n}\sum_{i=1}^{n}E_D\left[\frac{y_i}{\sqrt{n}}\frac{\int \nabla Q_d^U(Y_n,\boldsymbol{\omega})\varphi_\beta(\boldsymbol{\omega}\cdot\boldsymbol{x}_i)d\boldsymbol{\omega}}{P_d^U(Y_n)}\right.\\
&\quad+\frac{1}{2n}H_2(y_i)\frac{\int \nabla\nabla Q_d^U(Y_n,\boldsymbol{\omega})A_n(\boldsymbol{\omega})d\boldsymbol{\omega}}{P_d^U(Y_n)}\\
&\quad\left.-\frac{y_i^2}{2n}\left(\frac{\int \nabla Q_d^U(Y_n,\boldsymbol{\omega})\varphi_\beta(\boldsymbol{\omega}\cdot\boldsymbol{x}_i)d\boldsymbol{\omega}}{P_d^U(Y_n)}\right)^2\right]\\
&=E_{gen}-\frac{1}{n}E_D\left[\frac{\int \nabla Q_d^U(Y_n,\boldsymbol{\omega})Y_n(\boldsymbol{\omega})d\boldsymbol{\omega}}{P_d^U(Y_n)}\right]. \qquad \text{(A.36)}
\end{aligned}
$$

On the other hand, from equation A.30, the generalization error is written as

$$
E_{gen}=\frac{1}{2}\log\left(\frac{n+1}{n}\right)+E_DE_{p_oq}\left[\log P_d^U(D_n)\right]-E_DE_{p_oq}\left[\log P_d^U(D_{n+1})\right].
$$

From the fact that

$$
\lim_{n\to\infty}E_DE_{p_oq}\left[\log P_d^U(D_n)\right]<\infty,
$$

we finally get

$$
E_{gen}=\frac{1}{2n}.
$$

Proof of Theorem 9. Let us define

$$
Z_n=p(D)=\int|\xi|^d\prod_{i=1}^{n}p(y_i|\boldsymbol{x}_i,\xi,\boldsymbol{\omega})d\xi d\boldsymbol{\omega}, \qquad \text{(A.37)}
$$

$$
Z_{n+1}=p(y,\boldsymbol{x},D)=\int|\xi|^d\prod_{i=1}^{n+1}p(y_i|\boldsymbol{x}_i,\xi,\boldsymbol{\omega})d\xi d\boldsymbol{\omega}. \qquad \text{(A.38)}
$$

The Bayesian predictive distribution can then be written as

$$
\hat{p}_{\text{BAYES}}(\boldsymbol{x}|D)=\frac{Z_{n+1}}{Z_n}. \qquad \text{(A.39)}
$$

Similar to the cone model, we get

$$\begin{aligned} Z_n &= \frac{1}{\sqrt{2\pi}^n} \int |\xi|^d \exp\left\{-\frac{1}{2}\sum y_i^2 + \xi \sum y_i \varphi_i - \frac{1}{2}\sum \varphi_i^2\right\} d\xi d\boldsymbol{\omega} \\ &= \frac{\sqrt{2\pi}}{\sqrt{2\pi}^n \sqrt{n}^{d+1}} \exp\left\{-\frac{1}{2}\sum y_i^2\right\} \\ &\quad \times \int \frac{1}{\sqrt{A_n}^{d+1}} I_d\left(\frac{Y_n(\boldsymbol{\omega})}{\sqrt{A_n(\boldsymbol{\omega})}}\right) \exp\left\{\frac{1}{2}\frac{Y_n(\boldsymbol{\omega})^2}{A_n(\boldsymbol{\omega})}\right\} d\boldsymbol{\omega}. \end{aligned} \tag{A.40}$$

Therefore, the predictive distribution can be written as

$$\hat{p}_{\text{BAYES}}(y|\boldsymbol{x}, D) = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n}{n+1}}^{d+1} \exp\left\{-\frac{1}{2}y^2\right\} \frac{P_d^J(D_{n+1})}{P_d^J(D_n)}. \tag{A.41}$$

By using the same approaches as for the uniform prior, we can easily obtain the final results.

Proof of Theorem 10. From equation A.14, the generalization error is written as

$$\begin{aligned} E_{gen} &= \frac{d+1}{2} \log\left(\frac{n+1}{n}\right) + E_D E_{p_o q}\left[\log P_d^J(D_n)\right] \\ &\quad - E_D E_{p_o q}\left[\log P_d^J(D_{n+1})\right]. \end{aligned} \tag{A.42}$$

From the fact that

$$\lim_{n\to\infty} E_D E_{p_o q}\left[\log P_d^J(D_n)\right] < \infty, \tag{A.43}$$

we finally get

$$E_{gen} = \frac{d+1}{2n} \tag{A.44}$$

For the proof, the same derivation process as for the uniform case can be applied.

References

Akaho, S., & Kappen, H. J. (2000). Nonmonotonic generalization bias of gaussian mixture models. *Neural Computation*, *12*, 6, 1411–1428.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Appl. Comp., AC-19*, 716–723.

Amari, S. (1967). Theory of adaptive pattern classifiers. *IEEE Trans., EC-16*(3), 299–307.

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics. 27*, 77–87.

Amari, S. (1987). Differential geometry of a parametric family of invertible linear systems—Riemannian metric, dual affine connections and divergence. *Mathematical Systems Theory, 20*, 53–82.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation, 10*, 251–276.

Amari, S. (2003). New consideration on criteria of model selection. In L. Rutkowski & J. Kacprzyk (Eds.), *Neural networks and soft computing (Proceedings of the Sixth International Conference on Neural Networks and Soft Computing)* (pp. 25–30). Heidelberg: Physica Verlag.

Amari, S., & Burnashev, M. V. (2003). On some singularities in parameter estimation problems. *Problems of Information Transmission. 39*, 352–372.

Amari, S., & Murata, N. (1993). Statistical theory of learning curves under entropic loss criterion. *Neural Computation, 5*, 140–154.

Amari, S., & Nagaoka, H. (2000). *Information geometry*. New York: AMS and Oxford University Press.

Amari, S., & Nakahara, H. (2005). Difficulty of singularity in population coding. *Neural Computation, 17*, 839–858.

Amari, S., & Ozeki, T. (2001). Differential and algebraic geometry of multilayer perceptrons. *IEICE Trans., E84-A*, 31–38.

Amari, S., Ozeki, T., & Park, H. (2003). Learning and inference in hierarchical models with singularities. *Systems and Computers in Japan, 34*, 34–42.

Amari, S., Park, H., & Fukumizu, K. (2000). Adaptive method of realizing natural gradient learning for multilayer perceptrons. *Neural Computation, 12*, 1399–1409.

Amari, S., Park, H., & Ozeki, T. (2001). Statistical inference in nonidentifiable and singular statistical models. *J. of the Korean Statistical Society, 30*(2), 179–192.

Amari, S., Park, H., & Ozeki, T. (2002). Geometrical singularities in the neuromanifold of multilayer perceptrons. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems, 14* (pp. 343–350). Cambridge, MA: MIT Press.

Brockett, R. W. (1976). Some geometric questions in the theory of linear systems. *IEEE Trans. on Automatic Control, 21*, 449–455.

Chen, A. M., Lu, H., & Hecht-Nielsen, R. (1993). On the geometry of feedforward neural network error surfaces. *Neural Computation, 5*, 910–927.

Dacunha-Castelle, D., & Gassiat, É. (1997). Testing in locally conic models, and application to mixture models. *Probability and Statistics, 1*, 285–317.

Fukumizu, K. (1999). Generalization error of linear neural networks in unidentifiable cases. In O. Watanabe & T. Yokomori (Eds.), *Algorithmic learning theory: Proceedings of the 10th International Conference on Algorithmic Learning Theory (ALT'99)* (pp. 51–62). Berlin: Springer-Verlag.

Fukumizu, K. (2003). Likelihood ratio of unidentifiable models and multilayer neural networks. *Annals of Statistics, 31*(3), 833–851.

Fukumizu, K., & Amari, S. (2000). Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks, 13*, 317–327.

Hagiwara, K. (2002a). On the problem in model selection of neural network regression in overrealizable scenario. *Neural Computation, 14*, 1979–2002.

Hagiwara, K. (2002b). Regularization learning, early stopping and biased estimator. *Neurocomputing, 48*, 937–955.

Hagiwara, K., Hayasaka, T., Toda, N., Usui, S., & Kuno, K. (2001). Upper bound of the expected training error of neural network regression for a gaussian noise sequence. *Neural Networks, 14*, 1419–1429.

Hagiwara, K., Toda, N., & Usui, S. (1993). On the problem of applying AIC to determine the structure of a layered feed-forward neural network. *Proceedings of IJCNN* (Vol. 3, pp. 2263–2266). Nagoya, Japan.

Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. *Proc. Barkeley Conf. in Honor of J. Neyman and J. Kiefer, 2*, 807–810.

Hotelling, H. (1939). Tubes and spheres in n-spaces, and a class of statistical problems. *Amer. J. Math., 61*, 440–460.

Inoue, M., Park, H., & Okada, M. (2003). On-line learning theory of soft committee machines with correlated hidden units—Steepest gradient descent and natural gradient descent. *J. Phys. Soc. Jpn., 72*(4), 805–810.

Kang, K., Oh, J.-H., Kwon, S., & Park, Y. (1993). Generalization in a two-layer neural networks. *Phys. Rev. E, 48*(6), 4805–4809.

Kitahara, M., Hayasaka, T., Toda, N., & Usui, S. (2000). On the statistical properties of least squares estimators of layered neural networks (in Japanese). *IEICE Transactions, J86-D-II*, 563–570.

Kůrková, V., & Kainen, P. C. (1994). Functionally equivalent feedforward neural networks. *Neural Computation, 6*, 543–558.

Liu, X., & Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Annals of Statistics, 31*(3), 807–832.

Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Murata, N., Yoshizawa, S., & Amari, S. (1994). Network information criterion—determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks, 5*(6), 865–872.

Park, H., Amari, S., & Fukumizu, K. (2000). Adaptive natural gradient learning algorithms for various stochastic models. *Neural Networks, 13*, 755–764.

Park, H., Inoue, M., & Okada, M. (2003). On-line learning dynamics of multilayer perceptrons with unidentifiable parameters. Submitted to *J. Phys. A: Math. Gen., 36*, 11753–11764.

Rattray, M., & Saad, D. (1999). Analysis of natural gradient descent for multilayer neural networks. *Physical Review E, 59*, 4523–4532.

Rattray, M., Saad, D., & Amari, S. (1998). Natural gradient descent for on-line learning. *Physical Review Letters, 81*, 5461–5464.

Riegler, P., & Biehl, M. (1995). On-line backpropagation in two-layered neural networks. *J. Phys. A; Math. Gen., 28*, L507–L513.

Risssanen, J. (1986). Stochastic complexity and modeling. *Ann. Statist. 14*, 1080–1100.

Rosenblatt, F. (1961). *Principles of neurodynamics*. New York: Spartan.

Rüger, S. M., & Ossen, A. (1997). The metric of weight space. *Neural Processing Letters, 5*, 63–72.

Rumelhart, D., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error backpropagation. In D. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge, MA: MIT Press.

Saad, D., & Solla, A. (1995). On-line learning in soft committee machines. *Phys. Rev. E, 52*, 4225–4243.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Sussmann, H. J. (1992). Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks. 5*, 589–593.

Watanabe, S. (2001a). Algebraic analysis for non-identifiable learning machines. *Neural Computation, 13*, 899–933.

Watanabe, S. (2001b). Algebraic geometrical methods for hierarchical learning machines. *Neural Networks, 14*(8), 1409–1060.

Watanabe, S. (2001c). Algebraic information geometry for learning machines with singularities. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems, 13* (pp. 329–336). Cambridge, MA: MIT Press.

Watanabe, S., & Amari, S. (2003). Learning coefficients of layered models when the true distribution mismatches the singularities. *Neural Computation, 15*(5), 1013–1033.

Weyl, H. (1939). On the volume of tubes. *Amer. J. Math., 61*, 461–472.

Wu, S., Amari, S., & Nakahara, H. (2002). Population coding and decoding in a neural field: A computational study. *Neural Computation, 14*, 999–1026.

Wu, S., Nakahara, H., & Amari, S. (2001). Population coding with correlation and an unfaithful model. *Neural Computation, 13*, 775–797.

Yamazaki, K., & Watanabe, S. (2002). A probabilistic algorithm to calculate the learning curves of hierarchical learning machines with singularities. *Trans. on IEICE, J85-D-II*(3), 363–372.

Yamazaki, K., & Watanabe, S. (2003). Singularities in mixture models and upper bounds of stochastic complexity. *International Journal of Neural Networks, 16*(7), 1029–1038.

Received June 28, 2004; accepted May 26, 2005.