

# Asymptotic law of likelihood ratio for multilayer perceptron models

Joseph Rynkiewicz

CES-SAMOS-MATISSE

Université de Paris 1, 90 rue de Tolbiac, 750013 Paris

joseph.rynkiewicz@univ-paris1.fr

**Abstract.** We consider regression models involving multilayer perceptrons (MLP) with one hidden layer and a Gaussian noise. The data are assumed to be generated by a true MLP model and the estimation of the parameters of the MLP is done by maximizing the likelihood of the model. When the number of hidden units of the model is over-estimated, the Fischer information matrix of the model is singular and the asymptotic behavior of the LR statistic is unknown or can be divergent if the set of possible parameter is too large. This paper deals with this case, and gives the exact asymptotic law of the LR statistic. Namely, if the parameters of the MLP lie in a suitable compact set, we show that the LR statistic converges to the maximum of the square of a Gaussian process indexed by a class of limit score functions.

**Key words:** Multilayer Perceptron, Likelihood ratio statistic, Donsker class, Gaussian process

## 1 Introduction.

Feedforward neural networks are well known and are popular tools to deal with non-linear statistic models. We can describe MLP as a parametric family of probability density functions. If the noise of the regression model is Gaussian then it is well known that the maximum likelihood estimator is equal to the least-square estimator. Therefore, Gaussian likelihood is the usual assumption when we consider feedforward neural networks from a statistical viewpoint. White [12] reviews statistical properties of MLP estimation in detail. However he leaves an important question pending: the asymptotic behavior of the estimator when an MLP in use has redundant hidden units and the Fisher information matrix is singular. Amari et al. [1] give several examples of behavior of the LR statistic in such cases. Fukumizu [5] shows that, for unbounded parameters, the LR statistic can have an order lower bounded by  $O(\log(n))$  with  $n$  the number of observations instead of the classical convergence property to  $\chi^2$  law.

However, a fairly natural assumption is to consider that the parameters are bounded. Indeed, computer calculations always assume that numbers are bounded. Moreover a safe practice is to bound the parameters in order to avoid numerical problems. In such context, different situations can occur. In some

cases, such as mixture models, the LR statistic is tight and the calculation of the asymptotic distribution is possible ([9]). In other cases it may occur that even if the parameters are bounded the likelihood ratio diverges this is for example the case in hidden Markov models (Gassiat and Keribin ([6]). So the behavior of likelihood ratio in the case of MLPs with bounded parameters is still an open question.

In this paper, we derive the distribution of the likelihood ratio if the parameters are in a suitable compact set (i.e. bounded and closed). To obtain this result we use recent techniques introduced by Dacunha-Castelle and Gassiat [3] and Liu and Shao [9]. These techniques consist in finding a parameterization separating the identifiable part and the unidentifiable part of the parameter vector, then we can obtain an asymptotic development of the likelihood of the model which allows us to show that a set of generalized score functions is a Donsker class and to find the asymptotic distribution of the LR statistic. The paper is organized as follows. In section 2 we state the model and the main assumptions. Section 3 presents our main theorem and explains its meaning with a brief summary and a statement of significance of this work in the conclusion. Finally, we prove the theorem in the appendix.

## 2 The model.

We consider the model of regression for  $i \in \mathbb{N}^*$ :

$$Y_i = F_{\theta^0}(X_i) + \varepsilon_i, \quad (1)$$

where  $X_i \in \mathbb{R}^d$  are observed exogenous variables and  $Y_i \in \mathbb{R}$  is the variable to explain. The data  $(Y_i, X_i)$  are assumed to be generated by this true model. The noise  $(\varepsilon_i)_{i \in \mathbb{N}^*}$  is a sequence of independent and identically distributed (i.i.d.)  $\mathcal{N}(0, \sigma^2)$  variables.

### 2.1 The regression function.

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

$$F_{\theta}(x) = \beta + \sum_{i=1}^k a_i \phi(w_i^T x + b_i), \quad (2)$$

with  $\theta = (\beta, a_1, \dots, a_k, b_1, \dots, b_k, w_{10}, \dots, w_{1d}, \dots, w_{k0}, \dots, w_{kd}) \subset \mathbb{R}^{k \times (d+2)+1}$  the parameters of the model. The transfer function  $\phi$  will be assumed bounded and three times derivable. We assume also that the first, second and third derivatives of the transfer function  $\phi$ :  $\phi'$ ,  $\phi''$  and  $\phi'''$  are bounded. In order to simplify the presentation, we assume that the variance of the noise  $\sigma^2$  is known. Note that it is assumed that the true model (1) is included in the considered set of parameter  $\Theta$ . Let us define the true number of hidden units as the smallest integer  $k^0$  so that  $\theta^0 = (\beta^0, a_1^0, \dots, a_{k^0}^0, b_1^0, \dots, b_{k^0}^0, w_1^0, \dots, w_{k^0}^0)$  exists with  $F_{\theta^0}$  equal to the true regression function of model (1).

## 2.2 Parameterization of the model.

Let us write  $\|\cdot\|$  for the Euclidean norm. Let us consider the variable  $Z_i = (X_i, Y_i)$  where  $X_i$  and  $Y_i$  follow the probability law induced by the model (1). We assume that the law of  $X_i$  will be  $q(x)\lambda_d(x)$  with  $\lambda_d$  the Lebesgue measure on  $\mathbb{R}^d$  and  $q(x) > 0$  for all  $x \in \mathbb{R}^d$ . The likelihood of the observation  $z := (x, y)$  for a parameter vector  $\theta = (\beta, a_1, \dots, a_k, b_1, \dots, b_k, w_{11}, \dots, w_{1d}, \dots, w_{kd})$  will be written:

$$f_\theta(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-F_\theta(x))^2} q(x). \quad (3)$$

Let  $\eta > 0$  be a small constant and  $M$  a huge constant, the set of possible parameters will be

$$\Theta_k := \{\theta = (\beta, a_1, \dots, a_k, b_1, \dots, b_k, w_{10}, \dots, w_{1d}, \dots, w_{k0}, \dots, w_{kd}), \\ \forall 1 \leq i \leq k, \|w_i\| \geq \eta, \|a_i\| \geq \eta \text{ and } \|\theta\| \leq M\}.$$

*Constraints on the parameter set.* The constraint  $\|w_i\| \geq \eta$  is introduced in order to avoid the hidden unit from being constant like the bias  $\beta$ , instead of being a function of  $x$ . The constraint  $\|a_i\| \geq \eta$  forces the parameters of the hidden units to converge to one of the parameter vector  $w_j^0, j \in \{1, \dots, k^0\}$  when they maximize the likelihood. Finally, with the constraint  $\|\theta\| \leq M$ , the parameters are bounded and the set  $\Theta_k$  compact. Note that these constraints are very easy to set in practice.

The true density of the observation will be denoted  $f(z) := f_{\theta^0}(z)$ . The main goal of the parametric statistic is to give an estimation of the true parameter  $\theta_0$  thanks to the observations  $(z_1, \dots, z_n)$ . This can be done by maximizing the log-likelihood function :

$$l_n(\theta) := \sum_{i=1}^n \log f_\theta(z_i). \quad (4)$$

The parameter vectors  $\hat{\theta}_n$  realizing the maximum will be called Maximum Likelihood Estimator (MLE). However, the MLE belongs to a non-null dimension submanifold if the number of hidden units is overestimated. In the next section we will study the behavior of  $\sup_{\theta \in \Theta_k} \sum_{i=1}^n \log f_\theta(z_i) - \log f(z_i)$ , where  $k > k^0$ , which is the key point to guess the true architecture of the MLP model.

## 3 Asymptotic distribution of the LR statistic.

We will use the abbreviation  $Pg = \int gdP$  for an integrable function  $g$  and a measure  $P$ . We will define the  $L^2(P)$  norm as  $\|g\|_2 = \sqrt{Pg^2}$  and the map  $\Omega : L^2(P) \rightarrow L^2(P)$  as  $\Omega(g) = \frac{g}{\|g\|_2}$  if  $g \neq 0$ . The maximum of the log-likelihood will be denoted :

$$\lambda_n^k = \sup_{\theta \in \Theta_k} \sum_{i=1}^n \log f_\theta(z_i) - \log f(z_i). \quad (5)$$

Finally, let us note

$$e(z) := \frac{1}{\sigma^2} \left( y - \left( \beta^0 + \sum_{i=1}^{k^0} a_i^0 \phi(b_i^0 + w_i^{0T} x) \right) \right). \quad (6)$$

For what follows, we will assume the properties:

H-1 : The parameters of  $\Theta_k$  realizing the true regression function  $F_{\theta_0}$  lie in the interior of  $\Theta_k$ .

H-2 : Let  $k$  be an integer greater or equal to  $k^0$  and  $\delta_{(w_i, b_i)}$  be the indicator function of the parameters  $(w_i, b_i)$ . The model is identifiable in the weak following sense:

$$F_{\theta^0} = F_{\theta} \Leftrightarrow \beta^0 = \beta \text{ and } \sum_{i=1}^{k^0} a_i^0 \delta_{(w_i^0, b_i^0)} = \sum_{i=1}^k a_i \delta_{(w_i, b_i)}. \quad (7)$$

Note that, it is possible that some new constraint on the parameters have to be set to fulfill this assumption. For example, if the transfer function is the hyperbolic tangent (or any odd function), the constraints on the parameters  $a_i$  will be :  $a_i \geq \eta$ , in order to avoid a symmetry on the sign (because  $\tanh(-t) = -\tanh(t)$ ).

H-3 :  $E(\|X\|^6) < \infty$ .

H-4 : the functions of the set

$$\begin{aligned} & \left( 1, \left( x_j x_l \phi''(w_i^{0T} x + b_i^0) \right)_{1 \leq l \leq j \leq d, 1 \leq i \leq k^0}, \left( x_j \phi''(w_i^{0T} x + b_i^0) \right)_{1 \leq j \leq d, 1 \leq i \leq k^0} \right. \\ & \left. \phi''(w_i^{0T} x + b_i^0)_{1 \leq i \leq k^0}, \left( x_j \phi'(w_i^{0T} x + b_i^0) \right)_{1 \leq j \leq d, 1 \leq i \leq k^0} \right. \\ & \left. \left( \phi'(w_i^{0T} x + b_i^0) \right)_{1 \leq i \leq k^0}, \left( \phi(w_i^{0T} x + b_i^0) \right)_{1 \leq i \leq k^0} \right) \end{aligned}$$

are linearly independent in the Hilbert space  $L^2(q\lambda_d)$ .

We get then the following result:

**Theorem 1.** *Under the assumptions H-1, H-2, H-3 an H-4, a centered Gaussian process  $\{W_S, S \in \mathbb{F}^k\}$  with continuous sample path and covariance kernel  $P(W_{S_1} W_{S_2}) = P(S_1 S_2)$  exists so that*

$$\lim_{n \rightarrow \infty} 2\lambda_n^k = \sup_{S \in \mathbb{F}^k} (\max(W_S, 0))^2. \quad (8)$$

The index set  $\mathbb{F}^k$  is defined as  $\mathbb{F}^k = \cup_t \mathbb{F}_t^k$ , the union runs over any possible  $t = (t_0, \dots, t_{k^0}) \in \mathbb{N}^{k^0+1}$  with  $0 = t_0 < t_1 < \dots < t_{k^0} \leq k$  and

$$\begin{aligned} \mathbb{F}_t^k = & \left\{ \Omega \left( \gamma e(z) + \sum_{i=0}^{k^0} \epsilon_i e(z) \phi(w_i^{0T} x + b_i^0) \right. \right. \\ & + \sum_{i=0}^{k^0} e(z) \phi'(w_i^{0T} x + b_i^0) (\zeta_i^T x + \alpha_i) \\ & \left. \left. + \sum_{i=1}^{k^0} e(z) sg(a_i^0) \phi''(w_i^{0T} x + b_i^0) \left( \delta(i) \left( \sum_{j=t_{i-1}+1}^{t_i} \nu_j^{t_i T} x x^T \nu_j^t + \eta_j \nu_j^{t_i T} x + \eta_j^{t_i^2} \right) \right) \right) \right\} \\ & \gamma, \epsilon_1, \dots, \epsilon_{k^0}, \alpha_1, \dots, \alpha_{k^0}, \eta_1, \dots, \eta_{k^0} \in \mathbb{R}; \\ & \zeta_1, \dots, \zeta_{k^0}, \nu_1^t, \dots, \nu_{t_{k^0}}^t \in \mathbb{R}^{d+1} \end{aligned}$$

where  $e(z)$  is defined by (6),  $\delta(i) = 1$  if a vector  $\mathbf{q}$  exists so that:  
 $\sum_{j=t_{i-1}+1}^{t_i} q_j = 1$ ,  $\sum_{j=t_{i-1}+1}^{t_i} \sqrt{q_j} \nu_j^t = 0$  and  $\sum_{j=t_{i-1}+1}^{t_i} \sqrt{q_j} \eta_j = 0$ , otherwise  
 $\delta(i) = 0$ . The function  $sg$  is defined by  $sg(x) = 1$  if  $x > 0$  and  $sg(x) = -1$  if  
 $x < 0$ .

This theorem is proved in the appendix. Note that this theorem prove that the LR statistic is tight so penalized likelihood like BIC yields a consistent method to identify the minimal architecture of the true model, practical applications of such method can be found, for example, in Mangeas [8].

## 4 Conclusion.

We have computed the asymptotic distribution of the LR statistic for parametric MLP regression. Note that the assumption H-4 can be proved for sigmoidal transfert function by using a method similar to Fukumizu [4]. So this theorem can be applied to the most widely used transfer functions for MLP. Finally, the limiting distribution uses the framework proposed by Dacunha-Castelle and Gassiat [3]. It is known that the convergence to the limit is very slow, but this theorem shows that the LR statistic is tight, so information criteria such as the Bayesian information criteria (BIC) will be consistent in the sense that they will select the model with the true dimension  $k^0$  with probability 1, as the number of observations goes to infinite. This is the main practical application of the results obtained in this paper.

## 5 Appendix: Proof of Theorem 1.

Let  $s_\theta(z)$  be the generalized score function defined by

$$s_\theta(z) := \frac{\frac{f_\theta}{f}(z) - 1}{\|\frac{f_\theta}{f}(z) - 1\|_2} . \quad (9)$$

Firstly, we will get an asymptotic development of the generalized score when the model is over-parameterized. We will reparameterize the model using the same method as Liu and Shao [9] for the mixture models.

### 5.1 Reparameterization.

If  $\frac{f_\theta}{f} - 1 = 0$ , we have  $\beta = \beta^0$  and a vector  $t = (t_i)_{1 \leq i \leq k^0}$  exists so that  $0 = t_0 < t_1 < \dots < t_{k^0} \leq k$  and, up to permutations, we have  $w_{t_{i-1}+1} = \dots = w_{t_i} = w_i^0$ ,  $b_{t_{i-1}+1} = \dots = b_{t_i} = b_i^0$ ,  $\sum_{j=t_{i-1}+1}^{t_i} a_j = a_i^0$ . Let us define  $s_i = \sum_{j=t_{i-1}+1}^{t_i} a_j - a_i^0$  and, if  $\sum_{j=t_{i-1}+1}^{t_i} a_j \neq 0$ , let us write  $q_j = \frac{a_j}{\sum_{j=t_{i-1}+1}^{t_i} a_j}$ . If  $\sum_{j=t_{i-1}+1}^{t_i} a_j = 0$ ,  $q_j$  will be set to 0. We get then the reparameterization  $\theta = (\Phi_t, \psi_t)$  with

$$\Phi_t = \left( \beta, (w_j)_{j=1}^{t_{k^0}}, (b_j)_{j=1}^{t_{k^0}}, (s_i)_{i=1}^{k^0} \right), \quad \psi_t = \left( (q_j)_{j=1}^{t_{k^0}} \right) . \quad (10)$$

With this parameterization, for a fixed  $t$ ,  $\Phi_t$  is an identifiable parameter and all the non-identifiability of the model will be in  $\psi_t$ . Namely,  $\frac{f_\theta}{f}(z)$  will be equal to 
$$\frac{\exp\left(-\frac{1}{2\sigma^2}\left(y-\left(\beta+\sum_{i=1}^{k^0}(s_i+a_i^0)\sum_{j=t_{i-1}+1}^{t_i}q_j\phi(w_j^T x)\right)\right)^2\right)}{\exp\left(-\frac{1}{2\sigma_0^2}\left(y-\left(\beta^0+\sum_{i=1}^{k^0}a_i^0\phi(w_i^0 T x)\right)\right)^2\right)}$$
. So  $\frac{f_\theta}{f}(z) = 1$  if and only if

$$\Phi_t^0 = \left(\beta^0, \underbrace{w_1^0, \dots, w_1^0}_{t_1}, \dots, \underbrace{w_{k^0}^0, \dots, w_{k^0}^0}_{t_{k^0} - t_{k^0-1}}, \underbrace{b_1^0, \dots, b_1^0}_{t_1}, \dots, \underbrace{b_{k^0}^0, \dots, b_{k^0}^0}_{t_{k^0} - t_{k^0-1}}, \underbrace{0, \dots, 0}_{k^0}\right).$$

Now, the third derivative of the transfer function is bounded and a constant  $C$  exists so that we have the following inequalities:

$$\forall \theta_i, \theta_j, \theta_l \in \{b_1, \dots, b_k, w_{11}, \dots, w_{kd}\}, \sup_{\theta \in \Theta_k} \left\| \frac{\partial^3 F_\theta(X)}{\partial \theta_i \partial \theta_j \partial \theta_l} \right\| \leq C(1 + \|X\|^3). \quad (11)$$

So, thanks to the assumption H-3, the third order derivative of the function  $\frac{f_\theta}{f}(z)$  with respect to the components of  $\Phi_t$  will be dominated by a square integrable function. Then, by the Taylor formula with an integral remainder around the identifiable parameter  $\Phi_t^0$ , we get the following expansion for the likelihood ratio:

**Lemma 1.** *For a fixed  $t$ , let us write  $D(\Phi_t, \psi_t) := \left\| \frac{f(\Phi_t, \psi_t)}{f} - 1 \right\|_2$ . In the neighborhood of the identifiable parameter  $\Phi_t^0$ , the following approximation is true:*

$$\frac{f_\theta}{f}(z) = 1 + (\Phi_t - \Phi_t^0)^T f'_{(\Phi_t^0, \psi_t)}(z) + 0.5(\Phi_t - \Phi_t^0)^T f''_{(\Phi_t^0, \psi_t)}(z)(\Phi_t - \Phi_t^0) + o(D(\Phi_t, \psi_t)),$$

with

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

and

$$\begin{aligned} (\Phi_t - \Phi_t^0)^T f''_{(\Phi_t^0, \psi_t)}(z)(\Phi_t - \Phi_t^0) &= \\ &\left(1 - \frac{1}{e^2(z)}\right) \left( (\Phi_t - \Phi_t^0)^T f'_{(\Phi_t^0, \psi_t)}(z) f'_{(\Phi_t^0, \psi_t)}(z) (\Phi_t - \Phi_t^0) \right) \\ &+ e(z) \times \left( \sum_{i=1}^{k^0} \sum_{j=t_{i-1}+1}^{t_i} q_j (w_j - w_i^0)^T x x^T (w_j - w_i^0) a_i^0 \phi''(w_i^0 T x + b_i^0) \right. \\ &+ \sum_{i=1}^{k^0} \sum_{j=t_{i-1}+1}^{t_i} q_j (w_j - w_i^0)^T x (b_j - b_i^0) \phi''(w_i^0 T x + b_i^0) \\ &+ \sum_{i=1}^{k^0} \sum_{j=t_{i-1}+1}^{t_i} q_j (b_j - b_i^0)^2 \phi''(w_i^0 T x + b_i^0) \\ &+ \sum_{i=1}^{k^0} \sum_{j=t_{i-1}+1}^{t_i} q_j (w_j - w_i^0)^T x s_i \phi'(w_i^0 T x + b_i^0) \\ &\left. + \sum_{i=1}^{k^0} \sum_{j=t_{i-1}+1}^{t_i} q_j (b_j - b_i^0) s_i \phi'(w_i^0 T x + b_i^0) \right). \end{aligned}$$

The development, obtained by a straightforward calculation of the derivatives of  $\frac{f_\theta}{f}(z)$  with respect to the components of  $\Phi_t$  up to the second order.

Now, the convergence to a Gaussian process will be derived from the Donsker property of the set of generalized score functions  $\mathbb{S} = \{s_\theta(z), \theta \in \Theta_k\}$ . Let an  $\varepsilon$ -bracket  $[l, u]$  be a set of function  $h$  with  $l \leq h \leq u$  with  $\sqrt{P(l-u)^2} < \varepsilon$ . The bracketing number  $N_{[]}(\varepsilon, \mathbb{S}, L^2(f\lambda_{d+1}))$  is the minimum number of  $\varepsilon$ -brackets needed to cover  $\mathbb{S}$ . The entropy with bracketing is the logarithm of the bracketing number. It is well known (Van der Vaart [11]) that the class of functions  $\mathbb{S}$  will be Donsker if its entropy with bracketing grows with a slower order than  $\frac{1}{\varepsilon}^2$ . A sufficient condition for Donsker property is then that the bracketing number grows as a polynomial function of  $\frac{1}{\varepsilon}$ .

## 5.2 Polynomial bound for the growth of bracketing number.

Let us write  $D(\theta) := \left\| \frac{f(\theta)}{f} - 1 \right\|_2$ , for all  $\varepsilon > 0$ , the set of parameters can be divided in two sets:  $\mathbb{S}_\varepsilon$  and  $\mathbb{S}_0$  with

$$\mathbb{S}_\varepsilon = \{\theta \in \Theta_k \text{ so that } D(\theta) \geq \varepsilon\} \text{ and } \mathbb{S}_0 = \{\theta \in \Theta_k \text{ so that } D(\theta) < \varepsilon\}.$$

For  $\theta_1$  and  $\theta_2$  belonging to  $\mathbb{S}_\varepsilon$ , we get:

$$\begin{aligned} & \left\| \frac{\frac{f_{\theta_1} - 1}{f} - \frac{f_{\theta_2} - 1}{f}}{\left\| \frac{f_{\theta_1} - 1}{f} - 1 \right\|_2} - \frac{\frac{f_{\theta_2} - 1}{f} - \frac{f_{\theta_1} - 1}{f}}{\left\| \frac{f_{\theta_2} - 1}{f} - 1 \right\|_2} \right\|_2 = \left\| \frac{\frac{f_{\theta_1} - 1}{f} - 1}{\left\| \frac{f_{\theta_1} - 1}{f} - 1 \right\|_2} - \frac{\frac{f_{\theta_2} - 1}{f} - 1}{\left\| \frac{f_{\theta_1} - 1}{f} - 1 \right\|_2} + \frac{\frac{f_{\theta_2} - 1}{f} - 1}{\left\| \frac{f_{\theta_1} - 1}{f} - 1 \right\|_2} - \frac{\frac{f_{\theta_2} - 1}{f} - 1}{\left\| \frac{f_{\theta_2} - 1}{f} - 1 \right\|_2} \right\|_2 \\ & \leq \left\| \frac{\frac{f_{\theta_1} - 1}{f} - 1}{\left\| \frac{f_{\theta_1} - 1}{f} - 1 \right\|_2} - \frac{\frac{f_{\theta_2} - 1}{f} - 1}{\left\| \frac{f_{\theta_1} - 1}{f} - 1 \right\|_2} \right\|_2 + \left\| \frac{\frac{f_{\theta_2} - 1}{f} - 1}{\left\| \frac{f_{\theta_1} - 1}{f} - 1 \right\|_2} - \frac{\frac{f_{\theta_2} - 1}{f} - 1}{\left\| \frac{f_{\theta_2} - 1}{f} - 1 \right\|_2} \right\|_2 \\ & \leq 2 \frac{\left\| \frac{f_{\theta_1} - 1}{f} - \frac{f_{\theta_2} - 1}{f} \right\|_2}{\left\| \frac{f_{\theta_1} - 1}{f} - 1 \right\|_2} \leq 2 \frac{\left\| \frac{f_{\theta_1} - 1}{f} - \frac{f_{\theta_2} - 1}{f} \right\|_2}{\varepsilon}. \end{aligned}$$

Hence, on  $\mathbb{S}_\varepsilon$ , it is sufficient that  $\left\| \frac{f_{\theta_1} - 1}{f} - \frac{f_{\theta_2} - 1}{f} \right\|_2 < \frac{\varepsilon^2}{2}$  for

$$\left\| \frac{\frac{f_{\theta_1} - 1}{f} - 1}{\left\| \frac{f_{\theta_1} - 1}{f} - 1 \right\|_2} - \frac{\frac{f_{\theta_2} - 1}{f} - 1}{\left\| \frac{f_{\theta_2} - 1}{f} - 1 \right\|_2} \right\|_2 < \varepsilon.$$

Now,  $\mathbb{S}_\varepsilon$  is a parametric class. Since the derivatives of the transfer functions are bounded and  $E\|X\| < \infty$  a function  $m(z)$  exists, with  $E[m(z)] < \infty$ , so that

$$\forall \theta_i \in \{\beta, a_1, \dots, a_k, w_{10}, \dots, w_{1d}, \dots, w_{kd}\}, \left| \frac{\partial \frac{f_\theta}{f}}{\partial \theta_i}(z) \right| \leq m(z).$$

According to the example 19.7 of Van der Vaart [11], a constant  $K$  exists so that the bracketing number of  $\mathbb{S}_\varepsilon$  is lower than

$$K \left( \frac{\text{diam} \Theta_k}{\varepsilon^2} \right)^{k \times (d+2) + 1} = K \left( \frac{\sqrt{\text{diam} \Theta_k}}{\varepsilon} \right)^{k \times (2d+4) + 2}, \quad (12)$$

where  $\text{diam}\Theta_k$  is the diameter of the smallest sphere of  $\mathbb{R}^k$  including  $\Theta_k$ .

For  $\theta$  belonging to  $\mathbb{S}_0$ ,  $\frac{f_\theta(z)}{f} - 1$  is the sum of a linear combination of

$$\begin{aligned} V(z) := & \left( e(z), \left( e(z)x_j x_l \phi''(w_i^{0T}x + b_i^0) \right)_{1 \leq l \leq j \leq d, 1 \leq i \leq k^0}, \right. \\ & \left( e(z)x_j \phi''(w_i^{0T}x + b_i^0) \right)_{1 \leq j \leq d, 1 \leq i \leq k^0}, \\ & \left( e(z)\phi''(w_i^{0T}x + b_i^0) \right)_{1 \leq i \leq k^0}, \left( e(z)x_j \phi'(w_i^{0T}x + b_i^0) \right)_{1 \leq j \leq d, 1 \leq i \leq k^0}, \\ & \left. \left( e(z)\phi'(w_i^{0T}x + b_i^0) \right)_{1 \leq i \leq k^0}, \left( e(z)\phi(w_i^{0T}x + b_i^0) \right)_{1 \leq i \leq k^0} \right) \end{aligned}$$

and of a term whose  $L^2(f\lambda_{d+1})$  norm is negligible compared to the  $L^2(f\lambda_{d+1})$  norm of this combination when  $\varepsilon$  goes to 0. By assumption H-4, a strictly positive number  $m$  exists so that for any vector of norm 1 with components

$$\begin{aligned} C = & \left( c, c_1, \dots, c_{k^0 \times \frac{d(d+1)}{2}}, d_1, \dots, d_{k^0 \times d}, e_1, \dots, e_{k^0}, \right. \\ & \left. f_1, \dots, f_{k^0 \times d}, g_1, \dots, g_{k^0}, h_1, \dots, h_{k^0} \right) \end{aligned}$$

and  $\varepsilon$  sufficiently small:

$$\|C^T V(z)\|_2 > m + \varepsilon.$$

Since any function  $\frac{\frac{f_\theta}{f} - 1}{\|\frac{f_\theta}{f} - 1\|_2}$  can be written:

$$\frac{C^T V(z) + o(\|C^T V(z)\|_2)}{\|C^T V(z) + o(\|C^T V(z)\|_2)\|_2},$$

$\mathbb{S}_0$  belongs to the set of functions:

$$\left\{ D^T V(z) + o(1), \|D\|_2 \leq \frac{1}{m} \right\} \subset \left\{ D^T V(z) + \gamma, \|D\|_2 \leq \frac{1}{m}, |\gamma| < 1 \right\},$$

whose bracketing number is smaller or equal to  $O\left(\frac{1}{\varepsilon}\right)^{k^0 \times \left(\frac{d(d+1)}{2} + 2d + 3\right) + 2}$ .

This proves that the bracketing number of  $\mathbb{S}$  is polynomial, hence  $\mathbb{S}$  is a Donsker class.

### 5.3 Asymptotic index set.

Since the class of generalized score functions  $\mathbb{S}$  is a Donsker class the theorem follows from theorem 3.1 of Gassiat [7] or theorem 3.1 of Liu and Shao [9]. Following these authors, the set of limit score functions  $\mathbb{F}^k$  is defined as the set of functions  $d$  so that one can find a sequence  $g_n := f_{\theta_k^n}, \theta_k^n \in \Theta_k$  satisfying  $\|\frac{g_n - f}{f}\|_2 \rightarrow 0$  and  $\|d - s_{g_n}\|_2 \rightarrow 0$ , where  $s_{g_n} = \frac{\frac{g_n}{f} - 1}{\|\frac{g_n}{f} - 1\|_2}$ . Note that, for a particular sequence of maximum likelihood estimators  $(\theta^n)_{n \in \mathbb{N}}$ , the partition of the indices  $t = (t_0, \dots, t_{k^0}) \in \mathbb{N}^{k^0+1}$  can depend on  $n$ , but  $(\theta^n)_{n \in \mathbb{N}}$  will be the union of converging sub-sequences belonging the set of limit score functions.

Let us define the two principal behaviors for the sequences  $g_n$  which influence the form of functions  $d$  :

– If the second order term is negligible behind the first one :

$$\frac{f_{\theta_k^n}}{f}(z) - 1 = (\Phi_k^n - \Phi^0)^T f'_{(\Phi_t^0, \psi_k^n)}(z) + o(D(\Phi_k^n, \psi_k^n)) .$$

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

$$\begin{aligned} \frac{f_{\theta_k^n}}{f}(z) - 1 &= (\Phi_k^n - \Phi^0)^T f'_{(\Phi_t^0, \psi_k^n)}(z) + \\ &0.5(\Phi_k^n - \Phi_t^0)^T f''_{(\Phi_t^0, \psi_k^n)}(z)(\Phi_k^n - \Phi_t^0) + o(D(\Phi_k^n, \psi_k^n)) . \end{aligned}$$

In the first case, each sequence  $g_n$  is the finite union of convergent subsequences  $g_k(n)$  and for each subsequence a set  $t = (t_0, \dots, t_{k^0}) \in \mathbb{N}^{k^0+1}$  (with  $0 = t_0 < t_1 < \dots < t_{k^0} \leq k$ ) exists so that the limit functions  $d$  of  $sg_k(n)$  will be:

$$\begin{aligned} \mathbb{D}_1^t &= \Omega \left\{ \gamma e(z) + \sum_{i=0}^{k^0} \epsilon_i e(z) \phi(w_i^{0T} x + b_i^0) \right. \\ &\quad \left. + \sum_{i=0}^{k^0} e(z) \phi'(w_i^{0T} x + b_i^0) (\zeta_i^T x + \alpha_i) , \right. \\ &\quad \left. \gamma, \epsilon_1, \dots, \epsilon_{k^0}, \alpha_1, \dots, \alpha_{k^0} \in \mathbb{R} ; \zeta_1, \dots, \zeta_{k^0} \in \mathbb{R}^{d+1} \right\} . \end{aligned}$$

In the second case, each sequence  $g_n$  is the finite union of convergent subsequences  $g_k(n)$  and for each subsequence, an index  $i$  exists so that :

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

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

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

Hence, a set  $t = (t_0, \dots, t_{k^0}) \in \mathbb{N}^{k^0+1}$  exists, with  $0 = t_0 < t_1 < \dots < t_{k^0} \leq k$  so that the set of functions  $d$  will be:

$$\begin{aligned} \Omega \left( \gamma e(z) + \sum_{i=0}^{k^0} \epsilon_i e(z) \phi(w_i^{0T} x + b_i^0) + \sum_{i=0}^{k^0} e(z) \phi'(w_i^{0T} x + b_i^0) (\zeta_i^T x + \alpha_i) \right. \\ \left. + \sum_{i=1}^{k^0} e(z) sg(a_i^0) \phi''(w_i^{0T} x + b_i^0) \left( \delta(i) \left( \sum_{j=t_{i-1}+1}^{t_i} \nu_j^{tT} x x^T \nu_j^t + \eta_j \nu_j^{tT} x + \eta_j^{t2} \right) \right) \right) \\ \gamma, \epsilon_1, \dots, \epsilon_{k^0}, \alpha_1, \dots, \alpha_{k^0}, \eta_1, \dots, \eta_{k^0} \in \mathbb{R} ; \\ \zeta_1, \dots, \zeta_{k^0}, \nu_1^t, \dots, \nu_{t_{k^0}}^t \in \mathbb{R}^{d+1} \left. \right\} , \end{aligned}$$

where  $\delta(i) = 1$  if a vector  $\mathbf{q}$  exists with  $\sum_{j=t_{i-1}+1}^{t_i} q_j = 1$  and  $\sum_{j=t_{i-1}+1}^{t_i} \sqrt{q_j} \nu_j = 0$  and  $\sum_{j=t_{i-1}+1}^{t_i} \sqrt{q_j} \eta_j = 0$ , otherwise  $\delta(i) = 0$ . So, the limit functions  $d$  will belong to  $\mathbb{F}^k$ .

Conversely, for  $x \in L^2(\lambda_{d+1})$ , let  $d$  be an element of  $\mathbb{F}^k$ :

$$\begin{aligned} d &= \Omega \left( \gamma e(z) + \sum_{i=0}^{k^0} \epsilon_i e(z) \phi(w_i^{0T} x + b_i^0) + \sum_{i=0}^{k^0} e(z) \phi'(w_i^{0T} x + b_i^0) (\zeta_i^T x + \alpha_i) \right. \\ &\quad \left. + \sum_{i=1}^{k^0} e(z) sg(a_i^0) \phi''(w_i^{0T} x + b_i^0) \left( \delta(i) \left( \sum_{j=t_{i-1}+1}^{t_i} \nu_j^{tT} x x^T \nu_j^t + \eta_j \nu_j^{tT} x + \eta_j^{t2} \right) \right) \right) . \end{aligned}$$

As functions  $d$  belong to the Hilbert sphere, one of their components is not equal to 0. Let us assume that this component is  $\gamma$ , but the proof would be similar with any other component. The norm of  $d$  is 1, so any component of  $d$  is determined by the ratio:  $\frac{\epsilon_i}{\gamma}, \dots, \frac{1}{\gamma}\eta_{k^0}^t$ .

Then, we can chose  $\theta_k^n = (\beta^n, a_1^n, \dots, a_k^n, w_1^n, \dots, w_k^n, b_1^n, \dots, b_k^n)$  so that:

$$\begin{aligned} \forall i \in \{1, \dots, k^0\} & : \frac{s_i^n}{\beta_n - \beta^0} \xrightarrow{n \rightarrow \infty} \frac{\epsilon_i}{\gamma}, \\ \forall i \in \{1, \dots, k^0\} & : \sum_{j=t_{i-1}+1}^{t_i} \frac{q_j^n}{\beta_n - \beta^0} (w_j^n - w_j^0) \xrightarrow{n \rightarrow \infty} \frac{1}{\gamma} \zeta_i, \\ \forall i \in \{1, \dots, k^0\} & : \sum_{j=t_{i-1}+1}^{t_i} \frac{q_j^n}{\beta_n - \beta^0} (b_j^n - b_j^0) \xrightarrow{n \rightarrow \infty} \frac{1}{\gamma} \alpha_i, \\ \forall j \in \{1, \dots, t_{k^0}\} & : \frac{\sqrt{q_j^n}}{\beta_n - \beta^0} (w_j^n - w_j^0) \xrightarrow{n \rightarrow \infty} \frac{1}{\gamma} \nu_j, \\ \forall j \in \{1, \dots, t_{k^0}\} & : \frac{\sqrt{q_j^n}}{\beta_n - \beta^0} (b_j^n - b_j^0) \xrightarrow{n \rightarrow \infty} \frac{1}{\gamma} \eta_j, \end{aligned}$$

since  $\Theta_k$  contains a neighborhood of the parameters realizing the true regression function  $F_{\theta^0}$ . ■

## References

1. Amari, S., Park, H., Ozeki, T.: Singularities Affect Dynamics of Learning in Neuro-manifolds. *Neural computation*. 18, 1007–1065 (2006)
2. Cottrell, M., Girard, B., Girard, Y., Mangeas, M., Muller, C.: Neural Modeling for Time Series: a Statistical Stepwise Method for Weight Elimination. *IEEE Transaction on Neural Networks*. 6, 1355–1364 (1995)
3. Dacunha-Castelle, D., Gassiat, E.: Testing the Order of a Model Using Locally Conic Parameterization: Population Mixtures and Stationary ARMA process. *The Annals of Statistics*. 27, 1178–1209 (1999)
4. Fukumizu, K.: A Regularity Condition of the Information Matrix of a Multilayer Perceptron Network. *Neural networks*. 9, 871–879 (1996)
5. Fukumizu, K.: Likelihood Ratio of Unidentifiable Models and Multilayer Neural Networks. *The Annals of Statistics*. 31, 833–851 (2003)
6. Gassiat, E., Keribin, C.: The Likelihood Ratio Test for the Number of Components in a Mixture with Markov Regime. *ESAIM Probability and statistics*. 4, 25–52 (2000)
7. Gassiat, E.: Likelihood Ratio Inequalities with Applications to Various Mixtures. *Annales de l'Institut Henri Poincaré*. 38, 897–906 (2002)
8. Mangeas, M.: Neural Model Selection: How to Determine the Fittest Criterion. *Proceedings of ICANN'97*. 987–992 (1997)
9. Liu, X., Shao, Y.: Asymptotics for Likelihood Ratio Tests Under Loss of Identifiability. *The Annals of Statistics*. 31, 807–832 (2003)
10. Sussmann, H. J.: Uniqueness of the Weights for Minimal Feed-Forward Nets with a Given Input-Output Map. *Neural networks*. 5, 589–593 (1992)
11. Van der Vaart, A.W.: *Asymptotic Statistics*. Cambridge: Cambridge university Press (1998)
12. White, H.: *Artificial Neural Networks: Approximation and Learning Theory*. Oxford, Basil Blackwell (1992)
13. Yao, J.: On Least Square Estimation for Stable Nonlinear AR Processes. *The Annals of Institut of Mathematical Statistics*. 52, 316–331 (2000)