

Research Memorandum No. 858

November 07, 2002

A general upper bound of likelihood ratio
in binary regression

Kenji Fukumizu
and
Katsuyuki Hagiwara

A general upper bound of likelihood ratio in binary regression

Kenji Fukumizu* and Katsuyuki Hagiwara
Institute of Statistical Mathematics and Mie University

November 7, 2002

Abstract

We derive a general upper bound of the asymptotic order of the likelihood ratio test statistics (LRTS) for binary regression. It has been known that in some unidentifiable cases of nonlinear regression the LRTS may diverge to infinity asymptotically and have a $\log n$ lower bound for the sample size n . This paper shows an upper bound of $\log n$ under the very general assumptions of the finiteness of VC dimension of the regressor class and the boundedness of the assumed true regressor.

1 Introduction

The asymptotic distribution of the likelihood ratio test statistics (LRTS) for a large sample size is an important topic in theory and practice. It works as a criterion of model selection, and has been used extensively in many fields. The most well-known result on the asymptotics of LRTS is the convergence to the chi-square distribution under some regularity conditions. If we have a statistical model with a d -dimensional parameter and the null hypothesis of a probability P_0 in the model, the LRTS under the null hypothesis converges to the chi-square of the degree of freedom d . However, this is not necessarily the case, if the regularity conditions are not satisfied, and various asymptotic distributions can be seen depending on specific models. Chernoff (1954) gives a general expression of the LRTS by the projection onto the model. Shapiro (1988) derives mixture of chi-squares as a limiting distribution for the models with the parameter in a convex region.

It is known that the LRTS may have a larger order than the ordinary constant order. Among other works, Hartigan (1985) shows that the LRTS of the Gaussian mixture models with two components diverges to infinity asymptotically under the null hypothesis of one component. In change point problems (Csörgő and Horváth 1996), the asymptotic distribution of the LRTS for the

*Part of this work is done while the author is staying at University of California, Berkeley.

model with one change point against the null hypothesis of no change point is known to be of the order $\log \log n$, where n is the number of sample. These examples suggest that the local behavior of the maximum likelihood estimator cannot be described by a finite dimensional statistics, but the infinite degree of freedom must be taken into account to discuss LRTS in general cases. Fukumizu (2003) discusses divergence of LRTS from the viewpoint of singularities in statistical models, and derives a useful sufficient condition of such divergence.

When the LRTS diverges, the first concern on its behavior is the asymptotic order. The purpose of this memo is to show a general $O_p(\log n)$ upper bound on the LRTS for binary regression models under mild conditions on the class of regressors. To derive the bound, only the finiteness of VC-dimension and the boundedness of the hypothesized probability are required. This result is applicable to most of binary regression problems, and gives a universal upper bound of LRTS.

There have been some results on the $\log n$ order of LRTS. Hagiwara et al. (2001) discusses the lower bound of LRTS for Gaussian nonlinear regression of neural networks, which can approximate the point-mass function, and derive a $\log n$ lower bound of the LRTS for the null hypothesis of the constant zero regressor. This result is extended to a much wider class of neural networks in Fukumizu (2003), which covers general noise models including binary regression. Combined with the lower bound in Fukumizu (2003), the main theorem of this memo shows that the LRTS of some type of neural networks has exactly the $\log n$ order. In Hagiwara (2001), the upper bound $\log n$ has been previously derived for a type of Gaussian node model, whose location parameter is restricted in a set of input data. This memo is an extension of the idea in Hagiwara (2001), which uses the exponential inequality for large deviation.

The main method used in this paper is the VC-dimension and exponential inequality on the sum of i.i.d. variables, which have been used also in the field of empirical processes on a function class (Dudley (1984), Pollard (1984), van der Vaart and Wellner (1996)) and computational learning theory (Vapnik (1982), Vapnik (1998), Haussler (1992)). However, this paper has two major differences from the previous studies. In those studies, the main concern is the convergence of $\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^n f(X_i) - E_Q[f]|$ for i.i.d. sample X_i from Q and a function class \mathcal{F} . The finiteness of the VC-dimension replaces this supremum over infinite number of functions with the maximum over finite ones of a polynomial order with respect to n . For individual function f , one can obtain an upper bound with the exponential decay on n . While we also use this general scheme in our discussion, one difference is that we have to consider $\sum_{i=1}^n f(X_i)$ instead of $\frac{1}{n} \sum_{i=1}^n f(X_i)$, which is of the constant order, to discuss LRTS. We need a closer look on the convergence rate around the maximum likelihood estimator. The other difficulty is the unboundedness of the likelihood function. Unlike the standard theory of empirical processes and computational learning theory, we have to consider the supremum over unbounded functions to discuss the LRTS. To solve this problem, we divide the functions into the unbounded part and the bounded part, which depend on the sample size n .

This paper is organized as follows. In Section 2, we show the main result of

this paper. In Section 3, we show the preliminary lemmas, which are used in the proof of the main theorem. Section 4 gives the proof of the theorem.

2 Main theorem

Let $(\mathcal{X}, \mathfrak{B}, Q)$ be a probability space, and \mathcal{F} be a class of measurable functions from \mathcal{X} to $(0, 1)$. The Vapnik-Červonenkis dimension (VC-dimension, Vapnik 1982, Vapnik 1998) of the function class \mathcal{F} is defined by the VC-dimension of the subgraphs $\{\{(x, y) \mid y \leq p(x)\} \mid p \in \mathcal{F}\}$, and denoted by $\dim_{VC}\mathcal{F}$. Let $p_0 : \mathcal{X} \rightarrow (0, 1)$ be a measurable function, and P_0 be a probability on $\mathcal{X} \times \{0, 1\}$, which is defined by the probability Q and the conditional probability $P_0(\{1\} \mid x) = p_0(x)$. Throughout this memo, $(X_1, Y_1), \dots, (X_n, Y_n)$ is assumed to be an i.i.d. sample with the law P_0 . Given the sample, the (log) likelihood ratio test statistics (LRTS) is defined by

$$\sup_{p \in \mathcal{F}} L_n(p), \quad (1)$$

where

$$L_n(p) = \sum_{i=1}^n \left\{ Y_i \log \frac{p(X_i)}{p_0(X_i)} + (1 - Y_i) \log \frac{1 - p(X_i)}{1 - p_0(X_i)} \right\}. \quad (2)$$

The main result of this memo is to show the following $\log n$ upper bound of LRTS;

Theorem 1. *Assume that $\dim_{VC}\mathcal{F} < \infty$ and there exists $\delta > 0$ such that $\delta \leq p_0(x) \leq 1 - \delta$ for all $x \in \mathcal{X}$. Then, we have for $n \rightarrow \infty$*

$$\sup_{p \in \mathcal{F}} L_n(p) = O_p(\log n). \quad (3)$$

The finiteness of VC-dimension is a natural assumption to exclude such functions that can fit an arbitrary number of data points. The above theorem gives the universal upper bound of LRTS for binary regression models.

It is known that the LRTS of some binary regression models has a lower bound of $\log n$. Fukumizu (2003) shows that if the multilayer perceptron model has at least two more hidden units than the true regressor, the asymptotic order of the LRTS is at least $\log n$ for a wide class of noise models including logistic regression. Then, the above theorem shows that the order of the LRTS in that case is exactly $\log n$.

3 Preliminary lemmas

For $p \in \mathcal{F}$, a variable $R_n(p)$ is defined by

$$R_n(p) = \sum_{i=1}^n \left\{ 2Y_i \left(-1 + \sqrt{\frac{p(X_i)}{p_0(X_i)}} \right) + 2(1 - Y_i) \left(-1 + \sqrt{\frac{1 - p(X_i)}{1 - p_0(X_i)}} \right) \right\}. \quad (4)$$

For a_1 and a_2 in $(0, 1)$, let B_i ($i = 1, 2$) be probabilities on $\{0, 1\}$, defined by $B_i(\{1\}) = a_i$. The Hellinger distance between these probabilities, $H(a_1, a_2)$, is given by

$$H^2(a_1, a_2) = (\sqrt{a_1} - \sqrt{a_2})^2 + (\sqrt{1-a_1} - \sqrt{1-a_2})^2 = 2 - 2\sqrt{a_1 a_2} - 2\sqrt{(1-a_1)(1-a_2)}. \quad (5)$$

We use also $V(a_1; a_2)$ defined by

$$V(a_1; a_2) = 4a_1(1-a_1) \left\{ \sqrt{\frac{a_2}{a_1}} - \sqrt{\frac{1-a_2}{1-a_1}} \right\}^2. \quad (6)$$

In the following proofs, for notational simplicity, p_i and $p_{0,i}$ are often used for $p(X_i)$ and $p_0(X_i)$, respectively, and $\mathbf{X}_{(n)}$ for (X_1, \dots, X_n) .

Lemma 2. *The following properties hold for an arbitrary p ;*

(i) $L_n(p) \leq R_n(p)$.

(ii) $\mathbb{E}[R_n(p) \mid \mathbf{X}_{(n)}] = -\sum_{i=1}^n H^2(p(X_i), p_0(X_i))$.

(iii) $V(a_1; a_2) \leq 4H^2(a_1, a_2)$.

Proof. (i) and (ii) are trivial. Since $V(a_1; a_2)$ is the variance of a random variable W , which takes $2(-1 + \sqrt{a_2/a_1})$ with probability a_1 and $2(-1 + \sqrt{(1-a_2)/(1-a_1)})$ with probability $1-a_1$, the inequality $\text{Var}[W] \leq \mathbb{E}[W^2]$ gives (iii). \square

Lemma 3. *Assume that there exists $\delta > 0$ such that $\delta \leq p_0(x) \leq 1 - \delta$. Then, for any $p \in \mathcal{F}$, $T > 0$, and $n, m \in \mathbb{N}$, we have*

$$\text{Prob}(R_m(p) \geq T \log n \mid \mathbf{X}_{(n)}) \leq n^{-8\delta^2 T}. \quad (7)$$

Proof. From Lemma 2-(ii), the left hand side of eq.(7) is equal to

$$P_{n,m} := \text{Prob}(R_m(p) - \mathbb{E}[R_m(p) \mid \mathbf{X}_{(n)}] \geq T \log n + \sum_{i=1}^m H^2(p_i, p_{0,i}) \mid \mathbf{X}_{(n)}).$$

Since the difference of $2Y_i \left(-1 + \sqrt{\frac{p(X_i)}{p_0(X_i)}}\right) + 2(1-Y_i) \left(-1 + \sqrt{\frac{1-p(X_i)}{1-p_0(X_i)}}\right)$ for $Y_i = 0$ and 1 is given by

$$\Delta_i = 2 \left| \sqrt{\frac{p(X_i)}{p_0(X_i)}} - \sqrt{\frac{1-p(X_i)}{1-p_0(X_i)}} \right|,$$

Hoeffding's inequality leads

$$P_{n,m} \leq \exp \left\{ -\frac{2\{T \log n + \sum_{i=1}^m H^2(p_i, p_{0,i})\}^2}{\sum_{i=1}^m \Delta_i^2} \right\}.$$

From Lemma 2-(iii), we obtain $\Delta_i^2 \leq H^2(p_i, p_{0,i})/\delta^2$, which proves $P_{n,m} \leq e^{-8\delta^2 T \log n} = n^{-8\delta^2 T}$. \square

Lemma 4. Let p and q be functions from \mathcal{X} to $(0, 1)$. Then,

$$L_n(p) - L_n(q) \leq \sum_{i=1}^n \left\{ \frac{2}{\sqrt{q_i(1-q_i)}} H(p_i, q_i) - H^2(p_i, q_i) \right\}. \quad (8)$$

Proof. It is easy to see

$$\begin{aligned} L_n(p) - L_n(q) &\leq \sum_{i=1}^n \left\{ 2Y_i \left(-1 + \sqrt{\frac{p_i}{q_i}} \right) + 2(1 - Y_i) \left(-1 + \sqrt{\frac{1-p_i}{1-q_i}} \right) \right\} \\ &= 2 \sum_{i=1}^n (Y_i - q_i) \left(\sqrt{\frac{p_i}{q_i}} - \sqrt{\frac{1-p_i}{1-q_i}} \right) + 2 \sum_{i=1}^n \left\{ q_i \left(-1 + \sqrt{\frac{p_i}{q_i}} \right) + (1 - q_i) \left(-1 + \sqrt{\frac{1-p_i}{1-q_i}} \right) \right\} \\ &\leq 2 \sum_{i=1}^n \left| \sqrt{\frac{p_i}{q_i}} - \sqrt{\frac{1-p_i}{1-q_i}} \right| - \sum_{i=1}^n H^2(p_i, q_i). \end{aligned} \quad (9)$$

From Lemma 2-(iii), we see $\left| \sqrt{\frac{p_i}{q_i}} - \sqrt{\frac{1-p_i}{1-q_i}} \right| \leq H(p_i, q_i) / \sqrt{q_i(1-q_i)}$, which completes the proof. \square

Lemma 5. For $0 \leq a \leq 1$ and $0 \leq b < 1$,

$$H^2(a, b) \leq \frac{2}{1 - \sqrt{b}} \left(\sqrt{a} - \sqrt{b} \right)^2. \quad (10)$$

Proof. The function $F(x) = (\sqrt{1-x} - \sqrt{1-b}) / (\sqrt{x} - \sqrt{b})$ for $x \in [0, 1]$ is negative, monotonically decreasing, and continuous at all x including $x = b$. Then, we have $(\sqrt{1-a} - \sqrt{1-b})^2 \leq F(1)^2 (\sqrt{a} - \sqrt{b})^2$, which gives the assertion. \square

4 Proof of the main theorem

The proof is divided into three parts. Except in the very last part of the proof, the variable $\mathbf{X}_{(n)} = (X_1, X_2, \dots, X_n)$ is fixed. By abuse of notation, in discussing conditional probability given $\mathbf{X}_{(n)}$, we treat X_i as a value in \mathcal{X} rather than a random variable. Let $d := \dim_{VC} \mathcal{F}$. We use positive constants α , λ , and γ such that $\alpha > 2d/\delta^2$, $\lambda > \log(1/\delta)$, and $\gamma > 1 + 5\lambda/4$. The number of elements of a set A is denoted by $|A|$.

Given $\mathbf{X}_{(n)}$, a function class $\mathcal{F}(\mathbf{X}_{(n)})$ denotes the restriction of the functions in \mathcal{F} to $\{X_1, \dots, X_n\}$. For a function $p \in \mathcal{F}(\mathbf{X}_{(n)})$, define $J_p \subset \{1, \dots, n\}$ by

$$J_p := \left\{ j \in \{1, \dots, n\} \mid p(X_j) < \frac{1}{n^\lambda} \text{ or } p(X_j) \geq 1 - \frac{1}{n^\lambda} \right\},$$

and $s_j(p) \in \{0, 1\}$ for $j \in J_p$ by

$$s_j(p) := \begin{cases} 1 & \text{if } p(X_j) \geq 1 - \frac{1}{n^\lambda}, \\ 0 & \text{if } p(X_j) < \frac{1}{n^\lambda}. \end{cases}$$

The set of all the possible locations and labels is denoted by

$$\mathcal{V}_n := \{ \{(X_j, s_j(p))\}_{j \in J_p} \mid p \in \mathcal{F}(\mathbf{X}_{(n)}) \}. \quad (11)$$

For a function p and $z = (x, y) \in \mathcal{X} \times (0, 1)$, define the indicator $G(z; p)$ of the subgraph of p by $G(z; p) = 1$ if $y \leq p(x)$, and $G(z; p) = 0$ otherwise. For the $2n$ points $Z_i^+ = (X_i, 1 - 1/n^\lambda)$, $Z_i^- = (X_i, 1/n^\lambda)$ ($1 \leq i \leq n$), we can see the following three equivalence relations; $p(X_i) \geq 1 - 1/n^\lambda$ if and only if $G(Z_i^+; p) = G(Z_i^-; p) = 1$; $p(X_i) < 1/n^\lambda$ if and only if $G(Z_i^+; p) = G(Z_i^-; p) = 0$; and $1/n^\lambda \leq p(X_i) < 1 - 1/n^\lambda$ if and only if $G(Z_i^+; p) = 0$ and $G(Z_i^-; p) = 1$. From this fact, the cardinality of \mathcal{V}_n is the same as that of $\{(G(Z_i^+; p), G(Z_i^-; p))_{i=1}^n \in \{0, 1\}^{2n} \mid p \in \mathcal{F}(\mathbf{X}_{(n)})\}$. From $\dim_{VC} \mathcal{F}(\mathbf{X}_{(n)}) \leq d$, we have

$$|\mathcal{V}_n| \leq C_d (2n)^d \quad (12)$$

for $n > d/2$, where C_d is a universal constant depending only on d (see Theorem 4.3a, p.146, Vapnik (1998)).

In the following proof, for a subset A in $\{1, \dots, n\}$, we use

$$L_A(p) := \sum_{j \in A} \left\{ Y_j \frac{p(X_j)}{p_0(X_j)} + (1 - Y_j) \frac{1 - p(X_j)}{1 - p_0(X_j)} \right\}. \quad (13)$$

I). Let $N_n := \alpha \log n$. Then, we can prove for $n \rightarrow \infty$

$$\begin{aligned} \text{Prob} \left(\exists p \in \mathcal{F}(\mathbf{X}_{(n)}) \text{ such that } |J_p| \geq N_n, \frac{1}{|J_p|} |\{j \in J_p \mid Y_j \neq s_j(p)\}| < \frac{1}{\log |J_p|} \mid \mathbf{X}_{(n)} \right) \\ \rightarrow 0 \quad (14) \end{aligned}$$

uniformly on $\mathbf{X}_{(n)}$.

To see this, let $\mathcal{V}_{n,m} := \{ \{(X_j, t_j)\}_{j \in J} \in \mathcal{V}_n \mid |J| = m \}$ for $0 \leq m \leq n$. Then, the left hand side of eq.(14), which is denoted by P^I , is upper bounded by

$$P^I \leq \sum_{m=N_n}^n \sum_{\{(X_j, t_j)\}_{j \in J} \in \mathcal{V}_{n,m}} \text{Prob} \left(\frac{1}{m} |\{j \in J \mid Y_j \neq t_j\}| < \frac{1}{\log m} \mid \mathbf{X}_{(n)} \right). \quad (15)$$

For a fixed $\{(X_j, t_j)\}_{j \in J} \in \mathcal{V}_{n,m}$, define a random variable $U := \sum_{j \in J} |Y_j - t_j|$. Then, we have $U = |\{j \in J \mid Y_j \neq t_j\}|$ and $E_{P_0}[U \mid \mathbf{X}_{(n)}] = \sum_{j \in J} \{(1 - t_j)p_0(X_j) + t_j(1 - p_0(X_j))\} \geq m\delta$. If n is sufficiently large so that $1/\log N_n <$

$\delta/2$, using Hoeffding's inequality, we obtain for $N_n \leq m \leq n$

$$\begin{aligned}
& \text{Prob}\left(\frac{1}{m}|\{j \in J \mid Y_j \neq t_j\}| < \frac{1}{\log m} \mid \mathbf{X}_{(n)}\right) \\
&= \text{Prob}\left(U - \mathbb{E}[U \mid \mathbf{X}_{(n)}] < \frac{m}{\log m} - \mathbb{E}[U \mid \mathbf{X}_{(n)}] \mid \mathbf{X}_{(n)}\right) \\
&\leq \exp\left\{-2\left(\frac{m}{\log m} - \mathbb{E}[U \mid \mathbf{X}_{(n)}]\right)^2/m\right\} \leq \exp\{-m\delta^2/2\} \\
&\leq \exp\{-N_n\delta^2/2\} \leq n^{-\alpha\delta^2/2}.
\end{aligned} \tag{16}$$

Because the total number of summands in eq.(15) is equal to $|\mathcal{V}_n|$, from eqs.(12), (15) and (16), we have

$$P^I \leq 2^d C_d n^d n^{-\alpha\delta^2/2} = 2^d C_d n^{d-\alpha\delta^2/2}. \tag{17}$$

By the assumption of $\alpha > 2d/\delta^2$, the assertion is obtained.

II). We will show that there exists a universal constant B_d , which depends only on d , and a constant $\nu = \nu(\delta, d, \gamma) > 0$ such that

$$\text{Prob}\left(\exists p \in \mathcal{F}(\mathbf{X}_{(n)}) \text{ such that } L_{J_p^C}(p) \geq T \log n \mid \mathbf{X}_{(n)}\right) \leq B_d n^{-8\delta^2 T + \nu} \tag{18}$$

for all $T > 0$ and sufficiently large n , where $J_p^C = \{1, \dots, n\} - J_p$.

To see this, let $\mathcal{W}_n := \{K \subset \{1, \dots, n\} \mid \exists p \in \mathcal{F}(\mathbf{X}_{(n)}) \text{ such that } K = \{j \mid 1/n^\lambda \leq p(X_j) < 1 - 1/n^\lambda\}\}$, and $\mathcal{W}_{n,m} := \{K \in \mathcal{W}_n \mid |K| = m\}$ for $0 \leq m \leq n$. Obviously, the cardinality of \mathcal{W}_n is not greater than that of \mathcal{V}_n , and bounded by

$$|\mathcal{W}_n| \leq |\mathcal{V}_n| \leq 2^d C_d n^d. \tag{19}$$

For each $K \in \mathcal{W}_n$, let $\mathcal{F}_K(\mathbf{X}_{(n)})$ be a subclass defined by

$$\mathcal{F}_K(\mathbf{X}_{(n)}) := \left\{p \in \mathcal{F}(\mathbf{X}_{(n)}) \mid \left\{j \mid \frac{1}{n^\lambda} \leq p(X_j) < 1 - \frac{1}{n^\lambda}\right\} = K\right\}. \tag{20}$$

From $\mathcal{F}(\mathbf{X}_{(n)}) = \cup_{m=0}^n \cup_{K \in \mathcal{W}_{n,m}} \mathcal{F}_K(\mathbf{X}_{(n)})$, the probability in the left hand side of eq.(18), which is denoted by P^{II} , is bounded by

$$P^{II} \leq \sum_{m=0}^n \sum_{K \in \mathcal{W}_{n,m}} \text{Prob}\left(\exists p \in \mathcal{F}_K(\mathbf{X}_{(n)}) \text{ such that } L_K(p) \geq T \log n \mid \mathbf{X}_{(n)}\right). \tag{21}$$

Fix $K \in \mathcal{W}_{n,m}$. Let $\|\cdot\|_{2,K}$ be the L^2 norm for the uniform distribution on $\{X_j \mid j \in K\}$; that is,

$$\|\varphi\|_{2,K} = \sqrt{\frac{1}{m} \sum_{i \in K} \varphi(X_i)^2}.$$

The ε -covering number of a function class \mathcal{G} with respect to an L^2 norm $\|\cdot\|_2$ is defined by the smallest number of functions $\{f_h\} \subset L^2$ such that for every $g \in \mathcal{G}$ there exists f_h that satisfies $\|g - f_h\|_2 < \varepsilon$. It is denoted by $\mathcal{N}(\varepsilon, \mathcal{G}, \|\cdot\|_2)$. Since the VC-dimension of the function class $\mathcal{F}_{K,n}^{1/2} = \{\sqrt{p} \mid p \in \mathcal{F}_K(\mathbf{X}_{(n)})\}$ is not greater than d , the covering number $\mathcal{N}(s, \mathcal{F}_{K,n}^{1/2}, \|\cdot\|_{2,K})$ is no more than $C_d s^{-2d}$ for any $s > 0$ (Section 2.6, van der Vaart and Wellner (1996); see also Lemma 25, Pollard (1984)). Take $s = 1/n^\gamma$, and let $l_{n,K} := \mathcal{N}(1/n^\gamma, \mathcal{F}_{K,n}^{1/2}, \|\cdot\|_{2,K})$. Then, there exist $l_{n,K}$ elements $\{q^{(h)}\}$ in $\mathcal{F}_{K,n}^{1/2}$ such that for any $p \in \mathcal{F}_{K,n}^{1/2}$ one can find h which satisfies $\|\sqrt{p} - \sqrt{q^{(h)}}\|_{2,K} \leq 1/n^\gamma$. From Lemma 5, we obtain $H^2(p_i, q_i^{(h)}) \leq 4n^{\lambda/2}/n^{2\gamma}$ for $i \in K$. Then, Lemma 4 gives

$$L_K(p) - L_K(q^{(h)}) \leq 2n^\lambda \sum_{i \in K} H(p_i, q_i^{(h)}) \leq 4n^{-\gamma+1+5\lambda/4}.$$

Noting the assumption $\gamma > 1 + 5\lambda/4$, we have for sufficiently large n

$$\begin{aligned} & \text{Prob}(\exists p \in \mathcal{F}_K(\mathbf{X}_{(n)}) \text{ such that } L_K(p) > T \log n \mid \mathbf{X}_{(n)}) \\ & \leq \text{Prob}(1 \leq \exists h \leq l_{n,K}, L_K(q^{(h)}) \geq T \log n - 4n^{-\gamma+1+5\lambda/4} \mid \mathbf{X}_{(n)}) \\ & \leq \sum_{h=1}^{l_{n,K}} \text{Prob}(L_K(q^{(h)}) \geq (T-1) \log n \mid \mathbf{X}_{(n)}) \\ & \leq C_d n^{2\gamma d} n^{-8\delta^2(T-1)}. \end{aligned} \quad (22)$$

In the last inequality, we use Lemma 2-(i) and Lemma 3. Combination of eqs.(19), (21) and (22) proves the assertion.

III). We use $N_n = \alpha \log n$ as in I). For a given $\mathbf{X}_{(n)}$, let $\mathcal{F}^{[1]}(\mathbf{X}_{(n)})$ and $\mathcal{F}^{[2]}(\mathbf{X}_{(n)})$ be the classes defined by $\mathcal{F}^{[1]}(\mathbf{X}_{(n)}) := \{p \in \mathcal{F}(\mathbf{X}_{(n)}) \mid |J_p| \geq N_n\}$ and $\mathcal{F}^{[2]}(\mathbf{X}_{(n)}) := \mathcal{F}(\mathbf{X}_{(n)}) - \mathcal{F}^{[1]}(\mathbf{X}_{(n)})$, respectively. Then, we have

$$\text{Prob}\left(\sup_{p \in \mathcal{F}} L_n(p) > T \log n \mid \mathbf{X}_{(n)}\right) \leq P_1 + P_2, \quad (23)$$

where P_1 and P_2 are given by

$$P_i := \text{Prob}(\exists p \in \mathcal{F}^{[i]}(\mathbf{X}_{(n)}) \text{ such that } L_n(p) > T \log n \mid \mathbf{X}_{(n)}). \quad (24)$$

The probability P_1 is bounded by

$$\begin{aligned} P_1 & \leq \text{Prob}(\exists p \in \mathcal{F}(\mathbf{X}_{(n)}) \text{ such that } L_{J_p^c}(p) > T \log n \mid \mathbf{X}_{(n)}) \\ & \quad + \text{Prob}(\exists p \in \mathcal{F}(\mathbf{X}_{(n)}) \text{ such that } L_{J_p}(p) > 0 \mid \mathbf{X}_{(n)}) \\ & =: P_{11} + P_{12}. \end{aligned} \quad (25)$$

From II), the probability P_{11} is bounded by

$$P_{11} \leq B_d n^{-8\delta^2 T + \nu} \quad (26)$$

for sufficiently large n . From the fact

$$\begin{aligned} L_{J_p}(p) &\leq \sum_{j \in J_p, s_j(p)=Y_j} \log \frac{1}{\delta} + \sum_{j \in J_p, s_j(p) \neq Y_j} \log \frac{1/n^\lambda}{\delta} \\ &= |J_p| \log(1/\delta) - \lambda \log n \cdot |\{j \in J_p \mid s_j(p) \neq Y_j\}|, \end{aligned} \quad (27)$$

the condition $L_{J_p}(p) > 0$ means $|\{j \in J_p \mid s_j(p) \neq Y_j\}| < \frac{|J_p| \log(1/\delta)}{\lambda \log n}$. By the assumption $\lambda > \log(1/\delta)$, the fact I) shows

$$P_{12} \longrightarrow 0 \quad (n \rightarrow \infty) \quad (28)$$

uniformly on $\mathbf{X}_{(n)}$.

Next, we consider a bound of P_2 . By the fact

$$\sum_{j \in J_p} \left\{ Y_j \log \frac{p(X_j)}{p_0(X_j)} + (1 - Y_j) \log \frac{1 - p(X_j)}{1 - p_0(X_j)} \right\} \leq N_n \log \frac{1}{\delta},$$

the probability P_2 is bounded by

$$P_2 \leq \text{Prob} \left(\exists p \in \mathcal{F}(\mathbf{X}_{(n)}) \text{ such that } L_{J_p^c}(p) > (T - \alpha \log(1/\delta)) \log n \mid \mathbf{X}_{(n)} \right). \quad (29)$$

From II), we have

$$P_2 \leq B_d n^{-8(T - \alpha \log(1/\delta)) + \nu}. \quad (30)$$

Since all the previous arguments do not depend on a specific sample $\mathbf{X}_{(n)}$, eqs.(26), (28), and (30) complete the proof of the theorem.

Acknowledgements

This work is partially supported by the Grants-in-Aid for Scientific Research 13780181.

References

- Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics* 25, 573–578.
- Csörgő, M. and L. Horváth (1996). *Limit Theorems in Change-Point Analysis*. John Wiley and Sons.
- Dudley, R. M. (1984). A course on empirical processes. In *Lecture Notes in Mathematics, 1097. École d'Été de Probabilités de Saint Flour XII - 1982*, pp. 1–142. Springer.
- Fukumizu, K. (2003). Likelihood ratio of unidentifiable models and multilayer neural networks. *The Annals of Statistics* 31(3), To be published.

- Hagiwara, K. (2001). On the training error and generalization error of neural network regression without identifiability. In *Proceedings of the Fifth International Conference on Knowledge-Based Intelligent Information Engineering Systems and Allied Technologies*, Volume 2, pp. pp.1575–1579. IOS Press.
- Hagiwara, K., T. Hayasaka, N. Toda, S. Usui, and K. Kuno (2001). Upper bound of the expected training error of neural network regression for a gaussian noise sequence. *Neural Networks* 14(10), 1419–1429.
- Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, pp. 807–810.
- Haussler, D. (1992). Decision theoretic generalization of the pac model for neural net and other learning applications. *Information and Computation* 100, 78–150.
- Pollard, D. (1984). *Convergence of stochastic processes*. Springer.
- Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review* 56(1), 49–62.
- van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes*. Springer Verlag.
- Vapnik, V. N. (1982). *Estimation of Dependences Based on Empirical Data*. Springer.
- Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience.