The Annals of Statistics
2003, Vol. 31, No. 3, 833–851

LIKELIHOOD RATIO OF UNIDENTIFIABLE MODELS AND MULTILAYER NEURAL NETWORKS

BY KENJI FUKUMIZU

Institute of Statistical Mathematics

This paper discusses the behavior of the maximum likelihood estimator (MLE), in the case that the true parameter cannot be identified uniquely. Among many statistical models with unidentifiability, neural network models are the main concern of this paper. It is known in some models with unidentifiability that the asymptotics of the likelihood ratio of the MLE has an unusually larger order. Using the framework of locally conic models put forth by Dacunha-Castelle and Gassiat as a generalization of Hartigan's idea, a useful sufficient condition of such larger orders is derived. This result is applied to neural network models, and a larger order is proved if the true function is given by a smaller model. Also, under the condition that the model has at least two redundant hidden units, a $\log n$ lower bound for the likelihood ratio is derived.

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 the MLE is an important problem in estimation theory, and the asymptotic normality under some regularity conditions is well known. 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. There are many statistical models with unidentifiability, such as finite mixture models [Hartigan (1985)], autoregressive moving averages [Veres (1987)], reduced rank regression [Fukumizu (1999)], change point problems [Csörgő and Horváth (1997)] and hidden Markov models [Gassiat and Kéribin (2000)]. The behavior of the MLE in such models has not been clarified completely, and many statistical methods such as model selection need special considerations.

The main topic of this paper is the asymptotic order of the likelihood ratio (LR) test statistics of the MLE as the sample size n goes to infinity. It has been reported that the LR of some unidentifiable models has a larger order than $O_p(1)$, which is the order given by ordinary asymptotic theory. Among many studies, Hartigan (1985) discussed the normal mixture models with two components under the null hypothesis of one component, and showed that the LR has a larger order than $O_p(1)$. He conjectured also that the order is $\log\log n$, which has been solved by Bickel and Chernoff (1993) and Liu and Shao (2001). Gassiat and Kéribin

Received March 2001; revised May 2002.
AMS 2000 subject classifications. Primary 62F12; secondary 62F10.
Key words and phrases. Likelihood ratio, unidentifiable model, multilayer neural networks, locally conic model.

(2000) discussed a similar mixture model in a hidden Markov setting and showed divergence of the LR for a two-state model under the null hypothesis of one state.

In this paper, a useful sufficient condition of a larger order of LR will be shown by using the framework of locally conic models [Dacunha-Castelle and Gassiat (1997)] in which unidentifiability is regarded as a conic singularity in the statistical model embedded in the functional space of the probability densities. The sufficient condition of LR divergence is given by a functional property of the tangent cone at the singularity.

Another main result is the asymptotic order of the LR for multilayer neural network models. It is known that multilayer neural networks also have unidentifiability in the parameterization. By analysis of the functional properties of the tangent cone, divergence of the LR will be shown on condition that the model has redundant hidden units to realize the true function, and a lower bound of $\log n$ will be derived for the models with at least two redundant hidden units.

2. Divergence of likelihood ratio in locally conic models.

2.1. *Preliminaries.* A *statistical model* $S = \{f(z;\theta) \mid \theta \in \Theta\}$ is a set of probability density functions on a measure space $(\mathcal{Z}, \mathcal{B}, \mu)$, which is parameterized by a differentiable manifold (with boundary) Θ. We assume that $\operatorname{Supp} f(z;\theta)$ is invariant for all $\theta \in \Theta$. Given an i.i.d. sample $Z_1, \dots, Z_n$ generated by the *true probability density* $f_0(z)$, we consider the *likelihood ratio*, defined by

$$\sup_{\theta\in\Theta} L_n(\theta), \qquad \text{where } L_n(\theta) = \sum_{i=1}^{n} \log \frac{f(Z_i;\theta)}{f_0(Z_i)}, \tag{1}$$

in the maximum likelihood framework. The main topic of this paper is the asymptotic behavior of the LR as the number of samples n goes to infinity.

It is assumed that the true probability density is included in the model S. Let Θ_0 be the set of true parameters: $\Theta_0 = \{\theta \in \Theta \mid f(z;\theta_0)\mu = f_0(z)\mu\}$. We *do not* assume the uniqueness of θ_0, but say that the true parameter is *unidentifiable* if Θ_0 is a union of finitely many submanifolds of Θ and the dimension of at least one of the submanifolds is larger than zero. There are many important models in which the true parameter can be unidentifiable. Finite mixture models and multilayer neural networks are some examples. Suppose, for example, we have a mixture model with two components $f(z;a_1,a_2,b) = b\,g(z;a_1) + (1-b)\,g(z;a_2)$ and the true density $f_0(z) = g(z;a_0)$ for some a_0. Then the set of true parameters contains $\{(a_1,a_2,b) \mid a_1 = a_2 = a_0\} \cup \{(a_1,a_2,b) \mid b = 0, a_2 = a_0\} \cup \{(a_1,a_2,b) \mid b = 1, a_1 = a_0\}$, which is high dimensional. If the true parameter is unidentifiable, the LR does not follow the usual chi-square asymptotics, which requires uniqueness of the true parameter in the regularity conditions.

2.2. *Locally conic model and likelihood ratio.* If a statistical model is considered in the functional space of probability density functions, the set of true parameters corresponds to a single point. This point is a singularity in the model S if the dimensionality shrinks only at an exceptional parameter set with measure zero. The local property around the singularity will be better understood by introducing a convenient parameterization. Following Dacunha-Castelle and Gassiat (1997), with some modification, a locally conic model is used for discussing unidentifiability.

We write $\mathbb{R}_{\geq 0} = \{\beta \in \mathbb{R} \mid \beta \geq 0\}$. Let A_0 be a $(d-1)$-dimensional differentiable manifold (with boundary), let Θ be a submanifold in $A_0 \times \mathbb{R}_{\geq 0}$, let $S = \{f(z;\theta) \mid \theta \in \Theta\}$ be a statistical model and let $f_0(z)$ be an element in S. The parameter $\theta \in \Theta$ is decomposed as $\theta = (\alpha, \beta)$ for $\alpha \in A_0$ and $\beta \in \mathbb{R}_{\geq 0}$. The statistical model S is called *locally conic* at f_0 if the following conditions are satisfied.

CONDITION 1. The parameter space Θ includes $\Theta_0 := A_0 \times \{0\}$, and the set of the parameters to give f_0 is Θ_0; that is, $f(z;(\alpha,\beta))\mu = f_0(z)\mu \Longleftrightarrow \beta = 0$.

CONDITION 2. For each $\alpha \in A_0$, the set $\Theta(\alpha) := \{\beta \in \mathbb{R}_{\geq 0} \mid (\alpha, \beta) \in \Theta\}$ is a closed interval with open interior.

CONDITION 3. $f(z;(\alpha,\beta))$ is differentiable on β (right differentiable at 0) for each $\alpha \in A_0$ and $f_0\mu$-a.e. z. For each $\alpha \in A_0$ the Fisher information at f_0 is 1:

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

Intuitively, a locally conic model S is a union of one-dimensional submodels $S_\alpha = \{f(z;\alpha,\beta) \mid \beta \in \Theta(\alpha)\}$. If the dimension of A_0 is larger than zero, the parameter to give f_0 is unidentifiable, which is a singularity in the model. The score function of S_α at the origin,

$$v_\alpha(z) = \frac{\partial \log f(z;(\alpha,0))}{\partial \beta}, \tag{3}$$

can be looked at as a unit tangent vector along S_α. The family of score functions $C = \{v_\alpha \mid \alpha \in A_0\}$ generates the tangent cone at the singularity f_0. We call the set C the *basis of the tangent cone*, which will have a key importance in the following discussion. An example of a locally conic model is the multilayer neural network model, which will be shown in Section 3.

Let a model $S = \{f(z;(\alpha,\beta)) \mid (\alpha,\beta) \in \Theta\}$ be locally conic at $f_0 \in S$ and let $Z_1, \ldots, Z_n$ be i.i.d. random variables with law $f_0\mu$. Assume that all the submodels S_α satisfy the following regularity conditions on asymptotic normality. Conditions 1–3 are slight modifications of Wald's conditions for consistency [Wald (1949)] and Condition 4 assures asymptotic efficiency [Cramér (1946)]. For simplicity, we write each submodel as $\{g(z;\beta) \mid \beta \in V\}$, omitting the index α, where $V = \Theta(\alpha)$.

CONDITIONS ON ASYMPTOTIC NORMALITY (AN).

1. For any $\beta \in V$, the integral $E_{f_0\mu}[|\log g(z;\beta)|]$ is finite.
2. If $V = \mathbb{R}_{\geq 0}$, the function $H(z;t) = \sup_{\beta\geq t} \log g(z;\beta)$ satisfies $\lim_{t\to\infty} E_{f_0\mu}[H(z;t)] < \infty$ and there exists Δ such that $\int_\Delta f_0(z)\,d\mu > 0$ and $\lim_{t\to\infty} H(z;t) = -\infty$ for all $z \in \Delta$.
3. $\lim_{\rho\downarrow 0} E_{f_0\mu}[\sup_{|\beta'-\beta|\leq\rho} \log g(z;\beta')] < \infty$ for all $\beta \in V$.
4. The density $g(z;\beta)$ is three times differentiable on β for all z and

$$\lim_{\rho\downarrow 0} \int \sup_{0\leq\beta\leq\rho} \left|\frac{\partial^\nu g(x;\beta)}{\partial\beta^\nu}\right| d\mu < \infty \qquad (\nu = 1, 2),$$

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

Under the assumptions AN, by applying the standard asymptotic theory to each S_α, the LR in the model S can be decomposed into [Dacunha-Castelle and Gassiat (1997)]

$$\sup_{\theta\in\Theta} L_n(\theta) = \sup_{\alpha\in A_0} L_n(\alpha, \hat\beta_\alpha) = \sup_{\alpha\in A_0} \left\{\tfrac{1}{2} U_n(\alpha)^2 \cdot \mathbf{1}_{U_n(\alpha)\geq 0} + o_p(1)\right\}, \tag{4}$$

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

$$U_n(\alpha) = \frac{1}{\sqrt{n}} \sum_{i=1}^n v_\alpha(Z_i), \qquad v_\alpha(z) = \frac{\partial}{\partial\beta} \log f\bigl(z; (\alpha, 0)\bigr). \tag{5}$$

The function $v_\alpha(z)$ belongs to the basis of the tangent cone C. While the variable $U_n(\alpha)$ converges in law to the standard normal distribution for each $\alpha \in A_0$, we have to consider $U_n(\alpha)$ over all α to see the LR in S.

2.3. *Larger order of the likelihood ratio.* The LR can have a larger order than $O_p(1)$ if the function class of the tangent cone is "rich" enough. In this subsection, a useful sufficient condition of such an unusually larger order is derived. We generalize Hartigan's (1985) idea on a Gaussian mixture model by applying it to the general expression of (4) for locally conic models, which is originally used for deriving the asymptotic distribution of the LR under the assumption of the uniform convergence of U_n to a Gaussian process [Dacunha-Castelle and Gassiat (1997, 1999)].

Note that the marginal distribution of U_n in (4) on finite points $v_1, \ldots, v_m$ in C always converges to an m-dimensional normal distribution with covariance $E_P[v_i v_j]$. Two components of the limit are independent if 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 arbitrarily large value, since the maximum of m independent samples from the standard normal distribution is approximately $\sqrt{2\log m}$ for large m. Hartigan (1985) applied this idea to a normal mixture model with two components, calculating the covariance explicitly. Generalization of his idea leads us to the following theorem.

THEOREM 1. *Let a statistical model $S = \{f(z; (\alpha, \beta))\}$ be locally conic at $f_0 \in S$ and let $C = \{v_\alpha(z) = \frac{\partial}{\partial\beta} f(z; (\alpha, 0))\}$ be the basis of the tangent cone. Assume that for each $\alpha \in A_0$ the submodel $S_\alpha = \{f(z; \alpha, \beta) \mid \beta\}$ satisfies the conditions* AN. *If there exists a sequence $\{v_n\}_{n=1}^{\infty}$ in C such that $v_n \to 0$ in probability, then, for arbitrary $M > 0$, we have*

$$\lim_{n\to\infty} \operatorname{Prob}\left(\sup_{(\alpha,\beta)} L_n(\alpha, \beta) \le M \right) = 0. \tag{6}$$

REMARK. The regularity condition AN can be replaced by any other conditions for asymptotic normality, such as Le Cam (1970). The condition AN uses a classical one by Cramér, which will give an easy extension to derive a lower bound of the order of the LR in the next section.

PROOF OF THEOREM 1. Using the bound

$$\begin{aligned} |E_{f_0\mu}[v_m v_n]| &\le \int_{\{|v_n|\ge\varepsilon\}} |v_m v_n| f_0 \, d\mu + \int_{\{|v_n|<\varepsilon\}} |v_m v_n| f_0 \, d\mu \\ &\le \left(\int_{\{|v_n|\ge\varepsilon\}} |v_m|^2 f_0 \, d\mu \right)^{1/2} + \varepsilon \int |v_m| f_0 \, d\mu, \end{aligned}$$

we have $\lim_{n\to\infty} E[v_m v_n] = 0$ for arbitrary $m \in \mathbb{N}$. From this fact, for arbitrary $\varepsilon > 0$ and $K \in \mathbb{N}$, there exist $v(\alpha_1), \dots, v(\alpha_K) \in C$ such that $|E[v(\alpha_i)v(\alpha_j)]| < \varepsilon$ for different i and j. The rest of the proof is exactly the same as the argument in Hartigan (1985), which is omitted here. $\square$

The sufficient condition of the theorem is very easy to apply. For example, consider the Gaussian mixture model with two components

$$f(x; \mu, b) = b\phi(x; \mu) + (1 - b)\phi(x; 0),$$

where $\phi(x; \mu)$ is the probability density function of the normal distribution with mean μ and variance 1. We see that for $\mu \neq 0$,

$$f(x; \mu, b) = \beta \frac{\exp(\mu x - \mu^2/2) - 1}{\| \exp(\mu x - \mu^2/2) - 1 \|_{L^2(\phi_0)}} \phi(x; 0) + \phi(x; 0), \tag{7}$$

where $\beta = b\| \exp(\mu x - \mu^2/2) - 1 \|_{L^2(\phi_0)}$. This gives a locally conic parameterization at $\phi(x; 0)$. It is easy to see that

$$\frac{\exp(\mu x - \mu^2/2) - 1}{\| \exp(\mu x - \mu^2/2) - 1 \|_{L^2(\phi_0)}}$$

converges to zero in probability as $\mu \to \infty$. This gives another proof of Hartigan (1985).

3. Likelihood ratio of multilayer perceptrons.

3.1. *Unidentifiability in multilayer perceptrons.* The *multilayer perceptron* model with H hidden units [Rumelhart, Hinton and Williams (1986)] is defined by a family of functions

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

where $x \in \mathcal{X} = \mathbb{R}$, $s(t) = \tanh(t)$ and $\theta = (a_1, b_1, c_1, \dots, a_H, b_H, c_H, d) \in \mathbb{R}^{3H+1}$.

Learning in neural networks can be regarded as statistical estimation. Throughout this paper, we assume that the input sample X_i is i.i.d. with law $Q = q(x)\mu_{\mathbb{R}}$, where $\mu_{\mathbb{R}}$ is the Lebesgue measure on $\mathbb{R}$ and the integral $E_Q|\log q(x)|^2$ is finite. Let $\mathcal{Y}$ be a subset of $\mathbb{R}$, let $(\mathcal{Y}, \mathcal{B}_y, \mu_y)$ be a measure space and let $r(y|u)$ be a conditional probability density function of $y \in \mathcal{Y}$ given $u \in \mathbb{R}$. The statistical model of a multilayer perceptron $\mathcal{M}_H$ is defined by

$$f(z;\theta) = r\big(y|\varphi(x;\theta)\big)q(x), \tag{9}$$

where $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$. We assume that the noise model $r(y|u)$ satisfies the following assumptions.

CONDITIONS ON THE NOISE MODEL (NM1).

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

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

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

Popular choices of $r(y|u)$ are the additive Gaussian noise $(1/\sqrt{2\pi}\sigma) \times \exp\{-(1/2\sigma^2)(y-u)^2\}$ for continuous y and the logistic model $e^{uy}/(1+e^u)$ 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. Suppose, for example, we have a multilayer perceptron model with two hidden units and the true function $\varphi_0(x)$ given by a perceptron with only one hidden unit, say, $\varphi_0(x) = b_0 \tanh(a_0 x)$. Then, any parameter θ in the high-dimensional set $\{\theta \mid a_1 = a_0, b_1 = b_0, c_1 = b_2 = d = 0\} \cup \{\theta \mid a_1 = a_2 = a_0, c_1 = c_2 = d = 0,$

$b_1 + b_2 = b_0\}$ realizes the function $\varphi_0(x)$. It is known that the true parameter is unidentifiable if and only if the true function can be realized by a network with a smaller number of hidden units than the model [Sussmann (1992); Fukumizu and Amari (2000)].

A locally conic structure can be seen in this unidentifiability of multilayer perceptrons. Suppose we have the model $\mathcal{M}_H$ and the true function $\varphi_0(x)$, which is given by a multilayer perceptron with K $(0 \le K < H)$ hidden units,

$$\varphi_0(x) = \sum_{k=1}^{K} b_k^0 s(a_k^0 x + c_k^0) + d^0, \tag{10}$$

with $a_k \neq 0, b_k \neq 0$ $(1 \le k \le K)$ and $(a_k, b_k) \neq \pm(a_i, b_i)$ $(1 \le k < i \le K)$.

For later use, we define a submodel of $\mathcal{M}_H$ as

$$\psi(x;\omega) = \varphi_0(x) + \beta\{\eta s(\xi x + \zeta) + \delta\}, \tag{11}$$

where $\omega \in \{\omega = (\alpha, \beta) = ((\xi, \eta, \zeta, \delta), \beta) \mid \eta \neq 0, \xi \neq 0, (\xi, \zeta) \neq \pm(a_k^0, c_k^0)$ $(1 \le k \le K), \beta \ge 0\}$. We can see that the model $\{r(y|\psi(x;\omega))q(x)\}$ is locally conic at $f_0(z) = r(y|\varphi_0(x))q(x)$. In fact, because the functions $\{1, s(a_k^0 x + c_k^0), s(\xi x + \zeta) \mid 1 \le k \le K\}$ are linearly independent [see Fukumizu (1996)], β must be zero to satisfy $\psi(x;\omega) = \varphi_0(x)$. This shows the condition NM1-1 of the definition. Let $N(\alpha)$ be the $L^2(f_0\mu)$ norm of a tangent vector $\frac{\partial}{\partial\beta} \log f(x, y; (\alpha, 0))$. It is given by

$$N(\alpha)^2 = \int G(\varphi_0(x)) \left\{ \frac{\partial \psi(x; (\alpha, 0))}{\partial \beta} \right\}^2 q(x)\, dx,$$

where

$$\frac{\partial \psi(x; (\alpha, 0))}{\partial \beta} = \eta s(\xi x + \zeta) + \delta. \tag{12}$$

Since this partial derivative is not a constant zero, we have $0 < N(\alpha) < \infty$ for all α. Replacing β by $N(\alpha)\beta$, we have locally conic parameterization.

3.2. *Divergence of the likelihood ratio in multilayer perceptrons.* For applying Theorem 1 to the multilayer perceptron model, we need additional assumptions on the noise model $r(y|u)$ to ensure the Conditions AN. These assumptions are satisfied by many important noise models. It is easy to see that the Gaussian and logistic models satisfy them.

CONDITIONS ON THE NOISE MODEL (NM2).

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

$$\sup_{\xi, u \in K} E_{r(y|\xi)} |\log r(y|u)| < \infty$$

and

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

2. The density $r(y|u)$ is three times differentiable on u for all $y \in \mathcal{Y}$ and for any compact set $K \subset \mathbb{R}$, there exists $\rho > 0$ such that

$$\sup_{\xi \in K} \int \sup_{|\xi'-\xi| \leq \rho} \left| \frac{\partial^\nu r(y|\xi')}{\partial^\nu u} \right| dy < \infty \qquad (\nu = 1, 2)$$

and

$$\sup_{\xi \in K} E_{r(y|\xi)}\left[\sup_{|\xi'-\xi| \leq \rho} \left| \frac{\partial^3 \log r(y|\xi')}{\partial^3 u} \right| \right] < \infty.$$

THEOREM 2. *Assume that the model is the multilayer perceptron with H hidden units $\mathcal{M}_H$ and the true function is given by a network with K hidden units for $K < H$. Under the assumptions* NM1 *and* NM2 *on the noise model $r(y|u)$, we have, for arbitrary $M > 0$,*

$$\lim_{n \to \infty} \operatorname{Prob}\left(\sup_\theta L_n(\theta) \leq M \right) = 0. \tag{13}$$

REMARK. This theorem means that the LR is strictly larger than $O_p(1)$.

PROOF OF THEOREM 2. Let $\sigma(x; \xi, h)$ be a bounded, monotone decreasing function defined by

$$\sigma(x; \xi, h) = \frac{1}{2}\left\{ 1 + s\left(-\frac{1}{2}\xi(x-h) \right) \right\} = \frac{1}{1 + \exp\{\xi(x-h)\}} \tag{14}$$

and let $\{g(z; t, c, \beta)\}$ be a submodel of (11), given by

$$g(z; t, c, \beta) = r\left(y \middle| \varphi_0(x) + \beta \frac{1}{\sqrt{B(t, c)}} \sigma\left(x; c^2, t + \frac{1}{c} \right) \right) q(x), \tag{15}$$

where $B(t, c) = \int G(\varphi_0(x)) \sigma(x; c^2, t + \frac{1}{c})^2 \, dQ(x)$ and $\beta \in [0, 1]$. The basis of the tangent cone C consists of the functions

$$v(x, y; t, c) = \frac{1}{\sqrt{B(t, c)}} \frac{\partial \log r(y|\varphi_0(x))}{\partial u} \sigma\left(x; c^2, t + \frac{1}{c} \right). \tag{16}$$

From the boundedness of $\varphi_0(x)$ and $\sigma(x; \xi, h)$, it is straightforward to see that NM1 and NM2 imply the asymptotic normality AN.

Fix $A > 0$ such that $G(\varphi_0(x)) \geq A$ for all $x \in \mathbb{R}$. Let $F_Q(t)$ be the distribution function of the input probability Q and let $t_0 = \inf\{t \in \mathbb{R} \mid F_Q(t) > 0\} \in \mathbb{R} \cup \{-\infty\}$. From the fact that $\lim_{c \to \infty} \sigma(x; c^2, t + \frac{1}{c}) = \chi_{(-\infty, t]}(x)$, we have, for given t, $B(t,c) \geq \frac{A}{4} F_Q(t)$ for sufficiently large c. For any $t > t_0$ and $\delta > 0$, we have $\sigma(x; c^2, t + \frac{1}{c}) \leq F_Q(t)$ for all $x \geq t + \delta$ and sufficiently large c. Then we can choose sequences $t_n \downarrow t_0$, $\delta_n \downarrow 0$ and sufficiently large c_n such that $|v(x, y; t_n, c_n)| \leq \frac{2}{\sqrt{A}} |\frac{\partial \log r(y|\varphi_0(x))}{\partial u}| \sqrt{F_Q(t_n)}$ holds for all $x \geq t_n + \delta_n$ and y. Because $F_Q(t_n) \to 0$, we have $v(x, y; t_n, c_n) \to 0$ almost everywhere. $\square$

3.3. *Asymptotic order of the likelihood ratio in multilayer perceptrons.* We will derive a $\log n$ lower bound for the LR in the case $K \leq H - 2$. To show this bound, we will use n^γ ($\gamma > 0$) "almost independent" variables in the basis of the tangent cone, as described below. However, unlike Theorem 1, approximation by the multidimensional Gaussian distribution is not obvious, because the dimensionality goes to infinity along with n. Sazonov's (1968) result and Lemma 1 in the Appendix are used to solve this problem.

Let $\mathcal{W} = \{w(x; \xi, h, t) \mid \xi, t \in \mathbb{R}, h > 0\}$ be a family of functions given by

$$w(x; \xi, h, t) = \frac{1}{\sqrt{A(\xi, h, t)}} \frac{1}{2} \{s(\xi(x - t + h)) - s(\xi(x - t - h))\}, \tag{17}$$

where $A(\xi, h, t) = E_Q[G(\varphi_0(x)) \frac{1}{4} \{s(\xi(x - t + h)) - s(\xi(x - t - h))\}^2]$ is a normalization constant. Note that $\lim_{\xi \to \infty} \frac{1}{2} \{s(\xi(x - t + h)) - s(\xi(x - t - h))\} = \chi_{[t-h, t+h]}$ for any t and h. Using an argument similar to Section 3.1, we can easily prove that the function family

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

defines a locally conic submodel of $\mathcal{M}_H$. The basis of the tangent cone includes an arbitrary number of almost independent functions for any family of disjoint intervals.

First, a general result will be shown under the condition that the regressor class can approximate $\chi_I(x)$ for any interval $I \subset \mathbb{R}$. For the theorem, we need further assumptions on the noise model $r(y|u)$. In listing them, we do not avoid overlap with the former assumptions for simplicity. It is not difficult to verify the following assumptions for the Gaussian model and the logistic model.

CONDITIONS ON THE NOISE MODEL (NM3).

1. For any compact set $K \subset \mathbb{R}$, there exists a nonnegative function $\tau(s)$ on $[0, \infty)$ such that for some positive numbers A_i, δ_i ($i = 1, 2$) and T_0,

$$\tau(s) \geq A_1 s^{\delta_1} \qquad (0 \leq \forall s \leq T_0) \quad \text{and} \quad \tau(s) \geq A_2 s^{\delta_2} \qquad (\forall s > T_0)$$

hold, and a lower bound of the KL divergence is given by

$$E_{r(y|\xi)}\left[\log \frac{r(y|\xi)}{r(y|u)}\right] \geq \tau(|u - \xi|)$$

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

2. There exist a continuous function $\ell_2(\xi)$ and $\nu > 0$ such that

$$E_{r(y|\xi)}\left[\sup_{|u|\le R}\left|\frac{\partial \log r(y|u)}{\partial u}\right|^2\right] \le \ell_2(\xi)R^{\nu} \qquad \text{for all } R \ge 1.$$

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

$$\sup_{u\in K} E_{r(y|u)}\big[|\log r(y|u)|^2\big] < \infty, \qquad \sup_{u\in K} E_{r(y|u)}\left[\left|\frac{\partial \log r(y|u)}{\partial u}\right|^3\right] < \infty$$

and

$$\sup_{\xi,u\in K} E_{r(y|\xi)}\left[\left|\frac{\partial^2 \log r(y|u)}{\partial u^2}\right|^2\right] < \infty.$$

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

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

THEOREM 3. *Let $r(y|u)$ be a conditional density function of $y \in \mathcal{Y}$ given $u \in \mathbb{R}$, which satisfies the conditions* NM1, NM2 *and* NM3, *let $\varphi_0(x)$ be a bounded function on $\mathbb{R}$ and let $f_0(z) = r(y|\varphi_0(x))q(x)$ be a density function with respect to the measure $\mu = \mu_{\mathbb{R}} \times \mu_y$, where $z = (x, y)$. For a closed interval I, a nonnegative value $M(I)$ is defined by*

$$M(I) = \left\|\frac{\partial \log r(y|\varphi_0(x))}{\partial u}\chi_I(x)\right\|^2_{L^2(f_0\mu)} = \int_I G(\varphi_0(x))q(x)\,dx, \tag{18}$$

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

$$u_I(z) = \frac{1}{\sqrt{M(I)}}\frac{\partial \log r(y|\varphi_0(x))}{\partial u}\chi_I(x) \tag{19}$$

for I with $M(I) > 0$. Suppose $\mathcal{W} = \{w(x;\alpha) \mid \alpha \in A_0\}$ is a family of functions with the following conditions: *the function*

$$v(z;\alpha) = \frac{\partial \log r(y|\varphi_0(x))}{\partial u}w(x;\alpha) \tag{20}$$

satisfies $\|v(z;\alpha)\|_{L^2(f_0\mu)} = 1$ for all $\alpha \in A_0$, and there exist $a, b > 0$ such that for any $\varepsilon > 0$ and closed interval I with $M(I) > 0$, we can find $w(x;\alpha) \in \mathcal{W}$ which satisfies (i) $0 < w(x;\alpha) \le \frac{a}{\sqrt{M(I)}}$ *for all $x \in \mathbb{R}$,* (ii) $w(x;\alpha) \ge \frac{b}{\sqrt{M(I)}}$ *for all $x \in I$ and* (iii) $\|v(z;\alpha) - u_I(z)\|_{L^2(f_0\mu)} < \varepsilon$.

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

$$\liminf_{n\to\infty}\operatorname{Prob}\left(\frac{\sup_{\alpha,\beta} L_n(\alpha,\beta)}{\log n} \ge \delta\right) > 0. \tag{21}$$

REMARK. This theorem asserts that the order of the LR is at least $\log n$. The model $\{f(z;\alpha,\beta)\}$ is regarded as a locally conic model by using $f(z;\alpha_+,\beta) = r(y|\varphi_0(x)+\beta w(x;\alpha_+))$ and $f(z;\alpha_-,\beta) = r(y|\varphi_0(x)-\beta w(x;\alpha_-))$ for $\beta \in \mathbb{R}_{\geq 0}$. For simplicity, we take negative β into consideration.

Theorem 3 can be applied to multilayer perceptrons for $K \leq H-2$.

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

$$\liminf_{n\to\infty} \operatorname{Prob}\left(\frac{\sup_\theta L_n(\theta)}{\log n} \geq \delta\right) > 0. \tag{22}$$

PROOF OF THEOREM 3. From NM1-3 and the boundedness of $\varphi_0(x)$, we have $0 < M(\mathbb{R}) < \infty$. Fix $K > 0$ such that the $M([-K,K]) = \frac{M(\mathbb{R})}{2}$. For an arbitrary $m \in \mathbb{N}$, we can obtain a partition $\{I_k^{[m]} \mid k = 1,\dots,m\}$ of $[-K,K]$ such that $I_k^{[m]}$'s are closed intervals with disjoint interiors and $M(I_k^{[m]}) = \frac{M(\mathbb{R})}{2m}$ for all k. For each k $(1 \leq k \leq m)$, a unit score function $u_k^{[m]}(z)$ is defined by

$$\begin{aligned} u_k^{[m]}(z) &= \frac{\partial}{\partial\beta} \log r\left(y \,\middle|\, \varphi_0(x) + \beta \frac{1}{\sqrt{M(I_k^{[m]})}} \chi_{I_k}(x)\right)\Bigg|_{\beta=0} \\ &= \sqrt{\frac{2m}{M(\mathbb{R})}} \frac{\partial \log r(y|\varphi_0(x))}{\partial u} \chi_{I_k^{[m]}}(x). \end{aligned}$$

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 $H_3(x) = E_{r(y|\varphi_0(x))}|\frac{\partial \log r(y|\varphi_0(x))}{\partial u}|^3$. By NM1-3 and NM3-3, there exists $B > 0$ such that $H_3(x) \leq BG(\varphi_0(x))$ for all $x \in [-K,K]$. Then we obtain

$$E_{f_0\mu}|u_k^{[m]}(z)|^3 = \frac{1}{M(I_k^{[m]})^{3/2}} \int H_3(x)\chi_{I_k^{[m]}}(x) q(x)\,dx \leq \frac{\sqrt{2}B}{\sqrt{M(\mathbb{R})}}\sqrt{m}. \tag{23}$$

Let P_n and Q_m be the probabilities of the m-dimensional random vector $(\frac{1}{\sqrt{n}}\sum_{i=1}^n u_1^{[m]}(Z_i), \dots, \frac{1}{\sqrt{n}}\sum_{i=1}^n u_m^{[m]}(Z_i))$ and the m-dimensional normal distribution $N(0, I_m)$, respectively. Let $\mathcal{D}$ denote the family of all the convex measurable sets on $\mathbb{R}^m$. A Berry–Esseen-type inequality [Sazonov (1968)] gives

$$\sup_{\Delta\in\mathcal{D}} |P_n(\Delta) - Q_m(\Delta)| \leq \frac{Lm^4}{\sqrt{n}} \sum_{1\leq k\leq m} E_{f_0\mu}|u_k^{[m]}(Z)|^3, \tag{24}$$

where L is a universal constant. From (23) and (24), choosing $\Delta = [-\nu\sqrt{\log m}, \nu\sqrt{\log m}]^m$, we have for all n and m,

$$\left|\operatorname{Prob}\left(\max_{1\le k\le m}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n u_k^{[m]}(Z_i)\right| > \nu\sqrt{\log m}\right) - \operatorname{Prob}\left(V_m > \nu\sqrt{\log m}\right)\right|$$
$$\le C'\frac{m^{11/2}}{\sqrt{n}},$$

where V_m is the maximum of the absolute values of m i.i.d. samples from $N(0,1)$, and C' is a constant independent of n and m. If we choose $0 < \nu < \sqrt{2}$ and $m = [n^\gamma]$ for $0 < \gamma < \frac{1}{11}$, where $[x]$ is the largest integer that is not larger than x, the extreme value theory, for arbitrary $\varepsilon > 0$, tells us that

$$\lim_{n\to\infty}\operatorname{Prob}\left(\max_{1\le k\le m}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n u_k^{[m]}(Z_i)\right|^2 > \nu^2\gamma\log n\right) > 1-\varepsilon. \tag{25}$$

By the assumptions on $\mathcal{W}$, for any $\varepsilon, \delta > 0$, $m \in \mathbb{N}$ and k ($1 \le k \le m$), there exists $w_k^{[m]} \in \mathcal{W}$ such that (i) $0 < w_k^{[m]}(x) \le \tilde{a}\sqrt{m}$, (ii) $w_k^{[m]}(x) \ge \tilde{b}\sqrt{m}$ on $I_k^{[m]}$ and (iii) $E_{f_0\mu}|v_k^{[m]}(z) - u_k^{[m]}(z)|^2 < \varepsilon\delta^2/m$, where $v_k^{[m]}(z)$ is a function defined by (20) for $w_k^{[m]}(x)$, and $\tilde{a}, \tilde{b}$ are positive constants independent of ε, δ, m and k. Then, noting the fact that

$$\max_{1\le k\le m}\left|\sum_{i=1}^n v_k^{[m]}(Z_i)\right| \le \max_{1\le k\le m}\left|\sum_{i=1}^n\left(v_k^{[m]}(Z_i) - u_k^{[m]}(Z_i)\right)\right| + \max_{i\le k\le m}\left|\sum_{i=1}^n u_k^{[m]}(Z_i)\right|,$$

we obtain from Chebyshev's inequality that

$$\begin{aligned}
&\operatorname{Prob}\left(\left|\max_{1\le k\le m}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n u_k^{[m]}(Z_i)\right| - \max_{i\le k\le m}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n v_k^{[m]}(Z_i)\right|\right| \ge \delta\right)\\
&\quad\le \operatorname{Prob}\left(1 \le \exists k \le m,\quad \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(u_k^{[m]}(Z_i) - v_k^{[m]}(Z_i)\right)\right| \ge \delta\right)\\
&\quad\le m\frac{E_{f_0\mu}|u_k^{[m]}(z) - v_k^{[m]}(z)|^2}{\delta^2} < \varepsilon.
\end{aligned} \tag{26}$$

Combining (25) and (26), we have a series $\{w_k^{[m]}\}$ and $\gamma' > 0$ such that

$$\lim_{n\to\infty}\operatorname{Prob}\left(\max_{1\le k\le m}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n v_k^{[m]}(Z_i)\right|^2 > \gamma'\log n\right) > 1-2\varepsilon. \tag{27}$$

From NM1-3, there exist $c, d > 0$ such that $\frac{c}{m} \le Q(I_k^{[m]}) \le \frac{d}{m}$ holds for all m and k ($1 \le k \le m$). Then, by the choice of $\{w_k^{[m]}\}$, Lemma 1 in the Appendix asserts that there exists $\gamma_1 > 0$ such that for all $0 < \gamma < \gamma_1$ and $m = [n^\gamma]$ we obtain the asymptotic expansion of the LR,

$$\begin{aligned}(28)\qquad &\max_{1\le k\le m}\sup_{|\beta|\le 1}\sum_{i=1}^{n}\log\frac{f_k^{[m]}(Z_i;\beta)}{f_0(Z_i)}\\ &\quad=\left\{\max_{1\le k\le m}\frac{1}{2}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}v_k^{[m]}(Z_i)\right)^2\right\}(1+o_p(1)),\end{aligned}$$

where $f_k^{[m]}(z;\beta) = r(y|\varphi_0(x) + \beta w_k^{[m]}(x))q(x)$. Noting that the range of β can be restricted to obtain the lower bound, the proof is completed by combining (27) and (28). $\square$

PROOF OF COROLLARY 1. We show that the function class $\mathcal{W} = \{w(x;\xi, h,t) \mid \xi,h,t \in \mathbb{R}\}$ defined by (17) satisfies the assumption of Theorem 3. Let $\sigma(x;\xi,h,t) = s(\xi(x-t+h)) - s(\xi(x-t-h))$ and $I = [t-c,t+c]$. By NM1-3, $M(I)$ is positive. We can easily find sequences $h_n \searrow c$ and $\xi_n \to \infty$ such that (A) $\sigma(x;\xi_n,h_n,t) \le 2$ for all $x \in \mathbb{R}$, (B) $\sigma(x;\xi_n,h_n,t) \ge \frac{1}{2}$ for all $x \in I$ and (C) $|\sigma(x;\xi_n,h_n,t) - \chi_I(x)| \to 0$ for all $x \in \mathbb{R}$. From (A), (C) and the boundedness of $G(\varphi_0(x))$, $\frac{\partial \log r(y|\varphi_0(x))}{\partial u}\sigma(x;\xi_n,h_n,t)$ converges to $\frac{\partial \log r(y|\varphi_0(x))}{\partial u}\chi_I(x)$ in $L^2(f_0\mu)$. This gives the assumption (iii). Also, we have $\frac{1}{2}M(I) \le A(\xi_n,h_n,t) \le 2M(I)$ for sufficiently large n. Combining this with (A) and (B), we obtain (i) and (ii) by taking $a = 2\sqrt{2}$ and $b = \frac{1}{2\sqrt{2}}$. $\square$

The order $\log n$ was formerly obtained by Hagiwara, Kuno and Usui (2000). However, they considered only the least square loss function and used its special property. The approach in this paper extends their results and can be applied to various noise models, including binary output models.

As shown in the above discussions, the behavior of the LR 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, Toda, Usui and Hagiwara (1996) showed that if the network model has one hidden unit of the step function and the true function is constant zero, then the LR under Gaussian noise has the order of $\log\log n$, which is essentially the same as the result of a change point problem [Csörgő and Horváth (1997)].

4. Conclusion. Under the assumption that the true parameter is unidentifiable, the larger asymptotic order of likelihood ratio test statistics has been investigated. I have shown a useful sufficient condition of an unusually larger order of the LR, using the framework of locally conic models [Dacunha-Castelle and Gassiat (1997)]. This result has been applied to neural network models to show the divergence of the LR in redundant cases. Also, a $\log n$ lower bound for the likelihood ratio has been obtained under the assumption that there are at least two redundant hidden units to realize the true function.

APPENDIX

Lemmas used for the proof of Theorem 3.

LEMMA 1. *Let $\varphi_0(x)$ be a bounded function on $\mathbb{R}$, let $\mathcal{Y}$ be a subset of $\mathbb{R}$, let $\{r(y|\xi) \mid \xi \in \mathbb{R}\}$ be a family of probability density functions on a measure space $(\mathcal{Y}, \mathcal{B}_y, \mu_y)$, which satisfies* NM1, NM2 *and* NM3, *let $Q = q(x)\mu_{\mathbb{R}}$ be a probability on $\mathbb{R}$ with $E_Q|\log q(x)|^2 < \infty$ and let $f_0(z)\mu = r(y|\varphi_0(x))q(x)\mu_{\mathbb{R}}\mu_y$. For fixed positive constants a, b, c, d and a compact interval D, function classes $\mathcal{W}_m$ $(m \in \mathbb{N})$ are defined by $\mathcal{W}_m = \{w \in L^2(f_0\mu) \mid \|w\|_{L^2(f_0\mu)} = 1,\ 0 < w(x) \le a\sqrt{m}$ for all $x \in \mathbb{R}$, and there exists a closed interval $I \subset D$ such that $\frac{c}{m} \le Q(I) \le \frac{d}{m}$ and $w(x) \ge b\sqrt{m}$ on $I\}$. Given $\gamma > 0$, let $m_n = [n^\gamma]$ for $n \in \mathbb{N}$ and let $\mathcal{G}_\gamma$ be a family of sequences $\{\{w_k^{(n)}\}_{n\in\mathbb{N},\,1\le k\le m_n} \mid w_k^{(n)} \in \mathcal{W}_{m_n}\}$. Suppose we have i.i.d. random variables $(X_1, Y_1), \dots, (X_n, Y_n)$ with the law $f_0\mu$. Then there exists $\gamma_0 > 0$ such that for any $0 < \gamma \le \gamma_0$ and $\{w_k^{(n)}\} \in \mathcal{G}_\gamma$, we obtain, as n goes to infinity,*

$$
\begin{aligned}
&\max_{1\le k\le m_n} \sup_{|\beta|\le 1} \sum_{i=1}^n \log \frac{r(Y_i|\varphi_0(X_i) + \beta w_k^{(n)}(X_i))}{r(Y_i|\varphi_0(X_i))} \\
&\qquad = \left\{ \max_{1\le k\le m_n} \frac{1}{2}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n u_k^{(n)}(X_i, Y_i) \right)^2 \right\} \bigl(1 + o_p(1)\bigr),
\end{aligned} \tag{29}
$$

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

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

First, we will establish the uniform consistency of the MLE for β.

LEMMA 2. *Let $r(y|\xi)$, $q(x)$, $\varphi_0(x)$, f_0, and $\mathcal{W}_m$ be the same as in Lemma* 1. *For $m \in \mathbb{N}$, define $\mathcal{H}_m = \{\{w_k\}_{k=1}^m \mid w_k \in \mathcal{W}_m\}$. Let $\widehat{\beta}_k^{[m]}(\Xi)$ be the maximum likelihood estimator of the model $\{r(y|\varphi_0(x) + \beta w_k^{[m]}(x))q(x) \mid \beta \in [-1, 1]\}$ for $\Xi = \{w_k^{[m]}\}_{k=1}^m \in \mathcal{H}_m$, given the i.i.d. sample $(X_1, Y_1), \dots, (X_n, Y_n)$ with the law $f_0(z)\mu$. Then there exist $A, \lambda, \nu > 0$ such that*

$$
\operatorname{Prob}\left(\max_{1\le k\le m} |\widehat{\beta}_k^{[m]}(\Xi)| \ge \varepsilon \right) \le A \frac{m^\lambda}{n\varepsilon^\nu} \tag{30}
$$

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

PROOF. The proof is divided into three parts. In the first two parts, we discuss only one $w(x) \in \mathcal{W}_m$ and write $f^{[m]}(z; \beta) = r(y|\varphi_0(x) + \beta w(x))q(x)$

for simplicity. We define $g^{[m]}(z;\beta;\rho)$ for $\beta\in[-1,1]$ and $\rho>0$ by

$$g^{[m]}(z;\beta,\rho)=\sup_{|\beta'-\beta|\le\rho}\log f^{[m]}(z;\beta'). \tag{31}$$

A constant M is fixed so that $|\varphi_0(x)|\le M$ for all $x\in\mathbb{R}$.

(a) *Bounds of* $E_{f_0\mu}[g^{[m]}(z;\beta,\rho)]$. We will show that there exist $B,C,\gamma,\eta>0$ such that, for arbitrary $\delta>0$ and $\beta\in[-1,1]$, the inequalities

$$E_{f_0\mu}\big[g^{[m]}(z;\beta,\rho)\big]\le E_{f_0\mu}\big[\log f^{[m]}(z;\beta)\big]+\delta \tag{32}$$

and

$$E_{f_0\mu}\big|g^{[m]}(z;\beta,\rho)\big|^2\le Cm^{\gamma}+2\delta^2 \tag{33}$$

hold for $\rho\le B\delta m^{-\eta}$.

From NM3-2, we can find $\tau>0$, $\Psi(y)$ and $\ell_2(\xi)$ such that

$$\big|\log f^{[m]}(z;\beta)-\log f^{[m]}(z;\beta')\big|\le\Psi(y)w(x)|\beta-\beta'| \tag{34}$$

and $E_{r(y|\xi)}[|\Psi(y)|^2]\le\ell_2(\xi)(M+a\sqrt{m})^{\tau}$ hold for $\beta\in[-1,1]$. Using $\Gamma=E_Q[\ell_2(\varphi_0(x))]<\infty$, (34) shows that

$$E_{f_0\mu}\big[g^{[m]}(z;\beta,\rho)\big]\le E_{f_0\mu}\big[\log f^{[m]}(z;\beta)\big]+\rho a\sqrt{\Gamma m(M+a\sqrt{m})^{\tau}},$$

which implies (32) by choosing $\rho\le B\delta m^{-(\tau/4+1/2)}$ for some B. The second assertion is also easily obtained from (34) and NM3-3.

(b) *Lower bound of KL divergence*. We show that there exist $D>0$, $\xi>0$ and $\zeta\in\mathbb{R}$ such that the bound

$$\sup_{\varepsilon\le|\beta|\le1}E_{f_0\mu}\big[\log f^{[m]}(z;\beta)\big]\le E_{f_0\mu}[\log f_0(z)]-Dm^{\zeta}\varepsilon^{\xi} \tag{35}$$

holds for arbitrary $0<\varepsilon<1$ and $m\in\mathbb{N}$.

From NM3-1, for all $x\in I$ and β with $|\beta|\ge\varepsilon$, we have

$$E_{r(y|\varphi_0(x))}\big[\log r\big(y|\varphi_0(x)+\beta w(x)\big)-\log r\big(y|\varphi_0(x)\big)\big]\le -F\varepsilon^{\xi}\sqrt{m^{\sigma}}$$

for some $\xi,\sigma,F>0$. By integrating this on x with the probability Q,

$$E_{f_0\mu}\big[\log f^{[m]}(z;\beta)-\log f_0(z)\big]\le -F\varepsilon^{\xi}m^{\sigma/2}\frac{c}{m}$$

is obtained, which gives the assertion.

(c) *Uniform consistency*. We write $f_k^{[m]}(z;\beta)=r(y|\varphi_0(x)+\beta w_k^{[m]}(x))q(x)$. By fact (b), we have $E_{f_0\mu}[\log f_k^{[m]}(z;\beta)]-E_{f_0\mu}[\log f_0(z)]\le-4\delta_m$ for all β with $\varepsilon\le|\beta|\le1$ and $m\in\mathbb{N}$, where $\delta_m=\frac{1}{4}Dm^{\zeta}\varepsilon^{\xi}$. From fact (a), we have $E_{f_0\mu}[g^{[m]}(z;\beta,\rho_m)]\le E_{f_0\mu}[\log f(z;\beta)]+\delta_m$ for all $\beta\in[-1,1]$ and $\rho_m=B\delta_m m^{-\eta}$. Let $N_m\in\mathbb{N}$ be given by $N_m=[1/\rho_m]+2$. Note that there exist $G,t>0$

such that $N_m \le Gm^t\varepsilon^{-\xi}$. Dividing the set $[-1,-\varepsilon]\cup[\varepsilon,1]$ into N_m intervals $J_j=[\beta_j-\rho_m,\beta_j+\rho_m]$ $(1\le j\le N_m)$ with disjoint interiors, we have

$$E_{f_0\mu}\big[g^{[m]}(z;\beta_j,\rho_m)\big]\le E_{f_0\mu}[\log f_0(z)]-3\delta_m \tag{36}$$

for all j. Then, by Chebyshev's inequality, we have

$$\begin{aligned}
&\mathrm{Prob}\left(\exists k,\exists\beta\in[-1,-\varepsilon]\cup[\varepsilon,1],\ \frac{1}{n}\sum_{i=1}^{n}\log f_k^{[m]}(Z_i;\beta)\ge\frac{1}{n}\sum_{i=1}^{n}\log f_0(Z_i)\right)\\
&\quad\le mN_m\ \mathrm{Prob}\left(\frac{1}{n}\sum_{i=1}^{n}g^{[m]}(Z_i;\beta_j,\rho_m)>\frac{1}{n}\sum_{i=1}^{n}\log f_0(Z_i)\right)\\
&\quad\le mN_m\ \mathrm{Prob}\left(\frac{1}{n}\sum_{i=1}^{n}g^{[m]}(Z_i;\beta_j,\rho_m)-E_{f_0\mu}\big[g^{[m]}(Z;\beta_j,\rho_m)\big]>\delta_m\right) \qquad (37)\\
&\qquad+mN_m\ \mathrm{Prob}\left(\frac{1}{n}\sum_{i=1}^{n}\log f_0(Z_i)-E_{f_0\mu}[\log f_0(Z_i)]<-\delta_m\right)\\
&\quad\le Gm^{t+1}\varepsilon^{-\xi}\left\{\frac{V[g^{[m]}(z;\beta_j,\rho_m)]}{n\delta_m^2}+\frac{V[\log f_0(Z)]}{n\delta_m^2}\right\}.
\end{aligned}$$

From (33), (37) and NM3-3, there exist $A,\lambda>0$ so that

$$\mathrm{Prob}\big(\exists k,\hat\beta_k^{[m]}\in[-1,-\varepsilon]\cup[\varepsilon,1]\big)\le A\frac{m^\lambda}{n\varepsilon^{3\xi}},$$

which proves Lemma 2. □

PROOF OF LEMMA 1. From Lemma 2, the MLE $\widehat\beta_k^{(n)}$ of the model $f_k^{(n)}(z;\beta)=r(y|\varphi_0(x)+\beta w_k^{(n)}(x))q(x)$ satisfies the likelihood equation

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

for all $1\le k\le m_n$, with a probability converging to 1. By the standard argument of Taylor expansion, we obtain

$$\sum_{i=1}^{n}\log\frac{f_k^{(n)}(Z_i;\widehat\beta_k^{(n)})}{f_0(Z_i)}=\frac{\Big(\frac{1}{\sqrt n}\sum_{i=1}^{n}\frac{\partial\log f_k^{(n)}(Z_i;0)}{\partial\beta}\Big)^2}{-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\log f_k^{(n)}(Z_i;0)}{\partial\beta^2}}\left\{S_n^{(k)}-\frac{1}{2}T_n^{(k)}\right\}, \tag{38}$$

where $S_n^{(k)}$ and $T_n^{(k)}$ are defined by

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

and

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

with β_k^* and β_k^{**} between 0 and $\widehat{\beta}_k^{(n)}$. The proof of Lemma 1 is completed if we show, for arbitrary $\varepsilon > 0$,

$$\text{Prob}\left(\max_{1\le k\le m_n}\left|\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \log f_k^{(n)}(Z_i;\tilde{\beta}_k)}{\partial\beta^2} + 1\right| \ge \varepsilon\right) \to 0 \qquad (n\to\infty) \tag{39}$$

with $\tilde{\beta}_k = 0$, β_k^* or β_k^{**}.

By Taylor expansion, we have

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

where η is between 0 and $\tilde{\beta}_k$. Using $\frac{\partial^2 \log f_k^{(n)}(z;0)}{\partial\beta^2} = \frac{\partial^2 \log r(y;\varphi_0(x))}{\partial u^2}(w_k^{(n)}(x))^2$ and NM3-3, we have $B > 0$ such that the bound

$$E_{f_0\mu}\left[\left|\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \log f_k^{(n)}(Z_i;0)}{\partial\beta^2} + 1\right|^2\right] \le \frac{2+2Bm_n^2}{n}$$

holds for all $n \in \mathbb{N}$. Then, by Chebyshev's inequality, for $0 < \gamma < \frac{1}{3}$ and $m_n = [n^\gamma]$, we obtain

$$\text{Prob}\left(\max_{1\le k\le m_n}\left|\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \log f_k^{(n)}(Z_i;0)}{\partial\beta^2} + 1\right| > \frac{\varepsilon}{2}\right) \le 2m_n\frac{2+2Bm_n^2}{n\varepsilon} \to 0. \tag{40}$$

Take $d > 2$. From NM3-4, there exists $C > 0$ such that

$$E_{r(y|\varphi_0(x))}\left[\sup_{|\beta|\le m_n^{-d}}\left|\frac{\partial^3 \log r(y|\varphi_0(x)+\beta w_k^{(n)}(x))}{\partial u^3}\right|^2\right] \le C$$

holds for all $x \in \mathbb{R}$ and sufficiently large n. If we define

$$M_k^{(n)}(z) = \sup_{|\beta|\le m_n^{-d}}\left|\frac{\partial^3 \log f_k^{(n)}(z;\beta)}{\partial\beta^3}\right|,$$

we have

$$\text{Prob}\left(1 \le \exists k \le m_n, \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3 \log f_k^{(n)}(Z_i;\eta)}{\partial\beta^3}\tilde{\beta}_k\right| \ge \frac{\varepsilon}{2}\right)$$

$$(41)\qquad \begin{aligned} &\le \text{Prob}\left(\max_{1\le k\le m_n}|\widehat{\beta}_k| \ge \frac{1}{m_n^d}\right) \\ &\quad + \text{Prob}\left(\max_{1\le k\le m_n}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3 \log f_k^{(n)}(Z_i;\eta)}{\partial\beta^3}\right| \ge \frac{\varepsilon}{2}m_n^d\right) \\ &\le \text{Prob}\left(\max_{1\le k\le m_n}|\widehat{\beta}_k| \ge \frac{1}{m_n^d}\right) + m_n\,\text{Prob}\left(\frac{1}{n}\sum_{i=1}^{n}M_k^{(n)}(Z_i) \ge \frac{\varepsilon}{2}m_n^d\right). \end{aligned}$$

Since $E_{f_0\mu}[(M_k^{(n)}(z))^2] \le C(a\sqrt{m})^6$ from NM3-4, by Chebyshev's inequality, the second term is not greater than $4m_n E[M_k^{(n)}(z)^2]\varepsilon^{-2}m_n^{-2d} \le 4Ca^6 m_n^{4-2d}\varepsilon^{-2}$, which converges to zero for $d > 2$. From Lemma 2, there exist $A, \lambda, \nu > 0$ such that the first term of (41) is bounded by $Am_n^{\lambda+d\nu}/n$. This converges to zero for sufficiently small γ with $\gamma(\lambda + d\nu) < 1$ and $m_n = [n^\gamma]$. Thus, for such γ and m_n, we obtain

$$(42)\qquad \text{Prob}\left(1 \le \exists k \le m_n, \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3 \log f_k^{(n)}(Z_i;\eta)}{\partial\beta^3}\tilde{\beta}_k\right| \ge \frac{\varepsilon}{2}\right) \to 0$$

as $n \to \infty$. Equations (40) and (42) show (39) and complete the proof. □

Acknowledgments. I thank Dr. Y. Kano at Osaka University, Dr. S. Kuriki at the Institute of Statistical Mathematics, Dr. K. Hagiwara at Mie University and Dr. S. Amari at RIKEN Brain Science Institute for valuable discussions. I also thank the anonymous referees for their helpful comments.

REFERENCES

BICKEL, P. and CHERNOFF, H. (1993). Asymptotic distribution of the likelihood ratio statistic in a prototypical nonregular problem. In *Statistics and Probability*: *A Raghu Raj Bahadur Festschrift* (J. K. Ghosh, S. K. Mitra, K. R. Parthasarathy and B. P. Rao, eds.) 83–96. Wiley Eastern, New Delhi.

CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press.

CSÖRGŐ, M. and HORVÁTH, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley, New York.

DACUNHA-CASTELLE, D. and GASSIAT, E. (1997). Testing in locally conic models and application to mixture models. *ESAIM Probab. Statist.* **1** 285–317.

DACUNHA-CASTELLE, D. and GASSIAT, E. (1999). Testing the order of a model using locally conic parametrization: Population mixtures and stationary ARMA processes. *Ann. Statist.* **27** 1178–1209.

FUKUMIZU, K. (1996). A regularity condition of the information matrix of a multilayer perception network. *Neural Networks* **9** 871–879.

FUKUMIZU, K. (1999). Generalization error of linear neural networks in unidentifiable cases. In *Algorithmic Learning Theory. Lecture Notes in Artificial Intelligence* (O. Watanabe and T. Yokomori, eds.) **1720** 51–62. Springer, Berlin.

FUKUMIZU, K. and AMARI, S. (2000). Local minima and plateaus in hierarchical structures of multilayer perceptions. *Neural Networks* **13** 317–327.

GASSIAT, E. and KÉRIBIN, C. (2000). The likelihood ratio test for the number of components in a mixture with Markov regime. *ESAIM Probab. Statist.* **4** 25–52.

HAGIWARA, K., KUNO, K. and USUI, S. (2000). On the problem in model selection of neural network regression in overrealizable scenario. In *Proc. IEEE–INNS–ENNS International Joint Conference on Neural Networks* **6** 461–466. IEEE Comput. Soc., Los Alamitos, CA.

HARTIGAN, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proc. Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. M. Le Cam and R. A. Olshen, eds.) **2** 807–810. Wadsworth, Belmont, CA.

HAYASAKA, T., TODA, N., USUI, S. and HAGIWARA, K. (1996). On the least square error and prediction square error of function representation with discrete variable basis. In *Proc. Conference on Neural Networks for Signal Processing* **6** 72–81. IEEE, New York.

LE CAM, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.* **41** 802–828.

LIU, X. and SHAO, Y. (2001). Asymptotic distribution of the likelihood ratio test in a two-component normal mixture model. Technical report, Dept. Statistics, Columbia Univ.

RUMELHART, D. E., HINTON, G. E. and WILLIAMS, R. J. (1986). Learning internal representations by error propagation. In *Parallel Distributed Processing* (D. E. Rumelhart, J. L. McClelland and the PDP Research Group, eds.) **1** 318–362. MIT Press, Cambridge, MA.

SAZONOV, V. V. (1968). On the multi-dimensional central limit theorem. *Sankhyā Ser. A* **30** 181–204.

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

VERES, S. (1987). Asymptotic distributions of likelihood ratios for overparameterized ARMA processes. *J. Time Ser. Anal.* **8** 345–357.

WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20** 595–601.

INSTITUTE OF STATISTICAL MATHEMATICS
4-6-7 MINAMI-AZABU
MINATO-KU, TOKYO 106-8569
JAPAN
E-MAIL: fukumizu@ism.ac.jp