View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by HAL-Paris1

General bound of overfitting for MLP regression models.

Joseph Rynkiewicz

▶ **To cite this version:**

Joseph Rynkiewicz. General bound of overfitting for MLP regression models.. Neurocomputing, Elsevier, 2012, 90, pp.106-110. <10.1016/j.neucom.2011.11.028>. <hal-00655918>

HAL Id: hal-00655918

https://hal.archives-ouvertes.fr/hal-00655918

Submitted on 3 Jan 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

General bound of overfitting for MLP regression models.

Rynkiewicz, J.

Abstract

Multilayer perceptrons (MLP) with one hidden layer have been used for a long time to deal with non-linear regression. However, in some task, MLP's are too powerful models and a small mean square error (MSE) may be more due to overfitting than to actual modelling. If the noise of the regression model is Gaussian, the overfitting of the model is totally determined by the behavior of the likelihood ratio test statistic (LRTS), however in numerous cases the assumption of normality of the noise is arbitrary if not false. In this paper, we present an universal bound for the overfitting of such model under weak assumptions, this bound is valid without Gaussian or identifiability assumptions. The main application of this bound is to give a hint about determining the true architecture of the MLP model when the number of data goes to infinite. As an illustration, we use this theoretical result to propose and compare effective criteria to find the true architecture of an MLP.

1 Introduction

Feed-forward neural networks are well known and popular tools to deal with non-linear regression models. We can describe MLP models as a parametric family of regression functions. White [8] reviews statistical properties of MLP estimation in detail. However he leaves an important question pending i.e. the asymptotic behavior of the estimator when the MLP in use has redundant hidden units. If we assume that the noise is Gaussian it is well known that the Least square Estimator (LSE) and the maximum likelihood estimator (MLE) are equivalent and Amari et al. [1] give several examples of

the behavior of MLE in case of redundant hidden units. Moreover, if n is the number of observations, Fukumizu [2] shows that, for unbounded parameters and Gaussian noise, LRTS can have an order lower bounded by $O(\log(n))$ instead of the classical convergence property to a χ^2 law. In the same spirit, Hagiwara and Fukumizu [3] investigate relation between LRTS divergence and weight size in a simple neural networks regression problem. Hence, if parameters of MLP models are not bounded, these papers show that, for Gaussian noise, the overfitting is strong, even if the number of data is large. Note that this is no more the case if the parameters of the MLP are supposed to be a priori bounded.

But, even if Gaussian assumption for the noise is standard, it may be not suitable for some models. This assumption is false, for example, when the range of observations is known to be bounded, since Gaussian variables can be arbitrary large in absolute value, even if the probability of such events is small. Hence, we need a theory which gives evaluation of the overfitting of MLP regression without knowing the density of the noise and which works even if the model is not identifiable.

In this paper, we prove an inequality bounding the MSE difference between the true model and an over-determined model. This inequality shows that, under suitable assumptions, the asymptotic overfitting of the MSE is upper bounded by the maximum of the square of a Gaussian process. Moreover, this bound shows that suitably penalized MSE criteria allow to select asymptotically the true model. The paper is organized as follows: In section 2 we state the model, section 3 presents our main inequality and in section 4 we apply this inequality to select the optimal architecture of the MLP model. Finally, in section 5, a little experiment gives us some insight to apply this theoretical results in real life problems.

2 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 a parameter vector for 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 $\theta = (\beta, a_1, \cdots, a_k, b_1, \cdots, b_k, w_{11}, \cdots, w_{1d}, \cdots, w_{k1}, \cdots, w_{kd})$ the parameter vector of the model and ϕ a bounded transfer function, usually a sigmoïdal function. Note that we consider only real functions, extension to vectorial functions is straightforward but not discussed in this paper. Let $\Theta_k \subset \mathbb{R}^{k\times(d+2)+1}$ be a compact (i.e. closed and bounded) set of possible parameters, we consider regression model $\mathcal{S} = \{f_\theta(y, x),\ \theta \in \Theta_k\}$ with

$$Y = f_\theta(X) + \varepsilon \tag{1}$$

X is a random input variable and ε is the noise of the model. Let n be a strictly positive integer, we assume that the observed data $(x_1, y_1), \cdots, (x_n, y_n)$ come from a true model $(X_i, Y_i)_{i\in\mathbb{N}, i>0}$ of which the true regression function is f_{θ^0}, for an θ^0 (possibly not unique) in the interior of Θ_k.

2.1 Estimation of MLP regression model

The main goal of non-linear regression is to give an estimation of the true parameter θ^0 based on observations $((x_1, y_1), \cdots, (x_n, y_n))$. This can be done by minimizing the Mean Square Error (MSE) function:

$$E_n(\theta) := \frac{1}{n}\sum_{t=1}^{n}\left(y_t - f_\theta(x_t)\right)^2 \tag{2}$$

with respect to parameter vector $\theta \in \Theta_k$. The parameter vectors $\hat{\theta}_n$ realizing the minimum will be called Least Square Estimator (LSE). Note that parameters realizing the true distribution function may belong to a non-null dimension sub-manifold if the number of hidden units is overestimated. Suppose, for example, we have a multilayer perceptron with two hidden units and the true function f_{θ^0} is given by a perceptron with only one hidden unit, say $f_{\theta^0} = a_0 \tanh(w_0 x)$, with $x \in \mathbb{R}$. Then, any parameter θ in the set:

$$\{\theta \,|w_2 = w_1 = w_0, b_2 = b_1 = 0, a_1 + a_2 = a_0\}$$

realizes the function f_{θ^0}. Hence, classical statistical theory for studying the LSE can not be applied because it requires the identification of the parameters (up to a permutation).

In the next section, we will compare MSE of over-determined models against MSE of the true model :

$$\frac{1}{n}\sum_{t=1}^{n}\left(y_t - f_\theta(x_t)\right)^2 - \frac{1}{n}\sum_{t=1}^{n}\left(y_t - f_{\theta^0}(x_t)\right)^2 = E_n(\theta) - E_n(\theta^0). \tag{3}$$

3 A general bound for the MSE

For an square integrable function $g(X,Y)$ the L_2 norm is:

$$\|g(X,Y)\|_2 := \sqrt{\int g^2(x,y)dP(x,y)}. \tag{4}$$

Now, for $\lambda > 0$, let us define the generalized derivative function :

$$d_\theta^\lambda(X,Y) = \frac{\frac{e^{-\lambda(Y-f_\theta(X))^2} - e^{-\lambda(Y-f_{\theta^0}(X))^2}}{e^{-\lambda(Y-f_{\theta^0}(X))^2}}}{\|\frac{e^{-\lambda(Y-f_\theta(X))^2} - e^{-\lambda(Y-f_{\theta^0}(X))^2}}{e^{-\lambda(Y-f_{\theta^0}(X))^2}}\|_2} = \frac{e^{-\lambda\left((Y-f_\theta(X))^2-(Y-f_{\theta^0}(X))^2\right)} - 1}{\|e^{-\lambda\left((Y-f_\theta(X))^2-(Y-f_{\theta^0}(X))^2\right)} - 1\|_2} \tag{5}$$

and let us define $\left(d_\theta^\lambda\right)_-(x,y) = \min\left\{0, d_\theta^\lambda(x,y)\right\}$. Note that the generalized derivative function converges toward the derivative function if θ converges toward θ_0.

For now, let us assume that d_θ^λ is well defined, this point will be discuss later. We can state the following inequality:

Inequality:
for $\lambda > 0$,

$$\sup_{\theta\in\Theta_k} n\cdot\left(E_n(\theta^0) - E_n(\theta)\right) \leq \frac{1}{2\lambda}\sup_{\theta\in\Theta_k}\frac{\sum_{i=1}^n d_\theta^\lambda(x_i,y_i)}{\sum_{i=1}^n \left(d_\theta^\lambda\right)_-^2(x_i,y_i)} \tag{6}$$

Proof:
We have

$$\begin{aligned}
&n\cdot(E_n(\theta^0) - E_n(\theta)) = \\
&\tfrac{1}{\lambda}\textstyle\sum_{i=1}^n \log\left(1 + \|\frac{e^{-\lambda(Y-f_\theta(X))^2} - e^{-\lambda(Y-f_{\theta^0}(X))^2}}{e^{-\lambda(Y-f_{\theta^0}(X))^2}}\|_2 d_\theta^\lambda(x_i,y_i)\right) \\
&\leq \sup_{0\leq p\leq \|\frac{e^{-\lambda(Y-f_\theta(X))^2} - e^{-\lambda(Y-f_{\theta^0}(X))^2}}{e^{-\lambda(Y-f_{\theta^0}(X))^2}}\|_2} \tfrac{1}{\lambda}\textstyle\sum_{i=1}^n \log\left(1 + p d_\theta^\lambda(x_i,y_i)\right) \\
&\leq \sup_{p\geq 0}\tfrac{1}{\lambda}\left(p\textstyle\sum_{i=1}^n d_\theta^\lambda(x_i,y_i) - \tfrac{p^2}{2}\textstyle\sum_{i=1}^n \left(d_\theta^\lambda\right)_-^2(x_i,y_i)\right).
\end{aligned}$$

Since for any real number u, $\log(1+u) \leq u - \frac{1}{2}u_-^2$. Finally, replacing p by the optimal value, we found

$$n\cdot(E_n(\theta^0) - E_n(\theta)) \leq \tfrac{1}{2\lambda}\tfrac{\sum_{i=1}^n d_\theta^\lambda(x_i,y_i)}{\sum_{i=1}^n \left(d_\theta^\lambda\right)_-^2(x_i,y_i)}$$

■

This inequality allows to prove the tightness of $n \cdot (E_n(\theta^0) - E_n(\theta))$ under simple assumptions. It is used in the next section to prove consistency of an estimator of the number of hidden units using penalized MSE criterion.

4 Estimation of the number of hidden units.

Let k^0 be the minimal number of hidden units needed to realize the true regression function f_{θ^0}. In this section, the set Θ of possible parameters will be set to

$$\Theta = \cup_{k=1}^{K} \Theta_k,$$

where K is a, possibly huge, fixed constant: The maximum number of hidden units for MLP models. We define the minimum-penalized MSE estimator of k^0, as the minimizer $\hat{k}$ of

$$T_n(k) = \min_{\theta \in \Theta} \left(E_n(\theta) + a_n(k) \right) \tag{7}$$

Let us assume the following assumptions:

(A1) $a_n(.)$ is increasing, $n \cdot (a_n(k_1) - a_n(k_2))$ tends to infinity as n tends to infinity, for any $k_1 > k_2$ and $a_n(k)$ tends to 0 as n tends to infinity for any k.

(A2) It exists $\lambda > 0$ so that $\left\{ d_\theta^\lambda, \theta \in \Theta \right\}$ is a Donsker class (see van der Vaart [7]) and appendix.

We now have:
Theorem:
Under **(A1)** *and* **(A2)**, $\hat{k}$ *converges in probability to the true number of hidden units* k^0.
Proof:
By applying the inequality,

$$P(\hat{k} > k^0) \leq \sum_{k=k^0+1}^{K} P\left(T_n(k) \geq T_n(k^0)\right) =$$
$$\sum_{k=k^0+1}^{K} P\left(n \left(\sup_{\theta \in \Theta_{k^0}} E_n(\theta) - \sup_{\theta \in \Theta_k} E_n(\theta) \right) \geq n \left(a_n(k) - a_n(k^0) \right) \right) \leq$$
$$\sum_{k=k^0+1}^{K} P\left(\frac{1}{\lambda} \sup_{\theta \in \Theta_k} \frac{\sum_{i=1}^{n} d_\theta^\lambda(x_i, y_i)}{\sum_{i=1}^{n} \left(d_\theta^\lambda\right)_-^2 (x_i, y_i)} \geq n \left(a_n(k) - a_n(k^0) \right) \right)$$

Now, under **(A2)**

$$sup_{\theta\in\Theta_k}\frac{1}{n}\left(\sum_{i=1}^{n} d_\theta^\lambda(x_i,y_i)\right)^2 = O_P(1)$$

where, $O_p(1)$ means bounded in probability. Moreover, under **(A2)** the set $\left\{\left(d_\theta^\lambda(x_i,y_i)\right)^2\right\}$ is Glivenko-Cantelli (the set admits an uniform law of large numbers). Hence

$$\inf_{\theta\in\Theta_k}\frac{1}{n}\sum_{i=1}^{n}\left(d_\theta^\lambda(x_i,y_i)\right)_-^2 \overset{n\to\infty}{\longrightarrow} \inf_{\theta\in\Theta_k}\parallel\left(d_\theta^\lambda(X,Y)\right)_-\parallel_2^2$$

But $\inf_{\theta\in\Theta_k}\parallel\left(d_\theta^\lambda(X,Y)\right)_-\parallel_2 > 0$, since the random variable $d_\theta^\lambda(X,Y)$ is centered and $\|d_\theta^\lambda(X,Y)\|_2 = 1$. Then, we get :

$$\frac{1}{\lambda}\sup_{\theta\in\Theta_k}\frac{\sum_{i=1}^{n} d_\theta^\lambda(x_i,y_i)}{\sum_{i=1}^{n}\left(d_\theta^\lambda\right)_-^2(x_i,y_i)} = O_P(1)$$

and $P(\hat{k} > k^0)$ tends to 0 as n tends to infinity. Finally,

$$P(\hat{k} < k^0) \leq \sum_{k=1}^{k^0-1} P\left(\sup_{\theta\in\Theta_k}\frac{E_n(\theta)-E_n(\theta^0)}{n} \geq \frac{a_n(k)-a_n(k^0)}{n}\right)$$

and $\sup_{\theta\in\Theta_k}\frac{E_n(\theta)-E_n(\theta^0)}{n}$ converges in probability to

$$\sup_{\theta\in\Theta_k} E\left(E_n(\theta)-E_n(\theta^0)\right) < 0$$

since $k < k^0$, so $\hat{k} \overset{P}{\longrightarrow} k^0$ ■

The assumption **(A1)** is fairly standard for model selection, in the Gaussian case **(A1)** will be fulfilled by the BIC criterion. The assumption **(A2)** is more difficult to check. First we note:

$$\left(e^{-\lambda\left((Y-f_\theta(X))^2-(Y-f_{\theta^0}(X))^2\right)}-1\right)^2 =$$
$$e^{-2\lambda\left((Y-f_\theta(X))^2-(Y-f_{\theta^0}(X))^2\right)} - 2e^{-\lambda\left((Y-f_\theta(X))^2-(Y-f_{\theta^0}(X))^2\right)} + 1$$

So, d_θ^λ is well defined if $E\left[e^{-2\lambda\left((Y-f_\theta(X))^2-(Y-f_{\theta^0}(X))^2\right)}\right] < \infty$, but

$$\begin{array}{l}(Y-f_\theta(X))^2-(Y-f_{\theta^0}(X))^2= \\ (Y-f_{\theta^0}(X)+f_{\theta^0}(X)-f_\theta(X))^2-(Y-f_{\theta^0}(X))^2= \\ 2\varepsilon(f_{\theta^0}(X)-f_\theta(X))+(f_{\theta^0}(X)-f_\theta(X))^2\end{array}$$

where $\varepsilon = Y - f_{\theta^0}(X)$ is the noise of the model. Since an MLP function is bounded, d_θ^λ is well defined if $\lambda > 0$ exists such that $e^{\lambda|\varepsilon|} < \infty$ i.e. ε admits exponential moments. Finally, using the same techniques of reparameterization as in Rynkiewicz [6], assumption **(A2)** can be shown to be true for MLP models with sigmoïdal transfer functions, if the set of possible parameters Θ is compact.

5 A little experiment

The theoretical penalization terms of the previous section can be chosen among a wide range of functions (see condition **A1**). In the sequel, a little experiment is conducted to assess the right rate of penalization to guess the "true" architecture of a model.

Consider a simulated model:

$$Z_t = F_{\theta^0}(X_t, Y_t) + \varepsilon_t, t = 1, \cdots, n,$$

with $((X_1, Y_1), \cdots, (X_n, Y_n))$ i.i.d., $(X_t, Y_t) \sim \mathcal{N}(0_{\mathbb{R}^2}, 3 \cdot I_2)$, $(\varepsilon_1, \cdots, \varepsilon_n)$ i.i.d., $\varepsilon_t \sim \mathcal{U}[-1, 1]$, the uniform law in $[-1; 1]$ and

$$\begin{array}{l}F_{\theta^0}(x,y) = \tanh(6\cdot x - 2\cdot y) + 2\cdot\tanh(8 - x + 3\cdot y) \\ -3\cdot tanh(2 - 6\cdot x - 2\cdot y) + 1.5.\end{array} \tag{8}$$

Here, the true model is an MLP with 2 inputs, 3 hidden units and one output. In order to avoid too long time of computation, the number of hidden units is assumed to be between 1 and 10. We estimate the true architecture of the MLP according to (7).

First, let us write the log-likelihood of the data as if the density of the noise would be Gaussian :

$$\begin{array}{l}l_\theta\left(\begin{pmatrix}x_1\\y_1\\z_1\end{pmatrix},\cdots,\begin{pmatrix}x_n\\y_n\\z_n\end{pmatrix}\right) = -\frac{n}{2}\cdot\log(2\pi\sigma^2) \\ -\sum_{t=1}^n \frac{1}{2\sigma^2}\left(z_t - F_\theta(x_t,y_t)\right)^2 + \sum_{t=1}^n g(x_t,y_t)\end{array} \tag{9}$$

Here, we assume that the variance of the noise σ^2 is a known constant and that the density of the explicative variables (X, Y) is a function $g(.,.)$ independent of the parameter vector θ. Two classical criteria are:

$$\begin{array}{l} AIC \ : \ \sum_{t=1}^{n} \frac{1}{\sigma^2} \left(z_t - F_\theta(x_t, y_t)\right)^2 + 2 \cdot D + Cte \\ BIC \ : \ \sum_{t=1}^{n} \frac{1}{\sigma^2} \left(z_t - F_\theta(x_t, y_t)\right)^2 + D \cdot \log(n) + Cte \end{array} \tag{10}$$

where D is the size of the parameter vector (the dimension of the model or the number of weights of the MLP) and Cte a constant independent of the parameter θ.

The optimization of the log-likelihood is done with respect to the parameter θ, so maximimizing this quantity is equivalent to minimizing :

$$E_n(\theta) = \frac{1}{n} \cdot \sum_{t=1}^{n} \left(z_t - F_\theta(x_t, y_t)\right)^2 \tag{11}$$

Hence, an AIC like minimum-penalized MSE criterion would be:

$$\frac{1}{n} \sum_{t=1}^{n} \left(z_t - F_\theta(x_t, y_t)\right)^2 + \frac{2\sigma^2 D}{n}$$

and, a BIC like minimum-penalized MSE criterion would be:

$$\frac{1}{n} \sum_{t=1}^{n} \left(z_t - F_\theta(x_t, y_t)\right)^2 + \frac{\sigma^2 D \log(n)}{n}$$

Note, that these criteria involve the knowledge of the variance of the noise σ^2. In a first experiment we will use the true variance of the noise, then this problem will be addressed in the following sections.

In the following sections, the optimization of MLP is done with the BroydenFletcherGoldfarbShanno (BFGS) method. In order to avoid bad local minima, 10 random initializations of the weights are done for each estimation.

5.1 Model selection with σ^2 known

We will compare 4 criteria, from the least penalized (AIC like) to the most penalized (Very Strong Penalization), the following penalized criteria are assessed:

- AIC like: $\frac{1}{n}\sum_{t=1}^{n}\left(z_t - F_\theta(x_t, y_t)\right)^2 + \frac{2\sigma^2 D}{n}$
- BIC like: $\frac{1}{n}\sum_{t=1}^{n}\left(z_t - F_\theta(x_t, y_t)\right)^2 + \frac{\sigma^2 D \log(n)}{n}$
- SP (Strong Penalization): $\frac{1}{n}\sum_{t=1}^{n}\left(z_t - F_\theta(x_t, y_t)\right)^2 + \frac{\sigma^2 D\sqrt{n}}{n}$
- VSP (Very Strong Penalization): $\frac{1}{n}\sum_{t=1}^{n}\left(z_t - F_\theta(x_t, y_t)\right)^2 + \frac{\sigma^2 D n^{3/4}}{n}$

We simulate $n = 100$, $n = 500$ and $n = 1000$ data according to the true model (8), for each n the experiment is repeated 100 times.

The following architectures are chosen by the penalized criteria :

- n=100

	nb h. units	1	2	3	4	5	6	7	8	9	10
AIC like	models sel.	0	0	**47**	**35**	**10**	**2**	**3**	**3**	0	0
BIC like	models sel.	0	0	**100**	0	0	0	0	0	0	0
SP	models sel.	0	0	**100**	0	0	0	0	0	0	0
VSP	models sel.	0	**74**	**26**	0	0	0	0	0	0	0

- n=500

	nb h. units	1	2	3	4	5	6	7	8	9	10
AIC like	models sel.	0	0	**3**	**14**	**17**	**14**	**13**	**15**	**10**	**13**
BIC like	models sel.	0	0	**100**	0	0	0	0	0	0	0
SP	models sel.	0	0	**100**	0	0	0	0	0	0	0
VSP	models sel.	0	**1**	**99**	0	0	0	0	0	0	0

- n=1000

	nb h. units	1	2	3	4	5	6	7	8	9	10
AIC like	models sel.	0	0	0	0	0	0	0	**7**	**24**	**69**
BIC like	models sel.	0	0	**100**	0	0	0	0	0	0	0
SP	models sel.	0	0	**100**	0	0	0	0	0	0	0
VSP	models sel.	0	0	**100**	0	0	0	0	0	0	0

The BIC like criterion and the Strong Penalization chose always the true architecture whatever the number of data. According to the theory, AIC like criterion is not consistent (see condition **A1**) and the chosen architecture is always too large. The Very Strong penalization chose a too small architecture

when the number of data is small ($n = 100$), however it is a consistent criterion, so its behavior is correct for larger number of data ($n = 500$ and $n = 1000$). This good results assume that the true variance of the noise is known, but for regression models this is never the case. A first idea may be to replace the unknown variance by the estimated one, the is done in the next section.

5.2 Model selection using estimated variance $\hat{\sigma}^2$.

The estimated variance $\hat{\sigma}^2$ is the mean square error of the model:

$$\hat{\sigma}^2 := E_n(\hat{\theta}) = \frac{1}{n}\sum_{t=1}^{n}(z_t - F_{\hat{\theta}}(x_t, y_t))^2$$

computed for the least square estimator $\hat{\theta}$. Hence, the comparison is done with the penalized criteria :

- AIC like: $\frac{1}{n}\sum_{t=1}^{n}(z_t - F_\theta(x_t, y_t))^2 + \frac{2\hat{\sigma}^2 D}{n}$
- BIC like: $\frac{1}{n}\sum_{t=1}^{n}(z_t - F_\theta(x_t, y_t))^2 + \frac{\hat{\sigma}^2 D \log(n)}{n}$
- SP (Strong Penalization): $\frac{1}{n}\sum_{t=1}^{n}(z_t - F_\theta(x_t, y_t))^2 + \frac{\hat{\sigma}^2 D\sqrt{n}}{n}$
- VSP (Very Strong Penalization): $\frac{1}{n}\sum_{t=1}^{n}(z_t - F_\theta(x_t, y_t))^2 + \frac{\hat{\sigma}^2 D n^{3/4}}{n}$

The following architectures are chosen by the penalized criteria:

- n=100

	nb h. units	1	2	3	4	5	6	7	8	9	10
AIC like	models sel.	0	0	0	0	0	**1**	0	**4**	**27**	**68**
BIC like	models sel.	0	0	**22**	**6**	**3**	**1**	**3**	**5**	**15**	**45**
SP	models sel.	0	0	**85**	**3**	**1**	0	0	0	**1**	**10**
VSP	models sel.	0	0	**100**	0	0	0	0	0	0	0

- n=500

	nb h. units	1	2	3	4	5	6	7	8	9	10
AIC like	models sel.	0	0	0	**4**	**4**	**6**	**9**	**23**	**19**	**35**
BIC like	models sel.	0	0	**100**	0	0	0	0	0	0	0
SP	models sel.	0	0	**100**	0	0	0	0	0	0	0
VSP	models sel.	0	0	**100**	0	0	0	0	0	0	0

- n=1000

	nb h. units	1	2	3	4	5	6	7	8	9	10
AIC like	models sel.	0	0	0	0	**3**	**7**	**12**	**16**	**29**	**33**
BIC like	models sel.	0	0	**100**	0	0	0	0	0	0	0
SP	models sel.	0	0	**100**	0	0	0	0	0	0	0
VSP	models sel.	0	0	**100**	0	0	0	0	0	0	0

As usual the AIC like criterion misbehaves like in the previous section. But, for a small number of data ($n = 100$) the use of an estimation of the variance of the noise instead of the true one leads to overestimation of the number of hidden unit for the BIC like criterion and the strong penalization. The explanation is that the variance of the noise is underestimated for large number of hidden units and so the penalized criterion. This drawback disappears for larger number of data ($n = 500$ and $n = 1000$) because the estimation of the variance becomes better. Despite that the Very Strong Penalized criterion seems to guess the true architecture whatever the number of data, to plug estimated variance instead of the true one seems maybe to be a too naive approach for small number of data. As the goal of penalized criterion is to compare models, we could use the logarithm of the mean square error instead of the mean square error itself, hence the lack of the true variance is no more a problem because this number is simplified in the difference of the logarithms. This approach is studied in the next section.

5.3 Model selection using logarithm of mean square error

Choosing between two number of hidden units, k_1 and k_2 is the result of a comparison between $\min_{\theta_1 \in \Theta_{k_1}} (E_n(\theta_1) + a_n(k_1))$ on one hand and $\min_{\theta_2 \in \Theta_{k_2}} (E_n(\theta) + a_n(k_2))$ on the other hand, where $a_n(k)$ is the penalization term. So, the results of the comparison may be changed if we consider $C \cdot E_n(\theta)$ instead of $E_n(\theta)$ when C is a very big (or very small) constant. But, if we compare $\min_{\theta_1 \in \Theta_{k_1}} (\log(E_n(\theta_1)) + a_n(k_1))$ and
$\min_{\theta_2 \in \Theta_{k_2}} (\log(E_n(\theta_2)) + a_n(k_2))$ the results of the comparison is the same if we change $E_n(\theta)$ in $C \cdot E_n(\theta)$.

Moreover, $E_n(\theta)$ may be seen as an approximation of the variance of the noise σ^2, so $E_n(\theta) = \sigma^2 \cdot (1 - \varepsilon(\theta))$ and

$$\log(E_n(\theta)) \simeq \log(\sigma^2) - \varepsilon(\theta) + o(\varepsilon(\theta))$$

The term $\varepsilon(\theta)$ may be seen as the normalized term of overfitting of the model. Finally we get:

$$\min_{\theta\in\Theta}\left(\log\left(E_n(\theta)\right)+a_n(k_1)\right)-\min_{\theta\in\Theta}\left(\log\left(E_n(\theta)\right)+a_n(k_2)\right)\simeq \varepsilon(\theta_2)-\varepsilon(\theta_1)+a_n(k_2)-a_n(k_1)$$

and the penalization term plays fully is role of compensation of the "normalized" overfitting.

The results on our little experiment with these new criteria are the following:

- n=100

	nb h. units	1	2	3	4	5	6	7	8	9	10
AIC like	models sel.	0	0	**3**	**7**	**4**	**11**	**9**	**13**	**24**	**29**
BIC like	models sel.	0	0	**96**	**4**	0	0	0	0	0	0
SP	models sel.	0	0	**100**	0	0	0	0	0	0	0
VSP	models sel.	**73**	**27**	0	0	0	0	0	0	0	0

- n=500

	nb h. units	1	2	3	4	5	6	7	8	9	10
AIC like	models sel.	0	0	**3**	**6**	**8**	**13**	**18**	**10**	**17**	**25**
BIC like	models sel.	0	0	**99**	**1**	0	0	0	0	0	0
SP	models sel.	0	0	**100**	0	0	0	0	0	0	0
VSP	models sel.	0	**69**	**31**	0	0	0	0	0	0	0

- n=1000

	nb h. units	1	2	3	4	5	6	7	8	9	10
AIC like	models sel.	0	0	**1**	**2**	**7**	**13**	**16**	**21**	**19**	**21**
BIC like	models sel.	0	0	**100**	0	0	0	0	0	0	0
SP	models sel.	0	0	**100**	0	0	0	0	0	0	0
VSP	models sel.	0	**2**	**98**	0	0	0	0	0	0	0

We can see that this method yields very good results, whatever the number of data, for BIC like criterion and the Strong Penalization without knowing the true variance of the noise. Strong Penalization seems even to be a little better than BIC like criterion.

6 Conclusion

MLP regression is widely used and always a very competitive method (see Osowski et al. [5]), but theoretical justification is lacking for determining the true architecture and especially the number of hidden units. Indeed, the classical asymptotic theory fails when the model is not identifiable. In this paper, we prove an inequality showing that overfitting of MLP is moderate if the noise admits exponential moments and the parameters of the model are a priori bounded. This bound justifies the use of penalized criteria in order to fit the architecture of MLP models in the framework of regression without knowing the density of the noise. Hence, The user can select the true number of hidden units thanks to penalized criteria, of the form

$$\begin{array}{l} E_n(\theta) + a_n(k) \\ \text{or} \\ \ln(E_n(\theta)) + a_n(k) \end{array}$$

If the penalization term $a_n(k)$ is well calibrated ($\frac{C \cdot k}{n} < a_n(k) < C \cdot k$), the true number of hidden units will be automatically selected if n is large enough. A little experiment suggests that a good choice of penalization seems to be the middle of the possible range: $a_n(k) = \frac{C \cdot k}{\sqrt{n}}$. The use of the logarithm of the mean square error $E_n(\theta)$ is an easy way to avoid to know the true variance of the noise. A further question could be to know if this empirical finding for the tuning of the penalization term can be justified theoretically.

Note that, this paper was only concerned with the identification of the true model. The point is more to have an idea of the complexity of the model determining the data than to have a predictive model. However, if there are enough data, the true model will also be the best predictor. Hence, the spirit of the studied criterion is then very different from the approach used in Extreme Learning Machine (ELM) which provides a good predictive model at extremely fast learning speed. Indeed, in ELM, all the hidden node parameters are independent from the target functions or the training datasets so the weights between input and hidden node are no more considered as parameters. Such method is really not concerned by model identification but only by predictive power and speed of computing, for example it is possible to use ELM even if the relation between inputs and ouput is linear. Finally, If ELM seems to work well in pratice, the theoritical justification of the superiority of such method is still lacking.

Another issue may be when the unknow regression function is not represented by an MLP at all, in such case there is no more over-determination and the asymptotic behaviour of the model is more underfitting than overfitting. The overfitting occures only when we consider finite number of data and the theory to deal with such problem is the very difficult non-asymptotic statistical theory as in Massart [4]. As far as we know this theory gives no hints for choosing the number of hidden units for finite number of data.

Appendix

A Donsker class is a notion from the "empirical processes theory". This theory deals with "law of large number" and "asymptotic normality" for set of functions. Basically, a Donsker class is a set of functions which is "not too big".

Let $X_1, \cdots, X_n$ be a random sample from a probability distribution P. The empirical distribution is the discrete uniform measure on the observations. We denote it by $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$, where δ_x is the probability distribution that is degenerate at x. Given a function f, we write $\mathbb{P}_n f$ for the expectation of f under the empirical measure and Pf for the expectation under P. Thus

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \text{ and } Pf = \int f dP.$$

The empirical process evaluated at f is defined as $\mathbb{G}_n f = \sqrt{n}\,(\mathbb{P}_n f - Pf)$. Consider a set of functions $\mathcal{F}$ endowed with the L_2 norm $\|\cdot\|$ (see (4)). For every $\varepsilon > 0$, we define an ε-bracket by
$[l,\, u] = \{f \in \mathcal{F},\, l \leq f \leq u\}$ such that $\|u - l\| < \varepsilon$. The ε-bracketing entropy is

$$\mathcal{H}_{[\cdot]}\left(\varepsilon, \mathcal{F}, \|\cdot\|\right) = \ln\left(\mathcal{N}_{[\cdot]}\left(\varepsilon, \mathcal{F}, \|\cdot\|\right)\right),$$

where $\mathcal{N}_{[\cdot]}\left(\varepsilon, \mathcal{F}, \|\cdot\|\right)$ is the minimum number of ε-brackets necessary to cover $\mathcal{F}$. $\mathcal{N}_{[\cdot]}\left(\varepsilon, \mathcal{F}, \|\cdot\|\right)$ is also called "covering number".

The class $\mathcal{F}$ of functions is called Donsker if the covering number, which depends of the diameter ε of the balls, is smaller than order $e^{\frac{1}{\varepsilon^2}}$ when ε goes to 0. Then, the sequence of processes $\{\mathbb{G}_n f, f \in \mathcal{F}\}$ converges in distribution to a tight limit process. Finally, if $\mathcal{F} = \left\{d_\theta^\lambda, \theta \in \Theta\right\}$ is a Donsker class, the

sequence of processes

$$\left(\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} d_\theta^\lambda(X_i, Y_i)\right)\right)_{n\in\mathbb{N}}$$

converges in distribution to a tight random Gaussien process. We get then the key property needed for the demonstration of the theorem:

$$sup_{\theta\in\Theta_k}\frac{1}{n}\left(\sum_{i=1}^{n} d_\theta^\lambda(x_i, y_i)\right)^2 = O_P(1)$$

References

[1] S. Amari, H. Park and T. Ozeki, Singularities affect dynamics of learning in Neuromanifolds, *Neural computation*, 18, 1007–1065, MIT Press, 2006.

[2] K. Fukumizu, Likelihood ratio of unidentifiable models and multilayer neural networks *The Annals of Statistics*,31, 833–851, IMS, 2003.

[3] K. Hagiwara and K. Fukumizu, Relation between weight size and degree of over-fitting in neural network regression. *Neural networks*, 21, 48–58, Elsevier, 2008.

[4] P. Massart, *Concentration inequalities and model selection*, Springer, Berlin, 2007.

[5] S. Osowski, K. Siwek and T. Markiewicz, MLP and SVM networks - a Comparative study. *proceedings of the* 6th *Nordic Signal Processing Symposium* (NSPS 2004), pages 37-40, June 9-11, Espoo (Finland), 2004.

[6] J. Rynkiewicz, Consistent estimation of the architecture of multilayer perceptrons. In M. Verleysen, editor, *proceedings of the* 14th *European Symposium on Artificial Neural Networks* (ESANN 2006), d-side pub., pages 149-154, April 28-30, Bruges (Belgium), 2006.

[7] A.W. van der Vaart, *Asymptotic statistics*, Cambridge university Press, Cambridge, 1998.

[8] H. White, *Artificial Neural Networks: Approximation and Learning Theory.*, Basil Blackwell, Oxford, 1992.

[9] J. Yao, On least square estimation for stable nonlinear AR processes. *The Annals of Institut of Mathematical Statistics*,52, 316–331, Springer, 2000.