# Likelihood Ratio of Unidentifiable Models and Multilayer Neural Networks 

Kenji Fukumizu<br>Institute of Statistical Mathematics<br>4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan<br>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) $\Theta$ such that $S$ is given by $S=$ $\{f(z ; \theta) \mid \theta \in \Theta\}$. We call $\Theta$ as the parameter space. We assume throughout this paper that $\operatorname{Supp} f(z ; \theta)$ is invariant for all $\theta \in \Theta$, and $f(z ; \theta)$ is differentiable on $\theta$ 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 $\left\{Z_{i}\right\}_{i=1}^{n}$ is defined by

$$
\begin{equation*}
L_{n}(\theta)=\sum_{i=1}^{n} \log \frac{f\left(Z_{i} ; \theta\right)}{f_{0}\left(Z_{i}\right)} . \tag{1}
\end{equation*}
$$

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

$$
\begin{equation*}
L_{n}(\hat{\theta})=\sup _{\theta \in \Theta} L_{n}(\theta)=\sup _{\theta \in \Theta} \sum_{i=1}^{n} \log \frac{f\left(Z_{i} ; \theta\right)}{f_{0}\left(Z_{i}\right)} . \tag{2}
\end{equation*}
$$

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 $f\left(z ; \theta_{0}\right)=f_{0}(z)$. We do not assume the uniqueness of $\theta_{0}$, and denote the set of true parameters by $\Theta_{0}$; that is, $\Theta_{0}=\left\{\theta \in \Theta \mid f\left(z ; \theta_{0}\right) \mu=f_{0}(z) \mu\right\}$. Unless $\Theta_{0}$ 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 $\Theta_{0}$ is a union of finitely many submanifolds of $\Theta$, 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 $f\left(z ; a_{1}, a_{2}, b\right)$ be a mixture model defined by

$$
\begin{equation*}
f\left(z ; a_{1}, a_{2}, b\right)=b g\left(z ; a_{1}\right)+(1-b) g\left(z ; a_{2}\right), \tag{3}
\end{equation*}
$$

where $b \in[0,1]$. Suppose that the true density $f_{0}(z)$ is given by $g\left(z ; a_{0}\right)$ for some $a_{0}$. Then, the set of parameters to give $f_{0}(z)$ contains $\left\{\left(a_{1}, a_{2}, b\right) \mid\right.$ $a_{1}=a_{2}=a_{0}, b:$ arbitrary $\} \cup\left\{\left(a_{1}, a_{2}, b\right) \mid b=0, a_{2}=a_{0}, a_{1}:\right.$ arbitrary $\} \cup$ $\left\{\left(a_{1}, a_{2}, b\right) \mid b=1, a_{1}=a_{0}, a_{2}\right.$ : 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. Feedforward 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), $\Theta$ 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 $\beta$ for each $\alpha \in A_{0}$ and $f_{0} \mu$-almost every $z$.
2. Let $\Theta_{0}$ and $\Theta(\alpha)$ be subsets defined by $\Theta_{0}=\Theta \cap\left(A_{0} \times\{0\}\right)$ and

$$
\begin{gather*}
\Theta(\alpha)=\Theta \cap(\{\alpha\} \times \mathbb{R}) \text { for } \alpha \in A_{0} \text {, respectively. Then, } \\
\Theta=\bigcup_{\alpha \in A_{0}} \Theta(\alpha) \tag{4}
\end{gather*}
$$

3. The set of the parameters to give $f_{0}$ is $\Theta_{0}$; that is,

$$
\begin{equation*}
f(z ;(\alpha, \beta)) \mu=f_{0}(z) \mu \Longleftrightarrow \beta=0 \tag{5}
\end{equation*}
$$

4. For all $\alpha \in A_{0}$,

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

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 $S_{\alpha}$ at the origin,

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

can be looked as a unit tangent vector in the direction of $S_{\alpha}$ (see fig.1). The family of score functions $C=\left\{v_{\alpha} \mid \alpha \in A_{0}\right\}$ 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 $L^{2}\left(f_{0} \mu\right)$. 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

$$
\begin{equation*}
\varphi(x ; \theta)=\sum_{j=1}^{H} b_{j} s\left(a_{j} x+c_{j}\right)+d \tag{8}
\end{equation*}
$$



Figure 1: Locally conic model
where $x \in \mathcal{X}=\mathbb{R}, s(t)=\tanh (t)$, and $\theta=\left(a_{1}, \ldots, a_{H}, b_{1}, \ldots, b_{H}, c_{1}, \ldots, c_{H}, d\right)^{T} \in$ $\Theta_{H}=\mathbb{R}^{3 H+1}$. 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 $\left(\mathcal{Y}, \mathcal{B}_{y}, \mu_{y}\right)$ 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 $r(y \mid u)$ is of class $C^{1}$ on $u$ for all $y \in \mathcal{Y}$.
2. For different $u_{1}$ and $u_{2}$, we have $r\left(y \mid u_{1}\right) \mu_{y} \neq r\left(y \mid u_{2}\right) \mu_{y}$.
3. The Fisher information $G(u)$ of $r(y \mid u)$, defined by

$$
\begin{equation*}
G(u)=\int\left(\frac{\partial \log r(y \mid u)}{\partial u}\right)^{2} r(y \mid u) d \mu_{y}, \tag{9}
\end{equation*}
$$

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

$$
\begin{equation*}
\lim _{\rho \downarrow 0} E_{r(y \mid u)}\left[\sup _{\left|u^{\prime}-u\right| \leq \rho}\left|\frac{\partial \log r\left(y \mid u^{\prime}\right)}{\partial u}\right|\right]<\infty . \tag{10}
\end{equation*}
$$

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

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

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

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

$$
\begin{equation*}
r(y \mid u)=\frac{1}{\sqrt{2 \pi} \sigma} \exp \left\{-\frac{1}{2 \sigma^{2}}(y-u)^{2}\right\} \tag{12}
\end{equation*}
$$

for continuous $y$, and the binomial distribution model

$$
\begin{equation*}
r(y \mid u)=\frac{e^{u y}}{1+e^{u}} \tag{13}
\end{equation*}
$$

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 $\varphi_{0}(x)=b_{0} \tanh \left(a_{0} x\right)$, then for any parameter $\theta$ in the set $\left\{\theta \in \Theta_{2} \mid a_{1}=a_{0}, b_{1}=b_{0}, c_{1}=0, b_{2}=\right.$ $0, d=0, a_{2}, c_{2}:$ arbitrary $\} \cup\left\{\theta \in \Theta_{2} \mid a_{1}=a_{0}, b_{1}=b_{0}, c_{1}=0, a_{2}=\right.$ $\left.0, b_{2} \tanh \left(c_{2}\right)+d=0\right\}$ 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 $H$

[^0]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 $\Theta_{H}^{*}$ is defined by $\Theta_{H}^{*}=\{\theta=$ $\left(a_{1}, \ldots, a_{H}, b_{1}, \ldots, b_{H}, c_{1}, \ldots, c_{H}, d\right) \in \Theta_{H} \mid a_{j} \neq 0, b_{j} \neq 0(1 \leq j \leq$ $\left.H),\left(a_{j}, c_{j}\right) \neq \pm\left(a_{h}, c_{h}\right)(1 \leq j<h \leq H)\right\}$. Note that in $\Theta_{H}^{*}$ the parameters that correspond to the functions realizable by a smaller-sized network are eliminated (see [10]). For a parameter in $\Theta_{H}^{*}$, it is known ([13]) that the functions $\left\{1, s\left(a_{j} x+c_{j}\right), s^{\prime}\left(a_{j} x+c_{j}\right) x, s^{\prime}\left(a_{j} x+c_{j}\right) \mid 1 \leq j \leq H\right\}$ are linearly independent.

Given a function

$$
\begin{equation*}
\varphi_{0}(x)=\sum_{k=1}^{K} b_{k}^{0} s\left(a_{k}^{0} x+c_{k}^{0}\right)+d^{0} \tag{14}
\end{equation*}
$$

for $\theta_{0}=\left(a_{1}^{0}, \ldots, a_{K}^{0}, b_{1}^{0}, \ldots, b_{K}^{0}, c_{1}^{0}, \ldots, c_{K}^{0}, d^{0}\right) \in \Theta_{K}^{*}$, the parameter space is again restricted slightly to $\Theta_{H}^{* *}$ by $\Theta_{H}^{* *}=\left\{\theta \in \Theta_{H}^{*} \mid\left(a_{j}, c_{j}\right) \neq \pm\left(a_{k}^{0}, c_{k}^{0}\right)(1 \leq\right.$ $k \leq K, K+1 \leq j \leq H)\}$. This reduction does not matter in discussing the maximum likelihood estimation, because MLE lies in $\Theta_{H}^{* *}$ with probability one. Introduce a new parameterization by

$$
\begin{align*}
\beta & =\operatorname{sgn}\left(b_{K+1}\right) \sqrt{b_{K+1}^{2}+\cdots+b_{H}^{2}} \\
\xi_{k} & =\frac{a_{k}-a_{k}^{0}}{\beta}, \quad(1 \leq k \leq K), \quad \xi_{j}=a_{j}, \quad(K+1 \leq j \leq H) \\
\eta_{k} & =\frac{b_{k}-b_{k}^{0}}{\beta}, \quad(1 \leq k \leq K), \quad \eta_{j}=\frac{b_{j}}{\beta}, \quad(K+1 \leq j \leq H) \\
\zeta_{k} & =\frac{c_{k}-c_{k}^{0}}{\beta}, \quad(1 \leq k \leq K), \quad \zeta_{j}=c_{j}, \quad(K+1 \leq j \leq H) \\
\delta & =\frac{d-d^{0}}{\beta} \tag{15}
\end{align*}
$$

for $\theta \in \Theta_{H}^{* *}$, and define new parameter spaces $\Pi_{H}$ and $\Pi_{H}^{* *}$ by

$$
\begin{align*}
\Pi_{H}=\{\omega & =\left(\xi_{1}, \ldots, \xi_{H}, \eta_{1}, \ldots, \eta_{H}, \zeta_{1}, \ldots, \zeta_{H}, \delta, \beta\right) \mid \\
& a_{k}^{0}+\beta \xi_{k} \neq 0(1 \leq k \leq K), \xi_{j} \neq 0(K+1 \leq j \leq H), \\
& \left(a_{k}^{0}+\beta \xi_{k}, c_{k}^{0}+\beta \zeta_{k}\right) \neq \pm\left(a_{h}^{0}+\beta \xi_{h}, c_{h}^{0}+\beta \zeta_{h}\right)(1 \leq k<h \leq K), \\
& \left(a_{k}^{0}+\beta \xi_{k}, c_{k}^{0}+\beta \zeta_{k}\right) \neq \pm\left(\xi_{j}, \zeta_{j}\right)(1 \leq k \leq K, K+1 \leq j \leq H), \\
& \left(\xi_{j}, \zeta_{j}\right) \neq \pm\left(\xi_{i}, \zeta_{i}\right)(K+1 \leq j<i \leq H), \\
& \left(\xi_{j}, \zeta_{j}\right) \neq \pm\left(a_{k}^{0}, c_{k}^{0}\right)(1 \leq k \leq K, K+1 \leq j \leq H), \\
& b_{k}^{0}+\beta \eta_{k} \neq 0(1 \leq k \leq K), \sum_{j=K+1}^{H} \eta_{j}^{2}=1, \eta_{j} \neq 0(K+1 \leq j \leq H), \\
& \left.\eta_{K+1}>0, \beta \in \mathbb{R}\right\} \tag{16}
\end{align*}
$$

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

$$
\begin{gather*}
\psi(x ; \omega)=\sum_{k=1}^{K}\left(b_{k}^{0}+\beta \eta_{k}\right) s\left(\left(a_{k}^{0}+\beta \xi_{k}\right) x+\left(c_{k}^{0}+\beta \zeta_{k}\right)\right) \\
+\sum_{j=K+1}^{H} \beta \eta_{j} s\left(\xi_{j} x+\zeta_{j}\right)+\beta \delta \tag{17}
\end{gather*}
$$

It is easy to see that the $\Pi_{H}^{* *}$ and $\Theta_{H}^{* *}$ 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 $\left\{\psi(x ; \omega) \mid \omega \in \Pi_{H}\right\}$, when the maximum likelihood estimation is discussed.

Let $S_{H}=\left\{f(x, y ; \omega) \mid \omega \in \Pi_{H}\right\}$ be a statistical model defined by

$$
\begin{equation*}
f(x, y ; \omega)=r(y \mid \psi(x ; \omega)) q(x) . \tag{18}
\end{equation*}
$$

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, $f_{0}(x, y)=r\left(y \mid \varphi_{0}(x)\right) q(x)$. The model $S_{H}$ is a locally conic model, if $\alpha$ summarizes $\left(\xi_{1}, \ldots, \zeta_{H}, \delta\right)$ 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 $\Pi_{H}$ by the fact $\theta_{0} \in \Theta_{K}^{*}$ and $\left(\xi_{j}, \zeta_{j}\right) \neq \pm\left(a_{k}^{0}, c_{k}^{0}\right)$ 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 $s\left(\xi_{j} x+\zeta_{j}\right)(K+1 \leq j \leq H)$ is not contained in the linear hull of the functions $\left\{1, s\left(a_{k}^{0} x+c_{k}^{0}\right), s\left(\xi_{i} x+\zeta_{i}\right), s\left(\left(a_{k}^{0}+\beta \xi_{k}\right) x+\right.\right.$ $\left.\left.\left(c_{k}^{0}+\beta \zeta_{k}\right)\right) \mid 1 \leq k \leq K, K+1 \leq i \leq H, i \neq j\right\}$ by the definition of $\Pi_{H}$, the coefficients of $s\left(\xi_{j} x+\zeta_{j}\right)$ 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 $L^{2}\left(f_{0}(x, y) \mu_{\mathbb{R}} \mu_{y}\right)$-norm of a tangent vector $\frac{\partial}{\partial \beta} \log f(x, y ;(\alpha, 0))$. This is essentially determined by the partial derivative:

$$
\begin{align*}
& \frac{\partial \psi(x ;(\alpha, 0))}{\partial \beta}=\sum_{j=K+1}^{H} \eta_{j} s\left(\xi_{j} x+\zeta_{j}\right)+\delta \\
& +\sum_{k=1}^{K} \eta_{k} s\left(a_{k}^{0} x+c_{k}^{0}\right)+\sum_{k=1}^{K} b_{k}^{0} \xi_{k} s^{\prime}\left(a_{k}^{0} x+c_{k}^{0}\right) x+\sum_{k=1}^{K} b_{k}^{0} \zeta_{k} s^{\prime}\left(a_{k}^{0} x+c_{k}^{0}\right) \tag{19}
\end{align*}
$$

The $L^{2}$ norm is calculated as

$$
\begin{align*}
N(\alpha)^{2} & =\iint r\left(y \mid \varphi_{0}(x)\right) q(x)\left\{\frac{\partial r\left(y \mid \varphi_{0}(x)\right)}{\partial u} \frac{\partial \psi(x ;(\alpha, 0))}{\partial \beta}\right\}^{2} d x d \mu_{y} \\
& =\int G\left(\varphi_{0}(x)\right)\left\{\frac{\partial \psi(x ;(\alpha, 0))}{\partial \beta}\right\}^{2} q(x) d x \tag{20}
\end{align*}
$$

Since $\varphi_{0}(x)$ is bounded, so is $G\left(\varphi_{0}(x)\right)$ by the continuity of $G(u)$. From eq.(19), the function $\left\{\frac{\partial}{\partial \beta} \psi(x ;(\alpha, 0))\right\}^{2}$ is also bounded. Thus, $N(\alpha)$ is finite. Because the functions $1, s\left(\xi_{j} x+\zeta_{j}\right), s\left(a_{k}^{0} x+c_{k}^{0}\right), s^{\prime}\left(a_{k}^{0} x+c_{k}^{0}\right) x$, and $s^{\prime}\left(a_{k}^{0} x+c_{k}^{0}\right)$ $(1 \leq k \leq K, K+1 \leq j \leq H)$ 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 $\beta$, 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 $\beta$. If the maximum likelihood estimator $\hat{\beta}_{\alpha}$ in $S_{\alpha}$ exists for each $\alpha \in A_{0}$, the likelihood ratio of the MLE in $S$ is given by

$$
\begin{equation*}
\sup _{\theta \in \Theta} L_{n}(\theta)=\sup _{\alpha} L_{n}\left(\alpha, \hat{\beta}_{\alpha}\right) . \tag{21}
\end{equation*}
$$

Assume that each submodel $S_{\alpha}$ 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 $\{g(z ; \beta) \mid \beta \in V\}$, neglecting the index $\alpha$. 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

$$
\begin{equation*}
H_{+}(z ; t)=\sup _{\beta \geq t} \log g(z ; \beta) \quad \text { and } \quad H_{-}(z ; s)=\sup _{\beta \leq s} \log g(z ; \beta), \tag{22}
\end{equation*}
$$

respectively. Then,

$$
\begin{equation*}
\lim _{t \uparrow b_{0}} E_{f_{0} \mu}\left[H_{+}(z ; t)\right]<\infty \quad \text { and } \quad \lim _{s \downarrow a_{0}} E_{f_{0} \mu}\left[H_{-}(z ; s)\right]<\infty . \tag{23}
\end{equation*}
$$

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

$$
\begin{array}{ll}
\lim _{t \uparrow b_{0}} H_{+}(z ; t)=-\infty & \text { for all } z \in \Delta_{+}, \\
\lim _{s \downarrow a_{0}} H_{-}(z ; s)=-\infty & \text { for all } z \in \Delta_{-} . \tag{25}
\end{array}
$$

[^1]4. For all $\beta \in V$,
\[

$$
\begin{equation*}
\lim _{\rho \downarrow 0} E_{f_{0} \mu}\left[\sup _{\left|\beta^{\prime}-\beta\right| \leq \rho} \log g\left(z ; \beta^{\prime}\right)\right]<\infty . \tag{26}
\end{equation*}
$$

\]

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

$$
\begin{equation*}
\lim _{\rho \downarrow 0} E_{f_{0} \mu}\left[\sup _{|\beta| \leq \rho}\left|\frac{\partial^{3} \log g(z ; \beta)}{\partial \beta^{3}}\right|\right]<\infty \tag{27}
\end{equation*}
$$

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 $S_{\alpha}$ satisfies the conditions [AN], the standard argument using Taylor expansion leads to

$$
\begin{equation*}
L_{n}\left(\alpha, \hat{\beta}_{\alpha}\right)=\frac{1}{2} U_{n}(\alpha)^{2}+o_{p}(1) \tag{28}
\end{equation*}
$$

where $U_{n}(\alpha)$ is a random variable defined by

$$
\begin{equation*}
U_{n}(\alpha)=\frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_{\alpha}\left(Z_{i}\right) \tag{29}
\end{equation*}
$$

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

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

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 $\alpha$, it can be looked as an empirical process over $\alpha$ or $C$, and every marginal distribution on finite points converges to a multidimensional normal distribution. The likelihood ratio of MLE is given by

$$
\begin{equation*}
\sup _{\theta \in \Theta} L_{n}(\theta)=\sup _{\alpha \in A_{0}}\left\{\frac{1}{2} U_{n}(\alpha)^{2}+o_{p}(1)\right\} \tag{31}
\end{equation*}
$$

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 $\alpha$, the term can be eliminated from the supremum;

$$
\begin{equation*}
\sup _{\theta \in \Theta} L_{n}(\theta)=\sup _{\alpha}\left\{\frac{1}{2} U_{n}(\alpha)^{2}\right\}+o_{p}(1) \tag{32}
\end{equation*}
$$

Furthermore, if the stochastic process $U_{n}$ converges "nicely" to a Gaussian process $W$ over $C$, the limit of the supremum of $\left|U_{n}\right|$ 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

$$
\begin{equation*}
\sup _{\theta \in \Theta} L_{n}(\theta)=\sup _{\alpha} \frac{1}{2} W^{2}+o_{p}(1) . \tag{33}
\end{equation*}
$$

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 $E_{P}\left[v_{i} v_{j}\right]$. 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 $C=\left\{v_{\alpha}(z)=\frac{\partial}{\partial \beta} f(z ;(\alpha, 0))\right\}$ 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 $\left\{v_{n}\right\}_{n=1}^{\infty}$ in $C$ such that $v_{n} \rightarrow 0$ in probability, then, for arbitrary $M>0$, we have

$$
\begin{equation*}
\lim _{n \rightarrow \infty} \operatorname{Prob}\left(\sup _{(\alpha, \beta)} L_{n}(\alpha, \beta) \leq M\right)=0 \tag{34}
\end{equation*}
$$

Proof. From Proposition 1 below, for arbitrary $\varepsilon>0$ and $K \in \mathbb{N}$, there exist $v\left(\alpha_{1}\right), \ldots, v\left(\alpha_{K}\right) \in C$ such that $\left|E\left[v\left(\alpha_{i}\right) v\left(\alpha_{j}\right)\right]\right|<\varepsilon$ 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 $W=\left(W_{1}, \ldots, W_{K}\right)$ be a random vector following the limiting normal distribution of $\left(U_{n}\left(v_{\alpha_{1}}\right), \ldots, U_{n}\left(v_{\alpha_{K}}\right)\right)$, and $\Sigma$ be the variance-covariance matrix of $W$. Because the absolute value of every off-diagonal element in $\Sigma$ is less than $\varepsilon$, 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

$$
\begin{align*}
P\left(\max _{1 \leq i \leq K}\left|W_{i}\right| \leq M\right) & \leq \int_{[-M, M]^{K}} \frac{1}{\sqrt{(2 \pi)^{K}|\Sigma|}} e^{-\frac{1}{2(1+(K-1) \varepsilon} W^{T} W} d W \\
& \leq \frac{(1+(K-1) \varepsilon)^{K / 2}}{|\Sigma|^{1 / 2}} \int_{[-M, M]^{K}} \frac{1}{(2 \pi)^{K / 2}} e^{-\frac{1}{2} u^{T} u} d u \\
& \leq\left(\frac{1+(K-1) \varepsilon}{1-(K-1) \varepsilon}\right)^{K / 2}\{\Phi(M)-\Phi(-M)\}^{K} \tag{35}
\end{align*}
$$

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

$$
\begin{equation*}
P\left(\max _{1 \leq i \leq K}\left|W_{i}\right| \leq M\right)<\delta \tag{36}
\end{equation*}
$$

The convergence of $\left(U_{n}\left(\alpha_{1}\right), \ldots, U_{n}\left(\alpha_{K}\right)\right)$ to $W$ means $\lim _{n \rightarrow \infty} P\left(\max _{i}\left|U_{n}\left(\alpha_{i}\right)\right| \leq\right.$ $M)=P\left(W \in[-M, M]^{K}\right)$. 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 $\left\{v_{n}\right\}_{n=1}^{\infty}$ be a sequence in $L^{2}(P)$ such that $\left\|v_{n}\right\|_{L^{2}(P)}=$ 1 for all $n$, and $v_{n} \rightarrow 0$ in probability. Then, there exists a subsequence $\left\{v_{n(k)}\right\}_{k=1}^{\infty}$ that satisfies

$$
\begin{equation*}
E_{P}\left|v_{n(k)} v_{n(h)}\right|<\varepsilon \tag{37}
\end{equation*}
$$

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 $\int Y^{2} d P \leq K$ and $\int X_{n}^{2} d P \leq K$, and $X_{n}$ converges to 0 in probability. Then, we have

$$
\begin{equation*}
\lim _{n \rightarrow \infty} E\left|Y X_{n}\right|=0 . \tag{38}
\end{equation*}
$$

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

For each $n \in \mathbb{N}$, a measurable set $A_{n}$ is defined by

$$
\begin{equation*}
A_{n}=\left\{\omega \in \Omega| | Y \left\lvert\,>\frac{\varepsilon}{3 \sqrt{K}} \quad\right. \text { and } \quad\left|X_{n}\right|>\frac{\varepsilon}{3 K}|Y|\right\} . \tag{39}
\end{equation*}
$$

Because $X_{n} \rightarrow 0$ in probability and $A_{n} \subset\left\{\left|X_{n}\right|>\frac{\varepsilon^{2}}{9 K^{3 / 2}}\right\}$, we can find $n_{0} \in \mathbb{N}$ such that for all $n \geq n_{0}$ we have $P\left(A_{n}\right)<\delta$, hence $\int_{A_{n}} Y^{2} d P<\frac{\varepsilon^{2}}{9 K}$.

Since $A_{n}^{c} \subset\left\{\omega| | Y \left\lvert\, \leq \frac{\varepsilon}{3 \sqrt{K}}\right.\right\} \cup\left\{\left.\omega| | X_{n}\left|\leq \frac{\varepsilon}{3 K}\right| Y \right\rvert\,\right\}$, we obtain for all $n \geq n_{0}$

$$
\begin{align*}
& \int\left|Y X_{n}\right| d P=\int_{A_{n}}\left|Y X_{n}\right| d P+\int_{A_{n}^{\varepsilon}}\left|Y X_{n}\right| d P \\
& \leq\left(\int_{A_{n}} Y^{2} d P\right)^{1 / 2}\left(\int_{A_{n}} X_{n}^{2} d P\right)^{1 / 2}+\int_{\left\{|Y| \leq \frac{\varepsilon}{3 \sqrt{K}}\right\}}\left|Y X_{n}\right| d P+\int_{\left\{\left|X_{n}\right| \leq \frac{\varepsilon}{3 K}|Y|\right\}}\left|Y X_{n}\right| d P \\
& <\frac{\varepsilon}{3 \sqrt{K}} \sqrt{K}+\frac{\varepsilon}{3 \sqrt{K}} \int\left|X_{n}\right| d P+\frac{\varepsilon}{3 K} \int|Y|^{2} d P \\
& \leq \frac{\varepsilon}{3}+\frac{\varepsilon}{3 \sqrt{K}} \cdot \sqrt{K}+\frac{\varepsilon}{3 K} \cdot K=\varepsilon \tag{40}
\end{align*}
$$

In the last line, we use the fact $\int\left|X_{n}\right| d P \leq\left(\int\left|X_{n}\right|^{2} d P\right)^{1 / 2} \leq \sqrt{K}$.

## 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 $r(y \mid u)$ to ensure the asymptotic normality conditions [AN] on the one-dimensional models.

## Conditions on noise model (NM2)

1. For any compact set $K \subset \mathbb{R}, \sup _{\xi, u \in K} E_{r(y \mid \xi)}|\log r(y \mid u)|$ is finite.
2. Let $h_{+}(y \mid s)$ and $h_{-}(y \mid s)$ be functions defined by

$$
\begin{equation*}
h_{+}(y \mid s)=\sup _{u \geq s} \log r(y \mid u) \quad \text { and } \quad h_{-}(y \mid s)=\sup _{u \leq-s} \log r(y \mid u), \tag{41}
\end{equation*}
$$

respectively. For any compact set $K \subset \mathbb{R}$ and $s \in \mathbb{R}, \sup _{\xi \in K} E_{r(y \mid \xi)}\left[h_{ \pm}(y \mid s)\right]$ is finite.
3. For an arbitrary compact set $K \subset \mathbb{R}$, there exist $\Delta_{+}, \Delta_{-} \subset \mathcal{Y}$ and $B>0$ such that

$$
\begin{array}{ll}
\lim _{s \rightarrow \infty} h_{+}(y \mid s)=-\infty & \text { for all } y \in \Delta_{+}, \\
\lim _{s \rightarrow \infty} h_{-}(y \mid s)=-\infty & \text { for all } y \in \Delta_{-}, \tag{43}
\end{array}
$$

and

$$
\begin{equation*}
\int_{\Delta_{ \pm}} r(y \mid \xi) d y \geq B \quad \text { for }{ }^{\forall} \xi \in K . \tag{44}
\end{equation*}
$$

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

$$
\begin{equation*}
\lim _{\rho \downarrow 0} \sup _{\substack{\xi \in K \\ u \in K}} E_{r(y \mid \xi)}\left[\sup _{\left|u^{\prime}-u\right| \leq \rho} \log r\left(y \mid u^{\prime}\right)\right]<\infty . \tag{45}
\end{equation*}
$$

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

$$
\begin{equation*}
\lim _{\rho \downarrow 0} \sup _{\xi \in K} E_{r(y \mid \xi)}\left[\sup _{\left|\xi^{\prime}-\xi\right| \leq \rho}\left|\frac{\partial^{3} \log r\left(y \mid \xi^{\prime}\right)}{\partial^{3} u}\right|\right]<\infty . \tag{46}
\end{equation*}
$$

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 $r(y \mid u)$ 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 $g(z ; \beta)=r\left(y \mid w_{0}(x)+\beta w(x)\right) q(x)$, satisfies the conditions [AN].

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

Since $H_{+}(z ; t)=h_{+}\left(y \mid w_{0}(x)+t w(x)\right)+\log q(x)$ and for any $t$ there exists $s_{0}$ such that $w_{0}(x)+t w(x) \geq s_{0}$ for all $x$, we have $E_{f_{0} \mu}\left[H_{+}(z ; t)\right] \leq$ $E_{Q}\left[E_{r\left(y \mid w_{0}(x)\right)}\left[h_{+}\left(y \mid s_{0}\right)\right]+\log q(x)\right]$. 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 $\left|w_{0}(x)\right| \leq M$. 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 $\lim _{t \rightarrow \infty} H_{+}(z ; t)=\lim _{t \rightarrow \infty} h_{+}\left(y \mid w_{0}(x)+\right.$ $t w(x))+\log q(x)=-\infty$, and $\int_{\mathcal{X} \times \Delta_{+}} f_{0}(z) d \mu=E_{Q}\left[\int_{\Delta_{+}} r\left(y \mid w_{0}(x)\right)\right] \geq B$.

From [NM2]-4 and the boundedness of $w(x)$, for any $\beta$ there exists $\rho_{0}>0$ and $C$ such that $E_{r\left(y \mid w_{0}(x)\right)}\left[\sup _{\left|\beta^{\prime}-\beta\right| \leq \rho} \log r\left(y \mid w_{0}(x)+\beta^{\prime} w(x)\right)\right] \leq C$ holds for all $\rho \in\left(0, \rho_{0}\right]$ and $x \in \mathbb{R}$. This shows the condition [AN]-4. By a similar argument, $[\mathrm{NM}]-5$ 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 $r(y \mid u)$, we have for arbitrary $M>0$,

$$
\begin{equation*}
\lim _{n \rightarrow \infty} \operatorname{Prob}\left(\sup _{\theta} L_{n}(\theta) \leq M\right)=0 \tag{47}
\end{equation*}
$$

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

$$
\begin{equation*}
\sigma(x ; \xi, h)=\frac{1}{2}\left\{1+s\left(-\frac{1}{2} \xi(x-h)\right)\right\}=\frac{1}{1+\exp \{\xi(x-h)\}}, \tag{48}
\end{equation*}
$$

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

$$
\begin{equation*}
g(z ; t, c, \beta)=r\left(y \mid \varphi_{0}(x)+\beta w(x ; t, c)\right) q(x), \tag{49}
\end{equation*}
$$

where

$$
\begin{equation*}
w(x ; t, c)=\frac{1}{\sqrt{B(t, c)}} \sigma\left(x ; c^{2}, t+\frac{1}{c}\right) \tag{50}
\end{equation*}
$$

and $B(t, c)$ is a normalizing constant of $L^{2}\left(f_{0} \mu\right)$ norm given by

$$
\begin{equation*}
B(t, c)=\int G\left(\varphi_{0}(x)\right) \sigma\left(x ; c^{2}, t+\frac{1}{c}\right)^{2} d Q(x) \tag{51}
\end{equation*}
$$

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

$$
\begin{equation*}
v(x, y ; t, c)=\frac{1}{\sqrt{B(t, c)}} \frac{\partial \log r\left(y \mid \varphi_{0}(x)\right)}{\partial u} \sigma\left(x ; c^{2}, t+\frac{1}{c}\right) \tag{52}
\end{equation*}
$$

Let $a$ be a positive number that satisfies $G\left(\varphi_{0}(x)\right) \geq a$ for all $x \in \mathbb{R}$. Such $a$ exists because of the continuity of $G(u)$ and the boundedness of $\varphi_{0}$. 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 $t_{0}=\inf \left\{t \in \mathbb{R} \mid F_{Q}(t)>0\right\} \in$ $\mathbb{R} \cup\{-\infty\}$, we have $F_{Q}(t)>0$ for all $t>t_{0}$, and $\lim _{t \downarrow t_{0}} F_{Q}(t)=0$.

Since $\sigma\left(x ; c^{2}, t+\frac{1}{c}\right)$ is bounded and converges to $\chi_{(-\infty, t]}(x)$ at every $x$ for $c \rightarrow+\infty$, by Lebesgue's dominated convergence theorem, we have $\lim _{c \rightarrow \infty} B(t, c)=\int_{-\infty}^{t} G\left(\varphi_{0}(x)\right) d Q(x) \geq a F_{Q}(t)$. Hence, for each $t$ we can find $c_{t}^{(1)}$ such that $\sqrt{B(t, c)} \geq \frac{1}{2} \sqrt{a F_{Q}(t)}$ for all $c \geq c_{t}^{(1)}$.

For any $t>t_{0}$ and $\delta>0$, there exists $c_{t}^{(2)}(\delta)>0$ such that $\sigma\left(x ; c^{2}, t+\right.$ $\left.\frac{1}{c}\right) \leq F_{Q}(t)$ for all $x \geq t+\delta$ and $c \geq c_{t}^{(2)}(\delta)$. Then, if a sequence $\left(t_{n}, \delta_{n}, c_{n}\right)$ is chosen so that $t_{n} \downarrow t_{0}, \delta_{n} \downarrow 0$, and $c_{n} \geq \max \left\{c_{t_{n}}^{(1)}, c_{t_{n}}^{(2)}\left(\delta_{n}\right)\right\}$, the inequality

$$
\begin{equation*}
\left|v\left(x, y ; t_{n}, c_{n}\right)\right| \leq \frac{2}{\sqrt{a}}\left|\frac{\partial \log r\left(y \mid \varphi_{0}(x)\right)}{\partial u}\right| \sqrt{F_{Q}\left(t_{n}\right)} \tag{53}
\end{equation*}
$$

holds for all $x \geq t_{n}+\delta_{n}$ and $y$. Because $F_{Q}\left(t_{n}\right) \rightarrow 0$ and $t_{n}+\delta_{n} \downarrow t_{0}$ for $n \rightarrow \infty$, the sequence $v\left(x, y ; t_{n}, c_{n}\right)$ 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

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

where $A(\xi, h, t)$ is a normalization constant of $L^{2}\left(f_{0} \mu\right)$ norm given by

$$
\begin{align*}
A(\xi, h, t) & =E_{f_{0} \mu}\left[\left(\frac{\partial \log r\left(y \mid \varphi_{0}(x)\right)}{\partial u} \frac{s(\xi(x-t+h))-s(\xi(x-t-h))}{2}\right)^{2}\right] \\
& =E_{Q}\left[G\left(\varphi_{0}(x)\right) \frac{1}{4}\{s(\xi(x-t+h))-s(\xi(x-t-h))\}^{2}\right] . \tag{55}
\end{align*}
$$

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

$$
\begin{equation*}
\psi(x ; \xi, h, t, \beta)=\varphi_{0}(x)+\beta w(x ; \xi, h, t) \tag{56}
\end{equation*}
$$

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

$$
\begin{equation*}
v(z ; \xi, h, t)=\frac{\partial \log r\left(y \mid \varphi_{0}(x)\right)}{\partial u} w(x ; \xi, h, t) \tag{57}
\end{equation*}
$$

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

$$
\begin{equation*}
A\left(\xi_{n}, h_{n}, 0\right) \geq \int_{-\frac{1}{2} h_{n}}^{\frac{1}{2} h_{n}} G\left(\varphi_{0}(x)\right)\left(\frac{1}{2}\right)^{2} q(x) d x \geq \frac{a}{4} h_{n} \tag{58}
\end{equation*}
$$

Thus, for all $x$ with $|x| \geq \frac{3}{2} h_{n}$, we have

$$
\begin{align*}
\left|v\left(z ; \xi_{n}, h_{n}, 0\right)\right| & =\left|\frac{\partial \log r\left(y \mid \varphi_{0}(x)\right)}{\partial u}\right| \frac{1}{\sqrt{A\left(\xi_{n}, h_{n}, 0\right)}} \frac{1}{2}\{s(\xi(x+h))-s(\xi(x-h))\} \\
& \leq \frac{2 \sqrt{h_{n}}}{\sqrt{a}}\left|\frac{\partial \log r\left(y \mid \varphi_{0}(x)\right)}{\partial u}\right| \tag{59}
\end{align*}
$$

For all $x \neq 0$, the sequence $v\left(z ; \xi_{n}, h_{n}, 0\right)$ 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

$$
\begin{equation*}
M(I)=E_{f_{0} \mu}\left[\left(\frac{\partial \log r\left(y \mid \varphi_{0}(x)\right)}{\partial u}\right)^{2} \chi_{I}(x)\right]=\int_{I} G\left(\varphi_{0}(x)\right) q(x) d x \tag{60}
\end{equation*}
$$

and a function $u_{I}(z)$ by

$$
\begin{equation*}
u_{I}(z)=\frac{1}{\sqrt{M(I)}} \frac{\partial \log r\left(y \mid \varphi_{0}(x)\right)}{\partial u} \chi_{I}(x), \tag{61}
\end{equation*}
$$

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) $\left\|v(z ; \xi, h, t)-u_{I}(z)\right\|_{L^{2}\left(f_{0} \mu\right)} \leq \varepsilon$, 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 $\sigma(x ; \xi, h)=$ $\frac{1}{2}\left\{s\left(\xi_{n}\left(x+h_{n}\right)\right)-s\left(\xi_{n}\left(x-h_{n}\right)\right)\right\}$ for abbreviation.

In a similar way to the argument before the lemma, there exist sequences $h_{n} \downarrow c$ and $\xi_{n} \rightarrow \infty$ such that
[1] $\sigma\left(x ; \xi_{n}, h_{n}\right) \leq 2$ for all $x \in \mathbb{R}$,
[2] $\left|\sigma\left(x ; \xi_{n}, h_{n}\right)-\chi_{I}(x)\right| \rightarrow 0$ for all $x \in \mathbb{R}$, and
[3] $\sigma\left(x ; \xi_{n}, h_{n}\right) \geq \frac{1}{2}$ for all $x \in I$.
From [1], [2] and the boundedness of $G\left(\varphi_{0}(x)\right)$, by Lebesgue's dominated convergence theorem, we obtain

$$
\begin{equation*}
\left\|\frac{\partial \log r\left(y \mid \varphi_{0}(x)\right)}{\partial u} \sigma\left(x ; \xi_{n}, h_{n}\right)-\frac{\partial \log r\left(y \mid \varphi_{0}(x)\right)}{\partial u} \chi_{I}(x)\right\|_{L^{2}\left(f_{0} \mu\right)} \quad \rightarrow 0 \tag{62}
\end{equation*}
$$

as $n$ goes to infinity. Eq.(62) means also $A\left(\xi_{n}, h_{n}\right) \rightarrow M(I)>0$. Then, a simple argument shows the assertion (iii). From eq. (62), there exists $n_{0} \in \mathbb{N}$ such that $\frac{1}{2} M(I) \leq A\left(\xi_{n}, h_{n}\right) \leq 2 M(I)$ 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
$r(y \mid u)$. 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, $h_{ \pm}(y \mid s)$ 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 $A_{i}, \delta_{i}(i=$ 1,2 ) and $T_{0}$ exist so that

$$
\begin{equation*}
\tau(s) \geq A_{1} s^{\delta_{1}} \quad \text { for } 0 \leq s \leq T_{0}, \quad \tau(s) \geq A_{2} s^{\delta_{2}} \quad \text { for } s>T_{0} \tag{63}
\end{equation*}
$$

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

$$
\begin{equation*}
E_{r(u \mid \xi)}\left[\log \frac{r(y \mid \xi)}{r(y \mid u)}\right] \geq \tau(|u-\xi|), \tag{64}
\end{equation*}
$$

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

$$
\begin{array}{ll}
\limsup _{s \rightarrow \infty} \frac{h_{+}(y \mid s)}{s^{\gamma}}<0 & \text { for all } y \in \Delta_{+}, \\
\limsup _{s \rightarrow \infty} \frac{h_{-}(y \mid s)}{s^{\gamma}}<0 & \text { for all } y \in \Delta_{-}, \tag{66}
\end{array}
$$

and

$$
\begin{equation*}
\int_{\Delta_{ \pm}} r(y \mid \xi) d y \geq B \quad \text { for }{ }^{\forall} \xi \in K \tag{67}
\end{equation*}
$$

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

$$
\begin{equation*}
E_{r(y \mid \xi)}\left|h_{ \pm}(y \mid s)\right|^{2} \leq \ell_{1}(\xi) s^{\lambda} . \tag{68}
\end{equation*}
$$

4. There exist a continuous function $\ell_{2}(\xi)$ and $\nu>0$ such that for arbitrary $R \geq 1$

$$
\begin{equation*}
E_{r(y \mid \xi)}\left[\left.\sup _{|u| \leq R} \frac{\partial \log r(y \mid u)}{\partial u}\right|^{2}\right] \leq \ell_{2}(\xi) R^{\nu} . \tag{69}
\end{equation*}
$$

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

$$
\begin{equation*}
\sup _{\xi, u \in K} E_{r(y \mid \xi)}\left[\left|\frac{\partial^{2} \log r(y \mid u)}{\partial u^{2}}\right|^{2}\right]<\infty \quad \text { and } \quad \sup _{u \in K} E_{r(y \mid u)}\left[\left|\frac{\partial \log r(y \mid u)}{\partial u}\right|^{3}\right]<\infty . \tag{70}
\end{equation*}
$$

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

$$
\begin{equation*}
\lim _{\rho \downarrow 0} \sup _{\xi \in K} E_{r(y \mid \xi)}\left[\sup _{\left|\xi^{\prime}-\xi\right| \leq \rho}\left|\frac{\partial^{3} \log r\left(y \mid \xi^{\prime}\right)}{\partial u^{3}}\right|^{2}\right]<\infty \tag{71}
\end{equation*}
$$

The conditions [NM3] are satisfied by the Gaussian noise model and the binary output model. Consider the Gaussian distribution with variance one: $r(y \mid u)=\frac{1}{\sqrt{2 \pi}} \exp \left\{-\frac{1}{2}(y-u)^{2}\right\}$, 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 $h_{+}(y \mid s)=-\frac{1}{2}(y-s)^{s}-\log \sqrt{2 \pi}$ for $y<s$ and $h_{+}(y \mid s)=-\log \sqrt{2 \pi}$ for $y \geq s$, and a similar form for $h_{-}(y \mid s)$. Then, the condition 2 can be verified for any bounded interval $\Delta_{ \pm}$and $\gamma=2$, and the condition 3 is trivial. Because $\frac{\partial \log r(y \mid u)}{\partial u}=y-u$, for any $R>0$, we see $\sup _{|u| \leq R}\left|\frac{\partial \log r(y \mid u)}{\partial u}\right|^{2}$ is $(y+R)^{2}$ for $y \geq 0$ and $(y-R)^{2}$ for $y<0$. Hence, the condition 4 is satisfied for $\nu=2$.

For the binary output model $r(y \mid u)=\frac{e^{y u}}{1+e^{u}}$, the conditions 5 and 6 are trivial. The KL-divergence is $E_{r(y \mid \xi)}[\log r(y \mid \xi)-\log r(y \mid u)]=\frac{1}{1+e^{\xi}}(u-\xi)+$ $\log \left(1+e^{-u}\right)-\log \left(1+e^{-\xi}\right)=\frac{1}{1+e^{-\xi}}(\xi-u)+\log \left(1+e^{u}\right)-\log \left(1+e^{\xi}\right)$. For any $C>0$ and $\xi \in[-C, C]$, we have $\left.E_{r(y \mid \xi)}[\log r(y \mid \xi)-\log r(y \mid u)]>\frac{1}{1+e^{C}} \right\rvert\, u-$ $\xi \mid-\log \left(1+e^{C}\right)$. Then, for any $u$ with $|u-\xi| \geq 2\left(1+e^{C}\right) \log \left(1+e^{C}\right)$, a lower bound $E_{r(y \mid \xi)}[\log r(y \mid \xi)-\log r(y \mid u)]>\frac{1}{2\left(1+e^{C}\right)}|u-\xi|$ is obtained. By Taylor expansion of KL-divergence, there exists $a>0$ such that $E_{r(y \mid \xi)}[\log r(y \mid \xi)-$ $\log r(y \mid u)] \geq a(u-\xi)^{2}$ for all $\xi \in[-C, C]$ and $|u-\xi| \leq 2\left(1+e^{C}\right) \log \left(1+e^{C}\right)$. Thus, we can choose $T_{0}=2\left(1+e^{C}\right) \log \left(1+e^{C}\right)$, and define $\tau(s)$ by $\frac{1}{2\left(1+e^{C}\right)} s$ for $0 \leq s \leq T_{0}$ and $a(u-\xi)^{2}$ for $s>T_{0}$. This shows the condition 1. Since $h_{+}(y \mid s)=-(1-y) \log \left(1+e^{s}\right)$ and $h_{-}(y \mid s)=y \log \left(1+e^{s}\right)$, The condition 3 is straightforward. For the condition 2 , choose $\Delta_{+}=\{0\}$ and $\Delta_{-}=\{1\}$. Then, $\lim _{s \rightarrow \infty} \frac{h_{ \pm}(y \mid s)}{s}=-1$ for all $y \in \Delta_{ \pm}$. The condition 4 is trivial, since $\frac{\partial \log r(y \mid u)}{\partial u}=-\log \left(1+e^{-u}\right)$ is bounded.
Theorem 4. Let $r(y \mid u)$ 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 $f_{0}(x, y)=r\left(y \mid \varphi_{0}(x)\right) q(x)$. For a closed interval $I$, a non-negative value $M(I)$ is defined by

$$
\begin{equation*}
M(I)=\iint\left(\frac{\partial \log r\left(y \mid \varphi_{0}(x)\right)}{\partial u}\right)^{2} \chi_{I}(x) r\left(y \mid \varphi_{0}(x)\right) q(x) d \mu_{y} d x \tag{72}
\end{equation*}
$$

and a function $u_{I}(z)$ is defined by

$$
\begin{equation*}
u_{I}(z)=\frac{1}{\sqrt{M(I)}} \frac{\partial \log r\left(y \mid \varphi_{0}(x)\right)}{\partial u} \chi_{I}(x) \tag{73}
\end{equation*}
$$

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

$$
\begin{equation*}
v(z ; \alpha)=\frac{\partial \log r\left(y \mid \varphi_{0}(x)\right)}{\partial u} w(x ; \alpha) \tag{74}
\end{equation*}
$$

satisfies $\|v(z ; \alpha)\|_{L^{2}\left(f_{0} \mu\right)}=1$ 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) $\iint\left|v(z ; \alpha)-u_{I}(z)\right|^{2} r\left(y \mid \varphi_{0}(x)\right) q(x) d \mu_{y} d x<\varepsilon$.

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

$$
\begin{equation*}
\liminf _{n \rightarrow \infty} \operatorname{Prob}\left(\frac{\sup _{\alpha, \beta} L_{n}(\alpha, \beta)}{\log n} \geq \delta\right) \quad>0 \tag{75}
\end{equation*}
$$

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

$$
\begin{equation*}
\liminf _{n \rightarrow \infty} \operatorname{Prob}\left(\frac{\sup _{\theta} L_{n}(\theta)}{\log n} \geq \delta\right)>0 \tag{76}
\end{equation*}
$$

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 $\left\{I_{k}^{[m]} \mid k=1, \ldots, m\right\}$ of $[-K, K]$ such that $I_{k}^{[m]}$,s are closed intervals with disjoint interiors, and $M\left(I_{k}^{[m]}\right)=\frac{M(\mathbb{R})}{2 m}$ for all $k$. For each $k(1 \leq k \leq m$, a function $u_{k}^{[m]}(z)$ is defined by
$u_{k}^{[m]}(z)=\left.\frac{\partial}{\partial \beta} \log r\left(y \left\lvert\, \varphi_{0}(x)+\beta \frac{1}{\sqrt{M\left(I_{k}^{[m]}\right)}} \chi_{I_{k}}(x)\right.\right)\right|_{\beta=0}=\sqrt{\frac{2 m}{M(\mathbb{R})}} \frac{\partial \log r\left(y \mid \varphi_{0}(x)\right)}{\partial u} \chi_{I_{k}^{[m]}}(x)$.

This is a tangent vector of the locally conic one-dimensional model $r\left(y \mid \varphi_{0}(x)+\right.$ $\left.\beta \frac{1}{\sqrt{M\left(I_{k}^{[m]}\right)}} \chi_{I_{k}^{[m]}}(x)\right) q(x)$ 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 $H_{3}(x)=E_{f_{0} \mu}\left|\frac{\partial \log r\left(y \mid \varphi_{0}(x)\right)}{\partial u}\right|^{3}$. By the assumptions [NM1]-3, [NM3]-5, and the boundedness of $\varphi_{0}(x)$, there exists $B>0$ such that $H_{3}(x) \leq B G\left(\varphi_{0}(x)\right)$ for all $x \in[-K, K]$. Then, we obtain

$$
\begin{aligned}
E_{f_{0} \mu}\left|u_{k}^{[m]}(z)\right|^{3} & =\frac{1}{M\left(I_{k}^{[m]}\right)^{3 / 2}} \int H_{3}(x) \chi_{I_{k}^{[m]}} q(x) d x \\
& \leq \frac{B}{M\left(I_{k}^{[m]}\right)^{3 / 2}} \int G\left(\varphi_{0}(x)\right) \chi_{I_{k}^{[m]}} q(x) d x=\frac{\sqrt{2} B}{\sqrt{M(\mathbb{R})}} \sqrt{m} .(78)
\end{aligned}
$$

Let $P_{n}$ and $Q_{m}$ be the probability distribution of the $m$-dimensional random vector $\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_{1}^{[m]}\left(Z_{i}\right), \ldots, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_{m}^{[m]}\left(Z_{i}\right)\right)$, and of the $m$ dimensional normal distribution $N\left(0, I_{m}\right)$, respectively. Let $\mathcal{D}$ denote the family of all the convex measurable sets on $\mathbb{R}^{m}$. The Berry-Esseen-type inequality ([18]) gives

$$
\begin{equation*}
\sup _{\Delta \in \mathcal{D}}\left|P_{n}(\Delta)-Q_{m}(\Delta)\right| \leq \frac{L m^{4}}{\sqrt{n}} \sum_{1 \leq k \leq m} E_{f_{0} \mu}\left|u_{k}^{[m]}(Z)\right|^{3} \tag{79}
\end{equation*}
$$

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

$$
\begin{align*}
\left\lvert\, \operatorname{Prob}\left(\max _{1 \leq k \leq m} \left\lvert\, \frac{1}{\sqrt{n}}\right.\right.\right. & \left.\sum_{i=1}^{n} u_{k}^{[m]}\left(Z_{i}\right) \mid>\nu \sqrt{\log m}\right) \\
& -\operatorname{Prob}\left(\max _{1 \leq k \leq m}\left|V_{k}\right|>\nu \sqrt{\log m}\right) \left\lvert\, \leq C^{\prime} \frac{m^{11 / 2}}{\sqrt{n}}\right. \tag{80}
\end{align*}
$$

where $V_{k}(1 \leq k \leq m)$ are i.i.d sample from the standard normal distribution, and $C^{\prime}$ is a constant independent of $n$ and $m$. Let $[x]$ denote the largest integer that is not larger than $x$. If we set $m=\left[n^{\gamma}\right]$ 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 $\left\{\max _{1 \leq k \leq m}\left|V_{k}\right|>\nu \sqrt{\log m}\right\}$ converges to 1 for $0<\nu<\sqrt{2}$. Thus, for such $\nu$ and arbitrary $\varepsilon>0$, we have

$$
\begin{equation*}
\operatorname{Prob}\left(\max _{1 \leq k \leq m}\left|\frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_{k}^{[m]}\left(Z_{i}\right)\right|^{2}>\nu^{2} \gamma \log n\right) \quad>1-\varepsilon, \tag{81}
\end{equation*}
$$

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) $E_{f_{0} \mu}\left|v_{k}^{[m]}(z)-u_{k}^{[m]}(z)\right|^{2}<\frac{\varepsilon \delta^{2}}{m}$, 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 $\varepsilon, m$ and $k$. Then, using Chebyshev's inequality, we obtain

$$
\begin{align*}
& \operatorname{Prob}\left(\left|\max _{1 \leq k \leq m}\right| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_{k}^{[m]}\left(Z_{i}\right)\left|-\max _{i \leq k \leq m}\right| \frac{1}{\sqrt{n}} v_{k}^{[m]}\left(Z_{i}\right)| | \geq \delta\right) \\
& \leq \operatorname{Prob}\left(1 \leq{ }^{\exists} k \leq m, \quad\left|\frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_{k}^{[m]}\left(Z_{i}\right)-\frac{1}{\sqrt{n}} v_{k}^{[m]}\left(Z_{i}\right)\right| \geq \delta\right) \\
& \leq m \operatorname{Prob}\left(\left|\frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_{k}^{[m]}\left(Z_{i}\right)-\frac{1}{\sqrt{n}} v_{k}^{[m]}\left(Z_{i}\right)\right| \geq \delta\right) \\
& \leq m \frac{E_{f_{0} \mu}\left|u_{k}^{[m]}(z)-v_{k}^{[m]}(z)\right|^{2}}{\delta^{2}}<\varepsilon . \tag{82}
\end{align*}
$$

Combining eqs.(81) and (82), there exists $\gamma^{\prime}>0$ such that

$$
\begin{equation*}
\operatorname{Prob}\left(\max _{1 \leq k \leq m}\left|\frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_{k}^{[m]}\left(Z_{i}\right)\right|^{2}>\gamma^{\prime} \log n\right)>1-2 \varepsilon \tag{83}
\end{equation*}
$$

holds for sufficiently large $n$.
Since $M(I)=\int G\left(\varphi_{0}(x)\right) q(x) d x$, by the assumption [NM1]-3 and the boundedness of $\varphi_{0}(x)$, there exist $c, d>0$ such that $\frac{c}{m} \leq Q\left(I_{k}^{[m]}\right) \leq \frac{d}{m}$ 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 $\gamma$
satisfying $0<\gamma<\gamma_{1}$ and $m=\left[n^{\gamma}\right]$, the following asymptotic expansion of the likelihood ratio holds;

$$
\begin{equation*}
\max _{1 \leq k \leq m} \sup _{\beta} \sum_{i=1}^{n} \log \frac{f_{k}^{[m]}\left(Z_{i} ; \beta\right)}{f_{0}\left(Z_{i}\right)}=\left\{\max _{1 \leq k \leq m} \frac{1}{2}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_{k}^{[m]}\left(Z_{i}\right)\right)^{2}\right\}\left(1+o_{p}(1)\right), \tag{84}
\end{equation*}
$$

where $f_{k}^{[m]}(z ; \beta)=r\left(y \mid \varphi_{0}(x)+\beta w_{k}^{[m]}(x)\right) q(x)$. 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 $\mathbb{R}, r(y \mid \xi)$ be a probability density function on a measure space $\left(\mathcal{Y}, \mathcal{B}_{y}, \mu_{y}\right)$ with one-dimensional parameter $\xi$, which satisfies the assumptions [NM1], [NM2], and [NM3], and $Q=q(x) d x$ 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 $\left(X_{1}, Y_{1}\right),\left(X_{2}, Y_{2}\right), \ldots$ with probability $r\left(y \mid \varphi_{0}(x)\right) q(x) \mu_{\mathbb{R}} \mu_{y}$. 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

$$
\mathcal{W}_{m}=\{w: \mathbb{R} \rightarrow \mathbb{R} \mid 0<w(x) \leq a \sqrt{m} \text { for all } x \in \mathbb{R}, \text { and there exists }
$$

a closed interval $I \subset D$ such that $\frac{c}{m} \leq Q(I) \leq \frac{d}{m}$ and $w(x) \geq b \sqrt{m}$ on $\left.I\right\}$.
For $\gamma>0$, let $m_{n}$ be a natural number given by $m_{n}=\left[n^{\gamma}\right]$ for $n \in \mathbb{N}$, and $\mathcal{G}_{\gamma}$ be a family of sequences $\left\{\left\{w_{k}^{(n)}\right\}_{n \in \mathbb{N}, 1 \leq k \leq m_{n}} \mid w_{k}^{(n)} \in \mathcal{W}_{m_{n}}\right\}$. Then, there exists $\gamma_{0}>0$ such that for any $\gamma \leq \gamma_{0}$ and $\left\{w_{k}^{(n)}\right\} \in \mathcal{G}_{\gamma}$, we obtain, as $n$ goes to infinity,
$\max _{1 \leq k \leq m_{n}} \sup _{\beta} \sum_{i=1}^{n} \log \frac{r\left(Y_{i} \mid \varphi_{0}\left(X_{i}\right)+\beta w_{k}^{(n)}\left(X_{i}\right)\right)}{r\left(Y_{i} \mid \varphi_{0}\left(X_{i}\right)\right)}=\left\{\max _{1 \leq k \leq m_{n}} \frac{1}{2}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_{k}^{(n)}\left(X_{i}, Y_{i}\right)\right)^{2}\right\}\left(1+o_{p}(1)\right)$,
where $u_{k}^{(n)}(x, y)$ is a tangent vector given by

$$
\begin{equation*}
u_{k}^{(n)}(x, y)=\left.\frac{\partial \log r\left(y \mid \varphi_{0}(x)+\beta w_{k}^{(n)}(x)\right)}{\partial \beta}\right|_{\beta=0}=\frac{\partial \log r\left(y \mid \varphi_{0}(x)\right)}{\partial \xi} w_{k}^{(n)}(x) \tag{86}
\end{equation*}
$$

First, we will establish the uniform consistency of the maximum likelihood estimator of $\beta$.

Lemma 4. Let $r(y \mid \xi), q(x), \varphi_{0}(x)$, 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, $\mathcal{H}_{m}=$ $\left\{\left\{w_{k}\right\}_{k=1}^{m} \mid w_{k} \in \mathcal{W}_{m}\right\}$. For $\Xi=\left\{w_{k}^{[m]}\right\}_{k=1}^{m} \in \mathcal{H}_{m}$, we write $\widehat{\beta}_{k}^{[m]}(\Xi)$ for the maximum likelihood estimator of the model $r\left(y \mid \varphi_{0}(x)+\beta w_{k}^{[m]}(x)\right) q(x)$, given an i.i.d. sample $\left(X_{1}, Y_{1}\right), \ldots,\left(X_{n}, Y_{n}\right)$ with probability $r\left(y \mid \varphi_{0}(x)\right) q(x) \mu_{\mathbb{R}} \mu_{y}$. Then, there exist positive constants $A, \gamma$, and $\nu$ such that the inequality

$$
\begin{equation*}
\operatorname{Prob}\left(\max _{1 \leq k \leq m}\left|\widehat{\beta}_{k}^{[m]}(\Xi)\right| \geq \varepsilon\right) \leq A \frac{m^{\gamma}}{n \varepsilon^{\nu}} \tag{87}
\end{equation*}
$$

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 $f^{[m]}(z ; \beta)=r\left(y \mid \varphi_{0}(x)+\beta w(x)\right) q(x)$ 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

$$
\begin{gather*}
H_{+}^{[m]}(z ; t)=\sup _{\beta \geq t} \log f^{[m]}(z ; \beta), \quad H_{-}^{[m]}(z ; t)=\sup _{\beta \leq-t} \log f^{[m]}(z ; \beta),  \tag{88}\\
\text { and } g^{[m]}(z ; \beta, \rho)=\sup _{\left|\beta^{\prime}-\beta\right| \leq \rho} \log f^{[m]}\left(z ; \beta^{\prime}\right), \tag{89}
\end{gather*}
$$

respectively. A constant $M$ is taken so that $\left|\varphi_{0}(x)\right| \leq M$ for all $x \in \mathbb{R}$. The true probability is denoted by $f_{0}(z) \mu$, where $f_{0}(z)=r\left(y \mid \varphi_{0}(x)\right) q(x)$ and $\mu=\mu_{y} d x$.
(a) Bound of $E_{f_{0} \mu}\left[H_{ \pm}^{[m]}(z ; t)\right]$. 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

$$
\begin{equation*}
E_{f_{0} \mu}\left[H_{ \pm}^{[m]}(z ; t)\right] \leq E_{f_{0} \mu}\left[\log f_{0}(z)\right]-3 \delta \tag{90}
\end{equation*}
$$

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 $\int_{\Delta} r(y \mid u) d y \geq \Gamma_{1}$ for all $u \in[-M, M]$ and $\log r(y \mid u) \leq-\Lambda_{1}|u|^{\gamma_{1}}$ for all $y \in \Delta$ and $u \geq R_{0}$. If we set $R=\frac{R_{0}+M}{b \sqrt{m}}$, we have $\varphi_{0}(x)+t w(x) \geq R_{0}$ 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

$$
\begin{align*}
H_{+}^{[m]}(z ; t) & =h\left(y \mid \varphi_{0}(x)+t w(x)\right)+\log q(x) \\
& \leq-\Lambda_{1}\left|\varphi_{0}(x)+t w(x)\right|^{\gamma_{1}}+\log q(x) \\
& \leq-\Lambda_{1}(b \sqrt{m} t-M)^{\gamma_{1}}+\log q(x) \tag{91}
\end{align*}
$$

is obtained. From the assumption [NM3]-3, there exists $F_{1}>0$ such that $E_{r(y \mid \xi)}\left[h_{+}(y \mid-M)\right] \leq F_{1}$ for all $\xi \in[-M, M]$. Since $\varphi_{0}(x)+\beta w(x) \geq-M$ for $\beta>0$, we see that for $t \geq 0$

$$
\begin{equation*}
E_{f_{0} \mu}\left[H_{+}^{[m]}(z ; t)\right] \leq E_{Q}\left[E_{r\left(y \mid \varphi_{0}(x)\right)}\left[h_{+}(y \mid-M)\right]\right] \leq F_{1} . \tag{92}
\end{equation*}
$$

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

$$
\begin{align*}
E_{f_{0} \mu}\left[H_{+}^{[m]}(z ; t)\right]= & \int_{I \times \Delta} H_{+}(z ; t) f_{0}(z) d \mu+\int_{\mathcal{X} \times \mathcal{Y} \backslash I \times \Delta} H_{+}(z ; t) f_{0}(z) d \mu d x d y \\
\leq & -\Lambda_{1}(a \sqrt{m} t-M)^{\gamma_{1}} \int_{I} \int_{\Delta} r\left(y \mid \varphi_{0}(x)\right) d \mu_{y} q(x) d x \\
& +E_{f_{0} \mu}\left[\left(H_{+}(z ; t)\right)_{+}\right]+E_{Q}|\log q(x)| \\
\leq & -\Lambda_{1} \Gamma_{1} \frac{c}{m}(b \sqrt{m} t-M)^{\gamma_{1}}+F_{1}+E_{Q}|\log q(x)| \tag{93}
\end{align*}
$$

From this inequality and the fact $E_{f_{0} \mu}\left[\log f_{0}(z)\right]<\infty$, 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 $\overline{\lambda_{1}>0}$ there exist $B_{2}>0$ and $\lambda_{2}>0$ such that the inequality

$$
\begin{equation*}
E_{f_{0} \mu}\left[\left|H_{ \pm}^{[m]}\left(z ; B_{1} m^{\lambda_{1}}\right)\right|^{2}\right] \leq B_{2} m^{\lambda_{2}} \tag{94}
\end{equation*}
$$

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 $\left|\ell_{i}(\xi)\right| \leq \Gamma_{2}$ for all $\xi \in[-M, M]$. Then, the bound

$$
\begin{align*}
E_{f_{0} \mu}\left[\left|H_{ \pm}^{[m]}(z ; t)\right|^{2}\right] & =E_{Q}\left[E_{r\left(y \mid \varphi_{0}(x)\right)}\left[\left|h_{+}\left(y \mid \varphi_{0}(x) \pm t \beta w(x)\right)\right|^{2}\right]\right] \\
& \leq E_{Q}\left[\ell_{1}\left(\varphi_{0}(x)\right)^{2}\left|\varphi_{0}(x) \pm t \beta w(x)\right|^{2 \lambda}\right] \\
& \leq \Gamma_{2}^{2}(M+t a \sqrt{m})^{2 \lambda} \tag{95}
\end{align*}
$$

is obtained. The above assertion is straightforward, if we choose sufficiently large $B_{2}$ and $\lambda_{2} \geq 2 \lambda\left(\lambda_{1}+\frac{1}{2}\right)$.
(c) Bounds of $E_{f_{0} \mu}\left[g^{[m]}(z ; \beta, \rho)\right]$. We will show the fact that there exist $\overline{A_{3}, B_{3}>0, \lambda_{3}>0 \text {, and } \gamma_{3}>0}$, such that for arbitrary $R \geq 1, \delta>0$ and $\beta \in[-R, R]$ the inequalities

$$
\begin{equation*}
E_{f_{0} \mu}\left[g^{[m]}(z ; \beta, \rho)\right] \leq E_{f_{0} \mu}\left[\log f^{[m]}(z ; \beta)\right]+\delta \tag{96}
\end{equation*}
$$

and

$$
\begin{equation*}
E_{f_{0} \mu}\left[\left|g^{[m]}(z ; \beta, \rho)\right|^{2}\right] \leq B_{3}(\sqrt{m} R)^{\lambda_{3}}+2 \delta^{2} \tag{97}
\end{equation*}
$$

for $\rho \leq A_{3} \delta(\sqrt{m} R)^{-\gamma_{3}}$.
Because $\left|\varphi_{0}(x)+\beta w(x)\right| \leq M+a \sqrt{m} R$ for $\beta \in[-R, R]$, from the assumption [NM3]-4, we can find $\nu_{3}>0, \Psi(y)$, and $\ell_{2}(\xi)$ such that

$$
\begin{equation*}
\left|\log f^{[m]}(z ; \beta)-\log f^{[m]}\left(z ; \beta^{\prime}\right)\right| \leq \Psi(y) w(x)\left|\beta-\beta^{\prime}\right| \tag{98}
\end{equation*}
$$

and

$$
\begin{equation*}
E_{r(y \mid \xi)}\left[|\Psi(y)|^{2}\right] \leq \ell_{2}(\xi)(M+a \sqrt{m} R)^{\nu_{3}} \tag{99}
\end{equation*}
$$

are satisfied. Eq.(98) means $\left|g^{[m]}(z ; \beta, \rho)-\log f^{[m]}(z ; \beta)\right| \leq a \sqrt{m} \rho \Psi(y)$. Since $\Gamma_{3}=E_{Q}\left[\ell_{2}\left(\varphi_{0}(x)\right)\right]<\infty$, we obtain

$$
\begin{align*}
E_{f_{0} \mu}\left[g^{[m]}(z ; \beta, \rho)\right] & \leq E_{f_{0} \mu}\left[\log f^{[m]}(z ; \beta)\right]+\rho a \sqrt{m}\left\{\Gamma_{3}(M+a \sqrt{m} R)^{\nu_{3}}\right\}^{1 / 2} \\
& \leq E_{f_{0} \mu}\left[\log f^{[m]}(z ; \beta)\right]+F_{3} \rho(\sqrt{m} R)^{\frac{\nu_{3}}{2}+1} \tag{100}
\end{align*}
$$

for $F_{3}=a \Gamma_{3}^{1 / 2}(M+a)^{\nu_{3} / 2}$. If we choose $\rho$ as $\left.\rho \leq \delta\left\{F_{3}(\sqrt{m} R)^{\frac{\nu_{3}}{2}+1}\right)\right\}^{-1}$, the first assertion is satisfied.

For this choice of $\rho$, noting that $a^{2} m \rho^{2} E_{f_{0} \mu}\left[\Psi(y)^{2}\right] \leq \delta^{2}$, we further obtain

$$
\begin{align*}
E_{f_{0} \mu}\left[\left|g^{[m]}(z ; \beta, \rho)\right|^{2}\right] & \leq 2 E_{f_{0} \mu}\left|\log f^{[m]}(z ; \beta)\right|^{2}+2 E_{f_{0} \mu}\left[\Psi(y)^{2}\right] a^{2} m \rho^{2} \\
& \leq 4 E_{f_{0} \mu}\left[\left|\log f_{0}(z)\right|^{2}+\beta^{2} a^{2} m \Psi(y)^{2}\right]+2 E_{f_{0} \mu}\left[\Psi(y)^{2}\right] a^{2} m \rho^{2} \\
& \leq 4 E_{f_{0} \mu}\left|\log f_{0}(z)\right|^{2}+4 a^{2} m(M+a \sqrt{m} R)^{2}\left\{\Gamma_{3}(M+a \sqrt{m} R)^{\nu_{3}}\right\}+2 \delta^{2} \\
& \leq 4 E_{f_{0} \mu}\left|\log f_{0}(z)\right|^{2}+\Lambda_{3}(\sqrt{m} R)^{\nu_{3}+4}+2 \delta^{2} \tag{101}
\end{align*}
$$

for sufficiently large $\Lambda_{3}$, which depends only on $a, M$, and $\Gamma_{3}$. Because $E_{f_{0} \mu}\left|\log f_{0}(z)\right|^{2}<\infty$ from the assumption [NM3]-3, the second assertion is obtained.
(d) Lower bound of KL-divergence. We will show that there exist $B_{4}>0$, $\overline{\lambda_{4}>0}$, and $\nu_{4} \in \mathbb{R}$ such that the bound

$$
\begin{equation*}
E_{f_{0} \mu}\left[\log f^{[m]}(z ; \beta)\right] \leq E_{f_{0} \mu}\left[\log f_{0}(z)\right]-B_{4} m^{\nu_{4}} \varepsilon^{\lambda_{4}} \tag{102}
\end{equation*}
$$

holds for arbitrary $0<\varepsilon<1, m \in \mathbb{N}$, and $\beta$ with $|\beta| \geq \varepsilon$.
By the property of Kullback-Leibler divergence, we have

$$
\begin{equation*}
E_{r\left(y \mid \varphi_{0}(x)\right)}\left[\log r\left(y \mid \varphi_{0}(x)+\beta w(x)\right)-\log r\left(y \mid \varphi_{0}(x)\right)\right] \leq 0 \tag{103}
\end{equation*}
$$

for all $x$ and $\beta$. From the assumption [NM3]-1, there exist positive constants $A_{4}, C_{4}, \gamma_{4}, \delta_{4}$, and $T_{0}$ such that, if $|\beta w(x)| \leq T_{0}$, the inequality

$$
\begin{equation*}
E_{r\left(y \mid \varphi_{0}(x)\right)}\left[\log r\left(y \mid \varphi_{0}(x)+\beta w(x)\right)-\log r\left(y \mid \varphi_{0}(x)\right)\right] \leq-A_{4}|\beta w(x)|^{\gamma_{4}} \tag{104}
\end{equation*}
$$

holds, and if $|\beta w(x)|>T_{0}$, the inequality

$$
\begin{equation*}
E_{r\left(y \mid \varphi_{0}(x)\right)}\left[\log r\left(y \mid \varphi_{0}(x)+\beta w(x)\right)-\log r\left(y \mid \varphi_{0}(x)\right)\right] \leq-C_{4}|\beta w(x)|^{\delta_{4}} \tag{105}
\end{equation*}
$$

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

$$
\begin{equation*}
E_{r\left(y \mid \varphi_{0}(x)\right)}\left[\log r\left(y \mid \varphi_{0}(x)+\beta w(x)\right)-\log r\left(y \mid \varphi_{0}(x)\right)\right] \leq-A_{4}(b \varepsilon \sqrt{m})^{\gamma_{4}} \tag{106}
\end{equation*}
$$

or

$$
\begin{equation*}
E_{r\left(y \mid \varphi_{0}(x)\right)}\left[\log r\left(y \mid \varphi_{0}(x)+\beta w(x)\right)-\log r\left(y \mid \varphi_{0}(x)\right)\right] \leq-C_{4}(b \varepsilon \sqrt{m})^{\delta_{4}} \tag{107}
\end{equation*}
$$

is satisfied.
From eqs.(103), (106), and (106), for $\lambda_{4}=\max \left\{\gamma_{4}, \delta_{4}\right\}, \kappa_{4}=\min \left\{\gamma_{4}, \delta_{4}\right\}$, and some constant $F_{4}>0$, the bound

$$
\begin{equation*}
E_{f_{0} \mu}\left[\log f^{[m]}(z ; \beta)-\log f_{0}(z)\right] \leq-F_{4} \varepsilon^{\lambda_{4}} m^{\frac{\kappa_{4}}{2}} \frac{c}{m} \tag{108}
\end{equation*}
$$

is obtained, which means the assertion.
(e) Uniform consistency. We write $f_{k}^{[m]}(z ; \beta)=r\left(y \mid \varphi_{0}(x)+\beta w_{k}^{[m]}(x)\right) q(x)$. For a fixed $\delta>0$, take $B_{1}$ and $\lambda_{1}$ in the assertion (a), and denote $R_{m}=$ $B_{1} m^{\lambda_{1}}$. Because we have

$$
\begin{equation*}
\sup _{\beta \geq R_{m}} \sum_{i=1}^{n} \log f_{k}^{[m]}\left(Z_{i} ; \beta\right) \leq \sum_{i=1}^{n} H_{+}^{[m]}\left(Z_{i} ; R_{m}\right) \tag{109}
\end{equation*}
$$

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

$$
\begin{align*}
& \operatorname{Prob}\left(\exists k, \sup _{\beta>R_{m}} \frac{1}{n} \sum_{i=1}^{n} \log f_{k}^{[m]}\left(Z_{i} ; \beta\right) \geq \frac{1}{n} \sum_{i=1}^{n} \log f_{0}\left(Z_{i}\right)\right) \\
& \leq m \operatorname{Prob}\left(\frac{1}{n} \sum_{i=1}^{n} H_{+}^{[m]}\left(Z_{i} ; R_{m}\right)>\frac{1}{n} \sum_{i=1}^{n} \log f_{0}\left(Z_{i}\right)\right) \\
& \leq m\left\{\operatorname{Prob}\left(\frac{1}{n} \sum_{i=1}^{n} H_{+}^{[m]}\left(Z_{i} ; R_{m}\right)-E_{f_{0} \mu}\left[H_{+}^{[m]}\left(Z ; R_{m}\right)\right]>\delta\right)\right. \\
& \left.\quad+\operatorname{Prob}\left(\frac{1}{n} \sum_{i=1}^{n} f_{0}\left(Z_{i}\right)-E_{f_{0} \mu}\left[\log f_{0}(Z)\right]<-\delta\right)\right\} \\
& \leq m\left\{\frac{V\left[H_{+}^{[m]}\left(Z ; R_{m}\right)\right]}{n \delta^{2}}+\frac{V\left[\log f_{0}(Z)\right]}{n \delta^{2}}\right\} \tag{110}
\end{align*}
$$

We can obtain a similar bound for $\beta<-R_{m}$ also, using $H_{-}^{[m]}(z ; t)$. From the fact (b), the variance $V\left[H_{ \pm}^{[m]}\left(Z ; R_{m}\right)\right]$ 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

$$
\begin{equation*}
\operatorname{Prob}\left({ }^{\exists} k,\left|\hat{\beta}_{k}^{[m]}\right| \geq R_{m}\right) \leq A_{5} \frac{m^{\lambda_{5}}}{n} \tag{111}
\end{equation*}
$$

holds for all $m, n \in \mathbb{N}$.
By the fact (d), we have $E_{f_{0} \mu}\left[\log f_{k}^{[m]}(z ; \beta)\right]-E_{f_{0} \mu}\left[\log f_{0}(z)\right] \leq-4 \delta_{m}$ for all $\beta$ 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

$$
\begin{equation*}
E_{f_{0} \mu}\left[g^{[m]}\left(z ; \beta, \rho_{m}\right)\right] \leq E_{f_{0} \mu}[\log f(z ; \beta)]+\delta_{m} \tag{112}
\end{equation*}
$$

for all $\beta \in\left[-R_{m}, R_{m}\right]$ and $\rho_{m}=A_{3} \delta_{m} \frac{1}{\left(\sqrt{m} R_{m}\right)^{\gamma_{3}}}$. Let $N_{m}$ be a natural number given by $N_{m}=\left[\frac{R_{m}}{\rho_{m}}\right]+1$. Then, there exist positive constants $C_{5}$ and $\nu_{5}$ such that $N_{m} \leq C_{5} m^{\nu_{5}} \varepsilon^{-\lambda_{4}}$. Dividing the interval $\left[-R_{m},-\delta\right] \cup\left[\delta, R_{m}\right]$ into $N_{m}$ intervals $J_{j}=\left[\beta_{j}-\rho_{m}, \beta_{j}+\rho_{m}\right]\left(1 \leq j \leq N_{m}\right)$ with disjoint interior, we have

$$
\begin{equation*}
E_{f_{0} \mu}\left[g^{[m]}\left(z ; \beta_{j}, \rho_{m}\right)\right] \leq E_{f_{0} \mu}\left[\log f_{0}(z)\right]-3 \delta_{m} \tag{113}
\end{equation*}
$$

for each $j$. Then, we obtain

$$
\begin{align*}
& \operatorname{Prob}\left({ }^{\exists} \beta \in\left[-R_{m},-\varepsilon\right] \cup\left[\varepsilon, R_{m}\right], \frac{1}{n} \sum_{i=1}^{n} \log f_{k}^{[m]}\left(Z_{i} ; \beta\right) \geq \frac{1}{n} \sum_{i=1}^{n} \log f_{0}\left(Z_{i}\right)\right) \\
& \leq \operatorname{Prob}\left({ }^{\exists} k, 1 \leq{ }^{\exists} j \leq N_{m}, \frac{1}{n} \sum_{i=1}^{n} \sup _{\beta \in J_{j}} \log f_{k}^{[m]}\left(Z_{i} ; \beta\right) \geq \frac{1}{n} \sum_{i=1}^{n} \log f_{0}\left(Z_{i}\right)\right) \\
& \leq N_{m} \operatorname{Prob}\left(\frac{1}{n} \sum_{i=1}^{n} g^{[m]}\left(Z_{i} ; \beta_{j}, \rho_{m}\right)>\frac{1}{n} \sum_{i=1}^{n} \log f_{0}\left(Z_{i}\right)\right) \\
& \leq C_{5} m^{\nu_{5}+1} \varepsilon^{-\lambda_{4}}\left\{\operatorname{Prob}\left(\frac{1}{n} \sum_{i=1}^{n} g^{[m]}\left(Z_{i} ; \beta_{j}, \rho_{m}\right)-E_{f_{0} \mu}\left[g^{[m]}\left(z ; \beta_{j}, \rho_{m}\right)\right]>\delta_{m}\right)\right. \\
& \left.\quad+\operatorname{Prob}\left(\frac{1}{n} \sum_{i=1}^{n} \log f_{0}\left(Z_{i}\right)-E_{f_{0} \mu}\left[\log f_{0}(Z)\right]<-\delta_{m}\right)\right\} \\
& \leq C_{5} m^{\nu_{5}+1} \varepsilon^{-\lambda_{4}}\left\{\frac{V\left[g^{[m]}\left(z ; \beta_{j}, \rho_{m}\right)\right]}{n \delta_{m}^{2}}+\frac{V\left[\log f_{0}(Z)\right]}{n \delta_{m}^{2}}\right\} . \tag{114}
\end{align*}
$$

From the fact (c), we see that there exist $F_{5}>0$ and $\tau_{5}>0$ such that

$$
\begin{equation*}
\operatorname{Prob}\left({ }^{\exists} k, \hat{\beta}_{k}^{[m]} \in\left[-R_{m},-\varepsilon\right] \cup\left[\varepsilon, R_{m}\right]\right) \leq F_{5} \frac{m^{\tau_{5}}}{n \varepsilon^{3 \lambda_{4}}} \tag{115}
\end{equation*}
$$

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 $\left\{\max _{1 \leq k \leq m_{n}}\left|\widehat{\beta}_{k}^{(n)}\right|<\varepsilon\right\}$ converges to one. Then, the maximum likelihood estimator $\widehat{\beta}_{k}^{(n)}$ of the model $f_{k}^{(n)}(z ; \beta)=r\left(y \mid \varphi_{0}(x)+\beta w_{k}^{(n)}(x)\right) q(x)$ satisfies the likelihood equation

$$
\begin{equation*}
\sum_{i=1}^{n} \frac{\partial \log f_{k}^{(n)}\left(Z_{i} ; \widehat{\beta}_{k}^{(n)}\right)}{\partial \beta}=0 \tag{116}
\end{equation*}
$$

for all $1 \leq k \leq m_{n}$, with a probability which converges to one. Taylor expansion of eq.(116) and $L_{n}$ leads

$$
\begin{equation*}
\sum_{i=1} \frac{\partial \log f_{k}^{(n)}\left(Z_{i} ; 0\right)}{\partial \beta}+\sum_{i=1}^{n} \frac{\partial^{2} \log f_{k}^{(n)}\left(Z_{i} ; \beta_{k}^{*}\right)}{\partial \beta^{2}} \widehat{\beta}_{k}^{(n)} \tag{117}
\end{equation*}
$$

and
$\sum_{i=1}^{n} \log \frac{f_{k}^{(n)}\left(Z_{i} ; \widehat{\beta}_{k}^{(n)}\right)}{f_{0}\left(Z_{i}\right)}=\sum_{i=1}^{n} \frac{\partial \log f_{k}^{(n)}\left(Z_{i} ; 0\right)}{\partial \beta} \widehat{\beta}_{k}^{(n)}+\frac{1}{2} \sum_{i=1}^{n} \frac{\partial^{2} \log f_{k}^{(n)}\left(Z_{i} ; \beta_{k}^{* *)}\right.}{\partial \beta^{2}}\left(\widehat{\beta}_{k}^{(n)}\right)^{2}$
for some $\beta_{k}^{*}$ and $\beta_{k}^{* *}$ between 0 and $\widehat{\beta}_{k}^{(n)}$. A simple calculation shows

$$
\begin{equation*}
\sum_{i=1}^{n} \log \frac{f_{k}^{(n)}\left(Z_{i} ; \widehat{\beta}_{k}^{(n)}\right)}{f_{0}\left(Z_{i}\right)}=\frac{\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f_{k}^{(n)}\left(Z_{i} ; 0\right)}{\partial \beta}\right)^{2}}{-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^{2} \log f_{k}^{(n)}\left(Z_{i} ; 0\right)}{\partial \beta^{2}}}\left\{S_{n}^{(k)}-\frac{1}{2} T_{n}^{(k)}\right\}, \tag{119}
\end{equation*}
$$

where

$$
\begin{equation*}
S_{n}^{(k)}=\frac{\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^{2} \log f_{k}^{(n)}\left(Z_{i} ; 0\right)}{\partial \beta^{2}}}{\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^{2} \log f_{n}^{(n)}\left(Z_{i} ; \beta_{k}^{*}\right)}{\partial \beta^{2}}} \tag{120}
\end{equation*}
$$

and

$$
\begin{equation*}
T_{n}^{(k)}=\frac{\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^{2} \log f_{k}^{(n)}\left(Z_{i} ; 0\right)}{\partial \beta^{2}} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^{2} \log f_{k}^{(n)}\left(Z_{i} ; \beta_{k}^{* *}\right)}{\partial \beta^{2}}}{\left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^{2} \log f_{k}^{(n)}\left(Z_{i} ; \beta_{k}^{*}\right)}{\partial \beta^{2}}\right)^{2}} . \tag{121}
\end{equation*}
$$

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

$$
\begin{equation*}
\operatorname{Prob}\left(\max _{1 \leq k \leq m_{n}}\left|\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^{2} \log f_{k}^{(n)}\left(Z_{i} ; \tilde{\beta}_{k}\right)}{\partial \beta^{2}}+1\right| \geq \varepsilon\right) \quad \longrightarrow 0 \quad(n \rightarrow \infty) \tag{122}
\end{equation*}
$$

for $\tilde{\beta}_{k}=0, \widehat{\beta}_{k}^{*}$, and $\widehat{\beta}_{k}^{* *}$. In fact, $\max _{k}\left|S_{n}^{(k)}-1\right|$ and $\max _{k}\left|T_{n}^{(k)}-1\right|$ converge to 0 in probability from the above convergence.

By Taylor expansion, we have

$$
\begin{equation*}
\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^{2} \log f_{k}^{(n)}\left(Z_{i} ; \tilde{\beta}_{k}\right)}{\partial \beta^{2}}=\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^{2} \log f_{k}^{(n)}\left(Z_{i} ; 0\right)}{\partial \beta^{2}}+\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^{3} \log f_{k}^{(n)}\left(Z_{i} ; \eta\right)}{\partial \beta^{3}} \tilde{\beta}_{k}, \tag{123}
\end{equation*}
$$

where $\eta$ is between 0 and $\tilde{\beta}_{k}$. Because $\frac{\partial^{2} \log f_{k}^{(n)}(z ; 0)}{\partial \beta^{2}}=\frac{\partial^{2} \log r\left(y ; \varphi_{0}(x)\right)}{\partial u^{2}}\left(w_{k}^{(n)}(x)\right)^{2}$, from the assumption [NM3]-5 and the fact $\left|w_{k}^{(n)}(x)\right| \leq a \sqrt{m_{n}}$, there exists $B_{1}>0$ such that

$$
\begin{equation*}
E_{f_{0} \mu}\left[\left|\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^{2} \log f_{k}^{(n)}\left(Z_{i} ; 0\right)}{\partial \beta^{2}}+1\right|^{2}\right] \leq \frac{2+2 B_{1} m_{n}^{2}}{n} \tag{124}
\end{equation*}
$$

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

$$
\begin{align*}
& \operatorname{Prob}\left(\max _{1 \leq k \leq m_{n}}\left|\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^{2} \log f_{k}^{(n)}\left(Z_{i} ; 0\right)}{\partial \beta^{2}}+1\right|>\frac{\varepsilon}{2}\right) \\
& \leq m_{n} \operatorname{Prob}\left(\left|\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^{2} \log f_{k}^{(n)}\left(Z_{i} ; 0\right)}{\partial \beta^{2}}+1\right|>\frac{\varepsilon}{2}\right) \\
& \leq m_{n} \frac{2+2 B_{1} m_{n}^{2}}{n} \longrightarrow 0 . \tag{125}
\end{align*}
$$

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

$$
\begin{equation*}
E_{r\left(y \mid \varphi_{0}(x)\right)}\left[\sup _{|\beta| \leq m_{n}^{-d}}\left|\frac{\partial^{3} \log r\left(y \mid \varphi_{0}(x)+\beta w_{k}^{(n)}(x)\right)}{\partial u^{3}}\right|^{2}\right] \leq B_{2}, \tag{126}
\end{equation*}
$$

for all $n \geq n_{0}$ and $x \in \mathbb{R}$. Let $M_{k}^{(n)}(z)$ be a function defined by

$$
\begin{equation*}
M_{k}^{(n)}(z)=\sup _{|\beta| \leq m_{n}^{-d}}\left|\frac{\partial^{3} \log f_{k}^{(n)}(z ; \beta)}{\partial \beta^{3}}\right| . \tag{127}
\end{equation*}
$$

The $L^{2}$ norm of this function is bounded by

$$
\begin{align*}
E_{f_{0} \mu}\left[\left(M_{k}^{(n)}(z)\right)^{2}\right] & =E_{Q}\left[E_{r\left(y \mid \varphi_{0}(x)\right)}\left[\sup _{|\beta| \leq m_{n}^{-d}}\left|\frac{\partial^{3} \log r\left(y \mid \varphi_{0}(x)+\beta w_{k}^{(n)}(x)\right)}{\partial u^{3}}\right|^{2}\left\{w_{k}^{(n)}(x)\right\}^{6}\right]\right] \\
& \leq B_{2}(a \sqrt{m})^{6} . \tag{128}
\end{align*}
$$

Then, we obtain a bound of the probability

$$
\begin{align*}
& \operatorname{Prob}\left(1 \leq{ }^{\exists} k \leq m,\left|\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^{3} \log f_{k}^{(n)}\left(Z_{i} ; \eta\right)}{\partial \beta^{3}} \tilde{\beta}_{k}\right| \geq \frac{\varepsilon}{2}\right) \\
& \leq \operatorname{Prob}\left(\max _{1 \leq k \leq m_{n}}\left|\widehat{\beta}_{k}\right| \geq \frac{1}{m_{n}^{d}}\right) \\
& \quad+\operatorname{Prob}\left(\max _{1 \leq k \leq m_{n}}\left|\widehat{\beta}_{k}\right|<\frac{1}{m_{n}^{d}}, \text { and } 1 \leq{ }^{\exists} k \leq m_{n},\left|\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^{3} \log f_{k}^{(n)}\left(Z_{i} ; \eta\right)}{\partial \beta^{3}}\right| \geq \frac{\varepsilon}{2} m_{n}^{d}\right) \\
& \leq \operatorname{Prob}\left(\max _{1 \leq k \leq m_{n}}\left|\widehat{\beta}_{k}\right| \geq \frac{1}{m_{n}^{d}}\right)+m_{n} \operatorname{Prob}\left(\frac{1}{n} \sum_{i=1}^{n} M_{k}^{(n)}\left(Z_{i}\right) \geq \frac{\varepsilon}{2} m_{n}^{d}\right) . \tag{129}
\end{align*}
$$

From Lemma 4, there exist positive constants $A, \lambda$, and $\nu$ 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 $\frac{4 m_{n} E\left[M_{n}^{(n)}(z)^{2}\right]}{\varepsilon^{2} m_{n}^{2 d}} \leq \frac{4 B_{2} a^{6} m_{n}^{4-2 d}}{\varepsilon^{2}}$, which converges to zero because we take $d>2$. For such $d$ fixed, taking sufficiently small $\gamma$ such that $\gamma(\lambda+d \nu)<1$, we see that the first term of eq.(129) converges to zero for $m_{n}=\left[n^{\gamma}\right]$. Thus, we have for any sufficiently small $\gamma>0$ and $m_{n}=\left[n^{\gamma}\right]$,

$$
\begin{equation*}
\operatorname{Prob}\left(1 \leq{ }^{\exists} k \leq m_{n},\left|\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^{3} \log f_{k}^{(n)}\left(Z_{i} ; \eta\right)}{\partial \beta^{3}} \tilde{\beta}_{k}\right| \geq \frac{\varepsilon}{2}\right) \quad \longrightarrow \quad 0, \tag{130}
\end{equation*}
$$

as $n \rightarrow \infty$.
Combination of eqs.(123), (125) and (130) means eq.(122), and completes the proof.


[^0]:    ${ }^{1}$ 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].

[^1]:    ${ }^{2}$ 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]).

