

Statistical Analysis of Unidentifiable Models and its Application to Multilayer Neural Networks

Kenji Fukumizu

Institute of Statistical Mathematics

4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan

E-mail: fukumizu@ism.ac.jp

Abstract

This paper discusses the maximum likelihood estimator of the parametric model that has lack of identifiability in low dimensional subsets in the parameter space. Among many statistical models with unidentifiability, neural network models are the main concern of this paper. The unidentifiable true parameter is formulated as a conic singularity of the model embedded in an infinite dimensional space of probability density functions. Following Hartigan's idea, the likelihood ratio of the maximum likelihood estimator is described by the supremum of an empirical process over a set of functions. It has been known in some models the asymptotics of the likelihood ratio has an unusually larger order. A useful sufficient condition of the larger order is shown, and applied to neural networks. The order of the asymptotic likelihood ratio are shown in various cases of multilayer perceptrons.

1 Introduction

This paper discusses the asymptotic behavior of the maximum likelihood estimator (MLE) under the condition that the true parameter is unidentifiable. The asymptotics of MLE is an important problem in statistical estimation theory, and the asymptotic normality under some regularization conditions are well known ([1]). However, if the dimensionality of the set of true parameters is larger than one, the Fisher information matrix at a true parameter is singular, and the asymptotic normality is no longer satisfied. The asymptotic behavior of MLE in such unidentifiable situations has not been clarified completely.

We formulate the problem of unidentifiability as a conic singularity ([2]) in the set of a statistical model, embedded in the space of all the probability

density functions. In this formulation, the likelihood ratio of the MLE, with the true probability at the singularity, can be described by the supremum of an empirical process over the unit vectors in the tangent cone, which marginally converges to a Gaussian distribution. This empirical process shows very different behavior depending on the functional property of the tangent cone. One of the interesting features is the order of the likelihood ratio as the number of samples n goes to infinity. A model satisfying the regularity condition of the usual asymptotic theory has the likelihood ratio of the order $O_p(1/n)$. However, larger order have been reported in some unidentifiable models. Hartigan ([3]) discusses the normal mixture models with two components, and shows the likelihood ratio test statistics under the hypothesis of one component has larger order than $O_p(1/n)$. In neural networks, the order $O_p(\log n/n)$ has been derived in unidentifiable cases ([4]). A useful sufficient condition of such larger order than $O_p(1/n)$ will be given in the term of functional properties of the tangent cone. I will further derive the order of the likelihood ratio for some neural network models, with the true probability at the singularity, analyzing the functional properties of the tangent cone.

2 Unidentifiability and Locally Conic Models

2.1 Preliminaries

Let $(\mathcal{Z}, \mathcal{B}, \mu)$ be a measure space. A *statistical model* $S = \{f(z; \theta) \mid \theta \in \Theta\}$ is a family of probability density functions on $(\mathcal{Z}, \mathcal{B}, \mu)$, where the parameter space Θ is a domain in the d -dimensional Euclidean space \mathbb{R}^d . We assume that $f(z; \theta) > 0$ for all z and θ , and differentiable on θ for each $z \in \mathcal{Z}$. Suppose that the probability distribution of i.i.d. random variables Z_1, Z_2, \dots, Z_n is $f_0(z)\mu$ with the probability density function $f_0(z) > 0$. The function f_0 is called the *true probability density*. Given the random variables, the *likelihood ratio* of the model S with respect to $\{Z_i\}_{i=1}^n$ is defined by

$$L_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log \frac{f(Z_i; \theta)}{f_0(Z_i)}. \quad (1)$$

Note that $L_n(\theta)$ is normalized by $1/n$, so that it can be compared with the Kullback-Leibler divergence, which will be defined later. We consider the *maximum likelihood estimator* (MLE) $\hat{\theta}$ that attains the maximum of the

likelihood ratio, if it exists. From the definition, we have

$$L_n(\hat{\theta}) = \sup_{\theta \in \Theta} L_n(\theta) = \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log \frac{f(Z_i; \theta)}{f_0(Z_i)}. \quad (2)$$

For a density $f(z; \theta)$ in the model S , we define the *Kullback-Leibler divergence* of $f(z; \theta)$ from $f(z)$ by

$$D(\theta) = \int f(z) \log \frac{f_0(z)}{f(z; \theta)} d\mu(z). \quad (3)$$

The main topic of this paper is the behavior of the likelihood ratio and the Kullback-Leibler divergence of the maximum likelihood estimator under the asymptotic assumption, where the number of samples goes to infinity.

2.2 Unidentifiability of the true parameter

Throughout this paper, the true probability density function $f(z)$ is assumed to be included in the model $\{f(z; \theta) \mid \theta \in \Theta\}$. Then, there exists $\theta_0 \in \Theta$ such that $f(z; \theta_0) = f_0(z)$. We *do not* assume the uniqueness of θ_0 , and denote the set of true parameters by Θ_0 ; i.e. $\Theta_0 = \{\theta \in \Theta \mid f(z; \theta) = f_0(z)\}$. Unless Θ_0 consists a single point, the usual view of asymptotic convergence to a single true parameter does not hold.

We say that the true parameter is *unidentifiable*, if the set of true parameters Θ_0 is a union of submanifolds of Θ , and the dimension of at least one of the submanifolds is equal to or larger than one. There are many important statistical models with unidentifiability. A famous example is a finite mixture model. Let $g(z; a)$ be a probability density function on \mathcal{Z} with a variable parameter a , and $f(z; a_1, a_2, b)$ be a mixture model defined by

$$f(z; a_1, a_2, b) = b g(z; a_1) + (1 - b) g(z; a_2), \quad (4)$$

where $b \in [0, 1]$. Suppose the true density is given by $g(z; a_0)$ for some a_0 , then, the set of parameters to give $g(z; a_0)$ is $\{(a_1, a_2, b) \mid a_1 = a_2 = a_0, b : \text{arbitrary}\} \cup \{(a_1, a_2, b) \mid b = 0, a_2 = a_0, a_1 : \text{arbitrary}\} \cup \{(a_1, a_2, b) \mid b = 1, a_1 = a_0, a_2 : \text{arbitrary}\}$, which is high dimensional. The reduced rank problems ([5]) and the change point problem ([6]) are other examples of models with unidentifiability. Feed-forward neural network models, such as multilayer perceptrons ([7]), also have unidentifiability. We will mainly discuss the multilayer perceptron model in this paper.

Our main concern is to investigate how the likelihood ratio and the Kullback-Leibler divergence asymptotically behave in the case that the true

parameter is unidentifiable. As a comparison, I briefly review the well-known results for identifiable models. Under some regularity conditions, the asymptotic distribution of the likelihood ratio and the Kullback-Leibler divergence have the same value in the leading term;

$$L_n(\hat{\theta}) = D(\hat{\theta}) + o_p(1/n), \quad (5)$$

and the limiting distribution is given by

$$nL_n(\hat{\theta}) \xrightarrow[n \rightarrow \infty]{} \chi_d^2 \quad \text{in law,} \quad (6)$$

where χ_d^2 denotes the chi-square distribution of freedom d . In unidentifiable cases, even the order of the likelihood ratio can be different from $O_p(1/n)$, as I will discuss later.

2.3 Locally conic model

The unidentifiability is defined in terms of parameters. However, if we consider the set of probability density functions defined by the model, the set of true parameters corresponds a single point in the space of density functions. The point is singular in the set of density functions of the model, because the dimensionality shrinks only at the point. The property of the set of density functions around the singularity can be better understood, if we can find more convenient parameterization than the original one. Following Dacunha-Castelle & Gassiat ([2]), we utilize a conic singularity, with some modification, to describe the unidentifiability.

Let $\Theta \subset \mathbb{R}^d$ be an open set, $S = \{f(z; \theta) \mid \theta \in \Theta\}$ be a statistical model, and $f_0(z)$ be an element in S . The parameter $\theta \in \Theta$ is decomposed as $\theta = (\alpha, \beta)$ for $\alpha \in \mathbb{R}^{d-1}$ and $\beta \in \mathbb{R}$. The statistical model S is called locally conic at f_0 if the following conditions are satisfied;

1. $f(z; \theta)$ is C^∞ function of θ for almost every z .
2. Let $\Theta_0 = \Theta \cap (\mathbb{R}^{d-1} \times \{0\})$, $A_0 = \{\alpha \in \mathbb{R}^{d-1} \mid (\alpha, 0) \in \Theta_0\}$, and $\Theta(\alpha) = \Theta \cap (\{\alpha\} \times \mathbb{R})$ for each $\alpha \in A_0$. Then,

$$\Theta = \bigcup_{\alpha \in A_0} \Theta(\alpha). \quad (7)$$

3. The set of the parameters to give f_0 is Θ_0 ; that is,

$$f(z; (\alpha, \beta))\mu = f_0(z)\mu \iff \beta = 0. \quad (8)$$

4. $\frac{\partial}{\partial \beta} \log f(z; \alpha, \beta)$ is in $L^2(f_{(\alpha, \beta)}\mu)$, and

$$\left\| \frac{\partial}{\partial \beta} \log f(z; \alpha, 0) \right\|_{L^2(f_0\mu)} = 1 \quad (9)$$

for all $\alpha \in A_0$.

Unless Θ_0 is a single point, the parameter giving f_0 is not identifiable. Geometrically, a locally conic model S is a d -dimensional set with a singularity at f_0 in the space of probability density functions. For each $\alpha \in A_0$, the submodel $S_\alpha = \{f(z; \theta) \mid \theta \in \Theta(\alpha)\}$ is a one-dimensional, identifiable statistical model. The score function of S_α at the origin,

$$v_\alpha(z) = \frac{\partial \log f(z; (\alpha, 0))}{\partial \beta}, \quad (10)$$

can be looked as a unit tangent vector in the direction of S_α . The family of score functions $C = \{v_\alpha \mid \alpha \in A_0\}$ generates the tangent cone at the singularity f_0 . We call C *the basis of the tangent cone*.

The view of tangent vectors can be rigorously formulated if S is included in a maximal exponential model ([8]), which is an infinite dimensional Banach manifold. The basis of the tangent cone C has a key importance in the following discussion. In the definition, we require only that the functions in C are in $L^2(f_0\mu)$. They are not necessarily real tangent vectors in the Banach manifold.

2.4 Neural network as a locally conic model

A feed-forward neural network model is an example of a locally conic model. We mainly discuss multilayer perceptrons ([7]) in later sections. The *multi-layer perceptron* model with H hidden units is defined by a family of functions

$$\varphi(x; \theta) = \sum_{j=1}^H b_j s(a_j x + c_j) + d, \quad (11)$$

where $x \in \mathcal{X} = \mathbb{R}$, $s(t) = \tanh(t)$, and $\theta = (a_1, c_1, b_1, \dots, a_H, c_H, b_H, d)^T$. We discuss only one-dimensional input and output for simplicity.

We can regard learning in neural networks as statistical estimation. Assume a probability $Q = q(x)dx$ on \mathcal{X} for the distribution of the input sample X_i , and a conditional probability density function $r(y \mid u)$ of $y \in \mathcal{Y} = \mathbb{R}$ given $u \in \mathbb{R}$. Throughout this paper, we put the following assumption;

[Conditions on noise model (NM)]

1. The Fisher information $I(u)$ of $r(y|u)$, defined by

$$I(u) = \int \left(\frac{\partial \log r(y|u)}{\partial u} \right)^2 r(y|u) dy, \quad (12)$$

is positive, finite, and continuous for all $u \in \mathbb{R}$.

2. The integral

$$\int \left| \frac{\partial \log r(y|u)}{\partial u} \right|^3 r(y|u) dy \quad (13)$$

is finite and continuous for $u \in \mathbb{R}$.

Using the function $\varphi(x; \theta)$, we define a statistical model by

$$f(z; \theta) = r(y | \varphi(x; \theta))q(x), \quad (14)$$

where $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.

Popular choices of $r(y | u)$ are the additive Gaussian noise model

$$r(y | u) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y - u)^2\right\} \quad (15)$$

for continuous y , and the binomial distribution model

$$r(y | u) = \frac{e^{uy}}{1 + e^u} \quad (16)$$

for binary output $y \in \{0, 1\}$, which often appears in classification problems.

The true parameter can be unidentifiable in the multilayer perceptron model. We see it using the simplest case. Suppose we have the multilayer perceptron model with 2 hidden units, and the true function $\varphi_0(x)$ can be given by a perceptron with only one hidden unit. If $\varphi_0(x) = b_0 \tanh(a_0 x)$, then for any parameter θ in the set $\{\theta \in \Theta \mid a_1 = a_0, b_1 = b_0, c_1 = 0, b_2 = 0, d = 0, a_2, c_2 : \text{arbitrary}\}$ and $\{\theta \in \Theta \mid a_1 = a_0, b_1 = b_0, c_1 = 0, a_2 = 0, b_2 \tanh(c_2) + d = 0\}$ the function $\varphi(x; \theta)$ equals to the true function ¹. We can see that the set of true parameters is a high dimensional subset in the parameter space. It is known if the true function can be realized by

¹These two subsets do not give all the parameters to realize $\varphi_0(x)$. The whole set of the true parameters is shown in [11].

a network with smaller number of hidden units than the model, the true parameter is unidentifiable ([9],[10],[11]).

This unidentifiability of multilayer perceptrons can be formulated as a locally conic model. Suppose we have the multilayer perceptrons with H hidden units. Let $K \in \mathbb{N}$ be less than H , and $\varphi_0(x)$ be a function realizable by a multilayer perceptron with K hidden units.

Let $\Theta_H^* = \{\theta = (a_1, \dots, a_H, b_1, \dots, b_H, c_1, \dots, c_H, d) \in \mathbb{R}^{3H+1} \mid a_j \neq 0, b_j \neq 0 (1 \leq j \leq H), (a_j, c_j) \neq \pm(a_h, c_h) (1 \leq j < h \leq H)\}$ be the parameter space of the multilayer perceptrons with H hidden units. Note that we eliminate the parameters which correspond functions realizable by a smaller-sized network (see [9]). This modification does not matter in discussing the maximum likelihood estimation, because the maximum likelihood estimator lies in Θ_H^* with probability one. For a parameter in Θ_H^* , it is known ([12]) that the functions $\{1, s(a_j x + c_j), s'(a_j x + c_j)x, s'(a_j x + c_j) \mid 1 \leq j \leq H\}$ are linearly independent.

Given a function

$$\varphi_0(x) = \sum_{k=1}^K b_k^0 s(a_k^0 x + c_k^0) + d^0 \quad (17)$$

for $(a_k^0, b_k^0, c_k^0, d^0) \in \Theta_K^*$, we slightly modify the parameter space as $\Theta_H^{**} = \{\theta \in \Theta_H^* \mid (a_j, c_j) \neq \pm(a_k^0, c_k^0) (1 \leq k \leq K, K+1 \leq j \leq H)\}$, and introduce a new parameterization by

$$\begin{aligned} \beta &= \text{sgn}(b_{K+1}) \sqrt{b_{K+1}^2 + \dots + b_H^2}, \\ \xi_k &= \frac{a_k - a_k^0}{\beta}, \quad (1 \leq k \leq K), & \xi_j &= a_j, \quad (K+1 \leq j \leq H), \\ \eta_k &= \frac{b_k - b_k^0}{\beta}, \quad (1 \leq k \leq K), & \eta_j &= \frac{b_j}{\beta}, \quad (K+1 \leq j \leq H), \\ \zeta_k &= \frac{c_k - c_k^0}{\beta}, \quad (1 \leq k \leq K), & \zeta_j &= c_j, \quad (K+1 \leq j \leq H), \\ \delta &= \frac{d - d^0}{\beta}. \end{aligned} \quad (18)$$

for $\theta \in \Theta_H^{**}$. Define a new parameter space Π_H by

$$\begin{aligned} \Pi_H = \{ & \omega = (\xi_1, \dots, \xi_H, \eta_1, \dots, \eta_H, \zeta_1, \dots, \zeta_H, \delta, \beta) \mid \\ & a_k^0 + \beta\xi_k \neq 0 \ (1 \leq k \leq K), \ \xi_j \neq 0 \ (K+1 \leq j \leq H), \\ & (a_k^0 + \beta\xi_k, c_k^0 + \beta\zeta_k) \neq (a_h^0 + \beta\xi_h, c_h^0 + \beta\zeta_h) \ (1 \leq k < h \leq H), \\ & (a_k^0 + \beta\xi_k, c_k^0 + \beta\zeta_k) \neq \pm(\xi_j, \zeta_j) \ (1 \leq k \leq K, K+1 \leq j \leq H), \\ & (\xi_j, \zeta_j) \neq \pm(\xi_i, \zeta_i) \ (K+1 \leq j < i \leq H), \\ & (\xi_j, \zeta_j) \neq \pm(a_k^0, c_k^0) \ (1 \leq k \leq K, K+1 \leq j \leq H), \\ & b_k^0 + \beta\eta_k \neq 0 \ (1 \leq k \leq K), \ \sum_{j=K+1}^H \eta_j^2 = 1, \ \eta_j \neq 0 \ (K+1 \leq j \leq H), \\ & \eta_{K+1} > 0, \ \beta \in \mathbb{R} \} \end{aligned} \quad (19)$$

and $\Pi_H^{**} = \{\omega \in \Pi_H \mid \beta \neq 0\}$. Rewrite the multilayer perceptron using this parameterization;

$$\begin{aligned} \psi(x; \omega) = & \sum_{k=1}^K (b_k^0 + \beta\eta_k) s((a_k^0 + \beta\xi_k)x + (c_k^0 + \beta\zeta_k)) \\ & + \sum_{j=K+1}^H \beta\eta_j s(\xi_j x + \zeta_j) + \beta\delta. \end{aligned} \quad (20)$$

It is easy to see that the Π_H^{**} and Θ_H^{**} are diffeomorphic by the above correspondence, and $\varphi(x; \theta) = \psi(x; \omega)$ for the corresponding $\theta \in \Theta_H^{**}$ and $\omega \in \Pi_H^{**}$.

We write $\omega = (\alpha, \beta)$, summarizing $(\xi_1, \dots, \zeta_H, \delta)$ by α . By the fact $(a_k^0, b_k^0, c_K^0, d^0) \in \Theta_K^*$ and $(\xi_j, \zeta_j) \neq \pm(a_k^0, c_k^0)$, we can show that $\Pi_H^{**} = \cup_{(\alpha, 0) \in \Pi_{H,0}} \Pi_H(\alpha)$. Consider the family of functions $\{\psi(x; \omega) \mid \omega \in \Pi_H\}$. We can see that $\psi(x; \omega) = \varphi_0(x)$ if and only if $\omega \in \Pi_{H,0}$; that is, $\beta = 0$. The sufficiency is trivial. For the necessity, because the functions $\{1, s(a_k^0 x + c_k^0), s(\xi_j x + \zeta_j), s((a_k^0 + \beta\xi_k)x + (c_k^0 + \beta\zeta_k)) \mid K+1 \leq j \leq H, 1 \leq k \leq K\}$ are linearly independent by the definition of Π_H , we see that the coefficients of $s(\xi_j x + \zeta_j)$ must be zero to realize $\psi(x; \omega) = \varphi_0(x)$. This implies $\beta = 0$.

The basis of the tangent cone is essentially determined by the following

partial derivatives;

$$\begin{aligned} \frac{\partial \psi(x; (\alpha, 0))}{\partial \beta} &= \sum_{j=K+1}^H \eta_j s(\xi_j x + \zeta_j) + \delta \\ &+ \sum_{k=1}^K \eta_k s(a_k^0 x + c_k^0) + \sum_{k=1}^K b_k^0 \xi_k s'(a_k^0 x + c_k^0) x + \sum_{k=1}^K b_k^0 \zeta_k s'(a_k^0 x + c_k^0). \end{aligned} \quad (21)$$

Let $q(x)$ be a p.d.f. of x , such that $q(x)$ is absolute continuous with respect to the Lebesgue measure on \mathbb{R} . Let $r(y|u)$ be a conditional p.d.f. of y given u , such that $r(y|u_1)dy \neq r(y|u_2)dy$ for different u_1 and u_2 , and the Fisher information $I(u)$ is positive and continuous on u . Let $S_H = \{f(x, y; \omega) \mid \omega \in \Pi_H\}$ be a statistical model defined by $f(x, y; \omega) = r(y|\psi(x; \omega))q(x)$. The model S_H consists of probability density functions corresponding to $\varphi_0(x)$ and the functions realized by multilayer perceptrons with H hidden units and not by a smaller-sized network. The function $f_0(x, y)$ be a density function defined by $\varphi_0(x)$, that is, $f_0(x, y) = r(y|\varphi_0(x))q(x)$. We have the following proposition;

Proposition 1. *Let S_H be the statistical model of multilayer perceptrons with H hidden units defined as above, and f_0 be a density function in S_K where $0 \leq K < H$. Then, S_H is locally conic at f_0 .*

Proof. From what we have seen, the model S_H satisfies the conditions 1, 2, and 3 in the definition of a locally conic model. For the condition 4, let $N(\alpha)$ be the $L^2(f_0(x, y)dxdy)$ -norm of $\frac{\partial}{\partial \beta} \log f(x, y; (\alpha, 0))$. We have

$$\begin{aligned} N(\alpha)^2 &= \int \int r(y|\varphi_0(x))q(x) \left(\frac{r'(y|\varphi_0(x))}{r(y|\varphi_0(x))} \frac{\partial \psi(x; (\alpha, 0))}{\partial \beta} \right)^2 dxdy \\ &= \int I(\varphi_0(x)) \left(\frac{\partial \psi(x; (\alpha, 0))}{\partial \beta} \right)^2 q(x) dx. \end{aligned} \quad (22)$$

Since $\varphi_0(x)$ is bounded, we see $I(\varphi_0(x))$ is bounded. From eq.(21), the function $\left(\frac{\partial \psi(x; (\alpha, 0))}{\partial \beta} \right)^2$ is also bounded. Thus, $N(\alpha)$ is finite. Because the functions $1, s(\xi_j x + \zeta_j), s(a_k^0 x + c_k^0), s'(a_k^0 x + c_k^0) x,$ and $s'(a_k^0 x + c_k^0)$ are linearly independent (see [12]), the partial derivative $\frac{\partial}{\partial \beta} \psi(x; (\alpha, 0))$ is not constant zero. Therefore, $0 < N(\alpha) < \infty$ for all $\alpha \in A_0$. Using $N(\alpha)\beta$ instead of β , we have the normalized tangent vectors at $f_0(x, y)$. \square

3 Maximum likelihood estimation in locally conic models

3.1 MLE and supremum of a random process

Let $S = \{f(z; (\alpha, \beta)) \mid (\alpha, \beta) \in \Theta\}$ be a statistical model, which is locally conic at $f_0 \in S$. Suppose Z_1, Z_2, \dots, Z_n are i.i.d. random variables with the law $f_0\mu$. For each $\alpha \in A_0$, the submodel $S_\alpha = \{f(z; (\alpha, \beta)) \mid \beta \in \Theta(\alpha)\}$ is a smooth, one-dimensional model with a variable parameter β . Consider the maximum likelihood estimator $\hat{\beta}_\alpha$ in S_α , then, the likelihood ratio of the maximum likelihood estimator in S is given by

$$\sup_{\theta \in \Theta} L_n(\theta) = \sup_{\alpha} L_n(\alpha, \hat{\beta}_\alpha). \quad (23)$$

Fix α and concentrate S_α for a while. Assume that each submodel satisfy the regularity conditions of the asymptotic efficiency². The Taylor expansion leads us to

$$L_n(\alpha, \hat{\beta}_\alpha) = \frac{1}{2n} U_n(\alpha)^2 + o_p(1/n), \quad (24)$$

where $U_n(\alpha)$ is an empirical process defined by

$$U_n(\alpha) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n v_\alpha(Z_i)}{\sqrt{\frac{1}{n} \sum_{i=1}^n v_\alpha(Z_i)^2}}, \quad (25)$$

and $v_\alpha(z)$ is a function in the basis of the tangent cone C , defined by

$$v_\alpha(z) = \frac{\partial}{\partial \beta} \log f(z; (\alpha, 0)). \quad (26)$$

The denominator of $U_n(\alpha)$ converges to one almost surely and the numerator converges in law to the standard normal distribution for each $\alpha \in A_0$. If we consider the behavior of $U_n(\alpha)$ over all α , it can be looked as a stochastic process over α or C , and every marginal distribution on finite points converges to a multidimensional normal distribution.

The maximum likelihood estimation is, then, given by

$$\sup_{\theta \in \Theta} L_n(\theta) = \sup_{\alpha} \left\{ \frac{1}{2n} U_n(\alpha)^2 + o_p(1/n) \right\}. \quad (27)$$

²A set of conditions is found in Sen and Singer ([13], Theorem 5.2.1), which shows weaker conditions than the famous ones by Cramér ([1]). Another set of conditions is given in Dacunha-Castelle and Gassiat ([2]), also.

If the higher order term of $o_p(1/n)$ is bounded uniformly over α , we can eliminate the term from the supremum;

$$\sup_{\theta \in \Theta} L_n(\theta) = \sup_{\alpha} \left\{ \frac{1}{2n} U_n(\alpha)^2 \right\} + o_p(1/n). \quad (28)$$

Furthermore, if the stochastic process U_n converges "nicely" to a Gaussian process W over C , the limit of the supremum of $|U_n|$ can be replaced by the square of the supremum of $|W|$. Then, we obtain

$$\sup_{\theta \in \Theta} L_n(\theta) = \sup_{\alpha} \frac{1}{2n} W^2 + o_p(1/n). \quad (29)$$

Dacunha-Castelle and Gassiat ([2]) discuss this case, assuming that the function class $C = \{v_{\alpha}(z)\}$ is Donsker, which assures the nice convergence of U_n .

Let (Ω, \mathcal{A}, P) be a probability space, $(\mathcal{Z}, \mathcal{B})$ be a measurable space, and $Z_n : \Omega \rightarrow \mathcal{Z}$ ($n \in \mathbb{N}$) be i.i.d. random variables with the law P . A family of Borel measurable functions $\mathcal{F} \subset \{v : \mathcal{Z} \rightarrow \mathbb{R}\}$ is called *Donsker* if $E_P[v(Z)]$ and $E_P[v(Z)^2]$ exist for all $v \in \mathcal{F}$, the map $z \mapsto \sup_{v \in \mathcal{F}} |v(z)|$ is finite for every $z \in \mathcal{Z}$, and the \mathcal{F} -indexed empirical processes

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (v(Z_i) - E_P[v(Z)]), \quad (30)$$

as considered to be random elements with their values in the Banach space $\ell^{\infty}(\mathcal{F})$ of all the bounded functions on \mathcal{F} with sup norm, converge in law to a tight³ Borel measurable random element with its value in $\ell^{\infty}(\mathcal{F})$.

When discussing the stochastic process U_n in eq.(24), we will investigate both of Donsker and non-Donsker cases. For Donsker cases, Dacunha-Castelle and Gassiat ([2]) clarify the limiting distribution of likelihood ratio of the maximum likelihood estimator, as we will see in the next subsection, and apply the result to finite mixture models and ARMA models. In this paper, we will derive a relation between the likelihood ratio and the Kullback-Leibler divergence of the maximum likelihood estimator in Donsker cases, as a simple consequence of their result. In non-Donsker cases, a diversity of phenomena are seen, and even the order of the likelihood ratio can be different from the usual $O_p(1/n)$, as I have discussed in Section 1.

³Let \mathcal{X} be a topological space, and $(\mathcal{X}, \mathfrak{S})$ be the Borel measurable space. A Borel measurable random variable $Z : \Omega \rightarrow \mathcal{X}$ is called *tight* if for arbitrary ε there exist a compact set K in \mathcal{X} such that $P(Z \in K) \geq 1 - \varepsilon$.

3.2 Donsker cases

To apply the theory of convergence to a Gaussian process, we have to assure the uniformity over α of the small order in eq.(24). First, for the uniform consistency of $\hat{\beta}_\alpha$, we need the following uniform Wald conditions.

[Uniform Wald conditions (W)]

1. There exists a set E with $f(z)\mu$ -probability 1 such that for any z in E and any α ,

$$\lim_{|\beta| \rightarrow \infty} f(z; (\alpha, \beta)) = 0. \quad (31)$$

2. Consider the functions

$$F(z; \beta, \rho) := \sup_{|\beta' - \beta| \leq \rho} f(z; \beta', \alpha), \quad G(z; r) := \sup_{|\beta| \geq r} f(z; \beta, \alpha) \quad (32)$$

for $\rho > 0$ and $r > 0$, and define $F^*(z; \beta, \rho) = \max\{F(z; \beta, \rho), 1\}$ and $G^*(z; r) = \max\{G(z; r), 1\}$. Then, the following conditions hold;

$$\lim_{\rho \rightarrow +0} E_{f_0(z)\mu}[\log F^*(z; \beta, \rho)] < \infty, \quad \lim_{r \rightarrow \infty} E_{f_0(z)\mu}[\log G^*(z; r)] < \infty. \quad (33)$$

Using the same discussion in Wald ([14]), under the above conditions (W), the maximum likelihood estimator in the submodel $\hat{\beta}_\alpha$ converges to 0 in probability uniformly over α .

To assure the uniformly small order of $o_p(1/n)$, we further assume the following condition:

[Uniformity condition (U)]

Consider the functions

$$\begin{aligned} H_1(z; \beta, \rho) &:= \sup_{|\beta' - \beta| \leq \rho} \left| \frac{\frac{\partial}{\partial \beta} f(z; \beta', \alpha)}{f(z; \beta', \alpha)} \right|, & K_1(z; r) &:= \sup_{|\beta| \geq r} \left| \frac{\frac{\partial}{\partial \beta} f(z; \beta, \alpha)}{f(z; \beta, \alpha)} \right|, \\ H_2(z; \beta, \rho) &:= \sup_{|\beta' - \beta| \leq \rho} \left| \frac{\frac{\partial^2}{\partial \beta^2} f(z; \beta', \alpha)}{f(z; \beta', \alpha)} \right|, & K_2(z; r) &:= \sup_{|\beta| \geq r} \left| \frac{\frac{\partial^2}{\partial \beta^2} f(z; \beta, \alpha)}{f(z; \beta, \alpha)} \right|. \end{aligned} \quad (34)$$

Then, the following conditions hold for $i = 1, 2$;

$$\lim_{\rho \rightarrow +0} E_{f_0(z)\mu}[(H_i(z; \beta, \rho))^2] < \infty, \quad \lim_{r \rightarrow \infty} E_{f_0(z)\mu}[K_i(z; r)] < \infty. \quad (35)$$

The following theorem is due to Dacunha-Castelle and Gassiat ([2]).

Theorem 1. *Let a statistical model $S = \{f(z; (\alpha, \beta))\}$ be locally conic at $f_0(z)$. Assume (W) and (U) hold, and the family of functions $C = \{v_\alpha(z) = \frac{\partial}{\partial \beta} f(z; (\alpha, 0))\}$ is Donsker. then the supremum of the likelihood ratio converges in law as follows;*

$$n \sup_{(\alpha, \beta)} L_n(\alpha, \beta) \longrightarrow \frac{1}{2} \sup_{v \in C} W^2, \quad (36)$$

where W is a tight, Borel measurable Gaussian process over C , which is a limit of the empirical process U_n .

A sufficient condition of the Donsker is known ([15]). A class of functions \mathcal{F} is Donsker if (i) the envelop function $F(z) = \sup_{v \in \mathcal{F}} |v(z)|$ is P -(outer) square integrable, (ii) the square root of the uniform entropy number is integrable, and (iii) P -measurability on some function classes are satisfied.

In these three conditions, the measurability conditions are automatically satisfied if $\mathcal{F} = \{w(z; a)\}$ is parameterized by a separable metric space and $w(z; a)$ is continuous about a for all z . This is true for the basis of the tangent cone of a locally conic model. A sufficient condition for integrability of the uniform entropy number is that the VC-dimension of \mathcal{F} is finite. These are satisfied by the tangent cone of many locally conic models, such as neural networks.

Note that the condition (i) is satisfied if the integral of the square of $H_1(z; 0, \rho)$ is finite for a sufficiently small ρ . Therefore, we obtain the following corollary.

Corollary 1. *Let a statistical model $S = \{f(z; (\alpha, \beta))\}$ be locally conic at $f_0(z)$. Assume (W) and (U) hold, and the VC-dimension of $C = \{v_\alpha(z) = \frac{\partial}{\partial \beta} f(z; (\alpha, 0) \mid \alpha \in A_0\}$ is finite. Then, C is Donsker, and eq.(36) holds for a tight, Borel measurable Gaussian process W .*

In Donsker cases, we can derive a simple relation between the likelihood ratio and the Kullback-Leibler divergence, which is satisfied by regular models also.

Theorem 2. *Under the same assumptions as Theorem 1 or Corollary 1, $D(\hat{\theta})$ and $L_n(\hat{\theta})$ have the order of $O_p(1/n)$, and the relation*

$$D(\hat{\theta}) = L_n(\hat{\theta}) + o_p(1/n) \tag{37}$$

holds.

Proof. The standard argument of Taylor expansion of D with respect to β gives the second argument. Since W is a tight Gaussian process, the class C is necessarily totally bounded in $L^2(P)$, and almost all the sample paths $v \mapsto W(v)$ are uniformly $L^2(P)$ continuous (see van der Vaart and Wellner [15], Section 1.5). Then, the supremum of $|W|$ is finite almost surely. \square

We can not obtain the exact distribution of the likelihood ratio or Kullback-Leibler divergence in unidentifiable cases, while we know also the limiting distribution in regular cases. The distribution of $\sup |W|$ is very difficult in general. In non-Donsker cases, a clear relation between the KL-divergence and the likelihood ratio has not been known yet.

3.3 Non-Donsker cases

As we mentioned in Section 1, the likelihood ratio of the maximum likelihood estimator does not necessarily have the usual order $O_p(1/n)$, but can have a larger order, if the function class of the tangent cone is "rich" enough like normal mixtures and multilayer perceptrons.

We derive a useful sufficient condition of such an unusually larger order, extending Hartigan's idea. Note that a marginal of U_n on finite points v_1, \dots, v_m in C always converges to a multi-dimensional normal distribution. The covariance of the limit is given by

$$E_P[v_i v_j]. \tag{38}$$

The two components are independent if their covariance is zero. Suppose we can find arbitrary number of "almost" independent Gaussian random variables in C , then, the supremum of $U_n(\alpha)$ on such variables can take an arbitrary large value, since the maximum of m independent samples from the standard normal distribution is approximately $\sqrt{2 \log m}$ for large m . Hartigan ([3]) applied this idea to a normal mixture model with two components, calculating the covariance explicitly. An extension of this idea leads us to the following theorem;

Theorem 3. Let a statistical model $S = \{f(z; (\alpha, \beta))\}$ be locally conic at $f_0(z)$, and $C = \{v_\alpha(z) = \frac{\partial}{\partial \beta} f(z; (\alpha, 0))\}$ be the basis of the tangent cone. Suppose there exists a sequence $\{v_n\}_{n=1}^\infty$ in C such that $v_n \rightarrow 0$ in probability, then, for arbitrary $M > 0$, we have

$$\lim_{n \rightarrow \infty} \Pr\left(\sup_{(\alpha, \beta)} nL_n(\alpha, \beta) \leq M\right) = 0. \quad (39)$$

Proof. From Proposition 2 below, for arbitrary $\varepsilon > 0$ and $K \in \mathbb{N}$, there exist $v(\alpha_1), \dots, v(\alpha_K) \in C$ such that $|E[v(\alpha_i)v(\alpha_j)]| < \varepsilon$ for different i and j . The rest of the proof is accomplished in the same way as Hartigan ([3]), which will be shown below.

Let $W = (W_1, \dots, W_K)$ be a random vector following the limiting normal distribution of $(U_n(v_{\alpha_1}), \dots, U_n(v_{\alpha_K}))$, and Σ be the variance-covariance matrix of W . Because the absolute value of every off-diagonal element in Σ is less than ε , by Geršgorin's inequality ([16]), $(1 + (K - 1)\varepsilon)I_K \leq \Sigma \leq (1 - (K + 1)\varepsilon)I_K$. We obtain for arbitrary $M > 0$

$$\begin{aligned} P\left(\max_{1 \leq i \leq K} |W_i| \leq M\right) &\leq \int_{[-M, M]^K} \frac{1}{\sqrt{(2\pi)^K |\Sigma|}} e^{-\frac{1}{2(1+(K-1)\varepsilon)} W^T W} dW \\ &\leq \frac{(1+(K-1)\varepsilon)^{K/2}}{|\Sigma|^{1/2}} \int_{[-M, M]^K} \frac{1}{(2\pi)^{K/2}} e^{-\frac{1}{2} u^T u} du \\ &\leq \left(\frac{1+(K-1)\varepsilon}{1-(K-1)\varepsilon}\right)^{K/2} \{\Phi(M) - \Phi(-M)\}^K, \end{aligned} \quad (40)$$

where $\Phi(t)$ is the cumulative distribution function of the standard normal distribution. For any $\delta > 0$ and $M > 0$, there exists $K \in \mathbb{N}$ such that $\{\Phi(M) - \Phi(-M)\}^K < \frac{\delta}{2}$. For such K , we can find $\varepsilon > 0$ that satisfies $\frac{1+(K-1)\varepsilon}{1-(K-1)\varepsilon} < 2$. Then, we have

$$P\left(\max_{1 \leq i \leq K} |W_i| \leq M\right) < \delta. \quad (41)$$

The convergence of $(U_n(\alpha_1), \dots, U_n(\alpha_K))$ to W means $\lim_{n \rightarrow \infty} P(\max_i |U_n(\alpha_i)| \leq M) = P(W \in [-M, M]^K)$. This completes the proof. \square

On the covariance of the random variables with bounded L^2 norm, we have the following proposition.

Proposition 2. Let $\{v_n\}_{n=1}^\infty$ be a sequence in $L^2(P)$ such that $\|v_n\|_{L^2(P)} = 1$ for all n , and $v_n \rightarrow 0$ in probability. Then, for all n and $\varepsilon > 0$, there exists ℓ_0 such that

$$E_P |v_n v_\ell| < \varepsilon \quad (42)$$

for all $\ell \geq \ell_0$.

This is a direct consequence of the following proposition.

Proposition 3. *Let (Ω, \mathcal{B}, P) be a probability space, and Y, X_1, X_2, \dots be random variables. Suppose that $\int Y^2 dP$ and $\int X_n^2 dP$ are upper bounded by K , and X_n converges to 0 in probability. Then, we have*

$$\lim_{n \rightarrow \infty} E|Y X_n| = 0. \quad (43)$$

Proof. Let ε be an arbitrary positive number. Because $\int Y^2 dP < \infty$, there exists $\delta > 0$ such that $\int_{\Delta} Y^2 dP < \frac{\varepsilon^2}{9K}$ for any measurable set Δ with $P(\Delta) < \delta$.

For each n , define a set

$$A_n = \{\omega \in \Omega \mid |Y| > \frac{\varepsilon}{3\sqrt{K}} \text{ and } |X_n| > \frac{\varepsilon}{3K}|Y|\}. \quad (44)$$

Because $X_n \rightarrow 0$ in probability and $A_n \subset \{|X_n| > \frac{\varepsilon^2}{9K^{3/2}}\}$, we can find n_0 such that $P(A_n) < \delta$ for all $n \geq n_0$. Then, we have $\int_{A_n} Y^2 dP < \frac{\varepsilon^2}{9K}$ for $n \geq n_0$.

For $n \geq n_0$, we derive

$$\begin{aligned} \int |Y X_n| dP &= \int_{A_n} |Y X_n| dP + \int_{A_n^c} |Y X_n| dP \\ &\leq \left(\int_{A_n} Y^2 dP \right)^{1/2} \left(\int_{A_n} X_n^2 dP \right)^{1/2} + \int_{\{|Y| \leq \frac{\varepsilon}{3\sqrt{K}}\}} |Y X_n| dP + \int_{\{|X_n| \leq \frac{\varepsilon}{3K}|Y|\}} |Y X_n| dP \\ &\leq \frac{\varepsilon}{3\sqrt{K}} \sqrt{K} + \frac{\varepsilon}{3\sqrt{K}} \int |X_n| dP + \frac{\varepsilon}{3K} \int |Y|^2 dP \\ &\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3\sqrt{K}} \cdot \sqrt{K} + \frac{\varepsilon}{3K} \cdot K = \varepsilon \end{aligned} \quad (45)$$

In the last line, we use the fact $\int |X_n| dP \leq (\int |X_n|^2 dP)^{1/2} \leq \sqrt{K}$. \square

4 Maximum Likelihood Estimation of Multilayer Perceptrons

We apply the results in the previous section to multilayer perceptrons. For simplicity, we discuss networks without bias terms in the output unit:

$$\varphi(x; \theta) = \sum_{j=1}^H b_j s(a_j x + c_j). \quad (46)$$

Similar to Section 2.4, given the true function $\varphi_0(x) = \sum_{j=1}^K b_j^0 s(a_j^0 x + c_j^0)$, the locally conic parameterization of this model is given by

$$\psi(x; \alpha, \beta) = \sum_{k=1}^K (b_k^0 + \beta \eta_k) s((a_k^0 + \beta \xi_k)x + (c_k^0 + \beta \zeta_k)) + \sum_{j=K+1}^H \beta \eta_j s(\xi_j x + \zeta_j), \quad (47)$$

where $\alpha = (\xi_i, \eta_i, \zeta_i)$, and the basis of the tangent cone C consist of the functions

$$v_\alpha(x, y) = \frac{1}{\|\sqrt{I(\varphi_0(x))} \frac{\partial \psi}{\partial \beta}(x; (\alpha, 0))\|_{L^2(Q)}} \frac{r'(y|\varphi_0(x))}{r(y|\varphi_0(x))} \frac{\partial \psi(x; (\alpha, 0))}{\partial \beta}, \quad (48)$$

where

$$\begin{aligned} \frac{\partial \psi(x; (\alpha, 0))}{\partial \beta} &= \sum_{j=K+1}^H \eta_j s(\xi_j x + \zeta_j) + \sum_{k=1}^K \eta_k s(a_k^0 x + c_k^0) \\ &\quad + \sum_{k=1}^K b_k^0 \xi_k s'(a_k^0 x + c_k^0) x + \sum_{k=1}^K b_k^0 \zeta_k s'(a_k^0 x + c_k^0). \end{aligned} \quad (49)$$

It is easy to see that C includes a sequence converging to constant zero almost everywhere, if $H - K \geq 2$. In fact, we can find such a sequence by $\zeta = \zeta_{K+1} = -\zeta_{K+2} \rightarrow 0$, $\xi_{K+1}, \xi_{K+2} \rightarrow \infty$ for the element for the form

$$\frac{1}{A} \frac{r'(y|\varphi_0(x))}{r(y|\varphi_0(x))} \{s(\xi_{K+1}x + \zeta) - s(\xi_{K+2}x - \zeta)\}, \quad (50)$$

where A is a normalizing constant. Applying Theorem 3, we obtain the following

Theorem 4. *Assume the model is the multilayer perceptron model (46) with H hidden units, and the true function is given by a network with K hidden units. If $K \leq H - 2$, for arbitrary $M > 0$, we have*

$$\lim_{n \rightarrow \infty} \Pr \left(\sup_{(\alpha, \beta)} nL_n(\alpha, \beta) \leq M \right) = 0. \quad (51)$$

We can derive a tighter lower bound of the likelihood ratio in the above problem, by counting a number of almost independent random variables in C .

Theorem 5. *Under the same assumptions as Theorem 4, we have*

$$\sup_{\theta} L_n(\theta) \gtrsim O_p\left(\frac{\log n}{n}\right), \quad (52)$$

as n goes to infinity.

Sketch of the proof. Take a subset in C defined by

$$w(x, y; a, c, \delta) = \frac{1}{N} \frac{r'(y|\varphi_0(x))}{r(y|\varphi_0(x))} \{s(a(x - (c + \delta))) - s(a(x - (c - \delta)))\}, \quad (53)$$

where N is a normalizing constant. For a interval $I \subset \mathbb{R}$, define a function

$$u(x, y; I) = \frac{1}{M(I)} \frac{r'(y|\varphi_0(x))}{r(y|\varphi_0(x))} \chi_I(x), \quad (54)$$

where $\chi_I(x)$ is the characteristic function of I , and $M(I)$ is the normalizing constant given by

$$M(I) = \int \int \left(\frac{r'(y|\varphi_0(x))}{r(y|\varphi_0(x))} \right)^2 \chi_I(x) r(y|\varphi_0(x)) q(x) dy dx \quad (55)$$

When a goes to infinity, the function $w(x, y; a, c, \delta)$ converges to $u(x, y; [c - \delta, c + \delta])$ in $L^2(f_0\mu)$. Let $A = M(\mathbb{R})$, and take $K > 0$ such that $M([-K, K]) = \frac{A}{2}$. We can obtain a partition $\{I_k \mid k = 1, \dots, m\}$ of $[-K, K]$ such that $M(I_k) = \frac{A}{2m}$ for all k . Then, $u(x, y; I_k)$ are uncorrelated.

Consider the third moment of $|u(x, y; I_k)|$. Let $H_3(x)$ be a function defined by $H_3(x) = \int \left| \frac{r'(y|\varphi_0(x))}{r(y|\varphi_0(x))} \right|^3 r(y|\varphi_0(x)) dy$. By the assumption [NM] in Section 2.4, there exists $B > 0$ such that $H_3(x) \leq BI(\varphi_0(x))$ for all $x \in [-K, K]$. Then, we easily obtain

$$E_{f_0\mu} |u(x, y; I_k)|^3 \leq B \sqrt{\frac{2}{A}} \sqrt{m}. \quad (56)$$

Let $F_n^{(k)}(t)$ be the cumulative distribution function of the random variable $\frac{1}{\sqrt{n}} \sum_{i=1}^n u(Z_i; I_k)$ for $1 \leq k \leq m$. Take $m = n^\gamma$ for $0 < \gamma < 1$. By Berry-Esseen-type inequality ([17], Theorem 3), we obtain

$$\sup_{t \in \mathbb{R}} |F_n^{(k)}(t) - \Phi(t)| \leq \frac{LC}{n^{(1-\gamma)/2}} \quad (57)$$

for all $1 \leq k \leq m$, where L is a universal constant. Therefore, the variables $\frac{1}{\sqrt{n}} \sum_{i=1}^n u(Z_i; I_k)$ are uniformly close to the standard normal distribution for sufficiently large n .

From the extreme value theory, the supremum of the absolute values of m i.i.d. samples from the standard normal distribution is $2 \log m$. Since the $m = n^\gamma$ random variables $\frac{1}{\sqrt{n}} \sum_{i=1}^n u(Z_i; I_k)$ are arbitrary close to standard normal distribution, and mutually almost independent, we can prove that the supremum of $|U_n|^2$ over them is of the order $\log m = \gamma \log n$. \square

The order $O_p(\log n/n)$ has been formerly obtained by Hagiwara et al. ([4]). However, they assume the additive Gaussian noise model. Our approach extends their results. The above theorem can be applied to various noise models, including binary output models.

As we can see in the above discussions, the behavior of the likelihood ratio deeply depends on the functional property of the tangent cone C . If the multilayer perceptron model has only one redundant hidden unit, the behavior can be totally different.

As a comparison with the above result, we see the likelihood of the multilayer perceptrons model with one hidden unit for the constant zero target. In this case, the basis of the tangent cone is Donsker, and we can apply Theorem 2.

Theorem 6. *Assume the model is the multilayer perceptron model (46) with one hidden unit, and the true function is constant zero. Then, we have*

$$D(\hat{\theta}) = L_n(\hat{\theta}) + o_p(1/n), \quad \text{and} \quad L_n(\hat{\theta}) = O_p(1/n). \quad (58)$$

We omit the proof. \square

5 Conclusion

We have discussed an approach to investigate the behavior of the maximum likelihood estimator in the case that the true parameter is not identifiable. We have seen that the unidentifiability of parameters in a statistical model can be formulated by a conic singularity in many cases. Following the discussion of Dacunha-Castelle and Gassiat ([2]), we have formulated the likelihood ratio of the maximum likelihood estimator by the supremum of an empirical process, which converges to the standard normal distribution marginally. Rather than concentrating on Donsker cases, we have discussed non-Donsker cases, and derived a useful sufficient condition of an unusual

larger order of the likelihood ratio. We have applied these results on neural network models, and derived the lower bound of the likelihood ratio, assuming that the true function is constant zero.

The omitted proofs will be presented in a forthcoming paper.

Acknowledgements

I thank Prof. Kano in Osaka University, Prof. Kuriki in the Institute of Statistical Mathematics, and Prof. Amari in RIKEN Brain Science Institute for valuable discussions.

References

- [1] Cramér, H. (1946) *Mathematical Methods of Statistics*. Princeton University Press.
- [2] Dacunha-Castelle, D. and Gassiat, E. (1997) Testing in locally conic models, and application to mixture models *ESAIM Probability and Statistics*, **1**, 285–317.
- [3] Hartigan, J.A. (1985). A failure of likelihood asymptotics for normal mixtures. *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer*, vol.II, pp.807–810.
- [4] Hagiwara, K., Kuno K., and Usui S. (2000) On the problem in model selection of neural network regression in overrealizable scenario. *Proceeding of International Joint Conference of Neural Networks*.
- [5] Fukumizu, K. (1999) Generalization error of linear neural networks in unidentifiable cases. O.Watanabe and T.Yokomori (eds.) *Lecture Notes in Artificial Intelligence 1720, Algorithmic Learning Theory (Proceedings of the 10th International Conference on Algorithmic Learning Theory (ALT'99))*, pp.51-62. Springer-Verlag: Berlin.
- [6] Csörgö, M. and Horváth, L. (1996) *Limit Theorems in Change-Point Analysis*. John Wiley & Sons.
- [7] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) in: *Learning internal representations by error propagation*, eds. D.E. Rumelhart, J.L. McClelland and the PDP Research Group, Parallel distributed processing, Vol.1 (MIT Press, Cambridge) pp.318–362.

- [8] Pistone, G. and Sempi, C. (1995). An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *The Annals of Statistics*, 23(5):1543–1561.
- [9] Sussmann, H.J. (1992) Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, **5**, 589–593.
- [10] Chen, A. M., Lu, H., and Hecht-Nielsen, R. (1993). On the geometry of feedforward neural network error surfaces. *Neural Computation*, **5**, 910–927.
- [11] Fukumizu, K. and Amari, S. (2000) Local Minima and Plateaus in Hierarchical Structures of Multilayer Perceptrons. *Neural Networks*, **13**(3), 317–327.
- [12] Fukumizu, K. (1996) A regularity condition of the information matrix of a multilayer perceptron network. *Neural Networks*, **9**(5), 871–879.
- [13] Sen, P.K. and Singer, J.M. (1993) *Large sample methods in statistics*. Chapman & Hall.
- [14] Wald, A. (1949) Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, **20**, 595–601.
- [15] Van der Vaart, A.W. & Wellner, J.A. (1996). *Weak convergence and empirical processes*. Springer Verlag.
- [16] Horn, R. & Johnson, C. (1990). *Matrix Analysis*. Cambridge University Press.
- [17] Nagev, S.V. (1965). Some limit theorems for large deviations. *Theory of Probability and its Applications*, **10**(2), 214–235.