Likelihood Ratio of Unidentifiable Models and 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

February 19, 2001

Abstract

This paper discusses the behavior of the maximum likelihood estimator, when the true parameter cannot be identified uniquely. Among many statistical models with unidentifiability, neural network models are the main concern of this paper. The set of unidentifiable true parameters is formulated as a conic singularity of the model, which is embedded in an infinite dimensional space of probability density functions. It has been known in some models with unidentifiability the asymptotics of the likelihood ratio of MLE has an unusually larger order. Following Hartigan's idea, the likelihood ratio of MLE is described by the supremum of an empirical process over a set of functions, and a useful sufficient condition of such larger orders is derived. This result is applied to neural network models, and a larger order is observed if the true function is realized by a network with a smaller number of hidden units than the model. A stronger lower bound of the order of likelihood ratio is also derived on condition that there are at least two redundant hidden units to realize the true function.

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 is well known ([1]). However, if the dimensionality of the set of true parameters is larger than zero, the Fisher information matrix at a true

parameter is singular, and the asymptotic normality is no longer satisfied. The behavior of MLE in such unidentifiable situations has not been clarified completely.

There are many statistical models that have unidentifiability. Finite mixture models, ARMA, reduced rank regression, and change point problems are typical examples of such models. Because the asymptotics of the MLE is not simple, model selection needs special consideration on such models. It is known that feed-forward neural networks have also the problem of unidentifiability. The true parameter of a feed-forward neural network model is unidentifiable, if the true function is realized by a network with smaller number of hidden units than the model. In this paper, we mainly discuss the neural network model in investigating the behavior of MLE closely.

We formulate the problem of unidentifiability as a conic singularity ([2]) in the set of a statistical model, which is 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 well described by the supremum of an empirical process over the unit vectors in the tangent cone. This empirical process shows very different behavior depending on the functional property of the tangent cone, while each marginal variable converge to a Gaussian distribution.

One of the interesting features is the order of the likelihood ratio of MLE, as the sample-size n goes to infinity. A model satisfying the regularity condition of the usual asymptotic theory has the likelihood ratio of the order $O_p(1)$. However, larger orders 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 a larger order than $O_p(1)$. In neural networks, the lower bound $O_p(\log n)$ has been derived in unidentifiable cases ([4]). In this paper, a useful sufficient condition of such larger orders than $O_p(1)$ will be given in the term of functional properties of the tangent cone. This result covers many models of a larger order of the likelihood ratio. Furthermore, a stronger lower bound of the order for some neural network models will be derived, by the analysis of 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, and S be a set of probability density functions on $(\mathcal{Z}, \mathcal{B}, \mu)$. The set S is called a *statistical model* if there is

a differentiable manifold (with boundary) Θ such that S is given by $S = \{f(z;\theta) \mid \theta \in \Theta\}$. We call Θ as the *parameter space*. We assume throughout this paper that $\mathrm{Supp} f(z;\theta)$ is invariant for all $\theta \in \Theta$, and $f(z;\theta)$ is differentiable on θ for each $z \in \mathcal{Z}$.

Suppose that the probability distribution of i.i.d. random variables $Z_1, Z_2, \ldots, Z_n$ is $f_0(z)\mu$ with the probability density function $f_0(z)$, which has the same support as the model S. 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

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

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

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

The main topic of this paper is the behavior of the likelihood ratio of MLE 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 $f_0(z)$ 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; that is, $\Theta_0 = \{\theta \in \Theta \mid f(z;\theta_0)\mu = f_0(z)\mu\}$. Unless Θ_0 is 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 finitely many submanifolds of Θ, and the dimension of at least one of the submanifolds is larger than zero. There are many important statistical models in which the true parameter can be unidentifiable. One of the most famous examples 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

$$f(z;a_1,a_2,b) = b\,g(z;a_1) + (1-b)\,g(z;a_2), \tag{3}$$

where $b \in [0,1]$. Suppose that the true density $f_0(z)$ is given by $g(z;a_0)$ for some a_0. Then, the set of parameters to give $f_0(z)$ contains $\{(a_1,a_2,b) \mid a_1 = a_2 = a_0,\ b : \text{arbitrary}\} \cup \{(a_1,a_2,b) \mid b = 0, a_2 = a_0,\ a_1 : \text{arbitrary}\} \cup \{(a_1,a_2,b) \mid b = 1, a_1 = a_0,\ a_2 : \text{arbitrary}\}$, which is high dimensional. The reduced rank problems ([5]), ARMA model ([6]), and the change point problem ([7]) are other examples of models with unidentifiability. Feed-forward neural network models, such as multilayer perceptrons ([8]), are also among such models. We will mainly discuss the multilayer perceptron model in this paper.

Our main concern is to investigate how the likelihood ratio of MLE behaves on condition that the true parameter is unidentifiable. If the true parameter is identifiable, under some regularity conditions, the asymptotic distribution of the likelihood ratio of MLE converges in law to the chi-square distribution of freedom d. On the other hand, in unidentifiable cases, even the order of the likelihood ratio of MLE can be different from $O_p(1)$, as shown later.

2.3 Locally conic model

In the previous subsection, the unidentifiability was defined in terms of the parameters. However, if the space of probability density functions is considered, the set of true parameters corresponds to a single point in the space. The point is a singularity in the set of density functions defined by the model, if the dimensionality shrinks only at the point. The property of the set of density functions around the singularity will be better understood, if more convenient parameterization can be introduced than the original one. Following Dacunha-Castelle & Gassiat ([2]), with some modification, a conic singularity is utilized for describing the unidentifiability.

Let A_0 be a $(d-1)$-dimensional differentiable manifold (with boundary), Θ an open set in $A_0 \times \mathbb{R}$, and $S = \{f(z;\theta) \mid \theta \in \Theta\}$ be a statistical model. The parameter $\theta \in \Theta$ is decomposed as $\theta = (\alpha, \beta)$ for $\alpha \in A_0$ and $\beta \in \mathbb{R}$. Let a function $f_0(z)$ be an element in S. The statistical model S is called *locally conic* at f_0 if the following conditions are satisfied;

1. $f(z;(\alpha,\beta))$ is differentiable on β for each $\alpha \in A_0$ and $f_0\mu$-almost every z.

2. Let Θ_0 and $\Theta(\alpha)$ be subsets defined by $\Theta_0 = \Theta \cap (A_0 \times \{0\})$ and

$\Theta(\alpha) = \Theta \cap (\{\alpha\} \times \mathbb{R})$ for $\alpha \in A_0$, respectively. Then,

$$\Theta = \bigcup_{\alpha \in A_0} \Theta(\alpha). \tag{4}$$

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

$$f(z;(\alpha,\beta))\mu = f_0(z)\mu \iff \beta = 0. \tag{5}$$

4. For all $\alpha \in A_0$,

$$\left\| \frac{\partial \log f(z;\alpha,0)}{\partial \beta} \right\|_{L^2(f_0\mu)} = 1. \tag{6}$$

If the dimension of A_0 is larger than zero, the parameter giving f_0 is not identifiable. Intuitively, 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,

$$v_\alpha(z) = \frac{\partial \log f(z;(\alpha,0))}{\partial \beta}, \tag{7}$$

can be looked as a unit tangent vector in the direction of S_α (see fig.1). The family of score functions $C = \{v_\alpha \mid \alpha \in A_0\}$ generates the tangent cone at the singularity f_0. We call the set C *the basis of the tangent cone*, which has a key importance in the following discussion.

The view of tangent vectors can be rigorously formulated if S is included in a maximal exponential model ([9]), which is an infinite dimensional Banach manifold. In the definition, we only require that the functions in C are in $L^2(f_0\mu)$. They are not necessarily tangent vectors of the Banach manifold in the sense of Pistone and Sempi ([9]).

2.4 Neural network as a locally conic model

A feed-forward neural network model is an example of a locally conic model. This paper mainly discusses multilayer perceptrons ([8]). The *multilayer perceptron* model with H hidden units is defined by a family of functions

$$\varphi(x;\theta) = \sum_{j=1}^{H} b_j \, s(a_j x + c_j) + d, \tag{8}$$

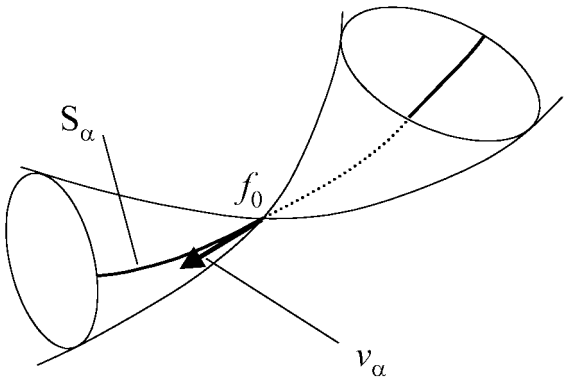

Figure 1: Locally conic model

where $x \in \mathcal{X} = \mathbb{R}$, $s(t) = \tanh(t)$, and $\theta = (a_1, \ldots, a_H, b_1, \ldots, b_H, c_1, \ldots, c_H, d)^T \in \Theta_H = \mathbb{R}^{3H+1}$. Only models with one-dimensional input and output is discussed for simplicity.

Learning in neural networks can be regarded as statistical estimation. Assume that the distribution of an input sample X_i is a probability Q on $\mathcal{X} = \mathbb{R}$. When the multilayer perceptron model is discussed, it is always assumed that Q is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$, which is written by $\mu_{\mathbb{R}}$, with the density function $q(x)$, and that the integral $E_Q|\log q(x)|$ is finite. Let $\mathcal{Y}$ be a subset of $\mathbb{R}$, and $(\mathcal{Y}, \mathcal{B}_y, \mu_y)$ be a measure space. Let $r(y \mid u)$ be a conditional probability density function of $y \in \mathcal{Y}$ given $u \in \mathbb{R}$. This is used for a noise model. Throughout this paper, we put the following assumptions;

[Conditions on noise model (NM1)]

1. The conditional density $r(y|u)$ is of class C^1 on u for all $y \in \mathcal{Y}$.
2. For different u_1 and u_2, we have $r(y|u_1)\mu_y \neq r(y|u_2)\mu_y$.
3. The Fisher information $G(u)$ of $r(y|u)$, defined by

$$G(u) = \int \Big(\frac{\partial \log r(y|u)}{\partial u} \Big)^2 r(y|u) d\mu_y, \tag{9}$$

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

4. For all $u \in \mathbb{R}$

$$\lim_{\rho \downarrow 0} E_{r(y|u)} \Big[\sup_{|u'-u| \leq \rho} \Big| \frac{\partial \log r(y|u')}{\partial u} \Big| \Big] < \infty. \tag{10}$$

The condition 4 assures the famous relation $E_{r(y|u)}[\frac{\partial^2 \log r(y|u)}{\partial u^2}] = -G(u)$ by Lebesgue's dominated convergence theorem.

Given the function $\varphi(x;\theta)$, the statistical model of multilayer perceptron is defined by

$$f(z;\theta) = r(y \mid \varphi(x;\theta))q(x), \tag{11}$$

where $z = (x,y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, with respect to the measure $\mu_{\mathbb{R}} \times \mu_y$.

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

$$r(y \mid u) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{-\frac{1}{2\sigma^2}(y-u)^2\Big\} \tag{12}$$

for continuous y, and the binomial distribution model

$$r(y \mid u) = \frac{e^{uy}}{1+e^u} \tag{13}$$

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

The true parameter can be unidentifiable in the multilayer perceptron model. It can be seen in the simplest case as follows. Suppose we have the multilayer perceptron model with 2 hidden units, and the true function $\varphi_0(x)$ is 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 $\{\theta \in \Theta_2 \mid a_1 = a_0, b_1 = b_0, c_1 = 0, b_2 = 0, d = 0, a_2, c_2 : \text{arbitrary}\} \cup \{\theta \in \Theta_2 \mid a_1 = a_0, b_1 = b_0, c_1 = 0, a_2 = 0, b_2 \tanh(c_2) + d = 0\}$ 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 that the true parameter is unidentifiable if and only if the true function can be realized by a network with smaller number of hidden units than the model ([10],[11],[12]).

This unidentifiability of multilayer perceptrons can be formulated as a locally conic model. Suppose we have the multilayer perceptrons with H

[1]These two subsets do not give all the parameters to realize $\varphi_0(x)$. The whole set of the true parameters is shown in [12].

hidden units. Let K be an integer such that $0 \leq K < H$, and $\varphi_0(x)$ be a function realizable by a multilayer perceptron with K hidden units.

A slightly restricted parameter space Θ_H^* is defined by $\Theta_H^* = \{\theta = (a_1, \ldots, a_H, b_1, \ldots, b_H, c_1, \ldots, c_H, d) \in \Theta_H \mid a_j \neq 0,\ b_j \neq 0\ (1 \leq j \leq H),\ (a_j, c_j) \neq \pm(a_h, c_h)\ (1 \leq j < h \leq H)\}$. Note that in Θ_H^* the parameters that correspond to the functions realizable by a smaller-sized network are eliminated (see [10]). For a parameter in Θ_H^*, it is known ([13]) 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

$$\varphi_0(x) = \sum_{k=1}^{K} b_k^0 \, s(a_k^0 x + c_k^0) + d^0 \tag{14}$$

for $\theta_0 = (a_1^0, \ldots, a_K^0, b_1^0, \ldots, b_K^0, c_1^0, \ldots, c_K^0, d^0) \in \Theta_K^*$, the parameter space is again restricted slightly to Θ_H^{**} by $\Theta_H^{**} = \{\theta \in \Theta_H^* \mid (a_j, c_j) \neq \pm(a_k^0, c_k^0)\ (1 \leq k \leq K, K+1 \leq j \leq H)\}$. This reduction does not matter in discussing the maximum likelihood estimation, because MLE lies in Θ_H^{**} with probability one. Introduce a new parameterization by

$$\begin{aligned}
&\beta = \mathrm{sgn}(b_{K+1})\sqrt{b_{K+1}^2 + \cdots + b_H^2}, \\
&\xi_k = \frac{a_k - a_k^0}{\beta}, \quad (1 \leq k \leq K), \qquad \xi_j = a_j, \quad (K+1 \leq j \leq H), \\
&\eta_k = \frac{b_k - b_k^0}{\beta}, \quad (1 \leq k \leq K), \qquad \eta_j = \frac{b_j}{\beta}, \quad (K+1 \leq j \leq H), \\
&\zeta_k = \frac{c_k - c_k^0}{\beta}, \quad (1 \leq k \leq K), \qquad \zeta_j = c_j, \quad (K+1 \leq j \leq H), \\
&\delta = \frac{d - d^0}{\beta}.
\end{aligned} \tag{15}$$

for $\theta \in \Theta_H^{**}$, and define new parameter spaces Π_H and Π_H^{**} by

$$\begin{aligned}\Pi_H = \{\omega = (\xi_1, \ldots, \xi_H, \eta_1, \ldots, \eta_H, \zeta_1, \ldots, \zeta_H, \delta, \beta) \mid \\ a_k^0 + \beta\xi_k \neq 0 \ (1 \leq k \leq K),\ \xi_j \neq 0 \ (K+1 \leq j \leq H), \\ (a_k^0 + \beta\xi_k, c_k^0 + \beta\zeta_k) \neq \pm(a_h^0 + \beta\xi_h, c_h^0 + \beta\zeta_h) \ (1 \leq k < h \leq K), \\ (a_k^0 + \beta\xi_k, c_k^0 + \beta\zeta_k) \neq \pm(\xi_j, \zeta_j) \ (1 \leq k \leq K, K+1 \leq j \leq H), \\ (\xi_j, \zeta_j) \neq \pm(\xi_i, \zeta_i) \ (K+1 \leq j < i \leq H), \\ (\xi_j, \zeta_j) \neq \pm(a_k^0, c_k^0) \ (1 \leq k \leq K, K+1 \leq j \leq H), \\ b_k^0 + \beta\eta_k \neq 0 \ (1 \leq k \leq K),\ \sum_{j=K+1}^{H} \eta_j^2 = 1,\ \eta_j \neq 0 \ (K+1 \leq j \leq H), \\ \eta_{K+1} > 0,\ \beta \in \mathbb{R}\} \end{aligned} \tag{16}$$

and $\Pi_H^{**} = \{\omega \in \Pi_H \mid \beta \neq 0\}$, respectively. The multilayer perceptron can be rewritten using this parameterization:

$$\begin{aligned}\psi(x;\omega) &= \sum_{k=1}^{K} (b_k^0 + \beta\eta_k)\, s\big((a_k^0 + \beta\xi_k)x + (c_k^0 + \beta\zeta_k)\big) \\ &\quad + \sum_{j=K+1}^{H} \beta\eta_j\, s(\xi_j x + \zeta_j) + \beta\delta. \end{aligned} \tag{17}$$

It is easy to see that the Π_H^{**} and Θ_H^{**} are diffeomorphic by the transform (15), and $\varphi(x;\theta) = \psi(x;\omega)$ holds for the corresponding $\theta \in \Theta_H^{**}$ and $\omega \in \Pi_H^{**}$. Thus, it suffice to consider $\{\psi(x;\omega) \mid \omega \in \Pi_H\}$, when the maximum likelihood estimation is discussed.

Let $S_H = \{f(x,y;\omega) \mid \omega \in \Pi_H\}$ be a statistical model defined by

$$f(x,y;\omega) = r(y|\psi(x;\omega))q(x). \tag{18}$$

The model S_H consists of probability density functions corresponding to $\varphi_0(x)$ and the functions given by $\varphi(x;\theta)$ for $\theta \in \Theta_H^{**}$. The function $f_0(x,y)$ be a density function defined by $\varphi_0(x)$, that is, $f_0(x,y) = r(y|\varphi_0(x))q(x)$. The model S_H is a locally conic model, if α summarizes $(\xi_1, \ldots, \zeta_H, \delta)$ and $\omega = (\alpha, \beta)$.

Theorem 1. *Let S_H be the statistical model of multilayer perceptrons with H hidden units defined by eqs. (17) and (18), and f_0 be a density function given by (14). Then, under the assumption [NM1], S_H is locally conic at f_0.*

Proof. Let A_0 be a set given by $A_0 = \{\alpha \mid (\alpha, 0)\}$, and $\Pi_H(\alpha)$ by $\Pi_H(\alpha) = \{(\alpha, \beta) \mid \beta \in \mathbb{R}\}$ for $\alpha \in A_0$. We can see $\Pi_H = \cup_{\alpha \in A_0} \Pi_H(\alpha)$, because for all $(\alpha, \beta) \in \Pi_H$, the point $(\alpha, 0)$ is also contained in Π_H by the fact $\theta_0 \in \Theta_K^*$ and $(\xi_j, \zeta_j) \neq \pm(a_k^0, c_k^0)$ for $K+1 \leq j \leq H$. We can also prove that $\psi(x;\omega) = \varphi_0(x)$ for all x if and only if $\omega \in \Pi_{H,0}$. The sufficiency is trivial. For the necessity, because $s(\xi_j x + \zeta_j)$ $(K+1 \leq j \leq H)$ is not contained in the linear hull of the functions $\{1, s(a_k^0 x + c_k^0), s(\xi_i x + \zeta_i), s((a_k^0 + \beta\xi_k)x + (c_k^0 + \beta\zeta_k)) \mid 1 \leq k \leq K, K+1 \leq i \leq H, i \neq j\}$ by the definition of Π_H, the coefficients of $s(\xi_j x + \zeta_j)$ in eq.(17) must be zero to realize $\psi(x;\omega) = \varphi_0(x)$. This implies $\beta = 0$. Thus, 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 $L^2(f_0(x,y)\mu_{\mathbb{R}}\mu_y)$-norm of a tangent vector $\frac{\partial}{\partial\beta}\log f(x,y;(\alpha,0))$. This is essentially determined by the partial derivative:

$$\frac{\partial\psi(x;(\alpha,0))}{\partial\beta} = \sum_{j=K+1}^{H} \eta_j s(\xi_j x + \zeta_j) + \delta$$
$$+ \sum_{k=1}^{K} \eta_k s(a_k^0 x + c_k^0) + \sum_{k=1}^{K} b_k^0 \xi_k s'(a_k^0 x + c_k^0) x + \sum_{k=1}^{K} b_k^0 \zeta_k s'(a_k^0 x + c_k^0). \quad (19)$$

The L^2 norm is calculated as

$$N(\alpha)^2 = \int\int r(y|\varphi_0(x))q(x)\Big\{\frac{\partial r(y|\varphi_0(x))}{\partial u}\frac{\partial\psi(x;(\alpha,0))}{\partial\beta}\Big\}^2 dx d\mu_y$$
$$= \int G(\varphi_0(x))\Big\{\frac{\partial\psi(x;(\alpha,0))}{\partial\beta}\Big\}^2 q(x)dx. \quad (20)$$

Since $\varphi_0(x)$ is bounded, so is $G(\varphi_0(x))$ by the continuity of $G(u)$. From eq.(19), the function $\big\{\frac{\partial}{\partial\beta}\psi(x;(\alpha,0))\big\}^2$ 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 $s'(a_k^0 x+c_k^0)$ $(1 \leq k \leq K, K+1 \leq j \leq H)$ are linearly independent (see [13]), the partial derivative $\frac{\partial}{\partial\beta}\psi(x;(\alpha,0))$ is not constant zero. Hence, the zero points of $\frac{\partial}{\partial\beta}\psi(x;(\alpha,0))$ has no accumulation points, and the probability of the set by Q is 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 $f_0(x,y)$. □

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 $Z_1, Z_2, \ldots, Z_n$ 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. If the maximum likelihood estimator $\hat{\beta}_\alpha$ in S_α exists for each $\alpha \in A_0$, the likelihood ratio of the MLE in S is given by

$$\sup_{\theta\in\Theta} L_n(\theta) = \sup_{\alpha} L_n(\alpha, \hat{\beta}_\alpha). \tag{21}$$

Assume that each submodel S_α satisfies some regularity conditions of the asymptotic normality. A set of conditions, which is essentially from Wald ([16]) and Cramér ([1]), is given as follows[2]. For simplicity, we write each submodel by $\{g(z;\beta)|\beta \in V\}$, neglecting the index α. The parameter set V is an open set in $\mathbb{R}$, and we write $a_0 = \inf\{\beta \mid \beta \in V\} \in \mathbb{R} \cup \{-\infty\}$ and $b_0 = \sup\{\beta \mid \beta \in V\} \in \mathbb{R} \cup \{\infty\}$.

[Conditions on asymptotic normality (AN)]

1. For any $\beta \in V$, the integral $E_{f_0\mu}[|\log g(z;\beta)|]$ is finite.
2. Let $H_+(z;t)$ and $H_-(z;s)$ be functions defined by

$$H_+(z;t) = \sup_{\beta\geq t} \log g(z;\beta) \quad \text{and} \quad H_-(z;s) = \sup_{\beta\leq s} \log g(z;\beta), \tag{22}$$

   respectively. Then,

$$\lim_{t\uparrow b_0} E_{f_0\mu}[H_+(z;t)] < \infty \quad \text{and} \quad \lim_{s\downarrow a_0} E_{f_0\mu}[H_-(z;s)] < \infty. \tag{23}$$

3. There exist Δ_+ and Δ_- such that $\int_{\Delta_\pm} f_0(z)d\mu > 0$ and

$$\lim_{t\uparrow b_0} H_+(z;t) = -\infty \quad \text{for all } z \in \Delta_+, \tag{24}$$

$$\lim_{s\downarrow a_0} H_-(z;s) = -\infty \quad \text{for all } z \in \Delta_-. \tag{25}$$

[2]Another set of conditions is found in van der Vaart ([14], Section 5.3), which is more refined than the famous ones by Cramér ([1]).

4. For all $\beta \in V$,

$$\lim_{\rho \downarrow 0} E_{f_0 \mu}\Big[\sup_{|\beta' - \beta| \leq \rho} \log g(z; \beta') \Big] < \infty. \tag{26}$$

5. The density $g(z;\beta)$ is three-times differentiable on β for all z, and

$$\lim_{\rho \downarrow 0} E_{f_0 \mu}\Big[\sup_{|\beta| \leq \rho} \Big| \frac{\partial^3 \log g(z;\beta)}{\partial \beta^3} \Big| \Big] < \infty. \tag{27}$$

The conditions 1–4 are slight modification of Wald's regularity conditions for the consistency of MLE $\widehat{\beta}_\alpha$ ([16]). The condition 5 assures asymptotic efficiency of $\widehat{\beta}_\alpha$ under the consistency assumption. If each submodel S_α satisfies the conditions [AN], the standard argument using Taylor expansion leads to

$$L_n(\alpha, \hat{\beta}_\alpha) = \frac{1}{2} U_n(\alpha)^2 + o_p(1), \tag{28}$$

where $U_n(\alpha)$ is a random variable defined by

$$U_n(\alpha) = \tfrac{1}{\sqrt{n}} \sum_{i=1}^{n} v_\alpha(Z_i), \tag{29}$$

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

$$v_\alpha(z) = \frac{\partial}{\partial \beta} \log f(z; (\alpha, 0)). \tag{30}$$

The variable $U_n(\alpha)$ 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 an empirical process over α or C, and every marginal distribution on finite points converges to a multidimensional normal distribution. The likelihood ratio of MLE is given by

$$\sup_{\theta \in \Theta} L_n(\theta) = \sup_{\alpha \in A_0} \left\{ \frac{1}{2} U_n(\alpha)^2 + o_p(1) \right\}. \tag{31}$$

Dacunha-Castelle and Gassiat ([2]) discuss the convergence of U_n, assuming the uniform convergence in the asymptotic normality and the empirical process. More precisely, if the higher order term of $o_p(1)$ in eq.(31) is bounded uniformly over α, the term can be eliminated from the supremum;

$$\sup_{\theta \in \Theta} L_n(\theta) = \sup_{\alpha} \left\{ \frac{1}{2} U_n(\alpha)^2 \right\} + o_p(1). \tag{32}$$

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 the supremum of $|W|$ (see Wellner & van der Vaart ([15]) and van der Vaart ([14]) for the detail). Then, we obtain

$$\sup_{\theta\in\Theta} L_n(\theta) = \sup_{\alpha} \frac{1}{2} W^2 + o_p(1). \tag{33}$$

Dacunha-Castelle & Gassiat propose a likelihood ratio test based on the supremum of the Gaussian process W.

Unlike Dacunha-Castelle & Gassiat ([2]), when discussing the stochastic process U_n in eq.(28), this paper will investigate non-uniform cases, in which the simplification in eqs. (32) and (33) does not hold. In non-uniform cases, the behavior of MLE is complex, and even the order of the likelihood ratio can be different from the usual $O_p(1)$, as I mentioned in Section 1.

3.2 Slower convergence in non-uniform cases

The likelihood ratio of MLE can have a larger order than $O_p(1)$, if the function class of the tangent cone is "rich" enough, as the cone in the normal mixture and multilayer perceptrons.

In this subsection, a useful sufficient condition of such an unusually larger order is derived, as an extension of Hartigan's idea ([3]). Note that the marginal distribution of U_n on finite points $v_1, \ldots, v_m$ in C always converges to a multi-dimensional normal distribution with the covariance $E_P[v_i v_j]$. Thus, two components of the limit are independent on condition that their covariance is zero. Suppose we can find an arbitrary number of "almost" uncorrelated 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 2. *Let a statistical model $S = \{f(z;(\alpha,\beta))\}$ be locally conic at $f_0 \in S$, and $C = \{v_\alpha(z) = \frac{\partial}{\partial\beta} f(z;(\alpha,0))\}$ be the basis of the tangent cone. Assume that for each $\alpha \in A_0$ the submodel $\{f(z;\alpha,\beta) \mid \beta\}$ satisfies the conditions of asymptotic normality [AN]. If there exists a sequence $\{v_n\}_{n=1}^{\infty}$ in C such that $v_n \to 0$ in probability, then, for arbitrary $M > 0$, we have*

$$\lim_{n\to\infty} \mathrm{Prob}\Big(\sup_{(\alpha,\beta)} L_n(\alpha,\beta) \le M\Big) = 0. \tag{34}$$

Proof. From Proposition 1 below, for arbitrary $\varepsilon > 0$ and $K \in \mathbb{N}$, there exist $v(\alpha_1), \ldots, v(\alpha_K) \in C$ 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 $W = (W_1, \ldots, W_K)$ be a random vector following the limiting normal distribution of $(U_n(v_{\alpha_1}), \ldots, U_n(v_{\alpha_K}))$, 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 ([17]), we have $(1 + (K-1)\varepsilon)I_K \leq \Sigma \leq (1 - (K+1)\varepsilon)I_K$. Then, for arbitrary $M > 0$, the inequality

$$\begin{aligned} P\big(\max_{1\leq i\leq K} |W_i| \ \leq M\big) &\leq \int_{[-M,M]^K} \frac{1}{\sqrt{(2\pi)^K |\Sigma|}} e^{-\frac{1}{2(1+(K-1)\varepsilon)} W^T W} dW \\ &\leq \frac{(1+(K-1)\varepsilon)^{K/2}}{|\Sigma|^{1/2}} \int_{[-M,M]^K} \frac{1}{(2\pi)^{K/2}} e^{-\frac{1}{2}u^T u} du \\ &\leq \left(\frac{1+(K-1)\varepsilon}{1-(K-1)\varepsilon}\right)^{K/2} \{\Phi(M) - \Phi(-M)\}^K \end{aligned} \tag{35}$$

holds, 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 $\left(\frac{1+(K-1)\varepsilon}{1-(K-1)\varepsilon}\right)^{K/2} < 2$. Then, eq.(35) leads

$$P\big(\max_{1\leq i\leq K} |W_i| \ \leq M\big) < \delta. \tag{36}$$

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

On the covariance of the random variables with bounded L^2 norm, we have the following proposition, which is used in the above proof.

Proposition 1. *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 $v_n \to 0$ in probability. Then, there exists a subsequence $\{v_{n(k)}\}_{k=1}^{\infty}$ that satisfies*

$$E_P|v_{n(k)}v_{n(h)}| < \varepsilon \tag{37}$$

for all different k and h.

This is a direct consequence of the following proposition.

Proposition 2. *Let $(\Omega, \mathcal{B}, P)$ be a probability space, and $Y, X_1, X_2, \ldots$ be random variables. Suppose there exists $K > 0$ such that $\int Y^2 dP \leq K$ and $\int X_n^2 dP \leq K$, and X_n converges to 0 in probability. Then, we have*

$$\lim_{n\to\infty} E|YX_n| = 0. \tag{38}$$

Proof. Let ε be any 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 $n \in \mathbb{N}$, a measurable set A_n is defined by

$$A_n = \{\omega \in \Omega \mid |Y| > \frac{\varepsilon}{3\sqrt{K}} \quad \text{and} \quad |X_n| > \frac{\varepsilon}{3K}|Y|\}. \tag{39}$$

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

Since $A_n^c \subset \{\omega \mid |Y| \leq \frac{\varepsilon}{3\sqrt{K}}\} \cup \{\omega \mid |X_n| \leq \frac{\varepsilon}{3K}|Y|\}$, we obtain for all $n \geq n_0$

$$\begin{aligned}
&\int |YX_n| dP = \int_{A_n} |YX_n| dP + \int_{A_n^c} |YX_n| dP \\
&\leq \Big(\int_{A_n} Y^2 dP\Big)^{1/2} \Big(\int_{A_n} X_n^2 dP\Big)^{1/2} + \int_{\{|Y| \leq \frac{\varepsilon}{3\sqrt{K}}\}} |YX_n| dP + \int_{\{|X_n| \leq \frac{\varepsilon}{3K}|Y|\}} |YX_n| dP \\
&< \frac{\varepsilon}{3\sqrt{K}}\sqrt{K} + \frac{\varepsilon}{3\sqrt{K}} \int |X_n| dP + \frac{\varepsilon}{3K} \int |Y|^2 dP \\
&\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3\sqrt{K}} \cdot \sqrt{K} + \frac{\varepsilon}{3K} \cdot K = \varepsilon
\end{aligned} \tag{40}$$

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

4 Likelihood Ratio of Multilayer Perceptrons

We apply the results in the previous section to the multilayer perceptron model, which is defined by eq.(8). We use the same notations as Section 2.4, giving the true function $\varphi_0(x)$ by eq.(14) and the locally conic parameterization by eq.(17).

We need additional assumptions on the noise model $r(y|u)$ to ensure the asymptotic normality conditions [AN] on the one-dimensional models.

Conditions on noise model (NM2)

1. For any compact set $K \subset \mathbb{R}$, $\sup_{\xi,u\in K} E_{r(y|\xi)}|\log r(y|u)|$ is finite.

2. Let $h_+(y|s)$ and $h_-(y|s)$ be functions defined by

$$h_+(y|s) = \sup_{u\geq s} \log r(y|u) \qquad \text{and} \qquad h_-(y|s) = \sup_{u\leq -s} \log r(y|u), \quad (41)$$

   respectively. For any compact set $K \subset \mathbb{R}$ and $s \in \mathbb{R}$, $\sup_{\xi\in K} E_{r(y|\xi)}[h_\pm(y|s)]$ is finite.

3. For an arbitrary compact set $K \subset \mathbb{R}$, there exist $\Delta_+, \Delta_- \subset \mathcal{Y}$ and $B > 0$ such that

$$\lim_{s\to\infty} h_+(y|s) = -\infty \quad \text{for all } y \in \Delta_+, \quad (42)$$

$$\lim_{s\to\infty} h_-(y|s) = -\infty \quad \text{for all } y \in \Delta_-, \quad (43)$$

   and

$$\int_{\Delta_\pm} r(y|\xi)dy \geq B \quad \text{for } {}^\forall\xi \in K. \quad (44)$$

4. For any compact set $K \subset \mathbb{R}$,

$$\lim_{\rho\downarrow 0} \sup_{\substack{\xi\in K\\ u\in K}} E_{r(y|\xi)}\Big[\sup_{|u'-u|\leq\rho} \log r(y|u')\Big] < \infty. \quad (45)$$

5. The density $r(y|u)$ is three-times differentiable on u for all $y \in \mathcal{Y}$, and for any compact set $K \subset \mathbb{R}$,

$$\lim_{\rho\downarrow 0} \sup_{\xi\in K} E_{r(y|\xi)}\Big[\sup_{|\xi'-\xi|\leq\rho} \Big|\frac{\partial^3 \log r(y|\xi')}{\partial^3 u}\Big|\Big] < \infty. \quad (46)$$

The above conditions are satisfied by many important noise models. In the case of the Gaussian noise model and binary output model, they can be checked easily. In fact, the conditions 1, 4, and 5 are easy. On the conditions 2 and 3, stronger conditions will be checked in Section 4.

The next lemma shows that the conditions [NM2] implies the asymptotic normality [AN] in some type of submodel in S_H.

Lemma 1. *Let $w_0(x)$ be a bounded function, $w(x)$ be a positive, bounded function, and $r(y|u)$ be a density function on $\mathcal{Y}$ which satisfies [NM1] and [NM2]. Then, the statistical model $\{g(z;\beta) \mid \beta \in \mathbb{R}\}$, which is defined by $g(z;\beta) = r(y|w_0(x) + \beta w(x))q(x)$, satisfies the conditions [AN].*

Proof. From [NM2]-1 and boundedness of $w(x)$ and $w_0(x)$, for each β there is $A > 0$ such that $E_{r(y|w_0(x))}|\log r(y|w_0(x) + \beta w(x))| \leq A$ for all $x \in \mathbb{R}$. The fact $E_Q|\log q(x)| < \infty$ implies the condition [AN]-1.

Since $H_+(z;t) = h_+(y|w_0(x) + tw(x)) + \log q(x)$ and for any t there exists s_0 such that $w_0(x) + tw(x) \geq s_0$ for all x , we have $E_{f_0\mu}[H_+(z;t)] \leq E_Q[E_{r(y|w_0(x))}[h_+(y|s_0)] + \log q(x)]$. The compactness of the range of $w_0(x)$ and the condition [NM2]-2 show the first assertion of [AN]-2. The second one is similar.

We will show only on H_+ for the assumption [AS]-3, because the proof on H_- is exactly the same. There exists $M > 0$ such that $|w_0(x)| \leq M$. Take $\Delta_+ \subset \mathcal{Y}$ and $B > 0$ in the assumption [NM2]-3 for a compact set $[-M, M]$. Then, for any $z \in \mathcal{X} \times \Delta_+$, we have $\lim_{t\to\infty} H_+(z;t) = \lim_{t\to\infty} h_+(y|w_0(x) + tw(x)) + \log q(x) = -\infty$, and $\int_{\mathcal{X}\times\Delta_+} f_0(z)d\mu = E_Q[\int_{\Delta_+} r(y|w_0(x))] \geq B$.

From [NM2]-4 and the boundedness of $w(x)$, for any β there exists $\rho_0 > 0$ and C such that $E_{r(y|w_0(x))}[\sup_{|\beta'-\beta|\leq\rho} \log r(y|w_0(x) + \beta' w(x))] \leq C$ holds for all $\rho \in (0, \rho_0]$ and $x \in \mathbb{R}$. This shows the condition [AN]-4. By a similar argument, [NM]-5 implies [AN]-5. □

Theorem 3. *Assume that the model is the multilayer perceptron model (8) with H hidden units, and the true function is given by a network with K hidden units for $K < H$. Under the assumptions [NM1] and [NM2] on the noise model $r(y|u)$, we have for arbitrary $M > 0$,*

$$\lim_{n\to\infty} \mathrm{Prob}\Big(\sup_{\theta} L_n(\theta) \leq M\Big) = 0. \tag{47}$$

Remark. This theorem means that the order of the likelihood ratio of MLE is strictly larger than $O_p(1)$.

Proof. For the lower bound, it suffice to consider a submodel in the locally conic parameterization eq.(17). Let $\sigma(x;\xi,h)$ be a bounded, monotone decreasing function given by

$$\sigma(x;\xi,h) = \frac{1}{2}\{1 + s(-\frac{1}{2}\xi(x-h))\} = \frac{1}{1+\exp\{\xi(x-h)\}}, \tag{48}$$

and $\{g(z;t,c)\}$ be a submodel defined by

$$g(z;t,c,\beta) = r(y|\varphi_0(x) + \beta w(x;t,c))q(x), \tag{49}$$

where

$$w(x;t,c) = \frac{1}{\sqrt{B(t,c)}}\sigma(x;c^2,t+\tfrac{1}{c}), \tag{50}$$

and $B(t,c)$ is a normalizing constant of $L^2(f_0\mu)$ norm given by

$$B(t,c) = \int G(\varphi_0(x))\sigma(x;c^2,t+\tfrac{1}{c})^2 dQ(x). \tag{51}$$

Because $\varphi_0(x)$ and $w(x;t,c)$ are bounded functions, from Theorem 2 and Lemma 1, we have only to show there is a sequence in the basis of the tangent cone C, which converges to zero in probability. The set C consists of the functions

$$v(x,y;t,c) = \frac{1}{\sqrt{B(t,c)}}\frac{\partial \log r(y|\varphi_0(x))}{\partial u}\sigma(x;c^2,t+\tfrac{1}{c}). \tag{52}$$

Let a be a positive number that satisfies $G(\varphi_0(x)) \geq a$ for all $x \in \mathbb{R}$. Such a exists because of the continuity of $G(u)$ and the boundedness of φ_0. Let $F_Q(t)$ be a distribution function of the input probability Q. From the assumption that Q is absolute continuous with respect to the Lebesgue measure, F_Q is continuous on $\mathbb{R}$. If we define $t_0 = \inf\{t \in \mathbb{R} \mid F_Q(t) > 0\} \in \mathbb{R} \cup \{-\infty\}$, we have $F_Q(t) > 0$ for all $t > t_0$, and $\lim_{t \downarrow t_0} F_Q(t) = 0$.

Since $\sigma(x;c^2,t+\frac{1}{c})$ is bounded and converges to $\chi_{(-\infty,t]}(x)$ at every x for $c \to +\infty$, by Lebesgue's dominated convergence theorem, we have $\lim_{c\to\infty} B(t,c) = \int_{-\infty}^{t} G(\varphi_0(x))dQ(x) \geq aF_Q(t)$. Hence, for each t we can find $c_t^{(1)}$ such that $\sqrt{B(t,c)} \geq \frac{1}{2}\sqrt{aF_Q(t)}$ for all $c \geq c_t^{(1)}$.

For any $t > t_0$ and $\delta > 0$, there exists $c_t^{(2)}(\delta) > 0$ such that $\sigma(x;c^2,t+\frac{1}{c}) \leq F_Q(t)$ for all $x \geq t+\delta$ and $c \geq c_t^{(2)}(\delta)$. Then, if a sequence (t_n,δ_n,c_n) is chosen so that $t_n \downarrow t_0$, $\delta_n \downarrow 0$, and $c_n \geq \max\{c_{t_n}^{(1)}, c_{t_n}^{(2)}(\delta_n)\}$, the inequality

$$|v(x,y;t_n,c_n)| \leq \frac{2}{\sqrt{a}}\Big|\frac{\partial \log r(y|\varphi_0(x))}{\partial u}\Big|\sqrt{F_Q(t_n)} \tag{53}$$

holds for all $x \geq t_n+\delta_n$ and y. Because $F_Q(t_n) \to 0$ and $t_n+\delta_n \downarrow t_0$ for $n \to \infty$, the sequence $v(x,y;t_n,c_n)$ converges to zero for all $x > t_0$ and y, which means almost everywhere convergence. □

If $K \leq H-2$, a different type of sequence can work for the proof of Theorem 3. Let $\mathcal{W} = \{w(x;\xi,h,t)\}$ be a family of functions defined by

$$w(x;\xi,h,t) = \frac{1}{\sqrt{A(\xi,h,t)}}\frac{1}{2}\{s(\xi(x-t+h)) - s(\xi(x-t-h))\}, \tag{54}$$

where $A(\xi, h, t)$ is a normalization constant of $L^2(f_0\mu)$ norm given by

$$A(\xi,h,t) = E_{f_0\mu}\Big[\Big(\frac{\partial \log r(y|\varphi_0(x))}{\partial u}\frac{s(\xi(x-t+h)) - s(\xi(x-t-h))}{2}\Big)^2\Big]$$
$$= E_Q\big[G(\varphi_0(x))\tfrac{1}{4}\{s(\xi(x-t+h)) - s(\xi(x-t-h))\}^2\big]. \qquad (55)$$

A subfamily of the multilayer perceptron in the locally conic parameterization is defined by

$$\psi(x;\xi,h,t,\beta) = \varphi_0(x) + \beta w(x;\xi,h,t). \qquad (56)$$

This is obtained by setting $\eta_i = \xi_i = \zeta_i = \delta = 0$ ($1 \le i \le k$ and $i \ge K+3$), $\xi_{K+1} = \xi_{K+2} = \xi$, $\zeta_{K+1} = -\zeta_{K+2} = h$, and $\eta_{K+1} = \eta_{K+2} = \frac{1}{2}$ in eq. (17). The basis of the tangent cone of the submodel $\{r(y|\psi(x;\xi,h,t,\beta))q(x)\}$ consists of the functions of the form

$$v(z;\xi,h,t) = \frac{\partial \log r(y|\varphi_0(x))}{\partial u} w(x;\xi,h,t). \qquad (57)$$

From the fact that $G(u)$ is positive and continuous, and that φ_0 is bounded, there exist $a, b > 0$ such that $a \le G(\varphi_0(x)) \le b$ for all $x \in \mathbb{R}$. For arbitrary $h > 0$ we can find $\delta(h) > 0$ so that for any $\xi \ge \delta(h)$, $\frac{1}{2}\{s(\xi(x+h)) - s(\xi(x-h))\}$ is larger than $\frac{1}{2}$ on $x \in [-\frac{h}{2}, \frac{h}{2}]$ and less than h on $x \notin [-\frac{3}{2}h, \frac{3}{2}h]$. Let $h_n > 0$ be a decreasing sequence which converges to zero. If ξ_n is taken so that $\xi_n \ge \delta(h_n)$, the normalization constant satisfies

$$A(\xi_n, h_n, 0) \ge \int_{-\frac{1}{2}h_n}^{\frac{1}{2}h_n} G(\varphi_0(x))\big(\tfrac{1}{2}\big)^2 q(x)dx \ge \frac{a}{4}h_n. \qquad (58)$$

Thus, for all x with $|x| \ge \frac{3}{2}h_n$, we have

$$|v(z;\xi_n,h_n,0)| = \Big|\frac{\partial \log r(y|\varphi_0(x))}{\partial u}\Big|\frac{1}{\sqrt{A(\xi_n,h_n,0)}}\frac{1}{2}\{s(\xi(x+h)) - s(\xi(x-h))\}$$
$$\le \frac{2\sqrt{h_n}}{\sqrt{a}}\Big|\frac{\partial \log r(y|\varphi_0(x))}{\partial u}\Big|. \qquad (59)$$

For all $x \ne 0$, the sequence $v(z;\xi_n,h_n,0)$ converges to zero from the fact $h_n \downarrow 0$,. This means almost everywhere convergence, since Q is absolutely continuous with respect to the Lebesgue measure.

The next lemma on the functional space $\mathcal{W}$ will be used in Corollary 1 after Theorem 4.

Lemma 2. *For a closed interval I, a non-negative value $M(I)$ is defined by*

$$M(I) = E_{f_0\mu}\Big[\Big(\frac{\partial \log r(y|\varphi_0(x))}{\partial u}\Big)^2 \chi_I(x)\Big] = \int_I G(\varphi_0(x))q(x)dx, \qquad (60)$$

and a function $u_I(z)$ by

$$u_I(z) = \frac{1}{\sqrt{M(I)}} \frac{\partial \log r(y|\varphi_0(x))}{\partial u} \chi_I(x), \qquad (61)$$

if $M(I) > 0$. Let $\mathcal{W}$ be defined as above. Under the assumptions [NM1], there exist $a, b > 0$ such that for an arbitrary $\varepsilon > 0$ and closed interval I with $M(I) > 0$, we can find a function $w(x; \xi, h, t) \in \mathcal{W}$, which satisfies (i) $0 < w(x; \xi, h, t) \leq \frac{a}{\sqrt{M(I)}}$ for all $x \in \mathbb{R}$, (ii) $w(x; \xi, h, t) \geq \frac{b}{\sqrt{M(I)}}$ for all $x \in I$, and (iii) $\|v(z; \xi, h, t) - u_I(z)\|_{L^2(f_0\mu)} \leq \varepsilon$, where $v(z; \xi, h, t)$ is given by eq. (57).

Proof. For notational simplicity, a proof will be given in the case of $I = [-c, c]$. The extension to the general case is straightforward. Write $w(x; \xi, h)$ and $v(z; \xi, h)$ for $w(x; \xi, h, 0)$ and $v(z; \xi, h, 0)$, respectively, and use $\sigma(x; \xi, h) = \frac{1}{2}\{s(\xi_n(x + h_n)) - s(\xi_n(x - h_n))\}$ for abbreviation.

In a similar way to the argument before the lemma, there exist sequences $h_n \downarrow c$ and $\xi_n \to \infty$ such that

[1] $\sigma(x; \xi_n, h_n) \leq 2$ for all $x \in \mathbb{R}$,

[2] $|\sigma(x; \xi_n, h_n) - \chi_I(x)| \to 0$ for all $x \in \mathbb{R}$, and

[3] $\sigma(x; \xi_n, h_n) \geq \frac{1}{2}$ for all $x \in I$.

From [1], [2] and the boundedness of $G(\varphi_0(x))$, by Lebesgue's dominated convergence theorem, we obtain

$$\Big\| \frac{\partial \log r(y|\varphi_0(x))}{\partial u} \sigma(x; \xi_n, h_n) - \frac{\partial \log r(y|\varphi_0(x))}{\partial u} \chi_I(x) \Big\|_{L^2(f_0\mu)} \to 0, \quad (62)$$

as n goes to infinity. Eq.(62) means also $A(\xi_n, h_n) \to M(I) > 0$. Then, a simple argument shows the assertion (iii). From eq. (62), there exists $n_0 \in \mathbb{N}$ such that $\frac{1}{2}M(I) \leq A(\xi_n, h_n) \leq 2M(I)$ for all $n \geq n_0$. Combining this with [1] and [3], we obtain (i) and (ii) in the assertion, by taking $a = 2\sqrt{2}$ and $b = \frac{1}{2\sqrt{2}}$. Note that a and b are taken so that they do not depend on I. □

In the case $K \leq H - 2$, we can derive a better lower bound of the likelihood ratio, by counting a number of almost independent random variables in C. However, we need to strengthen the assumptions on the noise model

$r(y|u)$. In listing the conditions, the most concise ones are not sought for, but the ones that can be easily checked are intended. Indeed, the following assumptions are verified easily for the Gaussian noise model and the binary output model, as shown later. In the following conditions, $h_{\pm}(y|s)$ is the same as in [NM2].

[Conditions on noise model (NM3)]

1. For an arbitrary compact set $K \subset \mathbb{R}$, there exists a non-negative function $\tau(s)$ defined on $[0, \infty)$ such that positive numbers A_i, δ_i $(i = 1, 2)$ and T_0 exist so that

$$\tau(s) \geq A_1 s^{\delta_1} \quad \text{for } 0 \leq s \leq T_0, \qquad \tau(s) \geq A_2 s^{\delta_2} \quad \text{for } s > T_0, \tag{63}$$

   and a lower bound of the KL-divergence is given by

$$E_{r(u|\xi)}\Big[\log \frac{r(y|\xi)}{r(y|u)}\Big] \geq \tau(|u - \xi|), \tag{64}$$

   for all $\xi \in K$ and $u \in \mathbb{R}$.

2. For an arbitrary compact set $K \subset \mathbb{R}$, there exist $\Delta_+, \Delta_- \subset \mathcal{Y}$ and $\gamma, B > 0$ such that

$$\limsup_{s \to \infty} \frac{h_+(y|s)}{s^\gamma} < 0 \quad \text{for all } y \in \Delta_+, \tag{65}$$

$$\limsup_{s \to \infty} \frac{h_-(y|s)}{s^\gamma} < 0 \quad \text{for all } y \in \Delta_-, \tag{66}$$

   and

$$\int_{\Delta_\pm} r(y|\xi) dy \geq B \quad \text{for } {}^\forall \xi \in K. \tag{67}$$

3. There exist a continuous function $\ell_1(\xi)$ and $\lambda > 0$ such that for all $s \geq 1$

$$E_{r(y|\xi)} |h_\pm(y|s)|^2 \leq \ell_1(\xi) s^\lambda. \tag{68}$$

4. There exist a continuous function $\ell_2(\xi)$ and $\nu > 0$ such that for arbitrary $R \geq 1$

$$E_{r(y|\xi)}\Big[\sup_{|u| \leq R} \Big| \frac{\partial \log r(y|u)}{\partial u} \Big|^2\Big] \leq \ell_2(\xi) R^\nu. \tag{69}$$

5. For any compact set $K \subset \mathbb{R}$,

$$\sup_{\xi,u\in K} E_{r(y|\xi)}\Big[\Big|\frac{\partial^2 \log r(y|u)}{\partial u^2}\Big|^2\Big] < \infty \quad \text{and} \quad \sup_{u\in K} E_{r(y|u)}\Big[\Big|\frac{\partial \log r(y|u)}{\partial u}\Big|^3\Big] < \infty. \tag{70}$$

6. For any compact set $K \subset \mathbb{R}$,

$$\limsup_{\rho\downarrow 0} \sup_{\xi\in K} E_{r(y|\xi)}\Big[\sup_{|\xi'-\xi|\le\rho}\Big|\frac{\partial^3 \log r(y|\xi')}{\partial u^3}\Big|^2\Big] < \infty. \tag{71}$$

The conditions [NM3] are satisfied by the Gaussian noise model and the binary output model. Consider the Gaussian distribution with variance one: $r(y|u) = \frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}(y-u)^2\}$, for simplicity. The conditions 5 and 6 are trivial. Because the KL-divergence is equal to $\frac{1}{2}(u-\xi)^2$, the condition 1 is satisfied. We have $h_+(y|s) = -\frac{1}{2}(y-s)^s - \log\sqrt{2\pi}$ for $y < s$ and $h_+(y|s) = -\log\sqrt{2\pi}$ for $y \ge s$, and a similar form for $h_-(y|s)$. Then, the condition 2 can be verified for any bounded interval $\Delta_\pm$ and $\gamma = 2$, and the condition 3 is trivial. Because $\frac{\partial \log r(y|u)}{\partial u} = y - u$, for any $R > 0$, we see $\sup_{|u|\le R}\big|\frac{\partial \log r(y|u)}{\partial u}\big|^2$ is $(y+R)^2$ for $y \ge 0$ and $(y-R)^2$ for $y < 0$. Hence, the condition 4 is satisfied for $\nu = 2$.

For the binary output model $r(y|u) = \frac{e^{yu}}{1+e^u}$, the conditions 5 and 6 are trivial. The KL-divergence is $E_{r(y|\xi)}[\log r(y|\xi) - \log r(y|u)] = \frac{1}{1+e^\xi}(u-\xi) + \log(1+e^{-u}) - \log(1+e^{-\xi}) = \frac{1}{1+e^{-\xi}}(\xi-u) + \log(1+e^u) - \log(1+e^\xi)$. For any $C > 0$ and $\xi \in [-C, C]$, we have $E_{r(y|\xi)}[\log r(y|\xi) - \log r(y|u)] > \frac{1}{1+e^C}|u-\xi| - \log(1+e^C)$. Then, for any u with $|u-\xi| \ge 2(1+e^C)\log(1+e^C)$, a lower bound $E_{r(y|\xi)}[\log r(y|\xi) - \log r(y|u)] > \frac{1}{2(1+e^C)}|u-\xi|$ is obtained. By Taylor expansion of KL-divergence, there exists $a > 0$ such that $E_{r(y|\xi)}[\log r(y|\xi) - \log r(y|u)] \ge a(u-\xi)^2$ for all $\xi \in [-C, C]$ and $|u-\xi| \le 2(1+e^C)\log(1+e^C)$. Thus, we can choose $T_0 = 2(1+e^C)\log(1+e^C)$, and define $\tau(s)$ by $\frac{1}{2(1+e^C)}s$ for $0 \le s \le T_0$ and $a(u-\xi)^2$ for $s > T_0$. This shows the condition 1. Since $h_+(y|s) = -(1-y)\log(1+e^s)$ and $h_-(y|s) = y\log(1+e^s)$, The condition 3 is straightforward. For the condition 2, choose $\Delta_+ = \{0\}$ and $\Delta_- = \{1\}$. Then, $\lim_{s\to\infty}\frac{h_\pm(y|s)}{s} = -1$ for all $y \in \Delta_\pm$. The condition 4 is trivial, since $\frac{\partial \log r(y|u)}{\partial u} = -\log(1+e^{-u})$ is bounded.

Theorem 4. *Let $r(y|u)$ be a conditional probability density function of $y \in \mathcal{Y}$ given $u \in \mathbb{R}$, which satisfies the conditions [NM1], [NM2], and [NM3],*

$\varphi_0(x)$ be a bounded function on $\mathbb{R}$, and $f_0(z)$ is a density function with respect to the measure $\mu = \mu_{\mathbb{R}} \times \mu_y$, which is defined by $f_0(x,y) = r(y|\varphi_0(x))q(x)$. For a closed interval I, a non-negative value $M(I)$ is defined by

$$M(I) = \int\int \Big(\frac{\partial \log r(y|\varphi_0(x))}{\partial u}\Big)^2 \chi_I(x) r(y|\varphi_0(x))q(x)d\mu_y dx, \tag{72}$$

and a function $u_I(z)$ is defined by

$$u_I(z) = \frac{1}{\sqrt{M(I)}} \frac{\partial \log r(y|\varphi_0(x))}{\partial u} \chi_I(x), \tag{73}$$

if $M(I) > 0$, where $z = (x,y)$. Suppose that $\mathcal{W} = \{w(x;\alpha) \mid \alpha \in A_0\}$ is a family of functions such that the the function

$$v(z;\alpha) = \frac{\partial \log r(y|\varphi_0(x))}{\partial u} w(x;\alpha) \tag{74}$$

satisfies $\|v(z;\alpha)\|_{L^2(f_0\mu)} = 1$ for all $\alpha \in A_0$. It is further assumed that there exist $a, b > 0$ such that for any $\varepsilon > 0$ and closed interval I with positive $M(I)$ we can find $w(x;\alpha) \in \mathcal{W}$ which satisfies
(i) $0 < w(x;\alpha) \leq \frac{a}{\sqrt{M(I)}}$ for all $x \in \mathbb{R}$,
(ii) $w(x;\alpha) \geq \frac{b}{\sqrt{M(I)}}$ for all $x \in I$, and
(iii) $\int\int |v(z;\alpha) - u_I(z)|^2 r(y|\varphi_0(x))q(x)d\mu_y dx < \varepsilon$.

Then, for the locally conic model $f(z;\alpha,\beta) = r(y|\varphi_0(x) + \beta w(x;\alpha))q(x)$ ($\alpha \in A_0$ and $\beta \in \mathbb{R}$), there exists $\delta > 0$ such that, given i.i.d. sample from $f_0\mu$, we have

$$\liminf_{n\to\infty} \mathrm{Prob}\Big(\frac{\sup_{\alpha,\beta} L_n(\alpha,\beta)}{\log n} \geq \delta\Big) \quad > \quad 0. \tag{75}$$

Remark. The above theorem asserts that the order of the likelihood ratio is at least $O_p(\log n)$.

From this theorem and Lemma 2, the following result on multilayer perceptrons is obtained.

Corollary 1. *Suppose that the model is the multilayer perceptron with H hidden units, and the true function is given by a network with K hidden units for $K \leq H-2$. Then, under the conditions [NM1], [NM2] and [NM3], there exists $\delta > 0$ such that*

$$\liminf_{n\to\infty} \mathrm{Prob}\Big(\frac{\sup_{\theta} L_n(\theta)}{\log n} \geq \delta\Big) \quad > \quad 0. \tag{76}$$

Proof of Theorem 4. From [NM1]-3 and the boundedness of $\varphi_0(x)$, the value $M(\mathbb{R})$ is positive and finite. Fix a positive number K such that $M([-K,K]) = \frac{M(\mathbb{R})}{2}$. Such K exists, since Q is absolutely continuous with respect to the Lebesgue measure. For an arbitrary $m \in \mathbb{N}$, we can obtain a partition $\{I_k^{[m]} \mid k = 1, \ldots, m\}$ of $[-K,K]$ such that $I_k^{[m]}$'s are closed intervals with disjoint interiors, and $M(I_k^{[m]}) = \frac{M(\mathbb{R})}{2m}$ for all k. For each k ($1 \leq k \leq m$, a function $u_k^{[m]}(z)$ is defined by

$$u_k^{[m]}(z) = \frac{\partial}{\partial \beta} \log r\Big(y \,\Big|\varphi_0(x) + \beta \frac{1}{\sqrt{M(I_k^{[m]})}} \chi_{I_k}(x)\Big)\Big|_{\beta=0} = \sqrt{\frac{2m}{M(\mathbb{R})}} \frac{\partial \log r(y|\varphi_0(x))}{\partial u} \chi_{I_k^{[m]}}(x). \tag{77}$$

This is a tangent vector of the locally conic one-dimensional model $r\Big(y\Big|\varphi_0(x) + \beta \frac{1}{\sqrt{M(I_k^{[m]})}} \chi_{I_k^{[m]}}(x)\Big) q(x)$ at the origin. Note that the functions $u_k^{[m]}(z)$ are uncorrelated under the probability $f_0\mu$.

Let $H_3(x)$ be a function defined by $H_3(x) = E_{f_0\mu}\Big|\frac{\partial \log r(y|\varphi_0(x))}{\partial u}\Big|^3$. By the assumptions [NM1]-3, [NM3]-5, and the boundedness of $\varphi_0(x)$, there exists $B > 0$ such that $H_3(x) \leq BG(\varphi_0(x))$ for all $x \in [-K,K]$. Then, we obtain

$$\begin{aligned} E_{f_0\mu}|u_k^{[m]}(z)|^3 &= \frac{1}{M(I_k^{[m]})^{3/2}} \int H_3(x) \chi_{I_k^{[m]}} q(x) dx \\ &\leq \frac{B}{M(I_k^{[m]})^{3/2}} \int G(\varphi_0(x)) \chi_{I_k^{[m]}} q(x) dx = \frac{\sqrt{2}B}{\sqrt{M(\mathbb{R})}} \sqrt{m}. \end{aligned} \tag{78}$$

Let P_n and Q_m be the probability distribution of the m-dimensional random vector $\big(\frac{1}{\sqrt{n}} \sum_{i=1}^n u_1^{[m]}(Z_i), \ldots, \frac{1}{\sqrt{n}} \sum_{i=1}^n u_m^{[m]}(Z_i)\big)$, and of the m-dimensional normal distribution $N(0, I_m)$, respectively. Let $\mathcal{D}$ denote the family of all the convex measurable sets on $\mathbb{R}^m$. The Berry-Esseen-type inequality ([18]) gives

$$\sup_{\Delta \in \mathcal{D}} |P_n(\Delta) - Q_m(\Delta)| \leq \frac{Lm^4}{\sqrt{n}} \sum_{1 \leq k \leq m} E_{f_0\mu}|u_k^{[m]}(Z)|^3, \tag{79}$$

where L is a universal constant. From eqs.(78) and (79), choosing $\Delta = [-\nu\sqrt{\log m}, \nu\sqrt{\log m}]^m$, we have for all n and m

$$\begin{aligned} \Big| \mathrm{Prob}\Big(\max_{1 \leq k \leq m} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n u_k^{[m]}(Z_i) \Big| > \nu\sqrt{\log m} \Big) \\ - \mathrm{Prob}\Big(\max_{1 \leq k \leq m} |V_k| > \nu\sqrt{\log m} \Big) \Big| \quad &\leq \quad C' \frac{m^{11/2}}{\sqrt{n}}, \end{aligned} \tag{80}$$

where V_k $(1 \leq k \leq m)$ are i.i.d sample from the standard normal distribution, and C' is a constant independent of n and m. Let $[x]$ denote the largest integer that is not larger than x. If we set $m = [n^\gamma]$ for $0 < \gamma < \frac{1}{11}$, the right hand side of eq.(80) converges to zero as n goes to infinity. From the extreme value theory, the probability of the event $\{\max_{1\leq k\leq m} |V_k| > \nu\sqrt{\log m}\}$ converges to 1 for $0 < \nu < \sqrt{2}$. Thus, for such ν and arbitrary $\varepsilon > 0$, we have

$$\mathrm{Prob}\Big(\max_{1\leq k\leq m}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} u_k^{[m]}(Z_i)\Big|^2 > \nu^2\gamma\log n\Big) \quad > 1-\varepsilon, \tag{81}$$

for sufficiently large n.

By the assumptions on $\mathcal{W}$, for arbitrary $\varepsilon, \delta > 0$, $m \in \mathbb{N}$, and k $(1 \leq k \leq m)$, there exists $w_k^{[m]} \in \mathcal{W}$ such that (i) $0 < w_k^{[m]}(x) \leq \tilde{a}\sqrt{m}$, (ii) $w_k^{[m]}(x) \geq \tilde{b}\sqrt{m}$ on I_k, and (iii) $E_{f_0\mu}\big|v_k^{[m]}(z) - u_k^{[m]}(z)\big|^2 < \frac{\varepsilon\delta^2}{m}$, where $v_k^{[m]}(z)$ is a function defined by eq.(74) for $w_k^{[m]}(x)$, and $\tilde{a}, \tilde{b}$ are positive constants independent of ε, m and k. Then, using Chebyshev's inequality, we obtain

$$\begin{aligned}
&\mathrm{Prob}\left(\left|\max_{1\leq k\leq m}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} u_k^{[m]}(Z_i)\Big| - \max_{i\leq k\leq m}\Big|\frac{1}{\sqrt{n}}v_k^{[m]}(Z_i)\Big|\right| \geq \delta\right)\\
&\leq \mathrm{Prob}\Big(1 \leq {}^{\exists}k \leq m, \quad \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} u_k^{[m]}(Z_i) - \frac{1}{\sqrt{n}}v_k^{[m]}(Z_i)\Big| \geq \delta\Big)\\
&\leq m\,\mathrm{Prob}\Big(\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} u_k^{[m]}(Z_i) - \frac{1}{\sqrt{n}}v_k^{[m]}(Z_i)\Big| \geq \delta\Big)\\
&\leq m\frac{E_{f_0\mu}|u_k^{[m]}(z) - v_k^{[m]}(z)|^2}{\delta^2} < \varepsilon.
\end{aligned} \tag{82}$$

Combining eqs.(81) and (82), there exists $\gamma' > 0$ such that

$$\mathrm{Prob}\left(\max_{1\leq k\leq m}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} v_k^{[m]}(Z_i)\Big|^2 > \gamma'\log n\right) > 1-2\varepsilon \tag{83}$$

holds for sufficiently large n.

Since $M(I) = \int G(\varphi_0(x))q(x)dx$, by the assumption [NM1]-3 and the boundedness of $\varphi_0(x)$, there exist $c, d > 0$ such that $\frac{c}{m} \leq Q(I_k^{[m]}) \leq \frac{d}{m}$ holds for all m and k $(1 \leq k \leq m)$. From this fact and the choice of $w_k^{[m]}$, Lemma 3 in Appendix asserts that there exists $\gamma_1 > 0$ such that for all positive γ

satisfying $0 < \gamma < \gamma_1$ and $m = [n^\gamma]$, the following asymptotic expansion of the likelihood ratio holds;

$$\max_{1\leq k\leq m} \sup_{\beta} \sum_{i=1}^{n} \log \frac{f_k^{[m]}(Z_i;\beta)}{f_0(Z_i)} = \Big\{ \max_{1\leq k\leq m} \frac{1}{2} \Big(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_k^{[m]}(Z_i) \Big)^2 \Big\} (1 + o_p(1)), \tag{84}$$

where $f_k^{[m]}(z;\beta) = r(y|\varphi_0(x) + \beta w_k^{[m]}(x))q(x)$. The proof is completed by combination of eqs.(83) and (84). □

The order $O_p(\log n)$ has been formerly obtained by Hagiwara et al. ([4]). However, they consider only the least square loss function, and use its special property. The approach in this paper extends their results. The above theorem can be applied to various noise models, including binary output models.

As shown 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. In fact, Hayasaka et al. ([19]) show that, if the network model has one hidden unit of step function, and the true function is constant zero, the likelihood ratio of MLE has the order of $O_p(\log\log n)$, under the least square loss function. This is essentially the same as the result of a change point problem ([7]).

5 Conclusion

An approach to investigate the behavior of MLE has been discussed on condition that the true parameter is unidentifiable. Following the discussion of Dacunha-Castelle and Gassiat ([2]), this paper has formulated the likelihood ratio of MLE by the supremum of an empirical process, which converges to the standard normal distribution marginally. Unlike Dacunha-Castelle and Gassiat ([2]), which concentrates on uniform convergence cases, non-uniform cases have been the main concern of this paper, and a useful sufficient condition of an unusually larger order of the likelihood ratio has been derived. These results have been applied to neural network models, and $O_p(\log n)$ lower bound of the likelihood ratio has been obtained, under the assumption that there are at least two redundant hidden units to realize the true function.

Acknowledgements

I thank Dr. Kano in Osaka University, Dr. Kuriki in the Institute of Statistical Mathematics, Dr. Hagiwara in Mie University, and Dr. 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] Veres, S. (1987) Asymptotic distributions of likelihood ratios for overparameterized ARMA processes. *Journal of Time Series Analysis*, **8**(3), 345–357.

[7] Csörgö, M. and Horváth, L. (1996) *Limit Theorems in Change-Point Analysis*. John Wiley & Sons.

[8] 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.

[9] 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.

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

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

[12] Fukumizu, K. and Amari, S. (2000) Local Minima and Plateaus in Hierarchical Structures of Multilayer Perceptrons. *Neural Networks*, **13**(3), 317–327.

[13] Fukumizu, K. (1996) A regularity condition of the information matrix of a multilayer perceptron network. *Neural Networks*, **9**(5), 871–879.

[14] Van der Vaart, A.W. (1998) *Asymptotic Statistics.* Cambridge University Press.

[15] Van der Vaart, A.W. & Wellner, J.A. (1996). *Weak convergence and empirical processes.* Springer Verlag.

[16] Wald, A. (1949) Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, **20**, 595–601.

[17] Horn, R. & Johnson, C. (1990). *Matrix Analysis.* Cambridge University Press.

[18] Sazonov, V.V. (1968). On the multi-dimensional central limit theorem. *Sankhya, Ser.A*, **30**, 181–204.

[19] Hayasaka, T., Toda, N., Usui, S., & Hagiwara K. (1996). On the least square error and prediction square error of function representation with discrete variable basis. it Proc. of Neural Networks for Signal Processing VI, 72–81.

A Lemmas used in the proof of Theorem 5

Lemma 3. *Let $\mathcal{Y}$ be a subset of $\mathbb{R}$, $r(y|\xi)$ be a probability density function on a measure space $(\mathcal{Y}, \mathcal{B}_y, \mu_y)$ with one-dimensional parameter ξ, which satisfies the assumptions [NM1], [NM2], and [NM3], and $Q = q(x)dx$ be*

a probability on $\mathcal{X} = \mathbb{R}$, which is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$ and $E_Q|\log q(x)| < \infty$. We have a bounded function $\varphi_0(x)$, and i.i.d. random variables $(X_1, Y_1), (X_2, Y_2), \dots$ with probability $r(y|\varphi_0(x))q(x)\mu_{\mathbb{R}}\mu_y$. Let a, b, c, and d be positive constants, and D be a compact interval. For $m \in \mathbb{N}$, a family of functions $\mathcal{W}_m$ is defined by

$$\mathcal{W}_m = \{w : \mathbb{R} \to \mathbb{R} \mid 0 < w(x) \le a\sqrt{m} \text{ for all } x \in \mathbb{R}, \text{ and there exists a closed interval } I \subset D \text{ such that } \tfrac{c}{m} \le Q(I) \le \tfrac{d}{m} \text{ and } w(x) \ge b\sqrt{m} \text{ on } I\}.$$

For $\gamma > 0$, let m_n be a natural number given by $m_n = [n^\gamma]$ for $n \in \mathbb{N}$, and $\mathcal{G}_\gamma$ be a family of sequences $\{\{w_k^{(n)}\}_{n\in\mathbb{N}, 1\le k\le m_n} \mid w_k^{(n)} \in \mathcal{W}_{m_n}\}$. Then, there exists $\gamma_0 > 0$ such that for any $\gamma \le \gamma_0$ and $\{w_k^{(n)}\} \in \mathcal{G}_\gamma$, we obtain, as n goes to infinity,

$$\max_{1\le k\le m_n} \sup_{\beta} \sum_{i=1}^{n} \log \frac{r(Y_i|\varphi_0(X_i) + \beta w_k^{(n)}(X_i))}{r(Y_i|\varphi_0(X_i))} = \Big\{ \max_{1\le k\le m_n} \frac{1}{2}\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_k^{(n)}(X_i, Y_i)\Big)^2 \Big\}(1 + o_p(1)), \tag{85}$$

where $u_k^{(n)}(x, y)$ is a tangent vector given by

$$u_k^{(n)}(x, y) = \frac{\partial \log r(y|\varphi_0(x) + \beta w_k^{(n)}(x))}{\partial \beta}\Big|_{\beta=0} = \frac{\partial \log r(y|\varphi_0(x))}{\partial \xi} w_k^{(n)}(x). \tag{86}$$

First, we will establish the uniform consistency of the maximum likelihood estimator of β.

Lemma 4. *Let $r(y|\xi)$, $q(x)$, $\varphi_0(x)$, and $\mathcal{W}_m$ be the same as in Lemma 3. For $m \in \mathbb{N}$, let $\mathcal{H}_m$ be the set of m functions in $\mathcal{W}_m$, that is, $\mathcal{H}_m = \{\{w_k\}_{k=1}^m \mid w_k \in \mathcal{W}_m\}$. For $\Xi = \{w_k^{[m]}\}_{k=1}^m \in \mathcal{H}_m$, we write $\widehat{\beta}_k^{[m]}(\Xi)$ for the maximum likelihood estimator of the model $r(y|\varphi_0(x) + \beta w_k^{[m]}(x))q(x)$, given an i.i.d. sample $(X_1, Y_1), \dots, (X_n, Y_n)$ with probability $r(y|\varphi_0(x))q(x)\mu_{\mathbb{R}}\mu_y$. Then, there exist positive constants A, γ, and ν such that the inequality*

$$\mathrm{Prob}\big(\max_{1\le k\le m} \big|\widehat{\beta}_k^{[m]}(\Xi)\big| \ge \varepsilon\big) \le A\frac{m^\gamma}{n\varepsilon^\nu} \tag{87}$$

holds for all $0 < \varepsilon < 1$, $n, m, \in \mathbb{N}$, and $\Xi \in \mathcal{H}_m$.

Proof. The proof is divided into five parts. In the first four parts, fix $w(x) \in \mathcal{W}_m$, and write $f^{[m]}(z;\beta) = r(y|\varphi_0(x)+\beta w(x))q(x)$ for $z=(x,y)$. We define $H_+^{[m]}(z;t)$, $H_-^{[m]}(z;t)$, and $g^{[m]}(z;\beta;\rho)$ for $t,\beta \in \mathbb{R}$ and $\rho > 0$ by

$$H_+^{[m]}(z;t) = \sup_{\beta \geq t} \log f^{[m]}(z;\beta), \qquad H_-^{[m]}(z;t) = \sup_{\beta \leq -t} \log f^{[m]}(z;\beta), \tag{88}$$

$$\text{and} \quad g^{[m]}(z;\beta,\rho) = \sup_{|\beta'-\beta|\leq\rho} \log f^{[m]}(z;\beta'), \tag{89}$$

respectively. A constant M is taken so that $|\varphi_0(x)| \leq M$ for all $x \in \mathbb{R}$. The true probability is denoted by $f_0(z)\mu$, where $f_0(z) = r(y|\varphi_0(x))q(x)$ and $\mu = \mu_y dx$.

(a) *Bound of* $E_{f_0\mu}[H_\pm^{[m]}(z;t)]$. First, we will show that for arbitrary $\delta > 0$, there exist $B_1, \lambda_1 > 0$ such that for all $m \in \mathbb{N}$ and $t \geq B_1 m^{\lambda_1}$ the inequality

$$E_{f_0\mu}[H_\pm^{[m]}(z;t)] \leq E_{f_0\mu}[\log f_0(z)] - 3\delta \tag{90}$$

holds. We will prove it only for $H_+^{[m]}(z;t)$, since the proof on $H_-^{[m]}(z;t)$ is exactly the same. From the assumption [NM3]-2, there exist $\Delta \subset \mathcal{Y}$, $\Gamma_1 > 0$, $\Lambda_1 > 0$, $\gamma_1 > 0$, and $R_0 > 0$ such that $\int_\Delta r(y|u)dy \geq \Gamma_1$ for all $u \in [-M, M]$ and $\log r(y|u) \leq -\Lambda_1|u|^{\gamma_1}$ for all $y \in \Delta$ and $u \geq R_0$. If we set $R = \frac{R_0+M}{b\sqrt{m}}$, we have $\varphi_0(x) + tw(x) \geq R_0$ for all $t \geq R$ and $x \in I$, where I is the interval on which $w(x) \geq b\sqrt{m}$. Then, for all $x \in I$, $y \in \Delta$, and $t \geq R$, the bound

$$\begin{aligned} H_+^{[m]}(z;t) &= h(y|\varphi_0(x) + tw(x)) + \log q(x) \\ &\leq -\Lambda_1|\varphi_0(x) + tw(x)|^{\gamma_1} + \log q(x) \\ &\leq -\Lambda_1(b\sqrt{m}t - M)^{\gamma_1} + \log q(x) \end{aligned} \tag{91}$$

is obtained. From the assumption [NM3]-3, there exists $F_1 > 0$ such that $E_{r(y|\xi)}[h_+(y|-M)] \leq F_1$ for all $\xi \in [-M, M]$. Since $\varphi_0(x) + \beta w(x) \geq -M$ for $\beta > 0$, we see that for $t \geq 0$

$$E_{f_0\mu}[H_+^{[m]}(z;t)] \;\leq\; E_Q[E_{r(y|\varphi_0(x))}[h_+(y|-M)]] \;\leq\; F_1. \tag{92}$$

For a real number r, we define $(r)_+$ by $(r)_+ = r$ if $r \geq 0$ and $(r)_+ = 0$ if

$r < 0$. Then, from eqs.(91) and (92), we obtain

$$\begin{aligned} E_{f_0\mu}[H_+^{[m]}(z;t)] &= \int_{I\times\Delta} H_+(z;t) f_0(z) d\mu + \int_{\mathcal{X}\times\mathcal{Y}\setminus I\times\Delta} H_+(z;t) f_0(z) d\mu dx dy \\ &\le -\Lambda_1 (a\sqrt{m}t - M)^{\gamma_1} \int_I \int_\Delta r(y|\varphi_0(x)) d\mu_y q(x) dx \\ &\quad + E_{f_0\mu}[(H_+(z;t))_+] + E_Q |\log q(x)| \\ &\le -\Lambda_1\Gamma_1 \frac{c}{m} (b\sqrt{m}t - M)^{\gamma_1} + F_1 + E_Q |\log q(x)|. \end{aligned} \tag{93}$$

From this inequality and the fact $E_{f_0\mu}[\log f_0(z)] < \infty$, we can conclude eq.(90) if we choose $\lambda_1 > \frac{1}{2}\gamma_1 - 1$ and sufficiently large $B_1 > 0$.

(b) *L^2 bound of $H_\pm^{[m]}(z;t)$.* Next, we will prove that for any $B_1 > 0$ and $\lambda_1 > 0$ there exist $B_2 > 0$ and $\lambda_2 > 0$ such that the inequality

$$E_{f_0\mu}[|H_\pm^{[m]}(z; B_1 m^{\lambda_1})|^2] \le B_2 m^{\lambda_2} \tag{94}$$

holds for any $m \in \mathbb{N}$.

To see this, let $\ell_1(\xi)$ and $\lambda > 0$ be in the assumption [NM3]-3, and $\Gamma_2 > 0$ satisfy $|\ell_i(\xi)| \le \Gamma_2$ for all $\xi \in [-M, M]$. Then, the bound

$$\begin{aligned} E_{f_0\mu}\left[|H_\pm^{[m]}(z;t)|^2\right] &= E_Q\left[E_{r(y|\varphi_0(x))}\left[|h_+(y|\varphi_0(x) \pm t\beta w(x))|^2\right]\right] \\ &\le E_Q\left[\ell_1(\varphi_0(x))^2 |\varphi_0(x) \pm t\beta w(x)|^{2\lambda}\right] \\ &\le \Gamma_2^2 (M + ta\sqrt{m})^{2\lambda} \end{aligned} \tag{95}$$

is obtained. The above assertion is straightforward, if we choose sufficiently large B_2 and $\lambda_2 \ge 2\lambda(\lambda_1 + \frac{1}{2})$.

(c) *Bounds of $E_{f_0\mu}[g^{[m]}(z;\beta,\rho)]$.* We will show the fact that there exist A_3, $B_3 > 0$, $\lambda_3 > 0$, and $\gamma_3 > 0$, such that for arbitrary $R \ge 1$, $\delta > 0$ and $\beta \in [-R, R]$ the inequalities

$$E_{f_0\mu}[g^{[m]}(z;\beta,\rho)] \le E_{f_0\mu}[\log f^{[m]}(z;\beta)] + \delta \tag{96}$$

and

$$E_{f_0\mu}[|g^{[m]}(z;\beta,\rho)|^2] \le B_3 (\sqrt{m}R)^{\lambda_3} + 2\delta^2 \tag{97}$$

for $\rho \le A_3 \delta (\sqrt{m}R)^{-\gamma_3}$.

Because $|\varphi_0(x) + \beta w(x)| \le M + a\sqrt{m}R$ for $\beta \in [-R, R]$, from the assumption [NM3]-4, we can find $\nu_3 > 0$, $\Psi(y)$, and $\ell_2(\xi)$ such that

$$|\log f^{[m]}(z;\beta) - \log f^{[m]}(z;\beta')| \le \Psi(y) w(x) |\beta - \beta'| \tag{98}$$

and

$$E_{r(y|\xi)}[|\Psi(y)|^2] \leq \ell_2(\xi)(M + a\sqrt{m}R)^{\nu_3} \tag{99}$$

are satisfied. Eq.(98) means $|g^{[m]}(z;\beta,\rho) - \log f^{[m]}(z;\beta)| \leq a\sqrt{m}\rho\Psi(y)$. Since $\Gamma_3 = E_Q[\ell_2(\varphi_0(x))] < \infty$, we obtain

$$\begin{aligned} E_{f_0\mu}[g^{[m]}(z;\beta,\rho)] &\leq E_{f_0\mu}[\log f^{[m]}(z;\beta)] + \rho a\sqrt{m}\{\Gamma_3(M + a\sqrt{m}R)^{\nu_3}\}^{1/2} \\ &\leq E_{f_0\mu}[\log f^{[m]}(z;\beta)] + F_3\rho(\sqrt{m}R)^{\frac{\nu_3}{2}+1}, \end{aligned} \tag{100}$$

for $F_3 = a\Gamma_3^{1/2}(M+a)^{\nu_3/2}$. If we choose ρ as $\rho \leq \delta\{F_3(\sqrt{m}R)^{\frac{\nu_3}{2}+1})\}^{-1}$, the first assertion is satisfied.

For this choice of ρ, noting that $a^2m\rho^2 E_{f_0\mu}[\Psi(y)^2] \leq \delta^2$, we further obtain

$$\begin{aligned} E_{f_0\mu}[|g^{[m]}(z;\beta,\rho)|^2] &\leq 2E_{f_0\mu}|\log f^{[m]}(z;\beta)|^2 + 2E_{f_0\mu}[\Psi(y)^2]a^2m\rho^2 \\ &\leq 4E_{f_0\mu}[|\log f_0(z)|^2 + \beta^2a^2m\Psi(y)^2] + 2E_{f_0\mu}[\Psi(y)^2]a^2m\rho^2 \\ &\leq 4E_{f_0\mu}|\log f_0(z)|^2 + 4a^2m(M + a\sqrt{m}R)^2\{\Gamma_3(M + a\sqrt{m}R)^{\nu_3}\} + 2\delta^2 \\ &\leq 4E_{f_0\mu}|\log f_0(z)|^2 + \Lambda_3(\sqrt{m}R)^{\nu_3+4} + 2\delta^2, \end{aligned} \tag{101}$$

for sufficiently large Λ_3, which depends only on a, M, and Γ_3. Because $E_{f_0\mu}|\log f_0(z)|^2 < \infty$ from the assumption [NM3]-3, the second assertion is obtained.

(d) *Lower bound of KL-divergence.* We will show that there exist $B_4 > 0$, $\lambda_4 > 0$, and $\nu_4 \in \mathbb{R}$ such that the bound

$$E_{f_0\mu}[\log f^{[m]}(z;\beta)] \leq E_{f_0\mu}[\log f_0(z)] - B_4m^{\nu_4}\varepsilon^{\lambda_4} \tag{102}$$

holds for arbitrary $0 < \varepsilon < 1$, $m \in \mathbb{N}$, and β with $|\beta| \geq \varepsilon$.

By the property of Kullback-Leibler divergence, we have

$$E_{r(y|\varphi_0(x))}[\log r(y|\varphi_0(x) + \beta w(x)) - \log r(y|\varphi_0(x))] \leq 0, \tag{103}$$

for all x and β. From the assumption [NM3]-1, there exist positive constants A_4, C_4, γ_4, δ_4, and T_0 such that, if $|\beta w(x)| \leq T_0$, the inequality

$$E_{r(y|\varphi_0(x))}\big[\log r(y|\varphi_0(x) + \beta w(x)) - \log r(y|\varphi_0(x))\big] \leq -A_4|\beta w(x)|^{\gamma_4} \tag{104}$$

holds, and if $|\beta w(x)| > T_0$, the inequality

$$E_{r(y|\varphi_0(x))}\big[\log r(y|\varphi_0(x) + \beta w(x)) - \log r(y|\varphi_0(x))\big] \leq -C_4|\beta w(x)|^{\delta_4} \tag{105}$$

holds. Since $w(x) \geq b\sqrt{m}$ for $x \in I$, for all $x \in I$ and β with $|\beta| \geq \varepsilon$, either

$$E_{r(y|\varphi_0(x))}\big[\log r(y|\varphi_0(x)+\beta w(x)) - \log r(y|\varphi_0(x))\big] \leq -A_4(b\varepsilon\sqrt{m})^{\gamma_4} \tag{106}$$

or

$$E_{r(y|\varphi_0(x))}\big[\log r(y|\varphi_0(x)+\beta w(x)) - \log r(y|\varphi_0(x))\big] \leq -C_4(b\varepsilon\sqrt{m})^{\delta_4} \tag{107}$$

is satisfied.

From eqs.(103), (106), and (106), for $\lambda_4 = \max\{\gamma_4, \delta_4\}$, $\kappa_4 = \min\{\gamma_4, \delta_4\}$, and some constant $F_4 > 0$, the bound

$$E_{f_0\mu}[\log f^{[m]}(z;\beta) - \log f_0(z)] \leq -F_4\varepsilon^{\lambda_4} m^{\frac{\kappa_4}{2}} \frac{c}{m} \tag{108}$$

is obtained, which means the assertion.

(e) *Uniform consistency.* We write $f_k^{[m]}(z;\beta) = r(y|\varphi_0(x)+\beta w_k^{[m]}(x))q(x)$. For a fixed $\delta > 0$, take B_1 and λ_1 in the assertion (a), and denote $R_m = B_1 m^{\lambda_1}$. Because we have

$$\sup_{\beta \geq R_m} \sum_{i=1}^{n} \log f_k^{[m]}(Z_i;\beta) \leq \sum_{i=1}^{n} H_+^{[m]}(Z_i;R_m), \tag{109}$$

for all m and k, eq.(90) and Chebyshev's inequality give

$$\begin{aligned}
&\mathrm{Prob}\Big(\exists k, \sup_{\beta > R_m} \frac{1}{n}\sum_{i=1}^{n} \log f_k^{[m]}(Z_i;\beta) \geq \frac{1}{n}\sum_{i=1}^{n} \log f_0(Z_i)\Big) \\
&\leq m\,\mathrm{Prob}\Big(\frac{1}{n}\sum_{i=1}^{n} H_+^{[m]}(Z_i;R_m) > \frac{1}{n}\sum_{i=1}^{n} \log f_0(Z_i)\Big) \\
&\leq m\Big\{\mathrm{Prob}\Big(\frac{1}{n}\sum_{i=1}^{n} H_+^{[m]}(Z_i;R_m) - E_{f_0\mu}[H_+^{[m]}(Z;R_m)] > \delta\Big) \\
&\qquad + \mathrm{Prob}\Big(\frac{1}{n}\sum_{i=1}^{n} f_0(Z_i) - E_{f_0\mu}[\log f_0(Z)] < -\delta\Big)\Big\} \\
&\leq m\left\{\frac{V[H_+^{[m]}(Z;R_m)]}{n\delta^2} + \frac{V[\log f_0(Z)]}{n\delta^2}\right\}.
\end{aligned} \tag{110}$$

We can obtain a similar bound for $\beta < -R_m$ also, using $H_-^{[m]}(z;t)$. From the fact (b), the variance $V[H_\pm^{[m]}(Z;R_m)]$ is bounded by by $B_2 m^{\lambda_2}$ for some $B_2 > 0$ and $\lambda_2 > 0$. This shows there exist $A_5 > 0$ and $\lambda_5 > 0$ such that

$$\mathrm{Prob}(^\exists k, |\hat{\beta}_k^{[m]}| \geq R_m) \leq A_5 \frac{m^{\lambda_5}}{n} \tag{111}$$

holds for all $m, n \in \mathbb{N}$.

By the fact (d), we have $E_{f_0\mu}[\log f_k^{[m]}(z;\beta)] - E_{f_0\mu}[\log f_0(z)] \leq -4\delta_m$ for all β with $|\beta| \geq \varepsilon$ and $m \in \mathbb{N}$, where $\delta_m = \frac{1}{4} B_4 m^{\nu_4} \varepsilon^{\lambda_4}$. From the fact (c), we have

$$E_{f_0\mu}[g^{[m]}(z;\beta,\rho_m)] \leq E_{f_0\mu}[\log f(z;\beta)] + \delta_m \tag{112}$$

for all $\beta \in [-R_m, R_m]$ and $\rho_m = A_3 \delta_m \frac{1}{(\sqrt{m} R_m)^{\gamma_3}}$. Let N_m be a natural number given by $N_m = [\frac{R_m}{\rho_m}] + 1$. Then, there exist positive constants C_5 and ν_5 such that $N_m \leq C_5 m^{\nu_5} \varepsilon^{-\lambda_4}$. Dividing the interval $[-R_m, -\delta] \cup [\delta, R_m]$ into N_m intervals $J_j = [\beta_j - \rho_m, \beta_j + \rho_m]$ $(1 \leq j \leq N_m)$ with disjoint interior, we have

$$E_{f_0\mu}[g^{[m]}(z;\beta_j,\rho_m)] \leq E_{f_0\mu}[\log f_0(z)] - 3\delta_m \tag{113}$$

for each j. Then, we obtain

$$\begin{aligned}
&\mathrm{Prob}\Big(^\exists \beta \in [-R_m, -\varepsilon] \cup [\varepsilon, R_m], \frac{1}{n}\sum_{i=1}^n \log f_k^{[m]}(Z_i;\beta) \geq \frac{1}{n}\sum_{i=1}^n \log f_0(Z_i)\Big) \\
&\leq \mathrm{Prob}\Big(^\exists k, 1 \leq {}^\exists j \leq N_m, \frac{1}{n}\sum_{i=1}^n \sup_{\beta \in J_j} \log f_k^{[m]}(Z_i;\beta) \geq \frac{1}{n}\sum_{i=1}^n \log f_0(Z_i)\Big) \\
&\leq N_m \,\mathrm{Prob}\Big(\frac{1}{n}\sum_{i=1}^n g^{[m]}(Z_i;\beta_j,\rho_m) > \frac{1}{n}\sum_{i=1}^n \log f_0(Z_i)\Big) \\
&\leq C_5 m^{\nu_5+1} \varepsilon^{-\lambda_4} \Big\{ \mathrm{Prob}\Big(\frac{1}{n}\sum_{i=1}^n g^{[m]}(Z_i;\beta_j,\rho_m) - E_{f_0\mu}[g^{[m]}(z;\beta_j,\rho_m)] > \delta_m\Big) \\
&\qquad + \mathrm{Prob}\Big(\frac{1}{n}\sum_{i=1}^n \log f_0(Z_i) - E_{f_0\mu}[\log f_0(Z)] < -\delta_m\Big)\Big\} \\
&\leq C_5 m^{\nu_5+1} \varepsilon^{-\lambda_4} \Big\{ \frac{V[g^{[m]}(z;\beta_j,\rho_m)]}{n\delta_m^2} + \frac{V[\log f_0(Z)]}{n\delta_m^2} \Big\}.
\end{aligned} \tag{114}$$

From the fact (c), we see that there exist $F_5 > 0$ and $\tau_5 > 0$ such that

$$\mathrm{Prob}\Big(^\exists k, \hat{\beta}_k^{[m]} \in [-R_m, -\varepsilon] \cup [\varepsilon, R_m]\Big) \leq F_5 \frac{m^{\tau_5}}{n\varepsilon^{3\lambda_4}} \tag{115}$$

for all $m, n \in \mathbb{N}$, and $0 < \varepsilon < 1$.

Combining eqs. (111) and (115), we have the assertion of Lemma 4. □

Proof of Lemma 3. From Lemma 4, for a fixed small $\varepsilon > 0$, the probability of the event $\{\max_{1\le k\le m_n} |\widehat{\beta}_k^{(n)}| < \varepsilon\}$ converges to one. Then, the maximum likelihood estimator $\widehat{\beta}_k^{(n)}$ of the model $f_k^{(n)}(z;\beta) = r(y|\varphi_0(x)+\beta w_k^{(n)}(x))q(x)$ satisfies the likelihood equation

$$\sum_{i=1}^{n} \frac{\partial \log f_k^{(n)}(Z_i;\widehat{\beta}_k^{(n)})}{\partial \beta} = 0 \tag{116}$$

for all $1 \le k \le m_n$, with a probability which converges to one. Taylor expansion of eq.(116) and L_n leads

$$\sum_{i=1}^{n} \frac{\partial \log f_k^{(n)}(Z_i;0)}{\partial \beta} + \sum_{i=1}^{n} \frac{\partial^2 \log f_k^{(n)}(Z_i;\beta_k^*)}{\partial \beta^2} \widehat{\beta}_k^{(n)} \tag{117}$$

and

$$\sum_{i=1}^{n} \log \frac{f_k^{(n)}(Z_i;\widehat{\beta}_k^{(n)})}{f_0(Z_i)} = \sum_{i=1}^{n} \frac{\partial \log f_k^{(n)}(Z_i;0)}{\partial \beta} \widehat{\beta}_k^{(n)} + \frac{1}{2} \sum_{i=1}^{n} \frac{\partial^2 \log f_k^{(n)}(Z_i;\beta_k^{**})}{\partial \beta^2} (\widehat{\beta}_k^{(n)})^2 \tag{118}$$

for some β_k^* and β_k^{**} between 0 and $\widehat{\beta}_k^{(n)}$. A simple calculation shows

$$\sum_{i=1}^{n} \log \frac{f_k^{(n)}(Z_i;\widehat{\beta}_k^{(n)})}{f_0(Z_i)} = \frac{\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f_k^{(n)}(Z_i;0)}{\partial \beta}\right)^2}{-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f_k^{(n)}(Z_i;0)}{\partial \beta^2}} \Big\{ S_n^{(k)} - \frac{1}{2} T_n^{(k)} \Big\}, \tag{119}$$

where

$$S_n^{(k)} = \frac{\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f_k^{(n)}(Z_i;0)}{\partial \beta^2}}{\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f_k^{(n)}(Z_i;\beta_k^*)}{\partial \beta^2}} \tag{120}$$

and

$$T_n^{(k)} = \frac{\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f_k^{(n)}(Z_i;0)}{\partial \beta^2} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f_k^{(n)}(Z_i;\beta_k^{**})}{\partial \beta^2}}{\left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f_k^{(n)}(Z_i;\beta_k^*)}{\partial \beta^2}\right)^2}. \tag{121}$$

The proof of Lemma 3 is completed if we show for arbitrary $\varepsilon > 0$

$$\mathrm{Prob}\Big(\max_{1\le k\le m_n}\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f_k^{(n)}(Z_i;\tilde{\beta}_k)}{\partial\beta^2}+1\Big|\ge\varepsilon\Big) \quad \longrightarrow 0 \quad (n\to\infty), \tag{122}$$

for $\tilde{\beta}_k = 0$, $\widehat{\beta}_k^*$, and $\widehat{\beta}_k^{**}$. In fact, $\max_k |S_n^{(k)} - 1|$ and $\max_k |T_n^{(k)} - 1|$ converge to 0 in probability from the above convergence.

By Taylor expansion, we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f_k^{(n)}(Z_i;\tilde{\beta}_k)}{\partial\beta^2} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f_k^{(n)}(Z_i;0)}{\partial\beta^2} + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3 \log f_k^{(n)}(Z_i;\eta)}{\partial\beta^3}\tilde{\beta}_k, \tag{123}$$

where η is between 0 and $\tilde{\beta}_k$. Because $\frac{\partial^2 \log f_k^{(n)}(z;0)}{\partial\beta^2} = \frac{\partial^2 \log r(y;\varphi_0(x))}{\partial u^2}(w_k^{(n)}(x))^2$, from the assumption [NM3]-5 and the fact $|w_k^{(n)}(x)| \le a\sqrt{m_n}$, there exists $B_1 > 0$ such that

$$E_{f_0\mu}\Big[\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f_k^{(n)}(Z_i;0)}{\partial\beta^2}+1\Big|^2\Big] \le \frac{2+2B_1 m_n^2}{n}, \tag{124}$$

holds for all $n \in \mathbb{N}$. Therefore, by Chebyshev's inequality, for $0 < \gamma < \frac{1}{3}$ we obtain

$$\begin{aligned}
&\mathrm{Prob}\Big(\max_{1\le k\le m_n}\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f_k^{(n)}(Z_i;0)}{\partial\beta^2}+1\Big|>\frac{\varepsilon}{2}\Big)\\
&\le m_n\,\mathrm{Prob}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f_k^{(n)}(Z_i;0)}{\partial\beta^2}+1\Big|>\frac{\varepsilon}{2}\Big)\\
&\le m_n\frac{2+2B_1 m_n^2}{n} \quad \longrightarrow \quad 0.
\end{aligned} \tag{125}$$

Let d be a positive number such that $d > 2$. From the assumption [NM3]-6, there exist $B_2 > 0$ and $n_0 \in \mathbb{N}$ such that

$$E_{r(y|\varphi_0(x))}\Big[\sup_{|\beta|\le m_n^{-d}}\Big|\frac{\partial^3 \log r(y|\varphi_0(x)+\beta w_k^{(n)}(x))}{\partial u^3}\Big|^2\Big] \le B_2, \tag{126}$$

for all $n \geq n_0$ and $x \in \mathbb{R}$. Let $M_k^{(n)}(z)$ be a function defined by

$$M_k^{(n)}(z) = \sup_{|\beta| \leq m_n^{-d}} \Big| \frac{\partial^3 \log f_k^{(n)}(z;\beta)}{\partial \beta^3} \Big|. \tag{127}$$

The L^2 norm of this function is bounded by

$$\begin{aligned} E_{f_0\mu}[(M_k^{(n)}(z))^2] &= E_Q\Big[E_{r(y|\varphi_0(x))}\Big[\sup_{|\beta| \leq m_n^{-d}} \Big| \frac{\partial^3 \log r(y|\varphi_0(x) + \beta w_k^{(n)}(x))}{\partial u^3} \Big|^2 \{w_k^{(n)}(x)\}^6 \Big] \Big] \\ &\leq B_2 (a\sqrt{m})^6. \end{aligned} \tag{128}$$

Then, we obtain a bound of the probability

$$\begin{aligned} &\mathrm{Prob}\Big(1 \leq {}^\exists k \leq m, \Big| \frac{1}{n} \sum_{i=1}^n \frac{\partial^3 \log f_k^{(n)}(Z_i;\eta)}{\partial \beta^3} \tilde{\beta}_k \Big| \geq \frac{\varepsilon}{2} \Big) \\ &\leq \mathrm{Prob}\Big(\max_{1 \leq k \leq m_n} |\widehat{\beta}_k| \geq \frac{1}{m_n^d} \Big) \\ &\quad + \mathrm{Prob}\Big(\max_{1 \leq k \leq m_n} |\widehat{\beta}_k| < \frac{1}{m_n^d}, \text{ and } 1 \leq {}^\exists k \leq m_n, \Big| \frac{1}{n} \sum_{i=1}^n \frac{\partial^3 \log f_k^{(n)}(Z_i;\eta)}{\partial \beta^3} \Big| \geq \frac{\varepsilon}{2} m_n^d \Big) \\ &\leq \mathrm{Prob}\Big(\max_{1 \leq k \leq m_n} |\widehat{\beta}_k| \geq \frac{1}{m_n^d} \Big) + m_n \, \mathrm{Prob}\Big(\frac{1}{n} \sum_{i=1}^n M_k^{(n)}(Z_i) \geq \frac{\varepsilon}{2} m_n^d \Big). \end{aligned} \tag{129}$$

From Lemma 4, there exist positive constants A, λ, and ν such that the first term of eq.(129) is bounded by $A \frac{m_n^{\lambda + d\nu}}{n}$. By Chebyshev's inequality, the second term is not greater than $\frac{4 m_n E[M_k^{(n)}(z)^2]}{\varepsilon^2 m_n^{2d}} \leq \frac{4 B_2 a^6 m_n^{4-2d}}{\varepsilon^2}$, which converges to zero because we take $d > 2$. For such d fixed, taking sufficiently small γ such that $\gamma(\lambda + d\nu) < 1$, we see that the first term of eq.(129) converges to zero for $m_n = [n^\gamma]$. Thus, we have for any sufficiently small $\gamma > 0$ and $m_n = [n^\gamma]$,

$$\mathrm{Prob}\Big(1 \leq {}^\exists k \leq m_n, \Big| \frac{1}{n} \sum_{i=1}^n \frac{\partial^3 \log f_k^{(n)}(Z_i;\eta)}{\partial \beta^3} \tilde{\beta}_k \Big| \geq \frac{\varepsilon}{2} \Big) \quad \longrightarrow \quad 0, \tag{130}$$

as $n \to \infty$.

Combination of eqs.(123), (125) and (130) means eq.(122), and completes the proof. □