Sankhya manuscript No.
(will be inserted by the editor)

Asymptotics for regression models under loss of identifiability

Joseph Rynkiewicz

Sankhya A, 78(2), pp 155-179, 2016

Abstract This paper discusses the asymptotic behavior of regression models under general conditions, especially if the dimensionality of the set of true parameters is larger than zero and the true model is not identifiable. Firstly, we give a general inequality for the difference of the sum of square errors (SSE) of the estimated regression model and the SSE of the theoretical true regression function in our model. A set of generalized derivative functions is a key tool in deriving such inequality. Under suitable Donsker condition for this set, we provide the asymptotic distribution for the difference of SSE. We show how to get this Donsker property for parametric models even though the parameters characterizing the best regression function are not unique. This result is applied to neural networks regression models with redundant hidden units when loss of identifiability occurs and gives some hints on how penalizing such models to avoid over-fitting.

Keywords regression models · Donsker class · loss of identifiability · multilayer neural networks

Mathematics Subject Classification (2000) 62H10 · 62F12

1 Introduction

This paper discusses the asymptotic behavior of the sum of square errors (SSE) for regression models under general conditions. The asymptotics of the SSE is an significant problem in estimation theory and, under some regularity conditions, the convergence toward a law proportional to a χ^2 law is well known if the true model is identifiable. However, if there is a loss of identifiability in parameters i.e. $f_\theta = f_{\theta'}$ for

Joseph Rynkiewicz
SAMM, Université de Paris 1
Tel.: +33-144078705
Fax: +33-144078704
E-mail: joseph.rynkiewicz@univ-paris1.fr

some $\theta \neq \theta'$, where f_θ is the regression function, the Hessian matrix of the regression model may be singular and the asymptotics of the SSE is unknown. There are many regression models with loss of identifiability such as reduced rank regression (Fukumizu (1999)), radial basis functions (Hagiwara (2002)) and multilayer neural networks models (White (1992)). The behavior of the SSE in such models has not been clarified completely and many statistical methods such as model selection need special considerations. Our results on SSE have not been obtained by computational learning theory (Anthony and Bartlett (1999), Vapnik (1998)) which focuses on estimation error rather than the degree of over-fitting. While this approach benefits from being free from the previously mentioned loss of identifiability problem, it is not suitable for the detailed analysis of over-fitting and estimated parameters. In this approach an upper bound is obtained for the estimation error and the bound is used to evaluate the accuracy of the estimated model in terms of generalization capability (Anthony and Bartlett (1999), Devroye et al (1996)). The main technique is to consider the worst case that is bounded by considering the supremum of the difference between the generalization error and the training error over all possible models. Although this simplifies the mathematical problem so that the detailed properties of specific estimated parameters are not required, it makes it difficult to analyze over-fitting.

This paper provides a general approach for deriving the asymptotic of the SSE in these types of regression models. Let $\mathscr{F}$ be the family of possible regression functions and suppose that we observe a random sample

$$(X_1, Y_1), \cdots, (X_n, Y_n),$$

from the distribution P of a vector (X, Y), with Y a real random variable. The regression model can be written as:

$$Y = f_0(X) + \varepsilon,\ E(\varepsilon | X) = 0,\ E\left(\varepsilon^2 | X\right) = \sigma^2 < \infty. \tag{1}$$

We assume that the true regression function f_0 belongs to the set $\mathscr{F}$:

$$f_0 = \arg\min_{f \in \mathscr{F}} \|Y - f(X)\|_2,$$

where

$$\|g(Z)\|_2 := \sqrt{\int g(z)^2 dP(z)}$$

is the $\mathscr{L}^2$ norm for an square integrable function g. This assumption may seem to be strong, but it is related to the choice of the explanatory variable X. Indeed, if X is poorly chosen (say X and Y are independents), then any statistical model with an intercept contains the true regression function.

A natural estimator of f_0 is the least square estimator (LSE) $\hat{f}$ that minimizes the SSE:

$$\hat{f} = \arg\min_{f \in \mathscr{F}} \sum_{t=1}^{n} (Y_t - f(X_t))^2. \tag{2}$$

$\hat{f}$ is expected to converge to the function f_0 under suitable conditions. If $\mathscr{F}$ is a parametric family and Θ is a set of possible parameters, $\mathscr{F} = \{f_\theta, \theta \in \Theta\}$, the LSE is the parameter $\hat{\theta}$ that minimizes

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sum_{t=1}^{n} (Y_t - f_\theta(X_t))^2. \tag{3}$$

Let us write Θ_0 the set of parameters realizing the best regression function f_0: $\forall\theta \in \Theta_0$, $f_\theta = f_0$. If the set $\mathscr{F}$ is large enough, it may be possible that the dimension of the interior of the set Θ_0 is larger than zero and various difficulties arise in analyzing the statistical properties of estimators of f_0. This is for example the case if $\mathscr{F}$ contains multilayer neural networks with redundant hidden units (see Fukumizu (2003)).

Under loss of identifiability of the parameters, the asymptotics for likelihood functions has been studied by Liu and Shao (2003) who improve the method of Dacunha-Castelle and Gassiat (1999) and Dacunha-Castelle and Gassiat (1997). The authors establish a general quadratic approximation of the log-likelihood ratio in a neighborhood of the true density, which is valid with or without loss of identifiability. In this paper, we will use a similar idea, but here we are interested in regression functions, not in density functions, so we will introduce generalized derivative functions:

$$d_f(x) = \frac{f(x) - f_0(x)}{\|f(X) - f_0(X)\|_2}, f \neq f_0. \tag{4}$$

Under some general regularity conditions, this paper shows that

$$\lim_{n\to\infty} \left(\sum_{t=1}^{n} (Y_t - f_0(X_t))^2 - \sum_{t=1}^{n} \left(Y_t - \hat{f}(X_t)\right)^2 \right) = \sigma^2 \sup_{d\in\mathscr{D}} \max\left\{W(d)^2; 0\right\}, \tag{5}$$

where $\mathscr{D}$ is the set of $\mathscr{L}^2$ limits of the generalized derivative functions d_f as $\|f(X) - f_0(X)\|_2 \to 0$ and $(W(d))_{d\in\mathscr{D}}$ a centered Gaussian process with covariance being the scalar product in $L^2(P)$. Such a result allows, for example, to fully explicit the asymptotic behavior of the SSE when regression functions are multilayer neural networks, even if $\mathscr{F}$ is too big and contains neural networks with redundant hidden units realizing the true regression function f_0. Let us recall that a feedforward neural network is defined as follows: Let $x = (x_1, \cdots, x_d)^T \in \mathbb{R}^d$ be the vector of inputs, $w_i := (w_{i1}, \cdots, w_{id})^T \in \mathbb{R}^d$ be the parameter vector of the hidden unit i and ϕ a sigmoid function. The function represented by the network with k hidden units can be written:

$$f_\theta(x) = \beta + \sum_{i=1}^{k} a_i \phi\left(w_i^T x + b_i\right),$$

with $\theta = (\beta, a_1, \cdots, a_k, b_1, \cdots, b_k, w_1, \cdots, w_k)$ the parameter vector of the model. In this case, under suitable assumptions, a centered Gaussian process $\{W(d), d \in \mathscr{D}\}$ with continuous sample paths exists so that

$$\lim_{n\to\infty} \left(\sum_{t=1}^{n} (Y_t - f_0(X_t))^2 - \sum_{t=1}^{n} (Y_t - f_\theta(X_t))^2 \right) = \sigma^2 \sup_{d\in\mathscr{D}} \max\left\{(W(d))^2; 0\right\}.$$

This result shows that the degree of over-fitting is bounded in probability, but depends on the size of the asymptotic set $\mathscr{D}$. In order to reduce the over-fitting, we will see that we need to control the size of the limit index functions $\mathscr{D}$. Our computation of the exact form of elements of $\mathscr{D}$ will show that this can be done by reducing the number of hidden units and limiting the size of input parameters $w_1, \cdots, w_k$.

All our results are a consequence of a very general inequality: For all regression functions $f \in \mathscr{F}$, $f \neq f_0$, if $\varepsilon_t := Y_t - f_0(X_t)$ is the noise for index t then

$$\sum_{t=1}^{n} (Y_t - f_0(X_t))^2 - \sum_{t=1}^{n} (Y_t - f(X_t))^2 \leq \frac{\left(\frac{\sum_{t=1}^{n} \varepsilon_t d_f(X_t)}{\sqrt{n}}\right)^2}{\frac{\sum_{t=1}^{n} \left(d_f(X_t)\right)^2}{n}} \tag{6}$$

Moreover, if $\mathscr{S} = \left\{d_f, f \in \mathscr{F}, f \neq f_0\right\}$ is a Donsker class, $\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varepsilon_t d_f(X_t)$ converges uniformly to some zero-mean Gaussian process and we get the previous limit result (5). Note that, even when the set $\mathscr{F}$ is a regular parametric bounded family, the function $\theta \mapsto d_{f_\theta}(x)$ may be not extendable by continuity in $\theta_0 \in \Theta_0$, hence the Donsker property of the set of generalized derivative functions has to be carefully studied. This problem also occurs for the generalized score functions S_θ of Liu and Shao (2003), although it was not mentioned by the authors.

The paper is organized as follows: Section 2 establishes the asymptotic distribution of the SSE for regression models if the set of generalized derivative functions $\mathscr{S}$ is Donsker. In the next section, we show how to get the Donsker property for $\mathscr{S}$ in the parametric case under loss of identifiability. As an example, section 4 characterizes the asymptotic distribution of regression using neural networks with redundant hidden units and gives some hints on how to select a good model using this distribution. The long proofs of our results are postponed to section 5.

2 Asymptotic distribution of the SSE

In this section we establish a quadratic approximation to SSE in a neighborhood of the true regression function f_0. Here, the set $\mathscr{F}$ is not assumed to be parametric, hence our results may be applied to a more general framework such as non-parametric regression models. For the sake of simplicity, we consider identically distributed independent variables, but all the following results can be easily generalized to geometrically mixing stationary sequence of random variables as in Olteanu and Rynkiewicz (2012) or Gassiat (2002). For example, our results may be applied to non-linear autoregressive models using multilayer neural networks as in Yao (2000). Under fairly general conditions (including the regularity conditions of this paper) the LSE is consistent, so the asymptotic distribution of SSE is determined by the local properties of the regression function in a small $\mathscr{L}^2$-neighborhood of the true regression function f_0.

Firstly, we present some definitions.

Definition 1 Let P be a probability measure.

– We will use the abbreviation $Pf = \int f dP$ for an integrable function f.

– For a square integrable function g,

$$\|g(Z)\|_2 := \sqrt{\int g(z)^2 dP(z)}$$

is the $\mathscr{L}^2$ norm.
– For a vector $x = (x_1, \cdots, x_k)$, let us write $|x| = \sqrt{x_1^2 + \cdots + x_k^2}$ for the Euclidean norm. The envelope function of a class of functions $\mathscr{F}$ is defined as

$$F(x) \equiv \sup_{f \in \mathscr{F}} |f(x)|.$$

– A family of random sequences

$$\{Y_n(g), g \in \mathscr{G}, n = 1, 2, \cdots\}$$

is said to be uniformly $O_P(1)$ if for every $\delta > 0$, there exist constants $M > 0$ and $N(\delta, M)$ such that

$$P\left(\sup_{g \in \mathscr{G}} |Y_n(g)| \leq M\right) \geq 1 - \delta$$

for all $n \geq N(\delta, M)$.
– A family of random sequences

$$\{Y_n(g), g \in \mathscr{G}, n = 1, 2, \cdots\}$$

is said to be uniformly $o_P(1)$ if for every $\delta > 0$ and $\varepsilon > 0$ there exists a constant $N(\delta, \varepsilon)$ such that

$$P\left(\sup_{g \in \mathscr{G}} |Y_n(g)| < \varepsilon\right) \geq 1 - \delta$$

for all $n \geq N(\delta, \varepsilon)$.

2.1 Upper bound for the SSE

We prove this lemma which gives a very general upper bound for the sum of square errors.

Lemma 1 *For all regression functions $f \in \mathscr{F}$ with $f \neq f_0$ and d_f defined in (4):*

$$\sum_{t=1}^{n} (Y_t - f_0(X_t))^2 - \sum_{t=1}^{n} (Y_t - f(X_t))^2 \leq \frac{\left(\frac{\sum_{t=1}^{n} \varepsilon_t d_f(X_t)}{\sqrt{n}}\right)^2}{\frac{\sum_{t=1}^{n} \left(d_f(X_t)\right)^2}{n}}.$$

Proof We have

$$\begin{array}{l}\sum_{t=1}^{n}\left(Y_t-f_0(X_t)\right)^2-\sum_{t=1}^{n}\left(Y_t-f(X_t)\right)^2= \\ \sum_{t=1}^{n}\left(Y_t-f_0(X_t)\right)^2-\left(Y_t-f(X_t)\right)^2= \\ \sum_{t=1}^{n}\left(Y_t-f_0(X_t)\right)^2-\left(Y_t-f_0(X_t)+f_0(X_t)-f(X_t)\right)^2= \\ \sum_{t=1}^{n}2\varepsilon_t\left(f(X_t)-f_0(X_t)\right)-\left(f(X_t)-f_0(X_t)\right)^2.\end{array}$$

Since $\|f_0(X_t)-f(X_t)\|_2$ is independent of the index t due to identical distribution for X_t's, let us write

$$A=\|f_0(X_t)-f(X_t)\|_2\times\sqrt{\sum_{t=1}^{n}\left(\frac{f(X_t)-f_0(X_t)}{\|f(X_t)-f_0(X_t)\|_2}\right)^2}$$

and

$$Z=\frac{\sum_{t=1}^{n}\varepsilon_t\frac{f(X_t)-f_0(X_t)}{\|f(X_t)-f_0(X_t)\|_2}}{\sqrt{\sum_{t=1}^{n}\left(\frac{f(X_t)-f_0(X_t)}{\|f(X_t)-f_0(X_t)\|_2}\right)^2}},$$

then remark that $0\leq(Z-A)^2\Leftrightarrow 2AZ-A^2\leq Z^2$ implies that

$$\sum_{t=1}^{n}\left(Y_t-f_0(X_t)\right)^2-\left(Y_t-f(X_t)\right)^2\leq\frac{\left(\frac{\sum_{t=1}^{n}\varepsilon_t\frac{f(X_t)-f_0(X_t)}{\|f(X_t)-f_0(X_t)\|_2}}{\sqrt{n}}\right)^2}{\frac{\sum_{t=1}^{n}\left(\frac{f(X_t)-f_0(X_t)}{\|f(X_t)-f_0(X_t)\|_2}\right)^2}{n}}.$$

■

2.2 Approximation of the SSE

Define the limit set of derivatives $\mathscr{D}$ as the set of functions $d\in L^2(P)$ such that one can find a sequence $(f_n)\in\mathscr{F}$ satisfying $\|f_n(X)-f_0(X)\|_2\xrightarrow[n\to\infty]{}0$ and $\|d-d_{f_n}\|_2\xrightarrow[n\to\infty]{}0$. With such (f_n), define, for all $t\in[0,1]$, $f_t=f_n$, where $n\leq\frac{1}{t}<n+1$. We thus have that, for any $d\in\mathscr{D}$, there exists a parametric path $(f_t)_{0\leq t\leq\alpha}$ with α a strictly positive real number, such that for any $t\in[0,\alpha]$, $f_t\in\mathscr{F}$, $t\mapsto\|f_t(X)-f_0(X)\|_2$ is continuous, tends to 0 as t tends to 0 and $\|d-d_{f_t}\|_2\to 0$ as t tends to 0. Using the reparameterization

$$\|f_u(X)-f_0(X)\|_2=u, \tag{7}$$

for any $d\in\mathscr{D}$, there exists a parametric path $(f_u)_{0\leq u\leq\alpha}$ such that:

$$\int\left(f_u-f_0-ud\right)^2dP=o(u^2). \tag{8}$$

Now, let us introduce some assumptions:

B-1 Let u be defined as (7), the map $u\mapsto P(Y-f_u(X))^2$ admits a second-order Taylor expansion with strictly positive second derivative $\frac{\partial^2P(Y-f_u(X))^2}{\partial u^2}$ at $u=0$.

B-2 The set of generalized derivative functions $\mathscr{S} = \left\{ d_f, f \in \mathscr{F}, f \neq f_0 \right\}$ is a Donsker class.

We have then the main theorem of this paper, whose proof is postponed to section 5:

Theorem 1 *Under (***B-1***) and (***B-2***)*

$$\sup_{f \in \mathscr{F}} \left(\sum_{t=1}^{n} (Y_t - f_0(X_t))^2 - (Y_t - f(X_t))^2 \right) = \sup_{d \in \mathscr{D}} \left(\max \left\{ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varepsilon_t d(X_t); 0 \right\} \right)^2 + o_P(1).$$

Even when the set of possible regression functions $\mathscr{F}$ is not parametric, this theorem proves the tightness of the SSE, if the set $\mathscr{S}$ is a Donsker class. Hence, the rate of convergence of the LSE $\hat{f}$ toward f_0 will be of order $\frac{1}{\sqrt{n}}$ which is the rate of parametric models.

Now, define $(W(d))_{d \in \mathscr{D}}$ the centered Gaussian process with covariance the scalar product in $L^2(P)$, an immediate application of Theorem 1 gives:

Corollary 1 *Under (***B-1***) and (***B-2***),*

$$\sup_{f \in \mathscr{F}} \left(\sum_{t=1}^{n} (Y_t - f_0(X_t))^2 - \sum_{t=1}^{n} (Y_t - f(X_t))^2 \right)$$

converges in distribution to

$$\sigma^2 \sup_{d \in \mathscr{D}} \left(\max \{ W(d); 0 \} \right)^2.$$

As we see, the Donsker property of the set of generalized derivative functions $\mathscr{S}$ is fundamental for the results above. van der Vaart (1998) gives several examples of Donsker class of functions, this property depends on the "size" of the class. A relatively simple way to measure the size of a class $\mathscr{S}$ is in terms of entropy. In the next section we will show how to measure it for parametric models under loss of identifiability.

3 Donsker property for $\mathscr{S}$

This section will give a framework for the demonstration of Donsker property for the set of generalized derivative functions $\mathscr{S}$ for parametric models with compact possible set of parameters and under loss of identifiability. Note that this framework could be easily adapted to likelihood ratio test and generalized score functions of Liu and Shao (2003).

First, we recall the notion of bracketing entropy. Consider the set $\mathscr{S}$ endowed with the norm $\|\cdot\|_2$. For every $\eta > 0$, we define an η-bracket by
$[l, u] = \{ f \in \mathscr{S}, l \leq f \leq u \}$ such that $\|u - l\|_2 < \eta$. The η-bracketing entropy is

$$\mathscr{H}_{[\cdot]}(\eta, \mathscr{S}, \|\cdot\|_2) = \ln \left(\mathscr{N}_{[\cdot]}(\eta, \mathscr{S}, \|\cdot\|_2) \right),$$

where $\mathscr{N}_{[\cdot]}(\eta,\mathscr{S},\|\cdot\|_2)$ is the minimum number of η-brackets necessary to cover $\mathscr{S}$. With the previous notations if

$$\int_0^1 \sqrt{\mathscr{H}_{[\cdot]}(\eta,\mathscr{S},\|\cdot\|_2)}d\eta < \infty,$$

then, according to the Theorem 19.5 of van der Vaart (1998), the set $\mathscr{S}$ is Donsker. Moreover, if the number of η-brackets necessary to cover $\mathscr{S}$, $\mathscr{N}_{[\cdot]}(\eta,\mathscr{S},\|\cdot\|_2)$, is a polynomial function of $\frac{1}{\eta}$, then $\mathscr{S}$ will be Donsker, so we will prove this sufficient condition.

In general, if a class of function

$$\mathscr{F} = \left\{ f_\theta, \theta \in \Theta \subset \mathbb{R}^D,\ \Theta compact \right\}$$

is parametric and regular, a function $G \in L^2(P)$ exists such that

$$\left| f_{\theta_1}(x) - f_{\theta_2}(x) \right| \leq |\theta_1 - \theta_2|\, G(x)$$

and according to van der Vaart (1998), if $0 < \eta < \mathrm{diam}\Theta$, a constant K exists such that

$$\mathscr{N}_{[\cdot]}(\eta,\mathscr{F},\|\cdot\|_2) \leq K \left(\frac{\mathrm{diam}\Theta}{\eta \|G\|_2} \right)^D.$$

However, even if the set $\mathscr{F}$ is parametric, compact and regular, the set

$$\mathscr{S} = \left\{ d_{f_\theta} = \frac{f_\theta - f_0}{\|f_\theta - f_0\|_2}, \theta \in \Theta, f_\theta \neq f_0 \right\}$$

is not regular, since $\theta \mapsto d_{f_\theta}(x)$ is, in general, not extendable by continuity in parameters realizing the best regression function f_0. Hopefully, we can show that the number of η-brackets necessary to cover $\mathscr{S}$ is a polynomial function of $\frac{1}{\eta}$ by another method, similar to Olteanu and Rynkiewicz (2012).

Let us assume:

C-1 A function $G \in L^2(P)$ exists such that for any f_{θ_1} and f_{θ_2} in $\mathscr{F}$

$$\left| f_{\theta_1}(x) - f_{\theta_2}(x) \right| \leq |\theta_1 - \theta_2|\, G(x). \tag{9}$$

C-2 A reparameterization $\theta \mapsto (\phi,\psi)$ exists such that for positive integers (q_0,q_1) and linearly independent functions $g_{\beta_i^0}$, $g'_{\beta_i^0}$, $g''_{\beta_i^0}$, $i = 1,...,q_0$, $g_{\beta_j}, j = 1,\cdots,q_1$ the difference of regression functions can be written:

$$\begin{aligned} &f_\theta - f_0 = \\ &f_{(\phi,\psi)} - f_0 = (\phi-\phi_0)^T \frac{\partial f_{(\phi_0,\psi)}}{\partial \phi} + \frac{1}{2}(\phi-\phi_0)^T \frac{\partial^2 f_{(\phi_0,\psi)}}{\partial \phi^2}(\phi-\phi_0) + o(\|f_{(\phi,\psi)} - f_0\|_2^2) \\ &= \sum_{i=1}^{q_0} \alpha_i g_{\beta_i^0} + \sum_{i=1}^{q_1} \nu_i g_{\beta_i} + \sum_{i=1}^{q_0} \delta_i^T g'_{\beta_i^0} + \sum_{i=1}^{q_0} \gamma_i^T g''_{\beta_i^0} \gamma_i + o\left(\|f_{(\phi,\psi)} - f_0\|_2^2\right). \end{aligned} \tag{10}$$

We can now state the result:

Proposition 1 *Under (***C-1***) and (***C-2***) a positive integer k exists so that the number of η-brackets $\mathscr{N}_{[\cdot]}(\eta,\mathscr{S},\|\cdot\|_2)$ covering $\mathscr{S}$ is $\mathscr{O}\left(\frac{1}{\eta}\right)^k$.*

Proof For proving that $\mathscr{N}_{[\cdot]}(\eta, \mathscr{S}, \|\cdot\|_2)$ is a polynomial function of $\frac{1}{\eta}$, we have to split $\mathscr{S}$ into two sets of functions: A set in a neighborhood of the true regression function f_0 and a second one at a distance at least η of f_0. For a sufficiently small $\eta > 0$, we consider $\mathscr{F}_\eta \subset \mathscr{F}$, a $\mathscr{L}^2$-neighborhood of f_0:
$\mathscr{F}_\eta = \{f_\theta \in \mathscr{F}, \|f_\theta - f_0\|_2 \leq \eta, f_\theta \neq f_0\}$. $\mathscr{S}$ is split into $\mathscr{S}_\eta = \left\{d_{f_\theta}, f_\theta \in \mathscr{F}_\eta\right\}$ and $\mathscr{S} \setminus \mathscr{S}_\eta$.

On $\mathscr{S} \setminus \mathscr{S}_\eta$, it can be easily seen that

$$\left\|d_{f_{\theta_1}} - d_{f_{\theta_2}}\right\|_2 \leq \frac{\left\|f_{\theta_1} - f_{\theta_2}\right\|_2}{\left\|f_{\theta_1} - f_0\right\|_2} + \left\| \frac{f_{\theta_2} - f_0}{\left\|f_{\theta_1} - f_0\right\|_2} - \frac{f_{\theta_2} - f_0}{\left\|f_{\theta_2} - f_0\right\|_2} \right\|_2$$

for every $f_{\theta_1}, f_{\theta_2} \in \mathscr{F} \setminus \mathscr{F}_\eta$. By (9), if $|\theta_1 - \theta_2| \leq \eta^3$, a positive constant C exists such that

$$\left\|f_{\theta_1} - f_{\theta_2}\right\|_2 \leq C\eta^3.$$

Then, by the definition of $\mathscr{S}_\eta$,

$$\begin{aligned}
&\left\| \frac{f_{\theta_2} - f_0}{\left\|f_{\theta_1} - f_0\right\|_2} - \frac{f_{\theta_2} - f_0}{\left\|f_{\theta_2} - f_0\right\|_2} \right\|_2 \\
&\leq \left\| \frac{f_{\theta_2} - f_0}{\left\|f_{\theta_2} - f_0\right\|_2 + C\eta^3} - \frac{f_{\theta_2} - f_0}{\left\|f_{\theta_2} - f_0\right\|_2} \right\|_2 \\
&\leq \frac{\left\|f_{\theta_2} - f_0\right\|_2}{\eta} \left(1 - \frac{1}{1 + C\eta^2}\right) = \left\|f_{\theta_2} - f_0\right\|_2 (C\eta + o(\eta))
\end{aligned}$$

and, if the set $\mathscr{F}$ is compact, a positive constant M exists so that

$$\left\|d_{f_{\theta_1}} - d_{f_{\theta_2}}\right\|_2 \leq C\eta^2 + \left\|f_{\theta_2} - f_0\right\|_2 (C\eta + o(\eta)) \leq M\eta.$$

Finally, we get:

$$\mathscr{N}_{[\cdot]}(\eta, \mathscr{S} \setminus \mathscr{S}_\eta, \|\cdot\|_2) = \mathscr{O}\left(\frac{1}{\eta^3}\right)^D = \mathscr{O}\left(\frac{1}{\eta}\right)^{3D}$$

where D is the dimension of parameter vectors of the model.

It remains to prove that the bracketing number is a polynomial function of $(\frac{1}{\eta})$ for $\mathscr{S}_\eta$. The idea is to reparameterize the model in a convenient manner which will allow a Taylor expansion around the identifiable part of the true value of the parameters, then, using this Taylor expansion, we can show that the bracketing number of $\mathscr{S}_\eta$ is a polynomial function of $\frac{1}{\eta}$. Indeed, according to the assumption **C-2** we have the approximation (10). Now, using the linear independence of functions g_{β_i}, $g_{\beta_i^0}$, $g'_{\beta_i^0}$, $g''_{\beta_i^0}$, for every vector $v = (\alpha_i, \delta_i, \gamma_i, i = 1, \cdots, q_0, \nu_j, j = 1, \cdots, q_1)$ of norm 1,

$$\left(v, (\beta_i)_{1 \leq i \leq q_1}\right) \mapsto \left\| \sum_{i=1}^{q_0} \alpha_i g_{\beta_i^0} + \sum_{i=1}^{q_1} \nu_i g_{\beta_i} + \sum_{i=1}^{q_0} \delta_i^T g'_{\beta_i} + \sum_{i=1}^{q_0} \gamma_i^T g''_{\beta_i} \gamma_i \right\|_2 > 0.$$

Using the compacity of sets

$$\mathscr{V} = \left\{v = (\alpha_i, \delta_i, \gamma_i, i = 1, \cdots, q_0, \nu_j, j = 1, \cdots, q_1), |v| = 1\right\}$$
and
$$\left\{(\beta_i)_{1 \leq i \leq q_1}\right\},$$

$m > 0$ exists so that for all $(\beta_i)_{1\leq i\leq q_1}$ and $v \in \mathscr{V}$,

$$\left\| \sum_{i=1}^{q_0} \alpha_i g_{\beta_i^0} + \sum_{i=1}^{q_1} v_i g_{\beta_i} + \sum_{i=1}^{q_0} \delta_i^T g'_{\beta_i} + \sum_{i=1}^{q_0} \gamma_i^T g''_{\beta_i} \gamma_i \right\|_2 \geq m.$$

At the same time, since

$$\left\| \frac{f_{(\phi_t,\psi_t)} - f_0}{\left\| f_{(\phi_t,\psi_t)} - f_0 \right\|_2} \right\|_2 = 1,$$

the Euclidean norm of coefficients $(\alpha_i, \delta_i, \gamma_i, i = 1, \cdots, q_0, v_i, i = 1, \cdots, q_1)$ in the development of $\frac{f_{(\phi_t,\psi_t)} - f_0}{\left\| f_{(\phi_t,\psi_t)} - f_0 \right\|_2}$ is upper bounded by $\frac{1}{m} + 1$. This fact implies that $\mathscr{S}_\eta$ can be included in

$$\mathscr{H} = \begin{array}{l} \left\{ \sum_{i=1}^{q_0} \left(\alpha_i g_{\beta_i^0} + \delta_i^T g'_{\beta_i^0} + \gamma_i^T g''_{\beta_i^0} \gamma_i \right) + \sum_{i=1}^{q_1} v_i g_{\beta_i} + C, \right. \\ \left. |(\alpha_i, \delta_i, \gamma_i, i = 1, \cdots, q_0, v_i, i = 1, \cdots, q_1)| \leq \frac{1}{m} + 1, |C| \leq \frac{1}{m} + 1 \right\} \end{array}$$

and a positive integer d exists so that $\mathscr{N}_{[\cdot]}(\eta, \mathscr{H}, \|\cdot\|_2) = \mathscr{O}\left(\frac{1}{\eta}\right)^d$. Finally, the positive integer k of the proposition will be equal to $\max(3D, d)$. ■

4 Application to regression with neural networks

Feedforward neural networks or multilayer perceptrons (MLP) are well known and popular tools to deal with non-linear regression models. White (1992) reviews the statistical properties of MLP estimation in detail, however, he eludes a significant point: The asymptotic behavior of the estimator when the MLP in use has redundant hidden units. When the noise of the regression model is assumed Gaussian, Amari et al (2006) provide several examples of the behavior of the likelihood ratio test statistic (LRTS) in such cases. Fukumizu (2003) shows that, for unbounded parameters, the LRTS can have an order lower bounded by $O(\log(n))$ with n the number of observations instead of the classical convergence property to a χ^2 law. Hagiwara and Fukumizu (2008) investigate relation between LRTS divergence and weight size in a simple neural networks regression problem with Gaussian noise. They show that the degree of over-fitting is strongly related to the size of the inputs weights, which is the reason why regularization techniques like "weight decay" (see Ripley (1996)), penalizing the model by the size of the parameters, work so well. In practice, the set of possible parameters of the MLP regression model is bounded and the behavior of LRTS and more generally the SSE is still unknown. In this section, we derive the distribution of the SSE if the parameters are in a compact (bounded and closed) set.

4.1 The model

Let $x = (x_1, \cdots, x_d)^T \in \mathbb{R}^d$ be the vector of inputs and
$w_i := (w_{i1}, \cdots, w_{id})^T \in \mathbb{R}^d$ be the parameter vector of the hidden unit i. The MLP function with k hidden units can be written :

$$f_\theta(x) = \beta + \sum_{i=1}^{k} a_i \phi\left(w_i^T x + b_i\right),$$

with $\theta = (\beta, a_1, \cdots, a_k, b_1, \cdots, b_k, w_1, \cdots, w_k)$ the parameter vector of the model. The transfer function ϕ will be assumed bounded and two times differentiable. We assume also that the first and second derivatives of the transfer function ϕ: $\phi^{'}$ and $\phi^{''}$ are bounded like for sigmoid functions, the most used transfer functions. Moreover, in order to avoid a symmetry on the signs of the parameters, we assume that, for $1 \leq i \leq k$, $a_i \geq 0$. Let $\Theta \subset \mathbb{R} \times \mathbb{R}^{+k} \times \mathbb{R}^{k\times(d+1)}$ be the compact set of possible parameters, the regression model (1) is then

$$Y = f_{\theta_0}(X) + \varepsilon,$$

with X a random vector and

$$\theta_0 = \left(\beta^0, a_1^0, \cdots, a_k^0, b_1^0, \cdots, b_k^0, w_1^0, \cdots, w_k^0\right)$$

a parameter vector such that $f_{\theta_0} = f_0$. Note that the set of parameters Θ_0 realizing the true regression function f_0 may belong to a non-null dimension sub-manifold if the number of hidden units is overestimated. Suppose, for example, that we have a multilayer perceptron with two hidden units and the true function f_0 is given by a perceptron with only one hidden unit, say $f_0 = a^0 \tanh(w^0 x)$, with $x \in \mathbb{R}$. Then, any parameter vector θ in the set:

$$\left\{\theta \,\middle|\, w_2 = w_1 = w^0, b_2 = b_1 = 0, a_1 + a_2 = a^0\right\}$$

realizes the function f_0. Hence, classical statistical theory for studying the LSE can not be applied because it requires the identification of the parameters (up to some permutations and sign symmetries) so that the Hessian matrix of the SSE with respect to the parameters will be definite positive in a neighborhood of the parameter vector realizing the true regression function. Let us denote k_0 the minimal number of hidden units to realize the true regression function f_0. We will compare the SSE of over-determined models against the true model :

$$\sum_{t=1}^{n} (Y_t - f_0(X_t))^2 - \sum_{t=1}^{n} (Y_t - f_\theta(X_t))^2,$$

when the loss of identifiability occurs (i.e. when $k > k_0$).

4.2 Asymptotic distribution of the difference of SSE

Let us give simple sufficient conditions for which the Donsker property holds for the set of generalized derivative functions. For any accumulation sequence of parameter θ_n leading to f_0, the assumption **H-1** allows the regression functions (f_{θ_n}) to be in a $\mathscr{L}^2$-neighborhood of f_0, in the same spirit of locally conic models of Dacunha-Castelle and Gassiat (1999). **H-1** will hold naturally with "weight decay" penalization. Moreover, if the probability distribution Q of the variable X admits a strictly positive density with respect to the Lebesgue measure, it is shown in section 5 that the assumption **H-3** will be true for the sigmoid transfer function.

H-1: Let $\{f_\theta, \theta \in \Theta\}$ be a set of MLP functions with bounded and two times differentiable transfer function ϕ. Assume that the first and second order derivatives (ϕ' and ϕ'') of this transfer function are also bounded. Moreover, assume that Θ is a closed ball of $\mathbb{R} \times \mathbb{R}^{+k} \times \mathbb{R}^{k\times(d+1)}$ for some positive integers (k,d), and its interior contains parameters realizing the true regression function f_0.
H-2: $E_Q(|X|^4) < \infty$.
H-3: Let k be a strictly positive integer, for distinct $(w_i, b_i)_{1\leq i\leq k}$ with $\forall i \in \{1,\cdots,k\}, |w_i| \neq 0$, the functions of the set

$$\begin{pmatrix} 1, \left(x_j x_l \phi''({w_i}^T x + b_i)\right)_{1\leq l\leq j\leq d,\ 1\leq i\leq k}, \left(x_j \phi''({w_i}^T x + b_i)\right)_{1\leq j\leq d,\ 1\leq i\leq k} \\ \phi''({w_i}^T x + b_i)_{1\leq i\leq k}, \left(x_j \phi'({w_i}^T x + b_i)\right)_{1\leq j\leq d,\ 1\leq i\leq k} \\ \left(\phi'({w_i}^T x + b_i)\right)_{1\leq i\leq k}, \left(\phi({w_i}^T x + b_i)\right)_{1\leq i\leq k} \end{pmatrix}$$

are linearly independent in the Hilbert space $\mathscr{L}^2(Q)$.

Then, We get the following result which is proven in section 5:

Theorem 2 *Let the map $\Omega : \mathscr{L}^2(Q) \to \mathscr{L}^2(Q)$ be defined as $\Omega(f) = \frac{f}{\|f\|_2}$. Under the assumptions* **H-1**, **H-2** *and* **H-3**, *a centered Gaussian process $\{W(d), d \in \mathscr{D}\}$ with continuous sample paths and a covariance kernel $P(W(d_1)W(d_2)) = P(d_1 d_2)$ exists so that*

$$\lim_{n\to\infty} \sum_{t=1}^{n} (Y_t - f_0(X_t))^2 - \sum_{t=1}^{n} (Y_t - f_\theta(X_t))^2 = \sigma^2 \sup_{d\in\mathscr{D}} \left(\max\{W(d);0\}\right)^2 .$$

The index set $\mathscr{D}$ is defined as $\mathscr{D} = \cup_t \mathscr{D}_t$, the union runs over any possible vector of integers $t = \left(t_1, \cdots, t_{k^0+1}\right) \in \mathbb{N}^{k^0+1}$ with $0 \leq t_1 \leq k - k^0 < t_2 < \cdots < t_{k^0+1} \leq k$ and

$$\begin{array}{l}
\mathscr{D}_t = \Big\{ \Omega \left(\gamma + \sum_{i=0}^{k^0} \varepsilon_i \phi(w_i^{0T} X + b_i^0) \right. \\
+ \sum_{i=0}^{k^0} \phi'(w_i^{0T} X + b_i^0)(\zeta_i^T X + \alpha_i) \\
+ \delta(i) \sum_{i=1}^{k^0} \phi''(w_i^{0T} X + b_i^0) \times \\
\left(\left(\sum_{j=t_i+1}^{t_{i+1}} \nu_j{}^T X X^T \nu_j + \eta_j \nu_j{}^T X + \eta_j{}^2 \right) \right) \\
\left. + \sum_{i=t_{k^0+1}+1}^{k} \mu_i \phi(w_i{}^T X + b_i) \right), \\
\gamma, \varepsilon_1, \cdots, \varepsilon_{k^0}, \alpha_1, \cdots, \alpha_{k^0}, \eta_{t_1}, \cdots, \eta_{t_{k^0+1}} \in \mathbb{R}, \\
\mu_{t_{k^0+1}+1}, \cdots, \mu_k \in \mathbb{R}^+; \zeta_1, \cdots, \zeta_{k^0}, \nu_{t_1}, \cdots, \nu_{t_{k^0+1}} \in \mathbb{R}^d, \\
(w_{k^0+1+1}, b_{k^0+1+1}), \cdots, (w_k, b_k) \in \Theta \backslash \left\{ (w_1^0, b_1^0), \cdots, (w_{k^0}^0, b_{k^0}^0) \right\} \Big\}.
\end{array}$$

$\delta(i) = 1$ if a vector $\mathbf{q}$ exists so that:
$q_j \geq 0$, $\sum_{j=t_i+1}^{t_{i+1}} q_j = 1$, $\sum_{j=t_i+1}^{t_{i+1}} \sqrt{q_j} \nu_j = 0$ and $\sum_{j=t_i+1}^{t_{i+1}} \sqrt{q_j} \eta_j = 0$, otherwise $\delta(i) = 0$.

This theorem shows that the degree of over-fitting is bounded in probability, but depends on the size of the asymptotic set $\mathscr{D}$. In order to reduce the over-fitting we then need to control the size of the limit functions in $\mathscr{D}$, this can be done by two different ways:

1. Reduce the number k of hidden units thanks to an information criterion like the BIC (see Schwarz (1978)).
2. Reduce the size of the inputs weights $(w_i, b_i)_{1 \leq i \leq k}$. Indeed, it is empirically known that the output of a large estimated network tends to have high curvature. This is caused by the over-fitting that occurs in the over-realizable case by the extreme values of the input weights (see Hagiwara and Fukumizu (2008)). However, note that the size of the weights has to be large enough so that Θ contains $(w_i^0, b_i^0)_{1 \leq i \leq k^0}$. So, we need to find a trade-off for this penalization.

In summary, Theorem 2 provides the following guidelines for the regularization of such models: Use both an information criterion to limit the number of hidden units and a penalization term proportional to the size of the inputs weights only. Fine tuning of these penalizations may be assessed thanks to cross-validation procedures (see Arlot and Celisse (2010)).

5 Proofs

5.1 Proofs of section 2

We prove here the Theorem 1. We have

$$\begin{array}{l}
\sum_{t=1}^n (Y_t - f_0(X_t))^2 - \sum_{t=1}^n (Y_t - f(X_t))^2 = \\
\sum_{t=1}^n (Y_t - f_0(X_t))^2 - (Y_t - f(X_t))^2 = 2\|f(X) - f_0(X)\|_2 \sum_{t=1}^n \varepsilon_t df(X_t) \\
- \|f(X) - f_0(X)\|_2^2 \sum_{t=1}^n df^2(X_t).
\end{array}$$

As soon as $\sum_{t=1}^n (Y_t - f_0(X_t))^2 - \sum_{t=1}^n (Y_t - f(X_t))^2 \geq 0$,

$$2\|f(X) - f_0(X)\|_2 \sum_{t=1}^n \varepsilon_t df(X_t) \geq \|f(X) - f_0(X)\|_2^2 \sum_{t=1}^n df^2(X_t)$$

so

$$\sup_{f\in\mathscr{F}, \sum_{t=1}^n (Y_t - f_0(X_t))^2 - \sum_{t=1}^n (Y_t - f(X_t))^2 \geq 0} \|f(X) - f_0(X)\|_2 \leq 2 \sup_{f\in\mathscr{F}} \max\left\{ \frac{\sum_{t=1}^n \varepsilon_t df(X_t)}{\sum_{t=1}^n df^2(X_t)}; 0 \right\}. \tag{11}$$

Since, $\mathscr{S}$ is Donsker

$$\sup_{f\in\mathscr{F}} \frac{1}{n} \left(\sum_{t=1}^n \varepsilon_t d_f(X_t) \right)^2 = O_P(1) \tag{12}$$

and S admits an envelope function F such that $P(F^2) < \infty$, $S^2 = \left\{ d_f^2, f \in \mathscr{F}, f \neq f_0 \right\}$ is Glivenko-Cantelli and

$$\sup_{f\in\mathscr{F}} \left| \frac{1}{n} \sum_{t=1}^n d_f^2(X_t) - 1 \right| = o_P(1). \tag{13}$$

Then, one may apply inequality (11) to obtain

$$\sup_{f\in\mathscr{F}, \sum_{t=1}^n (Y_t - f_0(X_t))^2 - \sum_{t=1}^n (Y_t - f(X_t))^2 \geq 0} \|f(X) - f_0(X)\|_2 = \frac{1}{\sqrt{n}} O_P(1). \tag{14}$$

By lemma 1,

$$\sup_{f\in\mathscr{F}} \sum_{t=1}^n (Y_t - f_0(X_t))^2 - \sum_{t=1}^n (Y_t - f(X_t))^2 \leq \sup_{f\in\mathscr{F}} \frac{\left(\max\left\{ \frac{\sum_{t=1}^n \varepsilon_t \frac{f_0(X_t) - f(X_t)}{\|f_0(X_t) - f(X_t)\|_2}}{\sqrt{n}}; 0 \right\} \right)^2}{\frac{\sum_{t=1}^n \left(\frac{f_0(X_t) - f(X_t)}{\|f_0(X_t) - f(X_t)\|_2} \right)^2}{n}}.$$

Using (13), we obtain that

$$\begin{aligned} &\sup_{f\in\mathscr{F}} \sum_{t=1}^n (Y_t - f_0(X_t))^2 - \sum_{t=1}^n (Y_t - f(X_t))^2 \\ &\leq \sup_{f\in\mathscr{F}} \left(\max\left\{ \frac{\sum_{t=1}^n \varepsilon_t \frac{f_0(X_t) - f(X_t)}{\|f_0(X_t) - f(X_t)\|_2}}{\sqrt{n}}; 0 \right\} \right)^2 + o_P(1). \end{aligned}$$

Let $\mathscr{F}_n = \left\{ f \in \mathscr{F} : \|f(X) - f_0(X)\|_2 \leq n^{-1/4} \right\}$. Using (14), we obtain that

$$\begin{aligned} &\sup_{f\in\mathscr{F}} \sum_{t=1}^n (Y_t - f_0(X_t))^2 - \sum_{t=1}^n (Y_t - f(X_t))^2 \\ &\leq \sup_{f\in\mathscr{F}_n} \left(\max\left\{ \frac{\sum_{t=1}^n \varepsilon_t \frac{f_0(X_t) - f(X_t)}{\|f_0(X_t) - f(X_t)\|_2}}{\sqrt{n}}; 0 \right\} \right)^2 + o_P(1). \end{aligned}$$

Let us write $\|d_f - \mathscr{D}\|_2 = \inf_{d\in\mathscr{D}} \|d_f - d\|_2$, we have $\sup_{f\in\mathscr{F}_n} \|d_f - \mathscr{D}\|_2 \xrightarrow[n\to\infty]{} 0$, thus, for a sequence u_n decreasing to 0, and with

$$\Delta_n = \left\{ d_f - d \ : \ f \in \mathscr{F}_n, \ d \in \mathscr{D}, \|d_f - d\|_2 \leq u_n \right\},$$

we obtain that

$$\begin{aligned}&\sup_{f\in\mathscr{F}}\sum_{t=1}^{n}\left(Y_t-f_0(X_t)\right)^2-\sum_{t=1}^{n}\left(Y_t-f(X_t)\right)^2\\&\leq\sup_{d\in\mathscr{D}}\left(\max\left\{\frac{\sum_{t=1}^{n}\varepsilon_t d(X_t)}{\sqrt{n}}+\sup_{\delta\in\Delta_n}\frac{\sum_{t=1}^{n}\varepsilon_t\delta(X_t)}{\sqrt{n}};0\right\}\right)^2+o_P(1).\end{aligned}$$

But, using the Donsker property, the definition of Δ_n and the property of asymptotic stochastic equicontinuity of empirical processes indexed by a Donsker class, we get:

$$\sup_{\delta\in\Delta_n}\frac{\sum_{t=1}^{n}\varepsilon_t\delta\left(X_t\right)}{\sqrt{n}}=o_P(1),$$

and

$$\begin{aligned}&\sup_{f\in\mathscr{F}}\sum_{t=1}^{n}\left(Y_t-f_0(X_t)\right)^2-\sum_{t=1}^{n}\left(Y_t-f(X_t)\right)^2\\&\leq\sup_{d\in\mathscr{D}}\left(\max\left\{\frac{\sum_{t=1}^{n}\varepsilon_t d(X_t)}{\sqrt{n}};0\right\}\right)^2+o_P(1).\end{aligned}\tag{15}$$

Since S admits a square integrable envelope function, a function m exists such that for u_1 and u_2 belonging to a parametric path converging to a limit function d:

$$\left|(y-f_{u_1}(x))^2-(y-f_{u_2}(x))^2\right|\leq m(x,y)|u_1-u_2|.$$

Moreover, along a path, the map

$$u\mapsto P(Y-f_u(X))^2$$

admits a second-order Taylor expansion with strictly positive second derivative $\frac{\partial^2 P(Y-f_u(X))^2}{\partial u^2}$ at $u=0$, and we can use classical normal asymptotic theorem for M-estimators (see Theorem 5.23 of van der Vaart (1998)) along this parametric paths, to obtain a sequence of finite subsets $\mathscr{D}_k$ increasing to $\mathscr{D}$ such that

$$\begin{aligned}&\sup_{f\in\mathscr{F}}\sum_{t=1}^{n}\left(Y_t-f_0(X_t)\right)^2-\sum_{t=1}^{n}\left(Y_t-f(X_t)\right)^2\\&\geq\sup_{d\in\mathscr{D}_k}\left(\frac{\sum_{t=1}^{n}\varepsilon_t d(X_t)}{\sqrt{n}}\right)^2+o_P(1)\end{aligned}$$

for any k. Therefore, equality holds in (15). ■

5.2 Proofs of section 4

We prove the assumption (**H-3**) for the for the sigmoid transfer function

$$\phi(t)=\frac{1}{1+e^{-}t}.$$

Note that MLP with sigmoid transfer functions or hyperbolic tangent transfer functions are equivalent, because a one-to-one correspondence between the two kinds of MLP exists as $\frac{1}{1+e^{-t}}=(1+\tanh(t/2))/2$. The proof is an extension of the results of Fukumizu (1996). We define the complex sigmoid function on $\mathbb{C}$ by $\phi(z)=\frac{1}{1+e^{-z}}$. The singularities of ϕ are

$$\left\{z\in\mathbb{C}\,\middle|\,z=(2n+1)\pi\sqrt{-1},n\in\mathbb{Z}\right\},$$

all of which are poles of order 1. Next, we review fundamental propositions in complex analysis.

Proposition 2 *Let ϕ be a holomorphic function on a connected open set D in $\mathbb{C}$ and p be a point in D. If a sequence $\{p_n\}_{n=1}^{\infty}$ exists in D so that $p_n \neq p, \lim_{n\to n} p_n = p$ and $\phi(p_n) = 0$ for all $n \in \mathbb{N}$ then $\phi(z) = 0$ for all $z \in D$.*

Proposition 3 *Let ϕ be a holomorphic function on a connected open set D in $\mathbb{C}$, and p be a point in D. Then the following equivalence relations hold:*

- *p is a removable singularity*

$$\Leftrightarrow \lim_{z\to p} f(z) \in \mathbb{C}.$$

- *p is a pole*

$$\Leftrightarrow \lim_{z\to p} |f(z)| = \infty.$$

- *p is an essential singularity*

$$\Leftrightarrow \lim_{z\to p} |f(z)| \text{ does not exist.}$$

Let $(\beta, a_1, \cdots, a_k, b_1, \cdots, b_k, w_1, \cdots, w_k)$ be a parameter vector such that $i \neq j \Rightarrow (b_i, w_i) \neq (b_j, w_j)$ and for all i , $|w_i| \neq 0$. By the lemma 3 of Fukumizu (1996), a basis of $\mathbb{R}^d$ $\left(x^{(1)}, \cdots, x^{(d)}\right)$ exists so that

1. For all $i \in \{1, \cdots, k\}$ and all $h \in \{1, \cdots, d\}$

$$w_i^T x^{(h)} \neq 0.$$

2. For all $i_1, i_2 \in \{1, \cdots, k\}$, $i_1 \neq i_2$ and all $h \in \{1, \cdots, d\}$

$$b_{i_1} + w_{i_1}^T x^{(h)} \neq \pm\left(b_{i_2} + w_{i_2}^T x^{(h)}\right).$$

For l, $1 \leq l \leq d$ and $i \in \{1, \cdots, k\}$ let us write $m_i^{(l)} := w_i^T x^{(l)}$. We set

$$S_i^{(l)} = \left\{ u \in \mathbb{C} \,\middle|\, u = \frac{(2n+1)\pi\sqrt{-1} - b_i}{m_i^{(l)}}, n \in \mathbb{Z} \right\}.$$

Clearly the points in $S_i^{(l)}$ are the singularities of $\phi\left(m_i^{(l)} u + b_i\right)$. Note that these points are poles of order 1 for

$$\phi(m_i^{(l)} u + b_i) = \frac{1}{1 + e^{-\left(m_i^{(l)} u + b_i\right)}},$$

of order 2 for

$$\phi'(m_i^{(l)} u + b_i) = -\frac{e^{-\left(m_i^{(l)} u + b_i\right)}}{\left(1 + e^{-\left(m_i^{(l)} u + b_i\right)}\right)^2}$$

and 3 for

$$\phi''(m_i^{(l)}u+b_i)=\frac{e^{-\left(m_i^{(l)}u+b_i\right)}}{\left(1+e^{-\left(m_i^{(l)}u+b_i\right)}\right)^2}+2\frac{e^{-2\left(m_i^{(l)}u+b_i\right)}}{\left(1+e^{-\left(m_i^{(l)}u+b_i\right)}\right)^3}.$$

Let be $D(l):=\mathbb{C}-\cup_{1\leq i\leq k}S_i^{(l)}$. Holomorphic functions on $D^{(l)}$ are defined as follows:

$$\begin{array}{l}\Psi^{(l)}(u):=\alpha_0+\sum_{i=1}^k\alpha_i\phi(m_i^{(l)}u+b_i)\\+\sum_{i=1}^k\sum_{j=1}^d\beta_{ij}\phi'(m_i^{(l)}u+b_i)x_j^{(l)}u+\sum_{i=1}^k\varepsilon_i\phi'(m_i^{(l)}u+b_i)\\+\sum_{i=1}^k\sum_{j,r=1,\ j\leq r}^d\gamma_{ijr}\phi''(m_i^{(l)}u+b_i)x_j^{(l)}x_r^{(l)}u^2\\+\sum_{i=1}^k\sum_{j=1}^d\eta_{ij}\phi''(m_i^{(l)}u+b_i)x_j^{(l)}u+\sum_{i=1}^k\rho_i\phi''(m_i^{(l)}u+b_i)\end{array}$$

The functions in the set

$$\begin{array}{l}\left(1,\left(x_jx_l\phi''({w_i}^Tx+b_i)\right)_{1\leq l\leq j\leq d,\ 1\leq i\leq k},\left(x_j\phi''({w_i}^Tx+b_i)\right)_{1\leq j\leq d,\ 1\leq i\leq k}\right.\\\phi''({w_i}^Tx+b_i)_{1\leq i\leq k},\left(x_j\phi'({w_i}^Tx+b_i)\right)_{1\leq j\leq d,\ 1\leq i\leq k}\\\left.\left(\phi'({w_i}^Tx+b_i)\right)_{1\leq i\leq k},\left(\phi({w_i}^Tx+b_i)\right)_{1\leq i\leq k}\right)\end{array}$$

are linearly independent if the following property is verified :

$$\forall u\in D(l),\Psi^{(l)}(u)=0\Leftrightarrow\text{all }\alpha_i,\varepsilon_i,\beta_{ij},\rho_i,\eta_{ij}\text{ and }\gamma_{ijr}\text{ are equal to 0.}$$

Let us assume that $\forall u\in D^{(l)},\Psi^{(l)}(u)=0$, then, by proposition 3, all the points in $S_i^{(l)}$ are removable singularities. Let us write

$$p_i^{(l)}:=\frac{\pi\sqrt{-1}-b_i}{m_i^{(l)}}\in S_i^{(l)}.$$

Clearly , for $1\leq i\leq k-1$, $p_k^{(l)}\notin S_i^{(l)}$, because for all $i_1,i_2\in\{1,\cdots,k\}$, $i_1\neq i_2$ and all $h\in\{1,\cdots,d\}$

$$b_{i_1}+w_{i_1}^Tx^{(h)}\neq\pm\left(b_{i_2}+w_{i_2}^Tx^{(h)}\right).$$

So, $\Psi^{(l)}(u)$ can be written as:

$$\begin{array}{l}\Psi^{(l)}(u)=\alpha_k\phi(m_i^{(l)}u+b_i)+\left(\sum_{i=1}^d\beta_{ki}x_i^{(l)}u+\varepsilon_k\right)\phi'(m_k^{(l)}u+b_k)\\+\left(\sum_{i,j=1,\ i\leq j}^d\gamma_{kij}x_i^{(l)}x_j^{(l)}u^2+\sum_{i=1}^d\eta_{ki}x_i^{(l)}u+\rho_k\right)\phi''(m_k^{(l)}u+b_k)\\+\Psi_{k-1}^{(l)}(u),\end{array}$$

where

$$\begin{array}{l}\Psi_{k-1}^{(l)}(u):=\alpha_0+\sum_{i=1}^{k-1}\alpha_i\phi(m_i^{(l)}u+b_i)\\+\sum_{i=1}^{k-1}\sum_{j=1}^d\beta_{ij}\phi'(m_i^{(l)}u+b_i)x_j^{(l)}u+\sum_{i=1}^{k-1}\varepsilon_i\phi'(m_i^{(l)}u+b_i)\\+\sum_{i=1}^{k-1}\sum_{j,r=1,\ j\leq r}^d\gamma_{ijr}\phi''(m_i^{(l)}u+b_i)x_j^{(l)}x_r^{(l)}u^2\\+\sum_{i=1}^{k-1}\sum_{j=1}^d\eta_{ij}\phi''(m_i^{(l)}u+b_i)x_j^{(l)}u+\sum_{i=1}^{k-1}\rho_i\phi''(m_i^{(l)}u+b_i).\end{array}$$

The point $p_k^{(l)}$ is a regular point of $\Psi_{k-1}^{(l)}(u)$ while $\phi(m_k^{(l)}u+w_{k0})$ has a pole of order 1 at $p_k^{(l)}$, $\phi^{'}(m_k^{(l)}u+w_{k0})$ has a pole of order 2 at $p_k^{(l)}$ and $\phi^{''}(m_k^{(l)}u+w_{k0})$ has a pole of order 3 at $p_k^{(l)}$. Since $p_k^{(l)}$ is a removable singularity of $\Psi^{(l)}(u)$, we have:

$$\begin{array}{l}\alpha_k=0,\ \varepsilon_k=0,\ \sum_{i=1}^d \beta_{ki}x_i^{(l)}=0 \text{ and}\\ \rho_k=0, \sum_{i=1}^d \eta_{ki}x_i^{(l)}=0, \sum_{i,j=1,\ i\leq j}^d \gamma_{kij}x_i^{(l)}x_j^{(l)}=0.\end{array}$$

As a result $\Psi^{(l)}(u)=\Psi_{k-1}^{(l)}(u)$. Applying the same argument successively to $p_{k-1}^{(l)},\cdots,p_1^{(l)}$, we obtain, for all $1\leq i\leq k$, $1\leq j\leq r\leq d$:

$$\begin{array}{c}\alpha_i=0,\\ \varepsilon_i=0,\\ \sum_{j=1}^d \beta_{ij}x_j^{(l)}=0,\\ \rho_i=0,\\ \sum_{j=1}^d \eta_{ij}x_j^{(l)}=0,\\ \sum_{j,r=1,\ j\leq r}^d \gamma_{ijr}x_j^{(l)}x_r^{(l)}=0,\\ \text{and } \alpha_0=0.\end{array}$$

Since $\left(x^{(1)},\cdots,x^{(d)}\right)$ form a basis of $\mathbb{R}^d$, we have $\beta_{ij}=0$ and $\eta_{ij}=0$ for all $1\leq i\leq k$ and $1\leq j\leq d$.

For γ_{ijr}, we get:

$$\sum_{j,r=1,\ j\leq r}^d \gamma_{ijr}x_j^{(l)}x_r^{(l)}=\sum_{r=1}^d\left(\sum_{j=1}^r \gamma_{ijr}x_j^{(l)}\right)x_r^{(l)}=0,$$

and, since $\left(x^{(1)},\cdots,x^{(d)}\right)$ form a basis of $\mathbb{R}^d$, for all $l\in\{1,\cdots,d\}$:

$$\begin{array}{c}\gamma_{i11}x_1^{(l)}=0,\\ \vdots\\ \sum_{j=1}^r \gamma_{ijr}x_j^{(l)}=0,\\ \vdots\\ \sum_{j=1}^d \gamma_{ijd}x_j^{(l)}=0\end{array}$$

and $\gamma_{ijr}=0$ for all $1\leq i\leq k$, $1\leq j\leq r\leq d$. This proves that the assumption **H-3** holds for sigmoid functions ■

Now, in order to prove the theorem 2, we have to check that the assumptions **C-1** and **C-2** of proposition 1 are true under the assumptions **H-1**, **H-2** and **H-3**. Then, we conclude thanks to theorem 1 and the computation of the set $\mathscr{D}$. Since f_θ are MLP functions, it is easy to see that assumption **H-1** implies assumption **C-1**. To prove **C-2**, we will get an asymptotic development of the generalized derivative functions.

Reparameterization. The idea is similar of the reparameterization of finite mixture models in Liu and Shao (2003). Under assumption **H-3**, if k_0 is the minimal number of hidden units to get the true function, the writing of f_0 with a neural network with k_0 hidden units is unique, up to some permutations:

$$f_0 = \beta^0 + \sum_{i=1}^{k_0} a_i^0 \phi\left(w_i^{0T} x + b_i^0\right). \tag{16}$$

Then, for a $\theta \in \Theta$, if $f_\theta = f_0$, a vector of integers $t = (t_i)_{1 \leq i \leq k^0+1}$ exists so that $0 \leq t_1 \leq k - k^0 < t_2 < \cdots < t_{k^0+1} \leq k$ and, up to permutations, we have $w_1 = \cdots = w_{t_1} = 0$ if $t_1 > 0$, $\left(w_{t_i+1} = \cdots = w_{t_{i+1}} = w_i^0\right)_{1 \leq i \leq k^0}$, $\left(b_{t_i+1} = \cdots = b_{t_{i+1}} = b_i^0\right)_{1 \leq i \leq k^0}$, $\left(\sum_{j=t_i+1}^{t_{i+1}} a_j = a_i^0\right)_{1 \leq i \leq k^0}$. Moreover, $\beta + \sum_{i=1}^{t_1} a_i \phi(b_i) = \beta^0$ if $t_1 > 0$ else $\beta = \beta_0$.

For $1 \leq i \leq k^0$, let us define $s_i = \sum_{j=t_i+1}^{t_{i+1}} a_j - a_i^0$ and, if $\sum_{t_i+1}^{t_{i+1}} a_j \neq 0$, let us write $q_j = \frac{a_j}{\sum_{t_i+1}^{t_{i+1}} a_j}$. If $\sum_{t_i+1}^{t_{i+1}} a_j = 0$, q_j will be set at 0. Moreover, let us write $\gamma = \beta + \sum_{i=1}^{t_1} a_i \phi(b_i) - \beta^0$ if $t_1 > 0$ else $\gamma = \beta - \beta_0$.

Then, we get the reparameterization $\theta \mapsto (\Phi_t, \psi_t)$ with

$$\begin{aligned}
\Phi_t &= \left(\gamma, (w_j)_{j=t_1}^{t_{k^0+1}}, (b_j)_{j=t_1}^{t_{k^0+1}}, (s_i)_{i=1}^{k^0}, (a_j)_{t_{k^0+1}+1}^{k}\right), \\
\psi_t &= \left((q_j)_{j=t_1}^{t_{k^0+1}}, (w_i, b_i)_{i=1+t_{k^0+1}}^{k}\right).
\end{aligned}$$

With this parameterization, for a fixed t, Φ_t is an identifiable parameter and all the non-identifiability of the model will be in ψ_t. Namely, f_θ will be equal to:

$$\begin{aligned}
f_\theta = {} & (\gamma + \beta^0) + \textstyle\sum_{i=1}^{k^0} (s_i + a_i^0) \sum_{j=t_{i-1}+1}^{t_i} q_j \phi(w_j^T x + b_j) \\
& + \textstyle\sum_{i=t_{k^0+1}+1}^{k} a_j \phi(w_i^T x + b_i).
\end{aligned}$$

So, for a fixed t, $f_{(\Phi_t^0, \psi_t)} = f_0$ if and only if

$$\begin{aligned}
\Phi_t^0 = {} & \\
(0, \; & \underbrace{w_1^0, \cdots, w_1^0}_{t_2 - t_1} \; , \cdots , \; \underbrace{w_{k^0}^0, \cdots, w_{k^0}^0}_{t_{k^0+1} - t_{k^0}}, \underbrace{b_1^0, \cdots, b_1^0}_{t_2 - t_1} , \cdots , \underbrace{b_{k^0}^0, \cdots, b_{k^0}^0}_{t_{k^0+1} - t_{k^0}}, \\
& \underbrace{0, \cdots, 0}_{k^0} \; \underbrace{0, \cdots, 0}_{k - t_{k^0+1}}).
\end{aligned}$$

Now, by **H-1**, the second derivative of the transfer function is bounded and a constant C exists so that we have the following inequalities:

$$\forall (\theta_i, \theta_j) \in \{b_1, \cdots, b_k, w_{11}, \cdots, w_{kd}\}^2, \; \sup_{\theta \in \Theta} \left\| \frac{\partial^2 f_\theta(X)}{\partial \theta_i \partial \theta_j} \right\| \leq C(1 + |X|^2).$$

So, thanks to assumption **H-2**, the second order derivative of the function $f_{(\Phi_t, \psi_t)}$ with respect to the components of Φ_t will be dominated by a square integrable function. Then, by assumption **H-3** and a Taylor expansion around the identifiable parameter Φ_t^0, we get the following expansion for the numerator of generalized derivative functions:

Lemma 2 *For a fixed t, in the neighborhood of the identifiable parameter* Φ_t^0*:*

$$f_{(\Phi_t,\psi_t)}(x) - f_0(x) = (\Phi_t - \Phi_t^0)^T f'_{(\Phi_t^0,\psi_t)}(x)$$
$$+0.5(\Phi_t - \Phi_t^0)^T f''_{(\Phi_t^0,\psi_t)}(x)(\Phi_t - \Phi_t^0) + o(\|f_{(\Phi_t,\psi_t)} - f_0\|_2^2),$$

with

$$(\Phi_t - \Phi_t^0)^T f'_{(\Phi_t^0,\psi_t)}(x) = \gamma + \sum_{i=1}^{k^0} s_i \phi(w_i^{0T} x + b_i^0)$$
$$+\sum_{i=1}^{k^0}\sum_{j=t_i+1}^{t_{i+1}} q_j \left(w_j - w_i^0\right)^T x a_i^0 \phi'(w_i^{0T} x + b_i^0)$$
$$+\sum_{i=1}^{k^0}\sum_{j=t_i+1}^{t_{i+1}} q_j \left(b_j - b_i^0\right) a_i^0 \phi'(w_i^{0T} x + b_i^0)$$
$$+\sum_{i=t_{k^0+1}+1}^{k} a_i \phi(w_i^T x + b_i)$$

and

$$(\Phi_t - \Phi_t^0)^T f''_{(\Phi_t^0,\psi_t)}(x)(\Phi_t - \Phi_t^0) =$$
$$\sum_{i=1}^{k^0}\sum_{j=t_i+1}^{t_{i+1}} q_j (w_j - w_i^0)^T x x^T (w_j - w_i^0) a_i^0 \phi''(w_i^{0T} x + b_i^0)$$
$$+\sum_{i=1}^{k^0}\sum_{j=t_i+1}^{t_{i+1}} q_j (w_j - w_i^0)^T x (b_j - b_i^0) \phi''(w_i^{0T} x + b_i^0)$$
$$+\sum_{i=1}^{k^0}\sum_{j=t_i+1}^{t_{i+1}} q_j (b_j - b_i^0)^2 \phi''(w_i^{0T} x + b_i^0)$$
$$+\sum_{i=1}^{k^0}\sum_{j=t_i+1}^{t_{i+1}} q_j (w_j - w_i^0)^T x s_i \phi'(w_i^{0T} x + b_i^0)$$
$$+\sum_{i=1}^{k^0}\sum_{j=t_i+1}^{t_{i+1}} q_j (b_j - b_i^0) s_i \phi'(w_i^{0T} x + b_i^0).$$

This development is obtained by a straightforward calculation of the derivatives of $f_{(\Phi_t,\psi_t)} - f_0$ with respect to the components of Φ_t up to the second order.

So, the numerator of generalized derivative functions can be written like (10) and the assumption **C-2** is true. The proposition 1 can be applied to this model and the polynomial bound for the growth of bracketing number shows the Donsker property of generalized derivative functions, hence the assumption **B-2** of theorem 1 is true. Moreover, under **C-2**, the map

$$\Phi_t \mapsto P(Y - f_{(\Phi_t,\psi_t)}(X))^2$$

admits a second-order Taylor expansion with strictly positive second derivative $\frac{\partial^2 P(Y - f_{(\Phi_t,\psi_t)}(X))^2}{\partial \Phi_t^2}$ at $\Phi_t = \Phi_t^0$, so the assumption **B-1** is also true and we can apply Theorem 1 and corollary 1.

Asymptotic index set D The set of limit score functions $\mathscr{D}$ is defined as the set of functions d so that one can find a sequence $(\Phi_n, \psi_n)_{n=1,\ldots}$ satisfying $\|f_{(\Phi_n,\psi_n)} - f_0\|_2 \to 0$ and $\|d - d_{f_{(\Phi_n,\psi_n)}}\|_2 \to 0$. This limit function depends on the development obtained in lemma 2.

Let us define the two principal behaviors for the sequences $f_{(\Phi_n,\psi_n)}$ which influence the form of functions d :

– If the second order term is negligible with respect to the first one:

$$f_{(\Phi_n,\psi_n)} - f_0 = (\Phi_n - \Phi^0)^T f'_{(\Phi_t^0,\psi_n)} + o(\|f_{(\Phi_n,\psi_n)} - f_0\|_2).$$

– If the second order term is not negligible with respect to the first one:

$$f_{(\Phi_n,\psi_n)} - f_0 = (\Phi_n - \Phi^0)^T f'_{(\Phi_i^0,\psi_n)} + \\ 0.5(\Phi_n - \Phi^0)^T f''_{(\Phi^0,\psi_n)}(\Phi_n - \Phi^0) + o(\|f_{(\Phi_n,\psi_n)} - f_0\|_2^2).$$

In the first case, a set $t = \left(t_1, \cdots, t_{k^0+1}\right)$ exists so that the limit function of $d_{f_{(\Phi_n,\psi_n)}}$ will be in the set:

$$\mathscr{D}_1 = \left\{ \Omega \left(\gamma + \sum_{i=1}^{k^0} \varepsilon_i \phi(w_i^{0T} X + b_i^0) + \sum_{i=1}^{k^0} \phi'(w_i^{0T} X + b_i^0)(\zeta_i^T X + \alpha_i) \right. \right. \\ \left. + \sum_{i=t_{k^0+1}+1}^{k} \mu_i \phi(w_i{}^T X + b_i) \right), \\ \gamma, \varepsilon_1, \cdots, \varepsilon_{k^0}, \alpha_1, \cdots, \alpha_{k^0} \in \mathbb{R}, \mu_{t_{k^0+1}+1}, \cdots, \mu_k \in \mathbb{R}^+; \\ \zeta_1, \cdots, \zeta_{k^0} \in \mathbb{R}^d, \\ \left. (w_{k^0+1+1}, b_{k^0+1+1}), \cdots, (w_k, b_k) \in \Theta \backslash \left\{ (w_1^0, b_1^0), \cdots, (w_{k^0}^0, b_{k^0}^0) \right\} \right\}$$

In the second case, an index i exists so that :

$$\sum_{j=t_i+1}^{t_{i+1}} q_j(v_j - w_i^0) = 0 \text{ and } \sum_{j=t_i+1}^{t_{i+1}} q_j(\eta_j - b_i^0) = 0,$$

otherwise, the second order term will be negligible compared to the first one. So

$$\sum_{j=t_i+1}^{t_{i+1}} \sqrt{q_j} \times \sqrt{q_j}(v_j - w_i^0) = 0 \text{ and } \sum_{j=t_i+1}^{t_{i+1}} \sqrt{q_j} \times \sqrt{q_j}(\eta_j - b_i^0) = 0.$$

Hence, a set $t = \left(t_1, \cdots, t_{k^0+1}\right)$ exists so that the set of functions d will be:

$$\mathscr{D}_2 = \left\{ \Omega \left(\gamma + \sum_{i=1}^{k^0} \varepsilon_i \phi(w_i^{0T} X + b_i^0) \right. \right. \\ + \sum_{i=1}^{k^0} \phi'(w_i^{0T} X + b_i^0)(\zeta_i^T X + \alpha_i) \\ + \delta(i) \sum_{i=1}^{k^0} \phi''(w_i^{0T} X + b_i^0) \times \\ \left(\left(\sum_{j=t_i+1}^{t_{i+1}} v_j{}^T X X^T v_j + \eta_j v_j{}^T X + \eta_j{}^2 \right) \right) \\ \left. + \sum_{i=t_{k^0+1}+1}^{k} \mu_i \phi(w_i{}^T X + b_i) \right), \\ \gamma, \varepsilon_1, \cdots, \varepsilon_{k^0}, \alpha_1, \cdots, \alpha_{k^0}, \eta_{t_1}, \cdots, \eta_{t_{k^0+1}} \in \mathbb{R}, \\ \mu_{t_{k^0+1}+1}, \cdots, \mu_k \in \mathbb{R}^+; \zeta_1, \cdots, \zeta_{k^0}, v_{t_1}, \cdots, v_{t_{k^0+1}} \in \mathbb{R}^d, \\ \left. (w_{k^0+1+1}, b_{k^0+1+1}), \cdots, (w_k, b_k) \in \Theta \backslash \left\{ (w_1^0, b_1^0), \cdots, (w_{k^0}^0, b_{k^0}^0) \right\} \right\}$$

where $\delta(i) = 1$ if a vector $\mathbf{q}$ exists so that
$q_j \geq 0, \sum_{j=t_i+1}^{t_{i+1}} q_j = 1, \sum_{j=t_i+1}^{t_{i+1}} \sqrt{q_j} v_j^t = 0$ and $\sum_{j=t_i+1}^{t_{i+1}} \sqrt{q_j} \eta_j = 0$, otherwise $\delta(i) = 0$.

Hence, the limit index set functions will belong to $\mathscr{D}$.

Conversely, let d be an element of $\mathscr{D}$, since function d is not null, one of its components is not equal to 0. Let us assume that this component is γ, but the proof would be similar with any other component. The norm of d is the constant 1, so any component of d is determined by the ratio: $\frac{\varepsilon_1}{\gamma}, \cdots, \frac{1}{\gamma} v_{k^0+1}$.

Then, since Θ contains a neighborhood of the parameters realizing the true regression function f_0, we can chose

$$\theta_n = (\beta^n, a_1^n, \cdots, a_k^n, w_1^n, \cdots, w_k^n, b_1^n, \cdots, b_k^n) \mapsto (\Phi_t^n, \psi_t^n)$$

so that:

$$\begin{aligned}
&\forall i \in \{1, \cdots, k^0\} \ : \ \tfrac{s_i^n}{\beta_n - \beta^0} \stackrel{n\to\infty}{\longrightarrow} \tfrac{\varepsilon_i}{\gamma}, \\
&\forall i \in \{1, \cdots, k^0\} \ : \ \textstyle\sum_{j=t_{i-1}+1}^{t_i} \tfrac{q_j^n}{\beta_n - \beta^0} \left(w_j^n - w_i^0\right) \stackrel{n\to\infty}{\longrightarrow} \tfrac{1}{\gamma}\zeta_i, \\
&\forall i \in \{1, \cdots, k^0\} \ : \ \textstyle\sum_{j=t_{i-1}+1}^{t_i} \tfrac{q_j^n}{\beta_n - \beta^0} \left(b_j^n - b_i^0\right) \stackrel{n\to\infty}{\longrightarrow} \tfrac{1}{\gamma}\alpha_i, \\
&\forall j \in \{t_1, \cdots, t_{k^0+1}\} \ : \ \tfrac{\sqrt{q_j^n}}{\beta_n - \beta^0} \left(w_j^n - w_i^0\right) \stackrel{n\to\infty}{\longrightarrow} \tfrac{1}{\gamma}\nu_j, \\
&\forall j \in \{t_1, \cdots, t_{k^0+1}\} \ : \ \tfrac{\sqrt{q_j^n}}{\beta_n - \beta^0} \left(b_j^n - b_i^0\right) \stackrel{n\to\infty}{\longrightarrow} \tfrac{1}{\gamma}\eta_j, \\
&\forall j \in \{t_{k^0+1} + 1, \cdots, k\} \ : \ \tfrac{\sqrt{q_j^n}}{\beta_n - \beta^0} a_j^n \stackrel{n\to\infty}{\longrightarrow} \tfrac{1}{\gamma}\mu_j.
\end{aligned}$$

■

References

Amari S, Park H, Ozeki T (2006) Singularities affect dynamics of learning in neuromanifolds. Neur comp 18:1007–1065

Anthony M, Bartlett P (1999) Neural network learning: Theoretical foundations. Cambridge University Press

Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. Stat Surveys 4:40–79

Dacunha-Castelle D, Gassiat E (1997) Testing in locally conic models and application to mixture models. ESAIM Probab Statist 1:285–317

Dacunha-Castelle D, Gassiat E (1999) Testing the order of a model using locally conic parametrization; population mixtures and stationary arma processes. Ann Statist 27(4):1178–1209

Devroye L, Gyorfi, Lugosi G (1996) A probabilistic theory of pattern recognition. Springer-Verlag

Doukhan P, Massart P, Rio E (1995) Invariance principles for absolutely regular empirical processes. Ann Inst Henri Poincar 31(2):393–427

Fukumizu K (1996) A regularity condition of the information matrix of a multilayer perceptron network. Neural networks 9(5):871–879

Fukumizu K (1999) Generalization error in linear neural networks in unidentifiable cases, Springer-Verlag, pp 51–62. No. 1720 in Lecture Notes in Artificial Intelligence

Fukumizu K (2003) Likelihood ratio of unidentifiable models and multilayer neural networks. Ann Statist 31(3):833–851

Gassiat E (2002) Likelihood ratio inequalities with applications to various mixture. Ann Inst Henri Poincar 38:897–906

Hagiwara K (2002) On the problem in model selection of neural networks regression on overrealizable scenario. Neural Computation 14:1979–2002

Hagiwara K, Fukumizu K (2008) Relation between weight size and degree of overfitting in neural network regression. Neural networks 21:48–58
Liu X, Shao Y (2003) Asymptotics for likelihood ratio tests under loss of identifiability. Ann Statist 31(3):807–832
Olteanu M, Rynkiewicz J (2012) Asymptotic properties of autoregressive regime-switching models. ESAIM: Probab Statist 16:25–47
Ripley B (1996) Pattern recognition and neural networks. Cambridge University Press
Schwarz G (1978) estimating the dimension of a model. Ann Stat 6:2:461–464
van der Vaart A (1998) Asymptotic statistics. Cambridge University Press
Vapnik V (1998) Statistical learning theory. John Wiley and Sons
White H (1992) Artificial neural networks. Blackwell
Yao J (2000) On least square estimation for stable nonlinear ar processes. Ann Inst Math Stat 52:316–331