# Asymptotic statistics for multilayer perceptron with ReLU hidden units

Rynkiewicz, J.[a]

[a]*SAMM, Universite Paris 1,*
*90 Rue de Tolbiac, 75013 Paris, France*

**Abstract**

In numerous tasks, deep networks are state of the art. However, they are still not well understood from a statistical point of view. In this article, we try to contribute to filling this gap, and we consider regression models involving deep multilayer perceptrons (MLP) with rectified linear (ReLU) functions for activation units. It is a difficult task to study the statistical properties of such models. The main reason is that in practice these models may be heavily overparameterized. For the sake of simplicity, we focus here on the sum of square errors (SSE) cost function which is the standard cost function for regression purposes. In this framework, we study the asymptotic behavior of the difference between the SSE of estimated models and the SSE of the theoretical best model. This behavior gives us information on the overfitting properties of such models. We use in this paper new methodology introduced to deal with models with a loss of identifiability, i.e. in the case that the true parameter cannot be identified uniquely. Hence, we don't have to assume that a unique parameter vector realizes the best regression function which seems to be a too strong assumption for heavily overparameterized models. Our results shed new light on the overfitting behavior of MLP models.

*Keywords:* regression models, loss of identifiability, deep neural networks, ReLU functions, Donsker class

## 1. Introduction

Deep-learning allows computational models that are composed of multiple processing layers to learn representations of data with various levels of abstraction (Lecun et al. [8]). The principle underlying these models is to

compute an objective function (cost function) that measures the error (or distance) between the actual output and the desired output. The machine then modifies its internal adjustable parameters (or weights) to reduce this error. Note that, even if the number of parameters is huge, the learning algorithms work very well for deep network, and poor local minima are rarely a problem with large networks. Regardless of the initial conditions, the system nearly always reaches solutions of very similar quality (see Lecun et al. [8]). After training, the users are interested in the performance of the system measured on a different set than the training set. This generalization ability is heavily related to the overtraining (or overfitting) of the machine on the training data. Neyshabur et al. [12] studied the overfitting from a non-asymptotic point of view, it means if the number of free parameters of the model is larger than the number of data. Our point of view is asymptotic; we consider that the number of data is larger than the number of free parameters of the network. Some simulations done in the last section illustrate the complementarity of the two points of view.

Many applications of deep learning use feedforward neural network architectures, which learn to map a fixed-size input to a fixed-size output. To go from one layer to the next, a set of units computes a weighted sum of their inputs from the previous layer and pass the result through a non-linear function. Currently, the most popular non-linear function is the rectified linear unit (ReLU), which is $f(z) = \max(z, 0)$ (see Lecun et al. [8]). Indeed Neural network with ReLU non-linearities have been highly successful for computer vision tasks and proved faster to train than standard sigmoid units (see Dahl et al. [4]). The ReLU have been used both as activation functions in standard neural nets and as units in restricted Boltzmann machines (see Nair and Hinton [11]), but they are more straightforward to incorporate into a standard feed-forward neural net. Even if these networks work very well in practice, very few theoretical results are available about such complex models even from an asymptotic point of view. We propose in this paper to contribute to filling this gap, and for the sake of simplicity, we will focus on quadratic cost function in the framework of regression. The present manuscript generalizes our previous results in Rynkiewicz [14] which only deal with single hidden layer networks with sigmoid transfer functions and Gaussian noise.

Let us assume that we observe a random sample of identically distributed independent variables $(X_1, Y_1), \cdots, (X_n, Y_n)$, from the distribution $P$ of a vector $(X, Y)$, with $Y$ a real random variable and $X$ a random vector in $\mathbb{R}^h$.

A regression model can be written as:

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

where $\theta^0$ is a parameter realizing the best theoretical regression function:

$$\theta^0 = \arg\min_{\theta\in\Theta} \|Y - f_\theta(X)\|_2. \tag{2}$$

Here,

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

is the $\mathcal{L}^2$-norm for an square integrable function $g$ and $\Theta$ the set of possible parameter vectors. According to the theory of probability, the true regression function is the conditional expectation of the target variable $Y$ with respect to the explicative variables $X$. In this paper, we will assume that the law of $X$ has a density strictly positive with respect to the Lebesgue measure (the law of $X$ does not degenerate) this implies that the true regression function will be unique. If the true regression function is in the set of possible models, it is then evident that it is unique and is equal to $f_{\theta^0}$. If the true regression function is not in the set of possible models, the nearest function of the true one in the set of possible functions, $f_{\theta^0}$ is the projection of the true function on this set (with respect to the $\mathcal{L}^2$-norm). If the set of possible functions is convex (i.e. if $f_1$ and $f_2$ are in the set then for $0 < a < 1$, $a \times f_1 + (1-a) \times f_2$ is in the set), then it is well known that this projection will be unique. Since the output layer of MLPs is linear is is very easy to guaranty that the set of possible models is convex. So, in both cases, the best regression function is unique and will be denoted $f_0$ in the following.

Note that, even if this cost function is designed for regression models, it may also be used for classification tasks (Bayes classifier) if the data to predict are in the set $\{0, 1\}$. Let us first recall some historical results on regression models with MLP.

*Known asymptotics for MLP regression models*

Historically, the first works on MLP focus on models with only one hidden layer because these networks have been shown to be universal approximators. Let $k$ be the number of units in the hidden layer and $\theta \in \mathbb{R}^D$ the parameter vector of the MLP model, where $D$ is the dimension of $\theta$. A natural estimator

of a function $f_0$ in model (1) is the MLP function parameterized by the least square estimator (LSE) $\hat{\theta}_n$ that minimizes the SSE:

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

If the smallest number of hidden units of an MLP function realizing the best theoretical regression function $f_0$ is $k^0 = k$ then, up to a finite set of permutations, the model will be identifiable, and White [17] has shown that the asymptotic behavior of the difference of SSE is the classic one:

$$\lim_{n\to\infty} \left( \sum_{t=1}^{n} (Y_t - f_0(X_t))^2 - \sum_{t=1}^{n} \left(Y_t - f_{\hat{\theta}_n}(X_t)\right)^2 \right) = \sigma^2 \chi^2(D). \tag{4}$$

The mean of the SSE of a model is an estimator of the expectation of the square error of this model (under the probability law of variables $(X_t, Y_t)$). The difference of SSE between the best and the estimated models is then an indicator of the artificial improvement we can get by using an overparameterized model. However, the asymptotic overfitting can be larger than in regular parametric models. Indeed, if $k$ is large enough, the assumption that the smallest number of hidden units of the best theoretical regression function is $k^0 = k$ seems very optimistic. For example, if the variable $X$ is not explanatory enough of $Y$, then the relationship between $X$ and $Y$ may be not so complicated, and the best regression function may be written as an MLP function with $k^0$ hidden units, with $k^0 < k$. In this case, an MLP with $k$ hidden units has redundant hidden units for realizing the best theoretical regression function $f_0$, and Fukumizu [5] has shown that the asymptotic overfitting may be much more substantial:

$$\lim_{n\to\infty} \left( \sum_{t=1}^{n} (Y_t - f_0(X_t))^2 - \sum_{t=1}^{n} \left(Y_t - f_{\hat{\theta}_n}(X_t)\right)^2 \right) = O(\ln(n)). \tag{5}$$

The main point in the result of Fukumizu [5] is the loss of identifiability for MLP with redundant hidden units so that the redundant hidden units can be dedicated to overfitting only. Suppose, for example, that for real inputs, we want to estimate an MLP with two hidden units:

$$f_\theta(x) = a_1 \phi(w_1 x + b_1) + a_2 \phi(w_2 x + b_2) + b_0,$$

with $x \in \mathbb{R}$, $\phi$ be a transfer function like $\phi(z) = \tanh(z)$, and
$\theta = (b_0, b_1, b_2, w_1, w_2, a_1, a_2)$ be the parameter vector of the MLP. Suppose also that the best function $f_0$ is given by an MLP with only one hidden unit, say $f_0 = a^0\phi(w^0 x)$. Then, any parameter vector $\theta$ in the set

$$\begin{array}{l} \{\theta \,|w_2 = w_1 = w^0, b_2 = b_1 = b_0 = 0, a_1 + a_2 = a^0\} \\ \cup \{\theta \,|w_1 = w^0, b_1 = b_0 = 0, a_1 = a^0, (w_2, b_2) \in \mathbb{R}^2, a_2 = 0\} \\ \cup \{\theta \,|w_2 = w^0, b_2 = b_0 = 0, a_2 = a^0, (w_1, b_1) \in \mathbb{R}^2, a_1 = 0\}, \end{array} \tag{6}$$

realizes the function $f_0$. The first set of the union is a consequence of the duplication of the hidden unit $\phi(w^0 x)$, the second and the third ones are consequences of the cancellation of the output weights. Hence, classical statistical theory for studying the asymptotic behavior of the SSE cannot be applied because it requires the identification of the parameters (up to some permutations and sign symmetries).

To get strong asymptotic overfitting, the parameters, of the estimated MLP, need to go to infinity. For example, Hagiwara et al. [6] investigate relations between overfitting and weights size in a simple neural networks regression problem with Gaussian noise. They show that the degree of overfitting is strongly related to the size of the input weights. That is the reason why regularization techniques like "weight decay" (see Ripley [15]), which penalizes the model by the size of the parameters, work so well and in practice. Following pioneering works of Dacunha-Castell and Gassiat [3] and Liu and Shao [9] on models with a loss of identifiability, Rynkiewicz [13] has shown that, if the set of possible parameters of the MLP regression model is bounded, the effect of redundant hidden units on overfitting is less significant. Hence, for an MLP with one hidden layer and sigmoidal transfer functions, a centered Gaussian process $\{W(d), d \in \mathcal{D}\}$ with continuous sample paths exists so that

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

This result shows that the degree of over-fitting is bounded in probability, but depends on the size of the asymptotic set of index functions $\mathcal{D}$. Our goal in this paper is to generalize the result of Rynkiewicz [13] to MLP functions with arbitrarily large number of hidden layers and ReLU activation functions. The paper is organized as follows: Firstly, we give general results for the behavior of the difference of the sum of square errors (SSE) of the estimated regression

model and the SSE of the theoretical best regression function. Then, we show that the conditions needed for such results are fulfilled by MLP with ReLU activation functions. An experimental study investigates the practical consequences of our theoretical results in the next section. Finally, we discuss the consequences of our findings on the overfitting behavior of MLP models.

## 2. Asymptotic behavior of the SSE for regression models

Firstly, we present some definitions.

- 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 $\mathcal{L}^2$ norm. In the following, this norm will be denoted $L^2(P)$.

- For a vector $x = (x_1, \cdots, x_l)$, let us write $|x| = \sqrt{x_1^2 + \cdots + x_l^2}$ for the Euclidean norm.
- A family of random sequences

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

  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

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

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

Let us introduce generalized derivative functions:

$$d_\theta(x) = \frac{f_\theta(x) - f_0(x)}{\|f_\theta(X) - f_0(X)\|_2},\ f_\theta \neq f_0. \tag{7}$$

These functions are the key tool for studying the asymptotic behavior of the difference of SSE between the estimated model and the best one. The generalized derivative functions are a convenient way to describe all possible paths

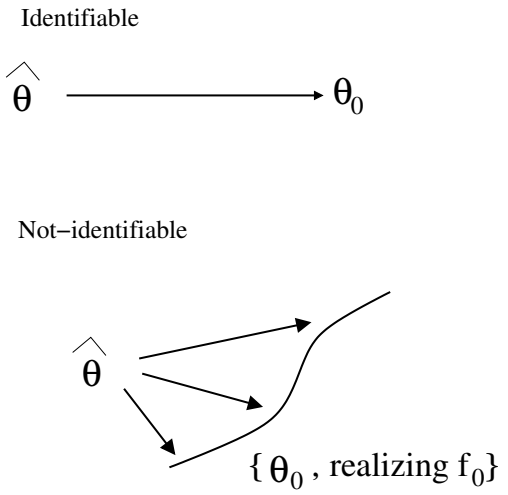


Figure 1: Set of possible paths for the estimated parameter vector $\hat{\theta}$.

for the estimated parameter vector $\hat{\theta}$ to converge toward the best one. The first lemma shows that the difference of the SSEs is bounded by a function of generalized derivative functions which will converge toward all the possible paths (see figure 1) and is proven in Rynkiewicz [13].

**Lemma 2.1.** *Let $(\varepsilon_t = Y_t - f_0(X_t))_{1\leq t\leq n}$ be the sequence of noise for the model (1), for all parameter $\theta \in \Theta$ with $f_\theta \neq f_0$ and $d_\theta$ defined in (7):*

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

The asymptotic behavior of the SSE will depend on the limit set of all the possible paths for $f_\theta$ to converge towards $f_0$, let us call this set $\mathcal{D}$. The limit set of derivatives $\mathcal{D}$ is the set of functions $d \in L^2(P)$ such that one can find a sequence $(\theta_n) \in \Theta$ satisfying $\|f_{\theta_n}(X) - f_0(X)\|_2 \xrightarrow[n\to\infty]{} 0$ and $\|d - d_{\theta_n}\|_2 \xrightarrow[n\to\infty]{} 0$. With such $(\theta_n)$, define, for all $t \in [0,1]$, $f_{\theta_t} = f_{\theta_n}$, where $n \leq \frac{1}{t} < n+1$. We thus have that, for any $d \in \mathcal{D}$, there exists a parametric path $(f_{\theta_t})_{0\leq t\leq \alpha}$ with $\alpha$ a strictly positive real number, such that for any $t \in [0,\alpha]$, $\theta_t \in \Theta$, $t \mapsto \|f_{\theta_t}(X) - f_0(X)\|_2$ is continuous, tends to 0 as $t$ tends to 0 and $\|d - d_{\theta_t}\|_2 \to 0$ as $t$ tends to 0. Using the reparameterization

$$\|f_{\theta_u}(X) - f_0(X)\|_2 = u, \tag{8}$$

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

$$\int (f_{\theta_u} - f_0 - ud)^2 \, dP = o(u^2). \tag{9}$$

Now, let us introduce some assumptions:

**B-1** Let $u$ be defined as (8), the map $u \mapsto P(Y - f_{\theta_u}(X))^2$ admits a second-order Taylor expansion with strictly positive second derivative $\frac{\partial^2 P(Y - f_{\theta_u}(X))^2}{\partial u^2}$ at $u = 0$.

**B-2** The set of generalized derivative functions $\mathcal{S} = \{d_\theta, \theta \in \Theta, f_\theta \neq f_0\}$ is a Donsker class (see van der Vaart [16] for the definition of Donsker class).

The assumption **B-1** means that the estimated parameter is going toward one of the best one along a smooth path. It is an assumption on the regularity of models which is generally fulfilled by parametric models (identifiable or not). Assumption **B-2** means that the set of possible directions for the convergence is not too big and we can apply the functional central limit theorem.

The following theorem, proven in Rynkiewicz [13], shows that with assumptions **B-1** and **B-2** the inequality of lemma 2.1 yields an approximation for the estimated model.

**Theorem 2.2.** *Under (***B-1***) and (***B-2***)*

$$\begin{aligned} &\sup_{\theta \in \Theta} \left( \sum_{t=1}^n (Y_t - f_0(X_t))^2 - (Y_t - f_\theta(X_t))^2 \right) = \\ &\sup_{d \in \mathcal{D}} \left( \max \left\{ \tfrac{1}{\sqrt{n}} \sum_{t=1}^n \varepsilon_t d(X_t); 0 \right\} \right)^2 + o_P(1). \end{aligned}$$

Now, define $(W(d))_{d \in \mathcal{D}}$ the centered Gaussian process with covariance the scalar product in $L^2(P)$, Gaussian processes can be seen as an infinite-dimensional generalization of multivariate normal distributions. It means that for a finite collection $(d_1, \cdots, d_n)$, $(W(d_1), \cdots, W(d_n))$ is a Gaussian vector with covariance function $Cov(W(d_i), W(d_j)) =< d_i, d_j >:= E(d_i d_j)$, so $< d_i, d_j >$ is a scalar product in $L^2(P)$. $d$ are functions belonging to the limit set of derivatives $\mathcal{D}$ where $\mathcal{D}$ describes all the possible paths for the convergence of the estimated parameter vector toward one of parameter vector realizing the best regression function. The following corollary is the asymptotic version of Theorem 2.2 when the number of observations goes toward infinity.

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

$$\sup_{\theta\in\Theta}\left(\sum_{t=1}^{n}\left(Y_t - f_0(X_t)\right)^2 - \sum_{t=1}^{n}\left(Y_t - f_\theta(X_t)\right)^2\right)$$

*converges in distribution to*

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

To get these results for MLP with ReLU activation functions, we have to check the assumptions (**B-1**) and (**B-2**) in this framework. Assumption (**B-1**) asserts that the model is, almost surely, regular along parametric paths. The difficult one is the assumption (**B-2**) on the Donsker property of the set of generalized derivative functions $\mathcal{S}$.

*Donsker property for* $\mathcal{S}$

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

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

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

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

then, according to the Theorem 19.5 of van der Vaart [16], the set $\mathcal{S}$ is Donsker. The intuition behind the Donsker property is that the number of $\eta$-brackets can go toward infinity, but not to fast. It can be seen as a measure of the infinite size a set of functions. For example, if the number of $\eta$-brackets necessary to cover $\mathcal{S}$, $\mathcal{N}_{[\cdot]}\left(\eta, \mathcal{S}, \|\cdot\|_2\right)$, is a polynomial function of $\frac{1}{\eta}$, then $\mathcal{S}$ will be Donsker (see van der Vaart [16]).

## 3. Application to MLP models

In the following, the structure of the networks refers to the way its units are arranged. It is specified by the dimension $d_0 = h$ of input $x$, the number

of layers $L$, and the number of units or width $d_l$ of each layer. Following the notations of Montufar et al. [10] an MLP is composed of layers which define functions $f_\theta : \mathbb{R}^h \to \mathbb{R}$ of the form

$$f_\theta(x) = f_{out} \circ \phi_L \circ f_L \circ \cdots \circ \phi_1 \circ f_1(x), \tag{10}$$

where $f_l$ is a linear preactivation function and $\phi_l$ is a nonlinear activation function. The parameter $\theta$ is composed of input weight matrices $\mathbf{W}_l \in \mathbb{R}^{d_l \times d_{l-1}}$ and bias vectors $\mathbf{b}_l \in \mathbb{R}^{d_l}$ for each layer $l \in \{1, \cdots, L\}$. The output of the $l$-th layer is a vector $x_l = (x_{l_1}, \cdots, x_{l,d_l})^T$ of activations $x_{l,i}$ of the units $i \in \{1, \cdots, d_l\}$ in that layer. This is computed from the activations of the preceding layer by $x_l = \phi_l(f_l(x_{l-1}))$. Given the activation $x_{l-1}$ of the units in the $(l-1)$-th layer, the preactivation of layer $l$ is given by:

$$f_l(x_{l-1}) = \mathbf{W}_l x_{l-1} + \mathbf{b}_l,$$

where $f_l = (f_{l,1}, \cdots, f_{l,d_l})^T$ is a vector of $\mathbb{R}^{d_l}$. The activation of the $i$-th unit in the $l$-th layer is given by

$$x_{l,i} = \phi_l(f_{l,i}(x_{l-1})).$$

From now, we assume that for all $l$, $\phi_l$ is the ReLU activation function. $f_{out}$ is just a linear function of the activation of the last layer $x_L$:

$$f_{out}(x_L) = \mathbf{W}_{out}^T x_{l-1} + \mathbf{b}_{out},$$

where $\mathbf{W}_{out} \in \mathbb{R}^{d_L}$ and $\mathbf{b}_{out} \in \mathbb{R}$.

We consider MLP with fixed structure. The parameter vector of an MLP is

$$\theta = (\mathbf{W}_1, \mathbf{b}_1, \cdots, \mathbf{W}_{out}, \mathbf{b}_{out}),$$

and we denote by $\Theta$, the set of possible parameters.

Let us introduce some assumptions for the regression models (1) with MLP functions:

**H-1**: Let $f_\theta, \theta \in \Theta$ be MLP functions with a fixed structure. We assume that $\Theta$ is a compact subset of $\mathbb{R}^D$ for some strictly positive integer $D$ and that the set $\{f_\theta, \theta \in \Theta\}$ is convex. Moreover, the set of parameters $\Theta_0$, realizing the best regression function, is assumed to be a subset of the interior of $\Theta$.

**H-2**: The explicative random vector $X$ admits a strictly positive density with respect to the Lebesgue measure of $\mathbb{R}^h$ and $P(|X|^2) < \infty$.

In this framework, we will show that the number of $\eta$-brackets necessary to cover the set of generalized derivative functions $\mathcal{S} = \{d_\theta, \theta \in \Theta, f_\theta \neq f_0\}$ is a polynomial function by considering a reparameterization of MLP functions. Namely, the previous parameterization of MLP models is not suitable to study the identifiability issues of these models, but it can be done by using the representation of these models in terms of piecewise linear continuous functions. Indeed a neural network with ReLU activation function can be seen as a linear spline in more flexible since the knots are placed using a training algorithm, and without the constraint of being equal to the functions it interpolates at certain points (see Hansson and Olsson [7]).

*Reparameterization*

Following the presentation of Montufar et al. [10], we remark that the ReLU functions have two types of behavior; they can be either constant 0 or linear, depending on their inputs. A hyperplane $\mathcal{H}_j = \left\{{\beta_j}^T x + \alpha_j = 0\right\}$ gives the boundary between these two behaviors, where $\beta_j = (\beta_{j1}, \cdots, \beta_{jh})^T \in \mathbb{R}^h$ and $\alpha_j \in \mathbb{R}$. The collection of all the hyperplanes coming from all units in an MLP forms a hyperplane arrangement. The hyperplanes in the arrangement split the input-space into several regions. Formally, a region $\mathcal{P}$ of a hyperplane arrangement $\{\mathcal{H}_1, \cdots, \mathcal{H}_n\}$ is the closure of a connected component of the complement $\mathbb{R}^h \backslash (\cup_j \mathcal{H}_j)$, i.e. a set of points delimited by these hyperplanes (possibly open towards infinity). Hence, for any region $\mathcal{P}_\mu \subset \mathbb{R}^h$ in a hyperplane arrangement, a set of parameter vectors $\mu = \{(\beta_1, \alpha_1), \cdots, (\beta_l, \alpha_l)\}$ exists such that $\mathcal{P}_\mu$ may be written as a finite intersection of half-spaces delimited by hyperplanes:

$$\mathcal{P}_\mu = \cap_{j \in 1, \cdots, l} \left\{x, {\beta_j}^T x + \alpha_j \geq 0\right\}. \tag{11}$$

Let us write $\mathbf{I}_\mathcal{P}$ the indicator function of the region $\mathcal{P}$, and denote by $N$ the total number of hidden units of the MLP. According to Montufar et al. [10], a integer $q \leq 2^N$ exists such that, for any $\theta \in \Theta$, $f_\theta$ can be written:

$$f_\theta(x) = \sum_{i=1}^{q} \left({\beta_i}^T x + \alpha_i\right) \times \mathbf{I}_{\mathcal{P}_{\mu(i)}}(x), \tag{12}$$

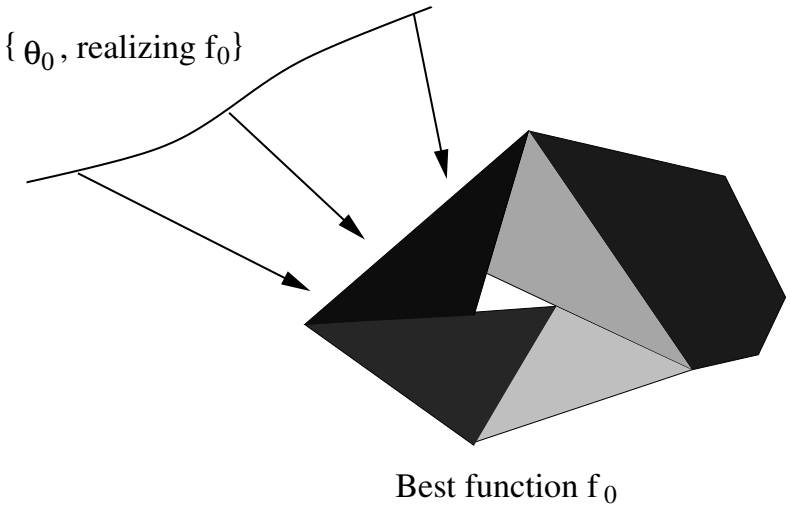


Figure 2: Set of parameters realizing the best function $f_0$.

with $(\beta_i, \alpha_i) \in \mathbb{R}^{h+1}$, $\mu(i)$ a set of parameters and $\left(\mathbf{I}_{\mathcal{P}_{\mu(1)}}(.), \cdots, \mathbf{I}_{\mathcal{P}_{\mu(q)}}(.)\right)$ are linearly independent indicator functions of regions of hyperplane arrangement. Note that, for some $\theta$, the integer $q$ is maybe too large, but we allow some parameter vectors $(\beta_i, \alpha_i) \in \mathbb{R}^{h+1}$ to be null vectors.

Now, for the best regression function $f_0$, $\left((\beta_1^0, \alpha_1^0), \cdots, (\beta_{q_0}^0, \alpha_{q_0}^0)\right)$ and a set of set of parameters $\{\mu_0(1), \cdots, \mu_0(q_0)\}$ exist such that $f_0$ may be written:

$$f_0(x) = \sum_{i=1}^{q_0} \left(\beta_i^{0T} x + \alpha_i^0\right) \times \mathbf{I}_{\mathcal{P}_{\mu_0(i)}}(x), \tag{13}$$

with $\left(\mathbf{I}_{\mathcal{P}_{\mu_0(1)}}(.), \cdots, \mathbf{I}_{\mathcal{P}_{\mu_0(q_0)}}(.)\right)$ linearly independent indicator functions of regions of hyperplane arrangement, and $q_0$ chosen as the smaller integer possible such that (13) realizes the best regression function. Note that a lot of sets of parameters $\mu_0(i)$ can define the region $\mathcal{P}_{\mu_0(i)}$, but since the regions $\left(\mathcal{P}_{\mu_0(1)}, \cdots, \mathcal{P}_{\mu_0(q_0)}\right)$ are fixed, we will denote them by $\left(\mathcal{P}_1^0, \cdots, \mathcal{P}_{q_0}^0\right)$. Now, the loss of identifiability (see figure 2) occurs in a way a little bit different than in the case of MLP with one hidden layer and sigmoid activation functions (see (6)). Namely, suppose, for example, that $q = 2$ , $q_0 = 1$, $x \in \mathbb{R}$, and that the best function $f_0$ is given by

$$f_0(x) = \left(\beta^0 x + \alpha^0\right) \times \mathbf{I}_{\mathcal{P}_0}(x).$$

Then, any parameter vector in the set:

$$\begin{array}{l}\{\mu_1, \mu_2, (\beta_1, \alpha_1), (\beta_2, \alpha_2)\} \text{ with} \\ \mathcal{P}_{\mu(1)} = \mathcal{P}_{\mu(2)} = \mathcal{P}_0, (\beta_2, \alpha_2) + (\beta_1, \alpha_1) = (\beta_0, \alpha_0), \text{ or} \\ \mathcal{P}_{\mu(1)} = \mathcal{P}_0, (\beta_1, \alpha_1) = (\beta_0, \alpha_0), (\beta_2, \alpha_2) = (0, 0), \text{ or} \\ \mathcal{P}_{\mu(2)} = \mathcal{P}_0, (\beta_2, \alpha_2) = (\beta_0, \alpha_0), (\beta_1, \alpha_1) = (0, 0), \text{ or} \\ \mathcal{P}_{\mu(1)} \cup \mathcal{P}_{\mu(2)} = \mathcal{P}_0, \mathcal{P}_{\mu(1)} \cap \mathcal{P}_{\mu(2)} \subset \{x, \beta_0^T x + \alpha_0 = 0\}, \\ (\beta_2, \alpha_2) = (\beta_1, \alpha_1) = (\beta_0, \alpha_0),\end{array} \tag{14}$$

realizes the function $f_0$. The first condition is a consequence of duplication of regions. The second and the third ones are a consequence of the cancellation of the weights. But, the fourth condition is now a consequence of the splitting of regions. Note that, if the MLP has enough redundant units, duplication and splitting of regions may be mixed (duplication of subsets of regions $(\mathcal{P}_{\mu_0(i)})_{1 \leq i \leq q_0}$ may be possible too). Using this representation of MLP functions we can show that the entropy with bracketing is a polynomial function of $\frac{1}{\eta}$. The method consists in finding a new parameterization of the MLP function $f_\theta$ which exhibits an identifiable parameter vector $\omega$ and putting other parameters in a parameter vector $\psi$. The following lemma is then the first step to show the assumption **B-2** in the particular case of MLP with ReLU transfer functions. It is proven in the appendix.

**Lemma 3.1.** *Under assumptions (***H-1***) and (***H-2***), a strictly positive integer $q$ and parametric functions $(g_{\psi_1}, \cdots, g_{\psi_q})$ exist so that we get the reparameterization: $\theta \mapsto (\omega, \psi) = (\omega_1, \cdots, \omega_q, \psi_1, \cdots, \psi_q)$, where $\omega_i \in \mathbb{R}^{h+1}$, and the difference of regression functions with the best one can be written:*

$$f_\theta(x) - f_0(x) := f_{(\omega,\psi)}(x) - f_0(x) = \sum_{i=1}^{q} (\omega_i - \omega_i^0)^T g_{\psi_i}(x). \tag{15}$$

*$\omega^0 = (\omega_1^0, \cdots, \omega_q^0)$ is a fixed parameter vector, and $f_{(\omega,\psi)}(x) = f_0(x)$ if and only if $\omega = \omega^0$.*

This lemma implies that the entropy with bracketing of $\mathcal{S}$ is polynomial. The following proposition which is a consequence of Lemma 3.1, shows then that assumption **B-2** is true for MLP with ReLU transfer functions. The proof of this proposition is postponed in the appendix.

**Proposition 1.** *A positive integer $Q$ and a constant $K$ exist so that the number of $\eta$-brackets $\mathcal{N}_{[.]}(\eta, \mathcal{S}, \|\cdot\|_2)$ covering $\mathcal{S}$ is smaller than $K\left(\frac{1}{\eta}^Q\right)$.*

The Donsker property for the set $\mathcal{S}$ of generalized derivative functions for MLP with ReLU transfer functions is then a consequence of Proposition 1.

Now, we deal with assumption (**B-1**). We remark that ReLU function has a distinguished (i.e., irregular) behavior at zero. So, functions $f_\theta = \sum_{i=1}^{q} \left({\beta_i}^T x + \alpha_i\right) \times \mathbf{I}_{\mathcal{P}_{\mu(i)}}(x)$ with $\mathcal{P}_\mu(i) = \cap_{j\in 1,\cdots,l} \left\{x, {\beta_j}^T x + \alpha_j \geq 0\right\}$, have a distinguished behavior at all inputs from any of hyperplanes $\mathcal{H}_j := \left\{x, {\beta_j}^T x + \alpha_j = 0\right\}$. But, under (**H-2**), the Lebesgue measure of the reunion $\mathcal{U}(\theta) = \cup H_j$ of all these hyperplanes is null. Now, since for all $\theta_u$ in $\Theta$:

$$\int_{\mathbb{R}^h} \left(f_{\theta_u} - f_0 - ud\right)^2 dP = \int_{\mathbb{R}^h \setminus \mathcal{U}(\theta_u)} \left(f_{\theta_u} - f_0 - ud\right)^2 dP, \tag{16}$$

the assumption (**B-1**) is true for MLP functions.

Thanks to equation (17) we can easily identify the asymptotic set of score function $\mathcal{D}$, and we get then the following theorem which is the transcription of Corollary 1 for the MLP case.

**Theorem 3.2.** *Let the map $\Omega : \mathcal{L}^2(Q) \to \mathcal{L}^2(Q)$ be defined as $\Omega(f) = \frac{f}{\|f\|_2}$. Under the assumptions* **H-1** *and* **H-2**, *a centered Gaussian process $\{W(d), d \in \mathcal{D}\}$ with continuous sample paths and a covariance kernel $P\left(W(d_1)W(d_2)\right) = P\left(d_1 d_2\right)$ exists so that*

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

*The index set $\mathcal{D}$ is defined as $\mathcal{D} = \cup_t \mathcal{D}_t$, where three integers $q_0 \leq q_1 \leq q$ exist so that the union runs over any possible vector of integers $t = (0, t_1, q) \in \mathbb{N}^3$ with $q_0 \leq t_1 \leq q_1$ and*

$$\mathcal{D}_t = \left\{\Omega\left(\sum_{i=0}^{t_1} \mathbb{I}_{\mathcal{P}_i^0}(X)(\zeta_i^T X + \alpha_i) + \sum_{i=t_1+1}^{q} \mathbb{I}_{\mathcal{P}_{\mu(i)}}(X)(\zeta_i^T X + \alpha_i)\right),\right.$$
$$\left.\zeta_1, \cdots, \zeta_q \in \mathbb{R}^h, \alpha_1, \cdots, \alpha_q \in \mathbb{R}\right\}.$$

*$\mathcal{P}_1^0, \cdots, \mathcal{P}_{q_1}^0$ are fixed regions of $\mathbb{R}^h$ and $\mathcal{P}_{\mu(q_0+1)}, \cdots, \mathcal{P}_{\mu(q)}$ are hyperplane arrangements of $\mathbb{R}^h$ delimited by finite sets of hyperplanes.*

This theorem shows that the degree of overfitting is bounded in probability, but depends on the size of the asymptotic set $\mathcal{D}$. Intuitively the set $\mathcal{D}$ is the degree of freedom of the estimated model when it is very near to the best model. The set $\mathcal{D}$ depends on the following hyperparameters: The number

of hidden neurons, the number of layer and the size of the parameters. The bigger the hyperparameters, the bigger the set $\mathcal{D}$. Hence, to limit the size of $\mathcal{D}$, we can reduce any of these tree hyperparameters, but they have to be large enough so that the best regression function $f_0$ belongs to the set of possible parameters.

## 4. An empirical investigation

In this section, we assess the effect of overparameterization on the training set and the influence of the form of the best regression function on it. Since, in practice, the data are often high dimensional, we chose to simulate inputs of size 2000. Let us write $0_{\mathbb{R}^{2000}}$ the null vector of $\mathbb{R}^{2000}$ and $I_{\mathbb{R}^{2000}}$ the identity matrix of $\mathbb{R}^{2000} \times \mathbb{R}^{2000}$. We trained fully connected feedforward networks with two hidden layers (three layer networks) and ReLU transfer functions on two sets of data:

1. For the first set, the input $X_t$ is a Gaussian random vector of size 2000, with each component centered, normalized and independent from each other: $X_t \sim \mathcal{N}(0_{\mathbb{R}^{2000}}, I_{\mathbb{R}^{2000}})$. The output $Y_t$ is a centered, normalized, Gaussian variable: $Y_t \sim \mathcal{N}(0,1)$, independent of $X_t$. We simulate a sample of independent vectors $(X_t, Y_t)$ of length 100000: $(X_t, Y_t)_{1 \leq t \leq 100000}$. In this case, the best regression function is $f_0(x) = 0$.
2. For the second set, we first create an MLP function $M_0$ with 2000 as input size, two hidden layers of size $2^3 = 8$ neurons each and randomly chosen weights between $-0.3$ and 0.3. The input $X_t$ is a Gaussian random vector of size 2000, with each component centered, normalized and independent from each other: $X_t \sim \mathcal{N}(0_{\mathbb{R}^{2000}}, I_{\mathbb{R}^{2000}})$. The output $Y_t$ is the sum of the MLP function $M_0(X_t)$ of the input, and a centered, normalized, Gaussian noise $\varepsilon_t \sim \mathcal{N}(0,1)$, independent of $X_t$: $Y_t = M_0(X_t) + \varepsilon_t$. We simulate a sample of independent vectors $(X_t, Y_t)$ of length 100000: $(X_t, Y_t)_{1 \leq t \leq 100000}$. In this case, the best regression function is the MLP function: $f_0(x) = M_0(x)$.

Moreover, for each of the two set of data, we simulate 100000 supplementary data for using it to assess the MSE of the models on a test set.

*Models trained.* On both sets, we trained 5 architectures with 2 hidden layers of the same sizes from $2^3$ to $2^7$, each time increasing the number of hidden units of each layer by factor 2. For the small architectures, $2^3$ or $2^4$ hidden

units, the amount of data is greater than the number of parameters of the MLP; it is the asymptotic framework where our results apply. For the great architectures, $2^6$ or $2^7$ hidden units, the number of parameters of the MLP is greater than the amount of data; it is the not-asymptotic framework, where our results do not apply, but we did it for comparing with the experiments of Neyshabur et al. [12]. Note that, since we generate the second data set with the best function using the smallest architecture, all the MLP of the second experiment can realize the best regression function.

For each experiment, we trained the network using Stochastic Gradient Descent (SGD) with mini-batch size 64, momentum 0.9 and fixed step size 0.01. We did not use any technic of regularization. We stopped the training when the number of epochs reached 1000. All the computations are done with Torch7 using a GPU.

*Evaluations.* For the trained architectures we give the number of hidden units of each hidden layer, the corresponding number of parameters, the MSE on the training data set and the MSE on the test data set. Note that, without overfitting, the MSE should be 1, hence the nearest the MSE of 1 the lesser the overfitting. We summarize the results for the two data sets on table 1. As expected by our results the training error depends not only

Table 1: Comparison of overtraining in function of architectures and data sets

| Nb of hidden units | Nb of parameters | MSE | Data set $f_0(x) = 0$ | Data set $f_0(x) = M_0(x)$ |
|---|---|---|---|---|
| $2^3$ | 16089 | training | 0.72 | 0.81 |
| | | test | 1.31 | 1.47 |
| $2^4$ | 32305 | training | 0.47 | 0.54 |
| | | test | 1.67 | 1.96 |
| $2^5$ | 65121 | training | 0.15 | 0.17 |
| | | test | 2.62 | 2.85 |
| $2^6$ | 132289 | training | 0.06 | 0.07 |
| | | test | 2.16 | 3.11 |
| $2^7$ | 272769 | training | 0.02 | 0.06 |
| | | test | 1.61 | 1.77 |

of the architecture of the MLP but also of the best regression function of the data set. Indeed, for all models, for the same number of parameters and data, the training error is closer to 1 when the best regression function is

more complicated. However, we can see that the relationship between the test error and the training error also depends on this best function and even if the learning error is closer to 1 for the second data set, the test error is more significant in this case. This fact, unexpected, remains to explain. Finally, we can see that when the number of parameters becomes greater than the amount of data, the test error seems to decrease when the amount of parameters increases, as in Neyshabur et al. [12]. However, the test error of the biggest model doesn't reach the test error of the smallest and less overparameterized model. Surprisingly, the behavior of the asymptotic case seems the inverse of the behavior of the not-asymptotic case.

## 5. Conclusion

For one hidden layer MLP, there are no differences from a statistical point of view between ReLU and sigmoidal transfer functions. However, few are known for theoretical properties of MLP with two hidden layers and more (Deep MLP). This paper gives an insight with such deep architecture when the transfer functions are ReLU functions.

In numerous statistical models, like identifiable parametric models, the overfitting depends solely on the complexity of the model in use. However, for MLP functions this is no more the case since it depends on the size of set $\mathcal{D}$ which is a function of the difference between the complexity of the MLP function in use and the complexity of the best regression function $f_0$ (the number of redundant hidden units). This fact explains the apparent contradiction noticed by some authors (cf Zhang et al. [18]), where an MLP does not overfit too much for a complex task but overfits a lot if you randomize the output data. Moreover, in the experiments, we have seen that the relationship between the overtraining on the learning set and the error on the test set is not obvious and seems also depends on the best regression function $f_0$. Finally, the behavior of the overfitting also relies on the comparison of the amount of data and number of parameter of the model and seems different in the asymptotic or not-asymptotic framework. It will be interesting to understand these surprising facts which we leave for future work.

## Appendix

## 6. Proof of lemma 3.1

If $f_\theta = f_0$, it exists (up to permutations):

- A vector of integers $t = (t_i)_{1 \leq i \leq q_0+1}$, so that $0 = t_1 < t_2 < \cdots < t_{q_0+1} \leq q$.
- A vector of integers $(n_i)_{1 \leq i \leq t_{q_0+1}}$
  and set of vectors of integers $\{(\eta_{i,1}, \cdots, \eta_{i,n_i})_{1 \leq i \leq t_{q_0+1}}\}$.
- Sets of regions
  $\left(\mathcal{P}_{\mu(\eta_{1+t_i,1})}, \cdots, \mathcal{P}_{\mu(\eta_{1+t_i,n_{1+t_i}})}, \cdots, \mathcal{P}_{\mu(\eta_{t_{i+1},1})}, \cdots, \mathcal{P}_{\mu(\eta_{t_{i+1},n_{t_{i+1}}})} \subset \mathcal{P}_i^0\right)_{1 \leq i \leq q_0}$,
  with
  $\mathcal{P}_{\mu(\eta_{i,j})} \cap \mathcal{P}_{\mu(\eta_{i,k})} \subset \{x, (\beta_i^0)^T x + \alpha_i^0 = 0\}$, if $j \neq k$, and
  $\left(\cup_{1+t_i \leq j \leq t_{i+1}, 1 \leq k \leq n_j} \mathcal{P}_{\mu(\eta_{j,k})} = \mathcal{P}_i^0\right)_{1 \leq i \leq q_0}$.
- Parameter vectors $\left(\sum_{j=t_i+1}^{t_{i+1}} (\beta_j^T x + \alpha_j) = (\beta_i^{0T} x + \alpha_i^0)\right)_{1 \leq i \leq k^0}$,

so that, we can write:

$$f_0(x) = f_\theta(x) = \sum_{i=1}^{q_0} \sum_{j=t_i+1}^{t_{i+1}} (\beta_j^T x + \alpha_j) \mathbf{I}_{\left\{\cup_{1 \leq k \leq n_j} \mathcal{P}_{\mu(\eta_{j,k})}\right\}}(x).$$

More generally, for any function $f_\theta$ defined by (12), let us define $s(\beta)_{ik} = \left(\sum_{j=1+t_i}^{t_{i+1}} \beta_{jk}\right)$ and, if $\sum_{j=1+t_i}^{t_{i+1}} \beta_{jk} \neq 0$, let us write $q(\beta)_{jk} = \frac{\beta_{jk}}{\sum_{j=1+t_i}^{t_{i+1}} \beta_{jk}}$. If $\sum_{j=1+t_i}^{t_{i+1}} \beta_{jk} = 0$, $q(\beta)_{jk}$ will be set at $\frac{1}{\sum_{j=1+t_i}^{t_{i+1}} 1}$. Let us write also $s(\alpha) = \left(\sum_{j=1+t_i}^{t_{i+1}} \alpha_j\right)$ and , if $\sum_{j=1+t_i}^{t_{i+1}} \alpha_j \neq 0$, let us write $q(\alpha)_j = \frac{\alpha_j}{\sum_{j=1+t_i}^{t_{i+1}} \alpha_j}$. If $\sum_{j=1+t_i}^{t_{i+1}} \alpha_j = 0$, $q(\alpha)_j$ will be set at $\frac{1}{\sum_{j=1+t_i}^{t_{i+1}} 1}$.
Moreover, let us write $q(\beta)_j = (q(\beta)_{j1}, \cdots, q(\beta)_{jd})^T$, $s(\beta)_i = (s_{i1}, \cdots, s_{id})^T$.

Now, for any $\theta \in \Theta$, a reparameterization $\theta \mapsto (\omega_t, \psi_t)$ exists with

$$\omega_t = \left((s(\beta)_i)_{i=1}^{q_0}, (s(\alpha)_i)_{i=1}^{q_0}, (\beta_i, \alpha_i)_{i=1+t_{q_0+1}}^{q}\right),$$
$$\psi_t = \left((q(\beta)_j)_{j=t_1}^{t_{q_0+1}}, (q(\alpha)_j)_{j=t_1}^{t_{q_0+1}}, \mu(\eta_{1,1}), \cdots, \mu(\eta_{t_{q_0+1},n_{t_{q_0+1}}}), \mu(1+t_{q_0+1}), \cdots, \mu(q)\right),$$

such that, if we write $diag(q(\beta)_j)$ the $d \times d$ matrix whose diagonal components

are the components of the vector $q(\beta)_j$ and zero elsewhere, we get:

$$\begin{array}{l} f_\theta(x) = \sum_{i=1}^{q_0} \sum_{j=t_i+1}^{t_{i+1}} \left(diag(q(\beta)_j) s(\beta)_i\right)^T x \mathbf{I}_{\left\{\cup_{1\leq k\leq n_j} \mathcal{P}_{\mu(\eta_{j,k})}\right\}}(x) + \\ \sum_{i=1}^{q_0} \sum_{j=t_i+1}^{t_{i+1}} s(\alpha)_i q(\alpha)_j \mathbf{I}_{\left\{\cup_{1\leq k\leq n_j} \mathcal{P}_{\mu(\eta_{j,k})}\right\}}(x) + \\ \sum_{i=t_{q_0+1}+1}^{q} \left({\beta_i}^T x + \alpha_i\right) \times \mathbf{I}_{\mathcal{P}_{\mu(i)}}(x) \\ = \sum_{i=1}^{q_0} s(\beta)_i^T \sum_{j=t_i+1}^{t_{i+1}} diag(q(\beta)_j) x \mathbf{I}_{\left\{\cup_{1\leq k\leq n_j} \mathcal{P}_{\mu(\eta_{j,k})}\right\}}(x) + \\ \sum_{i=1}^{q_0} s(\alpha)_i \sum_{j=t_i+1}^{t_{i+1}} q(\alpha)_j \mathbf{I}_{\left\{\cup_{1\leq k\leq n_j} \mathcal{P}_{\mu(\eta_{j,k})}\right\}}(x) + \\ \sum_{i=1+t_{q_0+1}}^{q} \left({\beta_i}^T x + \alpha_i\right) \times \mathbf{I}_{\mathcal{P}_{\mu(i)}}(x). \end{array} \tag{17}$$

With this parameterization, for a fixed $t$, $\omega_t$ is an identifiable parameter and all the non-identifiability of the model will be in $\psi_t$. Hence, if $0_{\mathbb{R}^{d+1}}$ is the vector with $d+1$ zeros, for a fixed $t$, $f_{\left(\omega_t^0, \psi_t\right)} = f_0$ if and only if

$$\omega_t^0 = \quad (\beta_1^0, \cdots, \beta_{q_0}^0, \quad \alpha_1^0, \cdots, \alpha_{q_0}^0, \quad \underbrace{0_{\mathbb{R}^{d+1}}, \cdots, 0_{\mathbb{R}^{d+1}}}_{q - t_{q_0+1}}).$$

Finally, let us write $q_1 = q - q_{t_{q_0+1}}$, then the lemma is proven with:
For $i \in \{1, \cdots, q_0\}$, $\omega_i = (s(\beta)_i, s(\alpha)_i)$,
for $i \in \{q_0+1, \cdots, q_1\}$, $\omega_i = (\beta_{1+t_{q_0+1}+i-(q_0+1)}, \alpha_{1+t_{q_0+1}+i-(q_0+1)})$,
for $i \in \{q_1, \cdots, q\}$, $\omega_i = 0_{\mathbb{R}^{d+1}}$,
for $i \in \{1, \cdots, q_0\}$,

$$g_{\psi_i}(x) = \left(\sum_{j=t_i+1}^{t_{i+1}} diag(q(\beta)_j) x \mathbf{I}_{\left\{\cup_{1\leq k\leq n_j} \mathcal{P}_{\mu(\eta_{j,k})}\right\}}(x), \sum_{j=t_i+1}^{t_{i+1}} q(\alpha)_j \mathbf{I}_{\left\{\cup_{1\leq k\leq n_j} \mathcal{P}_{\mu(\eta_{j,k})}\right\}}(x)\right)^T, \tag{18}$$

for $i \in \{q_0+1, \cdots, q_1\}$,

$$g_{\psi_i}(x) = \left(x \mathbf{I}_{\mathcal{P}_{\mu(1+t_{q_0+1}+i-(q_0+1))}}(x), \mathbf{I}_{\mathcal{P}_{\mu(1+t_{q_0+1}+i-(q_0+1))}}(x)\right)^T, \tag{19}$$

and, for $i \in \{q_1, \cdots, q\}$, $g_{\psi_i}(x)$ is a function in $\mathbb{R}^{h+1}$:

$$g_{\psi_i}(x) = \left(x \mathbf{I}_{\mathcal{P}_{\mu(i)}}(x), \mathbf{I}_{\mathcal{P}_{\mu(i)}}(x)\right), \tag{20}$$

such that all the indicator functions involved in the development of $f_\theta$, $\left(\mathbf{I}_{\left\{\cup_{1\leq k\leq n_1} \mathcal{P}_{\mu(\eta_{1,k})}\right\}}(x), \cdots, \mathbf{I}_{\mathcal{P}_{\mu(q)}}(.)\right)$ are linearly independent indicator functions of regions of hyperplane arrangement. Note that, the fixed parameter $\omega^0 = (\omega_1^0, \cdots, \omega_q^0)$ of the lemma is such that $\omega_i^0 = (\beta_i^0, \alpha_i^0)$ for $i \in \{1, \cdots, q_0\}$, and $\omega_i^0 = 0_{\mathbb{R}^{d+1}}$ for $i \in \{q_0+1, \cdots, q\}$.

## 7. Proof of proposition 1

With the notations of lemma 3.1 and its proof, let us consider the set of functions:

$$\left\{ f_{(\omega,\psi)} = f_0 + \sum_{i=1}^{q} (\omega_i - \omega_i^0)^T g_{\psi_i}, (\omega,\psi) \in \mathcal{F} = \mathcal{F}_\omega \times \mathcal{F}_\psi \right\},$$

where $\mathcal{F}$ is a compact subset of $\mathbb{R}^L$, for an strictly positive integer $L$. Then

$$\|f_{(\omega,\psi)} - f_0\|_2 = |\omega - \omega^0| \| \sum_{i=1}^{q} \frac{(\omega_i - \omega_i^0)}{|\omega - \omega^0|}^T g_{\psi_i} \|_2$$

Now,

$$\begin{Bmatrix} \sum_{i=1}^{q} \frac{(\omega_i - \omega_i^0)}{|\omega - \omega^0|}^T g_{\psi_i}, (\omega,\psi) \in \mathcal{F}, \omega \neq \omega^0 \end{Bmatrix} \cup$$
$$\begin{Bmatrix} \lim_{\omega \to \omega^0} \sum_{i=1}^{q} \frac{(\omega_i - \omega_i^0)}{|\omega - \omega^0|}^T g_{\psi_i}, (\omega,\psi) \in \mathcal{F} \end{Bmatrix} \subset \mathcal{V},$$

where

$$\mathcal{V} = \left\{ h_{v,\psi} = \sum_{i=1}^{q} v_i g_{\psi_i}, v = (v_1, \cdots, v_q), |v| = 1 \text{ and } \psi \in \mathcal{F}_\psi \right\}.$$

Note that for all $h_{v,\psi} \in \mathcal{V}$, $\|h_{v,\psi}\|_2 > 0$. Then, using the compacity of the set $\{v, |v| = 1\} \times \mathcal{F}_\psi$ and the continuity of $(v,\psi) \mapsto \|h_{v,\psi}\|_2$, $m > 0$ exists such that

$$\forall (v,\psi) \in \{v, |v| = 1\} \times \mathcal{F}_\psi, (v,\psi) \mapsto \|h_{v,\psi}\|_2 \geq m.$$

At the same time, since

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

a constant $C$ exists so that $\mathcal{S}$ can be included in the set of functions:

$$\mathcal{H} = \left\{ f_{(\gamma,\psi)} = \sum_{i=1}^{q} \gamma_i g_{\psi_i}, \gamma = (\gamma_1, \cdots, \gamma_q), |\gamma| \leq \frac{C}{m}, \psi_i \in \mathcal{F}_\psi \right\}.$$

According to the definitions (18), (19), and (20) of functions $g_{\psi_i}$, this set is a special case of piecewise polynomials functions whose VC-dimension is bounded (see Bartlett et al. [1]). So, according to van der Vaart [16], since VC-classes have polynomial covering number, a positive integer $Q$ exists so that $\mathcal{N}_{[\cdot]}\left(\eta, \mathcal{H}, \|\cdot\|_2\right) = O\left(\frac{1}{\eta}\right)^Q$.

## References


[1] Bartlett, P.L., Maiorov, V.V. and Meir, R., Almost linear vc-dimension bounds for piecewise polynomial networks. Neural Computation, 10(8) (1998) 2159-2173.

[2] Bengio, Y., Learning deep architectures for AI. Foundations and Trends in Machine Learning,2(1) (2009) 1127.

[3] Dacunha-Castelle, D. and Gassiat, E., Testing the order of a model using locally conic parametrization; population mixtures and stationary arma processes. Annals of Statistics, 27(4) (2003) 1178-1209.

[4] Dahl, G. E., Sainath, T. N., Hinton, G. E. Improving deep neural networks for lvcsr using rectified linear units and dropout. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, (2013) 8609-8613.

[5] Fukumizu, K., Likelihood ratio of unidentifiable models and multilayer neural networks, Annals of Statistics, 31(3) (2003) 833-851.

[6] Hagiwara, K. and Fukumizu K. Relation between weight size and degree of over-fitting in neural network regression. Neural Networks, 21 (2008) 48-58.

[7] Hansson, M. and Olsson C. Feedforward neural networks with ReLU activation functions are linear splines. (2017) Thesis, Lund University.

[8] LeCun, Yann, Bengio, Y. and Hinton, G., Deep learning, Nature, 521 (7553) (2015) 436-444.

[9] Liu, X. and Shao, Y., Asymptotics for likelihood ratio tests under loss of identifiability, Annals of Statistics, 31(3) (2003) 807-832.

[10] Montufar, G.F., Pascanu, R., Cho, K. and Bengio, Y., On the number of linear regions of deep neural networks, Advances in Neural Information Processing Systems, 27 (2014) 2924-2932.

[11] Nair, V. and Hinton, G. E., Rectified linear units improve restricted Boltzmann machines, 27th International Conference on Machine Learning (ICML-10), (2010) 807-814.

[12] Neyshabur, B., Li, Z., Bhojanapalli, S. LeCun, Y., Srebro, N. Towards Understanding the Role of Over-Parametrization in Generalization of Neural Networks. arXiv preprint arXiv:1805.12076, 2018.

[13] Rynkiewicz, J., Asymptotics for Regression Models Under Loss of Identifiability, Sankhya A, 78 (2) (2016) 155-179.

[14] Rynkiewicz, J., General bound of overfitting for MLP regression models, Neurocomputing, 90 (2012) 106-110.

[15] Ripley, B., Pattern recognition and neural networks, Cambridge university press (1996).

[16] van der Vaart, A.W., Asymptotic statistics, Cambridge university press (1998).

[17] White, H., Artificial neural networks, Blackwell (1992).

[18] Zhang, C. Bengio, S. Hardt, M. Recht, B. and Vinyals, O., Understanding deep learning requires rethinking generalization. ICLR 2017 (2017).