A general upper bound of likelihood ratio for regression

Kenji Fukumizu*
Institute of Statistical Mathematics

Katsuyuki Hagiwara
Nagoya Institute of Technology

July 4, 2003

Abstract

This paper discusses the likelihood ratio test statistics (LRTS) in regression problems, and derives a general upper bound of the asymptotic order of LRTS for the sample size n. In some cases of estimation, where the true parameter is not identifiable, the LRTS diverges to infinity asymptotically. It is also known that the LRTS of some nonlinear models has a lower bound of the order $\log n$. This paper shows $\log n$ gives an upper bound of the asymptotic order of regression under very general assumptions, which are satisfied by many practical probability models including the Gaussian noise and the binary regression.

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 has been used for a basis of many statistical methods such as hypothesis test and model selection. The most well-known result on the asymptotics of LRTS is its convergence to the chi-square distribution under some regularity conditions. If we have a statistical model with a d-dimensional parameter and assume 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, if the regularity conditions do not hold, the convergence to the chi-square is not guaranteed, and various results on the asymptotic distribution have been obtained for specific cases. Among other works, Hotelling (1939) analyzes LRTS of nonlinear regression models for a finite sample size using a geometrical method. Chernoff (1954) gives a general expression of the LRTS by the conic approximation of a model. Mixture of chi-squares is known as a limiting distribution for a class of

*Part of this work was done while the author was visiting University of California, Berkeley.

models, in which the neighborhood of the true parameter can be approximated by a convex cone (Shapiro 1988).

It is also known that the LRTS may have a larger asymptotic order than the ordinary constant order $O_p(1)$, when the sample size n goes to infinity. 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. Bickel and Chernoff (1993) and Liu and Shao (2001) derive the asymptotic distribution of this LRTS, which has the order of $\log\log n$. In a change point problem, where the model assumes the existence of a change point against the null hypothesis of no change point, the asymptotic distribution of the LRTS is known to be of the order $\log\log n$ (Csörgő and Horváth 1996). These examples suggest that one cannot describe the local behavior of the maximum likelihood estimator by finite dimensional sufficient statistics, and must incorporate the infinite degree of freedom in general. In this line of research, Fukumizu (2003) considers divergence of LRTS from the viewpoint of infinite number of orthogonal score functions around a singularity 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 paper is to show a general upper bound $O_p(\log n)$ for the LRTS in regression models. This bound is derived under mild conditions on the class of regression functions and the probability model. Thus the result is generally applicable to many practical regression problems, including the Gaussian noise model and binary regression. The asymptotic order is not only the first step for the exact distribution, but it will be meaningful for discussing statistical problems on models that show divergence of LRTS; it can be used, for example, to design the ratio of a penalty term in the penalized likelihood approach.

There have been some existing results on the $\log n$ order of LRTS. Hagiwara et al. (2001) discuss LRTS for a type of Gaussian nonlinear regression defined by neural networks, which can approximate the point-mass function, and derive a lower bound of $\log n$ for the null hypothesis that the regressor is constant zero and the samples are i.i.d. normal random variables. Fukumizu (2003) extends this $O_p(\log n)$ lower bound to a much wider class of nonlinear regression, which essentially focuses on neural networks. The result covers an arbitrary bounded function as the true regression, and requires only mild conditions on probability models. Combined with the lower bound in Fukumizu (2003), the main theorems of this paper show that the LRTS of some type of nonlinear regression has exactly the $\log n$ order. In Hagiwara (2001), the upper bound $O_p(\log n)$ has been previously obtained for a special type of radial basis function model, in which the location parameters are restricted at the sample points of the covariate. This paper pursuits the upper bound of $\log n$ by extending the idea in Hagiwara (2001), which uses the exponential inequality for large deviation.

The main mathematical technique used in this paper is exponential inequalities on the supremum of the sum of independent variables over a function class. Such inequalities often appear in the field of empirical processes (Dudley (1984), Pollard (1984), van der Vaart and Wellner (1996)) and computational learning

theory (Vapnik (1982), Vapnik (1998), Haussler (1992)). As we show in Lemma 3, the $\log n$ upper bound is easily obtained for the log likelihood of an individual probability density function. A typical course for obtaining an upper bound over a function class is to replace the supremum over infinite number of functions with the maximum over finite ones, which can be taken by assuming finiteness of covering number. While we also follow this general scheme in our discussion, one difficulty is the unboundedness of the log likelihood function. In general, a finite covering can be taken only for a function class which admits a uniform bound for all the functions. To solve this problem, we develop a method of dividing the function class into the unbounded part and the bounded part, which depend on the sample size n, and show that the unbounded part does not contribute significantly to the value of maximum likelihood.

2 Main theorems

Let $(\mathcal{X}, \mathfrak{B}, \mu)$ be a measure space, $\varphi_0 : \mathcal{X} \to \mathbb{R}$ be a measurable function, and $\mathcal{F}$ be a class of measurable functions from $\mathcal{X}$ to $\mathbb{R}$. Suppose that $p(y|u)$ is a parametric probability density function on $\mathbb{R}$ with respect to Borel measure, where $u \in \mathbb{R}$ is a parameter. Given $\mathbf{X}_n = \{X_i\}_{i=1}^n \subset \mathcal{X}^n$, we have independent random variables Y_i $(1 \leq i \leq n)$, each of which follows the law

$$Y_i \sim p(y|\varphi_0(X_i))dy.$$

For a given sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, the (log) likelihood ratio test statistics (LRTS) is defined by

$$\sup_{\varphi \in \mathcal{F}} L_n(\varphi), \tag{1}$$

where

$$L_n(\varphi) = \sum_{i=1}^n \log \frac{p(Y_i|\varphi(X_i))}{p(Y_i|\varphi_0(X_i))}. \tag{2}$$

The variables X_i may be either random or deterministic.

We use the Vapnik-Čhervonenkis dimension (VC-dimension, Vapnik 1982, Vapnik 1998) to restrict the complexity of the function class. Let $\mathcal{C}$ be a class of subsets of a set Ω. The VC-dimension of $\mathcal{C}$ is the largest integer m such that there are $z_1, \ldots, z_m$ in Ω for which $\{(\mathbf{1}_A(z_1), \ldots, \mathbf{1}_A(z_m)) \in \{0,1\}^m \mid A \in \mathcal{C}\} = \{0,1\}^m$ is satisfied, where $\mathbf{1}_A(z)$ is the indicator function of a set A. The VC-dimension of a function class $\mathcal{F}$ is defined by the VC-dimension of the subgraphs $\{\{(x,y) \in \mathcal{X} \times \mathbb{R} \mid y \leq \varphi(x)\} \mid \varphi \in \mathcal{F}\}$, and denoted by $\dim_{VC}\mathcal{F}$.

To state the main theorem, we need some assumptions on the probability model $p(y|u)$;

Assumption (A)

(A-I). For any $B > 0$, there exist a function $A_1(y;u)$ and a constant $\alpha > 0$ such that the inequality

$$\log p(y|u_2) - \log p(y|u_1) \leq A_1(y;u_1)(u_2 - u_1) \tag{3}$$

holds for all $u_1, u_2 \in \mathbb{R}$, and

$$\overline{\lim_{R\to\infty}} \sup_{u_0\in[-B,B]} E_{Y|u_0}\big[\sup_{|u|\le R}|A_1(Y;u)|\big]R^{-\alpha} < +\infty \tag{4}$$

is satisfied, where $E_{Y|u}$ denotes the expectation of Y with respect to the probability $p(y|u)dy$.

(A-II). For any $B > 0$, there exist constants $C > 0$, $\beta > 1$ and a function $A_2(y;u)$ such that the inequality

$$\log p(y|u) - \log p(y|u_0) \le A_2(y;u_0)(u-u_0) - C|u-u_0|^\beta \tag{5}$$

holds for all $u_0 \in [-B,B]$ and $u \in \mathbb{R}$, and the function $A_2(y;u)$ satisfies either one of the following two conditions;

(a) for $\gamma > 1$, which is given by $\frac{1}{\beta} + \frac{1}{\gamma} = 1$,

$$\sup_{u_0\in[-B,B]} E_{Y|u_0}[|A_2(Y;u_0)|^\gamma] < +\infty, \tag{6}$$

(b) there is $\delta > 0$ such that

$$\sup_{\{u_i^{(0)}\}\subset[-B,B]} E\Big[\max_{1\le i\le n}|A_2(Y_i;u_i^{(0)})|\Big] < n^\delta, \tag{7}$$

where Y_i follows the law $p(y_i|u_i^{(0)})dy$.

Theorem 1. *Assume that* $\dim_{VC}\mathcal{F} < \infty$ *and there exists* $B_0 > 0$ *such that* $|\varphi_0| \le B_0$. *If Assumption (A) is satisfied, then there exist constants* $T > 0$ *and* $a > 0$ *such that the bound*

$$\mathrm{Prob}\Big(\sup_{\varphi\in\mathcal{F}} L_n(\varphi) > T\log n \;\Big|\; \mathbf{X}_n\Big) \le n^{-a}$$

holds for any $\mathbf{X}_n$ *and sufficiently large* n.

While Assumption (A) covers many probability models for practical regression problems, binary regression is an example which does not satisfy (A-II). For binary regression, in which the variable Y takes values in $\{0,1\}$, under the assumption that the conditional probabilities of $Y = 1$ is within the interval $(0,1)$, the generic form of the probability model is the logistic model:

$$p(y|u) = \frac{e^{yu}}{1+e^u}.$$

The log likelihood satisfies

$$\log\frac{p(y|u_2)}{p(y|u_1)} = y(u_2-u_1) - \log\frac{1+e^{u_2}}{1+e^{u_1}},$$

in which the second term of the right hand side is asymptotically linear for a large u_2. Thus, we cannot find $\beta > 1$ in the assumption (A-II).

As we show in Theorem 2, however, the same statement as Theorem 1 holds for binary regression without Assumption (A).

Theorem 2. *Assume that the range of a function in $\mathcal{F}$ is $(0,1)$, $\dim_{VC}\mathcal{F} < \infty$, and there exists $B_0 > 0$ such that $|\varphi_0| \le B_0$. If the variable Y takes values in $\{0,1\}$ and the probability model is given by the logistic model, then there exist constants $T > 0$ and $a > 0$ such that the bound*

$$\mathrm{Prob}\Big(\sup_{\varphi\in\mathcal{F}} L_n(\varphi) > T\log n \;\Big|\; \mathbf{X}_n\Big) \le n^{-a}$$

holds for any $\mathbf{X}_n$ and sufficiently large n.

In the above theorems, the finiteness of VC-dimension is a natural assumption to exclude such function classes that can fit an arbitrary number of data points without errors. Thus the above theorems provide the universal upper bound of LRTS for regression models.

Note that the logistic model satisfies (A-I), which is used in the proof of the theorems as a common assumption. The assumption (A-II) works for preventing a function with a very large absolute value from contributing significantly to the likelihood function. For binary logistic regression, it is more difficult to exclude the contribution of such functions; the larger the value of $\varphi(X_i)$ is, the better it fits the sample with $Y_i = 1$. Thus, for binary regression, we need more elaborate discussion using VC-dimension of $\mathcal{F}$ to derive the bound.

A class of probabilities which satisfies Assumption (A) is given by an exponential family. Suppose the probability model $p(y|u)$ is an exponential family

$$p(y|u) = \exp\{\eta(y)u + \tau(x) - \psi(u)\}.$$

By the convexity of the cummulant generating function $\psi(u)$, we have

$$\begin{aligned}\log\frac{p(y|u_2)}{p(y|u_1)} &= \eta(y)(u_2-u_1) - (\psi(u_2)-\psi(u_1)) \\ &\le (\eta(y)-\psi'(u_1))(u_2-u_1).\end{aligned}$$

If we assume $\psi'(u)$ is bounded by a polynomial order, that is, if there exist $\alpha > 0$ and $D > 0$ such that $|\psi'(u)| \le D|u|^\alpha$ for all u, then, by defining $A_1(y;u) = \eta(y) - \psi'(u)$, we see $p(y|u)$ satisfies the condition of (A-I). If further $\psi(u)$ admits

$$\psi(u_2) - \psi(u_1) \le \psi'(u_1)(u_2-u_1) + F(u_1)|u_2-u_1|^\beta$$

for some continuous function $F(u)$ and constant $\beta > 1$, the assumption (A-II) is satisfied; in fact, (A-II)-(a) holds, because the moment of $\eta(y)$ for any order exists as a continuous function on u. The normal distribution is one of such probabilities that satisfy those assumptions, as the cummulant generating function $\psi_G(u) = u^2/2$ admits

$$\psi_G(u_2) - \psi_G(u_1) = \psi_G'(u_1)(u_2-u_1) + \frac{1}{2}(u_2-u_1)^2.$$

3 Proof of the theorems

First, we show a simple lemma on the bound of the log likelihood for a single probability density function.

Lemma 3. *Let m be a natural number, and $p_{0,1}(y), \dots, p_{0,m}(y)$ and $p_1(y), \dots, p_m(y)$ be probability density functions on a measure space $(\Omega, \mathcal{B}, \mu)$. Suppose $Y_1, \dots, Y_m$ are independent samples from $p_{0,1}\mu, \dots, p_{0,m}\mu$, respectively. Then, for an arbitrary $T > 0$ and a natural number n, we have*

$$\mathrm{Prob}\Big(\sum_{i=1}^{m} \log \frac{p_i(Y_i)}{p_{0,i}(Y_i)} \geq T \log n\Big) \leq n^{-T}.$$

Proof. From Chebyshev's inequality with the exponential function, the probability is upper bounded by

$$e^{-T \log n} \prod_{i=1}^{m} E_{p_{0,i}}\Big[\frac{p_i(Y_i)}{p_{0,i}(Y_i)}\Big] = n^{-T}.$$

□

We use the ε-covering number $\mathcal{N}(\varepsilon, \mathcal{G}, \|\cdot\|_2)$ of a function class $\mathcal{G}$ with respect to an L^2 norm $\|\cdot\|_2$. The ε-covering number is defined by the smallest number of functions $\{f_h\} \subset L^2$ such that for every $g \in \mathcal{G}$ there exits f_h that satisfies $\|g - f_h\|_2 < \varepsilon$. It is known that if $d = \dim_{VC}\mathcal{G}$ is finite and $|g| \leq B$ for any $g \in \mathcal{G}$, the ε-covering number $\mathcal{N}(\varepsilon, \mathcal{G}, \|\cdot\|_2)$ is no more than $H_d(B/\varepsilon)^{2d}$ for any $\varepsilon > 0$, where H_d is a universal constant (Section 2.6, van der Vaart and Wellner (1996); see also Lemma 25, Pollard (1984)).

We show theorems 1 and 2 in the same proof except when we use the assumption (A-II).

Proof of Theorems 1 and 2. We take and fix positive constants τ and λ so that they satisfy $\lambda > 1 + \delta/(\beta - 1)$ and $\tau > 1 + \alpha\lambda$, where α, β, and δ are given by the assumptions of the theorems. In the following, we fix $\mathbf{X}_n$, and regard $\mathcal{F}$ as a class of functions from $\mathbf{X}_n$ to $\mathbb{R}$. Note that all the constants taken in the proof do not depend on $\mathbf{X}_n$. We define a norm $\|\cdot\|_2$ on the functions on $\mathbf{X}_n$ by

$$\|\varphi\|_2 = \sqrt{\frac{1}{n}\textstyle\sum_{i=1}^{n} \varphi(X_i)^2},$$

which is the L^2 norm with respect to the uniform probability measure. Obviously, $\|\varphi\|_2 \leq 1/n^r$ implies $|\varphi(X_i)| \leq 1/n^{r-1/2}$ for all $1 \leq i \leq n$.

For a function $\varphi : \mathcal{X} \to \mathbb{R}$, a function $b_n(\varphi)$ on $\mathbf{X}_n$ is defined by

$$b_n(\varphi)(X_i) = \begin{cases} n^\lambda & \text{if } \varphi(X_i) \geq n^\lambda, \\ \varphi(X_i) & \text{if } -n^\lambda \leq \varphi(X_i) < n^\lambda, \\ -n^\lambda & \text{if } \varphi(X_i) < -n^\lambda. \end{cases}$$

A function class $\tilde{\mathcal{F}}_n(\mathbf{X}_n)$ on $\mathbf{X}_n$ is defined by

$$\tilde{\mathcal{F}}_n(\mathbf{X}_n) := \big\{\psi : \mathbf{X}_n \to \mathbb{R} \;\big|\; \text{there exists } \varphi \in \mathcal{F} \text{ such that } \psi = b_n(\varphi)\big\}.$$

It is easy to see $d := \dim_{VC}\tilde{\mathcal{F}}_n(\mathbf{X}_n) \le \dim_{VC}\mathcal{F} < \infty$. Thus, there are ℓ_n functions $\{\psi_n^{[k]} \mid k = 1, \ldots, \ell_n\}$ on $\mathbf{X}_n$ such that for an arbitrary $\psi \in \tilde{\mathcal{F}}_n(\mathbf{X}_n)$ there exists $\psi_n^{[k]}$ with $\|\psi - \psi_n^{[k]}\|_2 \le 1/n^{\tau+1/2}$. Since any function in $\tilde{\mathcal{F}}_n(\mathbf{X}_n)$ is bounded by n^λ, we have

$$\ell_n \le \mathcal{N}(1/n^{\tau+1/2}, \tilde{\mathcal{F}}_n(\mathbf{X}_n), \|\cdot\|_2) \le H_d n^{2d(\tau+1/2)\lambda}. \tag{8}$$

For a function $\varphi : \mathbf{X}_n \to \mathbb{R}$, we define I_φ and J_φ by

$$I_\varphi = \{i \in \{1, \ldots, n\} \mid -n^\lambda \le \varphi(X_i) < n^\lambda\}$$

and

$$J_\varphi = \{1, \ldots, n\} - I_\varphi,$$

respectively. Thus, the following upper bound is obvious;

$$\begin{aligned}
&\mathrm{Prob}\Big(\sup_{\varphi\in\mathcal{F}} L_n(\varphi) \ge T\log n \;\Big|\; \mathbf{X}_n\Big) \\
&\le \mathrm{Prob}\Big(\sup_{\varphi\in\mathcal{F}} \sum_{i\in I_\varphi} \log\frac{p(Y_i|\varphi(X_i))}{p(Y_i|\varphi_0(X_i))} \ge \frac{T}{2}\log n \;\Big|\; \mathbf{X}_n\Big) \\
&\quad + \mathrm{Prob}\Big(\sup_{\varphi\in\mathcal{F}} \sum_{i\in J_\varphi} \log\frac{p(Y_i|\varphi(X_i))}{p(Y_i|\varphi_0(X_i))} \ge \frac{T}{2}\log n \;\Big|\; \mathbf{X}_n\Big) \quad =: P^I + P^{II}.
\end{aligned}$$

(i) Bound of P^I

Under the assumption (A-I), which are satisfied by both the theorems, we will prove that there exist $T > 0$ and $\xi > 0$ such that the inequality

$$P^I \le n^{-\xi}$$

holds for any $\mathbf{X}_n$ and sufficiently large n.

Let $\mathcal{I}_n$ be a family of indices defined by

$$\mathcal{I}_n = \{I \subset \{1, \ldots, n\} \mid \text{ there exists } \varphi \in \mathcal{F} \text{ such that } I = I_\varphi\}.$$

For a function φ and $z = (x, y) \in \mathcal{X} \times \mathbb{R}$, let $G(z;\varphi)$ be the indicator function of the subgraph of φ; that is, $G(z;\varphi) = 1$ if $y \le \varphi(x)$, and $G(z;\varphi) = 0$ otherwise. Then, for the $2n$ points $Z_i^+ = (X_i, n^\lambda)$, $Z_i^- = (X_i, -n^\lambda)$ $(1 \le i \le n)$, we can see the following three equivalence relations; $\varphi(X_i) \ge n^\lambda$ if and only if $G(Z_i^+;\varphi) = G(Z_i^-;\varphi) = 1$; $\varphi(X_i) < -n^\lambda$ if and only if $G(Z_i^+;\varphi) = G(Z_i^-;\varphi) = 0$; and $-n^\lambda \le \varphi(X_i) < n^\lambda$ if and only if $G(Z_i^+;\varphi) = 0$ and $G(Z_i^-;\varphi) = 1$. From this fact, the cardinality of $\mathcal{I}_n$ is the same as that of the set $\{(G(Z_i^+;\varphi), G(Z_i^-;\varphi))_{i=1}^n \in \{0,1\}^{2n} \mid \varphi \in \mathcal{F}\}$. By the fact $\dim_{VC}\mathcal{F} = d < \infty$, we have

$$|\mathcal{I}_n| \le K_d(2n)^d \tag{9}$$

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

From the inequality

$$\begin{aligned}
&\sup_{\varphi\in\mathcal{F}} \sum_{i\in I_\varphi} \log \frac{p(Y_i|\varphi(X_i))}{p(Y_i|\varphi_0(X_i))} \\
&= \sup_{\varphi\in\mathcal{F}} \min_{1\le k\le \ell_n} \Big\{ \sum_{i\in I_\varphi} \log \frac{p(Y_i|\psi_n^{[k]}(X_i))}{p(Y_i|\varphi_0(X_i))} + \sum_{i\in I_\varphi} \log \frac{p(Y_i|\varphi(X_i))}{p(Y_i|\psi_n^{[k]}(X_i))} \Big\} \\
&\le \max_{I\in\mathcal{I}_n} \max_{1\le k\le \ell_n} \sum_{i\in I} \log \frac{p(Y_i|\psi_n^{[k]}(X_i))}{p(Y_i|\varphi_0(X_i))} + \sup_{\varphi\in\mathcal{F}} \min_{1\le k\le \ell_n} \sum_{i\in I_\varphi} \log \frac{p(Y_i|\varphi(X_i))}{p(Y_i|\psi_n^{[k]}(X_i))},
\end{aligned}$$

the upper bound of P^I is provided by

$$\begin{aligned}
P^I &\le |\mathcal{I}_n|\, \ell_n \max_{\substack{1\le k\le \ell_n \\ I\in\mathcal{I}_n}} \mathrm{Prob}\Big(\sum_{i\in I} \log \frac{p(Y_i|\psi_n^{[k]}(X_i))}{p(Y_i|\varphi_0(X_i))} \ge \frac{T}{4}\log n \;\Big|\; \mathbf{X}_n \Big) \\
&\quad + \mathrm{Prob}\Big(\sup_{\varphi\in\mathcal{F}} \min_{1\le k\le \ell_n} \sum_{i\in I_\varphi} \log \frac{p(Y_i|\varphi(X_i))}{p(Y_i|\psi_n^{[k]}(X_i))} \ge \frac{T}{4}\log n \;\Big|\; \mathbf{X}_n \Big) \\
&=: P^{I,1} + P^{I,2}.
\end{aligned}$$

From Eqs.(8), (9), and Lemma 3, we have

$$P^{I,1} \le H_d K_d 2^d\, n^{2d(\tau+1/2)\lambda + d - T/4}.$$

For a sufficiently large T, the exponent of n is a negative constant.

Next, we derive an upper bound of $P^{I,2}$ using assumption (A-I). Because φ and $b_n(\varphi)$ have the same value at X_i for $i \in I_\varphi$, for an arbitrary $\varphi \in \mathcal{F}$ there exists k_φ with $1 \le k_\varphi \le \ell_n$ such that $|\varphi(X_i) - \psi_n^{[k_\varphi]}(X_i)| \le 1/n^\tau$ for all $i \in I_\varphi$. By taking such k_φ, we obtain

$$\begin{aligned}
\sup_{\varphi\in\mathcal{F}} \min_{1\le k\le \ell_n} \sum_{i\in I_\varphi} \log \frac{p(Y_i|\varphi(X_i))}{p(Y_i|\psi_n^{[k]}(X_i))} &\le \sup_{\varphi\in\mathcal{F}} \sum_{i\in I_\varphi} A_1(Y_i; \psi_n^{[k_\varphi]}(X_i))(\varphi(X_i) - \psi^{[k_\varphi]}(X_i)) \\
&\le \sum_{i=1}^n S_{n^\lambda}(Y_i) \frac{1}{n^\tau},
\end{aligned}$$

where $S_R(y) := \sup_{|u|\le R} |A_1(y;u)|$. From the assumption (A-I) and Chebyshev's inequality, the upper bound of $P^{I,2}$ is given by

$$P^{I,2} \le \mathrm{Prob}\Big(\sum_{i=1}^n S_{n^\lambda}(Y_i) \ge n^\tau \;\Big|\; \mathbf{X}_n \Big) \le \frac{\sum_{i=1}^n E[S_{n^\lambda}(Y_i)]}{n^\tau} \le C' \frac{n^{\alpha\lambda+1}}{n^\tau},$$

where C' is a constant. Because τ is taken so that $\tau > \alpha\lambda + 1$, the exponent of n is a negative constant.

(ii) Bound of P^{II} under Assumption (A-II)

From the assumption (A-II), if the inequality

$$\sum_{i\in J_\varphi} \log \frac{p(Y_i|\varphi(x_i))}{p(Y_i|\varphi_0(x_i))} > 0$$

holds, we have

$$\sum_{i\in J_\varphi} |A_2(Y_i;\varphi_0(X_i))||\varphi(X_i)-\varphi_0(X_i)| > C\sum_{i\in J_\varphi} |\varphi(X_i)-\varphi_0(X_i)|^\beta. \qquad (10)$$

First, suppose (A-II)-(a) holds. By Hölder's inequality, Eq.(10) means

$$\Big(\sum_{i\in J_\varphi} |A_2(Y_i;\varphi_0(X_i))|^\gamma\Big)^{1/\gamma}\Big(\sum_{i\in J_\varphi} |\varphi(X_i)-\varphi_0(X_i)|^\beta\Big)^{1/\beta} > C\sum_{i\in J_\varphi} |\varphi(X_i)-\varphi_0(X_i)|^\beta.$$

If n is sufficiently large so that $n > \max\{B_0^{1/(\lambda-1)}, 2\}$, we see $|\varphi(X_i)-\varphi_0(X_i)| > (n-1)n^{\lambda-1} \geq n^{\lambda-1}$ for all $i \in J_\varphi$. Thus, the above inequality leads

$$\sum_{i\in J_\varphi} |A_2(Y_i;\varphi_0(X_i))|^\gamma > C^\gamma |J_\varphi| n^{\beta(\lambda-1)}.$$

By Chebyshev's inequality, we obtain

$$\begin{aligned} P^{II} &\leq \mathrm{Prob}\Big(\sum_{i\in J_\varphi} |A_2(Y_i;\varphi_0(X_i))|^\gamma > C^\gamma |J_\varphi| n^{\beta(\lambda-1)} \;\Big|\; \mathbf{X}_n\Big) \\ &\leq \frac{\sum_{i\in J_\varphi} E_{Y_i|\varphi_0(X_i)}[|A_2(Y_i;\varphi_0(X_i))|^\gamma]}{C^\gamma n^{\beta(\lambda-1)}|J_\varphi|} \quad \leq \frac{D}{C^\gamma} n^{-\beta(\lambda-1)}, \end{aligned}$$

where $D = \sup_{|u|\leq B_0} E_{Y|u}[|A_2(Y;u)|^\gamma] < \infty$. In the last line, the exponent of n is a negative constant, since we take $\lambda > 1$.

Next, assume (A-II)-(b). From Eq.(10), we have

$$\max_{i\in J_\varphi} |A_2(Y_i;\varphi_0(X_i))| \sum_{i\in J_\varphi} |\varphi(X_i)-\varphi_0(X_i)| > C\sum_{i\in J_\varphi} |\varphi(X_i)-\varphi_0(X_i)|^\beta.$$

Then, Hölder's inequality shows

$$\max_{i\in J_\varphi} |A_2(Y_i;\varphi_0(X_i))|\Big(\sum_{i\in J_\varphi} |\varphi(X_i)-\varphi_0(X_i)|^\beta\Big)^{1/\beta} |J_\varphi|^{1/\gamma} > C\sum_{i\in J_\varphi} |\varphi(X_i)-\varphi_0(X_i)|^\beta.$$

By a similar argument to the previous case, the bound

$$\max_{i\in J_\varphi} |A_2(Y_i;\varphi_0(X_i))| > Cn^{(\beta-1)(\lambda-1)}$$

must be satisfied for sufficiently large n. Thus, by Chebyshev's inequality and the assumption (A-II)-(b), we obtain

$$P^{II} \leq \frac{E\big[\max_{1\leq i\leq n}|A_2(Y_i;\varphi_0(X_i))|\big]}{Cn^{(\beta-1)(\lambda-1)}} \leq \frac{n^{\delta}}{Cn^{(\beta-1)(\lambda-1)}}.$$

Since $\beta > 1$ and $\lambda > 1 + \delta/(\beta-1)$, the exponent of n is a negative constant.

(iii) Bound of P^{II} for binary regression

Fix $\kappa > 0$ so that $\kappa \leq 1/(1+e^{B_0}) \leq e^{B_0}/(1+e^{B_0}) \leq 1-\kappa$. Take $\zeta > 2d/\kappa^2$, and let $N_n := \zeta \log n$. By partitioning $\mathcal{F}$, we have

$$\sup_{\varphi\in\mathcal{F}} \sum_{i\in J_\varphi} \log \frac{p(Y_i|\varphi(X_i))}{p(Y_i|\varphi_0(X_i))} \leq \sup_{\substack{\varphi\in\mathcal{F}\\ |J_\varphi|\leq N_n}} \sum_{i\in J_\varphi} \log \frac{p(Y_i|\varphi(X_i))}{p(Y_i|\varphi_0(X_i))} + \sup_{\substack{\varphi\in\mathcal{F}\\ |J_\varphi|> N_n}} \sum_{i\in J_\varphi} \log \frac{p(Y_i|\varphi(X_i))}{p(Y_i|\varphi_0(X_i))}.$$

From the inequality $\log \frac{p(Y_i|\varphi(X_i))}{p(Y_i|\varphi_0(X_i))} \leq \log(1/\kappa)$ for all $1 \leq i \leq n$, the first term in the right hand side is upper bounded by $\zeta \log(1/\kappa)\log n$. Thus, for $T > 2\zeta\log(1/\kappa)$, the probability P^{II} is bounded by

$$P^{II} \leq \mathrm{Prob}\left(\sup_{\substack{\varphi\in\mathcal{F}\\ |J_\varphi|> N_n}} \sum_{i\in J_\varphi} \log \frac{p(Y_i|\varphi(X_i))}{p(Y_i|\varphi_0(X_i))} \geq 0 \;\Big|\; \mathbf{X}_n \right). \tag{11}$$

We define a label $t_j(\varphi)$ for $\varphi \in \mathcal{F}$ and $j \in J_\varphi$ by

$$t_j(\varphi) = \begin{cases} 1 & \text{if } \varphi(X_j) \geq n^{\lambda} \\ 0 & \text{if } \varphi(X_j) < -n^{\lambda}. \end{cases}$$

Using these labels, we obtain

$$\begin{aligned} \sum_{i\in J_\varphi} \log \frac{p(Y_i|\varphi(X_i))}{p(Y_i|\varphi_0(X_i))} &= \sum_{\substack{i\in J_\varphi\\ t_i(\varphi)=Y_i}} \log \frac{p(Y_i|\varphi(X_i))}{p(Y_i|\varphi_0(X_i))} + \sum_{\substack{i\in J_\varphi\\ t_i(\varphi)\neq Y_i}} \log \frac{p(Y_i|\varphi(X_i))}{p(Y_i|\varphi_0(X_i))} \\ &\leq \sum_{\substack{i\in J_\varphi\\ t_i(\varphi)=Y_i}} \log \frac{1}{\kappa} + \sum_{\substack{i\in J_\varphi\\ t_i(\varphi)\neq Y_i}} \log \frac{1/(1+e^{n^{\lambda}})}{\kappa} \\ &= |J_\varphi|\log(1/\kappa) - \big|\{i\in J_\varphi \mid t_i(\varphi)\neq Y_i\}\big| \cdot \log(1+e^{n^{\lambda}}). \end{aligned}$$

Combined with Eq.(11), this leads

$$P^{II} \leq \mathrm{Prob}\Big(\text{there exits } \varphi\in\mathcal{F} \text{ such that } |J_\varphi| \geq N_n \text{ and } \frac{|\{i\in J_\varphi \mid t_i(\varphi)\neq Y_i\}|}{|J_\varphi|} < \frac{\log(1/\kappa)}{n^{\lambda}} \;\Big|\; \mathbf{X}_n\Big). \tag{12}$$

Let $\mathcal{V}_n$ be the set of points, which is defined by

$$\mathcal{V}_n = \{(X_j, t_j(\varphi))_{j \in J_\varphi} \mid \varphi \in \mathcal{F}\}.$$

By a similar argument to the derivation of the bound of $|\mathcal{I}_n|$, we see

$$|\mathcal{V}_n| \leq 2^d \, K_d \, n^d \tag{13}$$

for sufficiently large n. For a fixed $\Gamma = (X_j, t_j)_{j \in J_\Gamma} \in \mathcal{V}_n$, where J_Γ is the index set corresponding to Γ, we define a random variable U_Γ by

$$U_\Gamma = |\{j \in J_\Gamma \mid Y_j \neq t_j\}|,$$

where Y_j follows the law $p(y|\varphi_0(X_j))dy$ independently. The expectation of U_Γ is given by

$$E[U_\Gamma | \mathbf{X}_n] = \sum_{j \in J_\Gamma} t_j p(0|\varphi_0(X_j)) + \sum_{j \in J_\Gamma} (1 - t_j) p(1|\varphi_0(X_j)) \geq |J_\Gamma| \kappa.$$

Note that $|J_\Gamma| \geq N_n = \zeta \log n$ is assumed for $\Gamma \in \mathcal{V}_n$. If n is sufficiently large so that $\frac{1}{2}|J_\Gamma|\kappa > \frac{\log(1/\kappa)}{|J_\Gamma|^{\lambda-1}}$, by Hoeffding's inequality we obtain

$$\begin{aligned}
&\mathrm{Prob}\Big(\frac{|\{j \in J_\Gamma \mid Y_j \neq t_j\}|}{|J_\Gamma|} < \frac{\log(1/\kappa)}{n^\lambda} \;\Big|\; \mathbf{X}_n\Big) \\
&\leq \mathrm{Prob}\Big(U_\Gamma - E[U_\Gamma|\mathbf{X}_n] < \frac{\log(1/\kappa)}{|J_\Gamma|^{\lambda-1}} - E[U_\Gamma|\mathbf{X}X
\end{aligned}$$

References

Bickel, P. and H. Chernoff (1993). Asymptotic distribution of the likelihood ratio statisitcs in a prototypical non regular problems. In J. K. Ghosh, S. K. Mitra, K. R. Parthasarathy, and B. L. S. P. Rao (Eds.), *Statistics and Probability : A Raghu Raj Bahadur Festschrift*, pp. 83–96.

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 dÉté 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 Stastistics 31*(3), in press.

Hagiwara, K. (2001). On the training error and generalization error of neural network regression without identifiablity. 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.

Hotelling, H. (1939). Tubes and spheres in n-spaces, and a class of statistical problems. *American Journal of Mathematics 61*(2), 440–460.

Liu, X. and Y. Shao (2001). Asymptotic distribution of the likelihood ratio test in a two-component normal mixture model. Technical report, Department of Statistics, Columbia University.

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.