ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Asymptotic statistics for multilayer perceptron with ReLu hidden units

Joseph Rynkiewicz [1]

Université Paris I - SAMM
90 rue de tolbiac, Paris - France

Abstract. We consider regression models involving multilayer perceptrons (MLP) with rectified linear unit (ReLu) functions for hidden units. It is a difficult task to study statistical properties of such models for several reasons: A first difficulty is that these activation functions are not differentiable everywhere, a second reason is also that in practice these models may be heavily overparametrized. In general, the estimation of the parameters of the MLP is done by minimizing a cost function, we focus here on the sum of square errors (SSE) which is the standard cost function for regression purpose. In this framework, we can characterize the asymptotic behavior of the SSE of estimated models which give information on the possible overfitting of such models. This task is done using recent methodology introduced to deal with models with a loss of identifiability which is very flexible. So, we don't have to assume that a true model exits or that a finite set of parameters realize the best regression function.

1 Introduction

Feed-forward neural networks are well known and popular tools. These networks have gained in popularity since the surge of Deep Learning which provides outstanding practical results. Deep neural networks combine a cascade of multiple layers of non linear processing units and the ReLu function is now one of the most popular activation functions for such networks (see Lecun et al.[4]). Even if these networks work very well in practice, very few theoretical results are available about such complex models. We propose in this paper to fill a little bit this gap, hence we focus only on shallow networks with only one hidden layer, but we deal with ReLu activation functions for the hidden layer. We may hope, that our methodology may be extended to more complex networks. This paper is organized as follows: Firstly, we give a general inequality for the difference of the sum of square errors (SSE) of the estimated regression model and the SSE of the theoretical best regression function in our model. A set of generalized derivative functions is a key tool in deriving such inequality. Under suitable conditions, checked by MLP with ReLu hidden units, we provide the asymptotic distribution for the difference of SSE even if these models are not differentiable everywhere and if the parameters characterizing the best regression function are not unique and belong to an infinite set.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

2 The model

For an observation $x \in \mathbb{R}^d$, an MLP function with k hidden units can be written:

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

with $\theta = (\beta, a_1, \cdots, a_k, b_1, \cdots, b_k, w_{11}, \cdots, w_{1d}, \cdots, w_{kd}) \subset \mathbb{R}^{2k+1+k\times d}$ the parameter vector of the model, and $w_i := (w_{i1}, \cdots, w_{id})^T$. Let us denote $\Theta \subset \mathbb{R}^{2k+1+k\times d}$ the possible bounded set of parameters. The transfer function ϕ will be assumed to be a ReLu function: $\phi(z) = \max(0, z)$ for $z \in \mathbb{R}$. Note that this function is not differentiable with respect to $z = 0$. We observe a random sample of independent and identically distributed random vectors: $(X_1, Y_1), \cdots, (X_n, Y_n)$, from the distribution P of a vector (X, Y), with Y a real random variable. The regression model can be written as:

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

where f_0 is the best regression function which belongs to the set $\{f_\theta, \theta \in \Theta\}$:

$$f_0 = \arg\min_{\theta\in\Theta} \|Y - f_\theta(X)\|_2,$$

where, a general random variable Z,

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

is the $\mathcal{L}^2$ norm for ageneral square integrable function g. Let us write Θ_0 the set of parameters realizing the best regression function f_0: $\forall\theta \in \Theta_0,\ f_\theta = f_0$. Note that we do not assume that Θ_0 is a finite set which means that loss of identifiability can occur, this is the case if the MLP has redundant hidden units (see Fukumizu [1] or Rynkiewicz [5]). A natural estimator of f_0 is the least square estimator (LSE) $f_{\hat\theta}$ that minimizes the SSE:

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

$f_{\hat\theta}$ is expected to converge to the function f_0 under suitable conditions. Now, let us introduce generalized derivative functions:

$$d_\theta(x) = \frac{f_\theta(x) - f_0(x)}{\|f_\theta(X) - f_0(X)\|_2},\ f_\theta \neq f_0. \tag{3}$$

Note that these functions are always defined even if the functions f_θ are not differentiable everywhere. We give now the main results of this paper.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

2.1 Upper bound for the SSE

This lemma is proven in Rynkiewicz [5], it gives a very general upper bound for the sum of square errors.

Lemma 2.1 *Let ε_t be the error $Y_t - f_0(X_t)$, for all regression functions $f_\theta, \theta \in \Theta$ with $f_\theta \neq f_0$ and d_θ defined in (3), then*

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

Using this lemma, we can then give the asymptotic behavior of the SSE under fairly general assumptions.

2.2 Approximation of the SSE

First, we recall that 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 $n \geq N(\delta, \varepsilon)$. Define the limit set of derivatives $\mathcal{D}$ as 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 (θ_n), define, for all $t \in [0, 1]$, $f_t = f_{\theta_n}$, 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 α a strictly positive real number, such that for any $t \in [0, \alpha]$, $t \mapsto \|f_{\theta_t}(X) - f_0(X)\|_2$ 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

$$\|f_u(X) - f_0(X)\|_2 = u, \tag{4}$$

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

$$\int \left(f_u - f_0 - ud\right)^2 dP = o(u^2). \tag{5}$$

Now, let us introduce some assumptions:

B-1 Let u be defined as (4), the map $u \mapsto P(Y - f_u(X))^2$ admits a second-order Taylor expansion with strictly positive second derivative $\frac{\partial^2 P(Y - f_u(X))^2}{\partial u^2}$ at $u = 0$.

B-2 The set of generalized derivative functions $\mathcal{S} = \{d_\theta, \theta \in \{\Theta \backslash \Theta_0\}\}$ is a Donsker class (see van der Vaart [6], for definition of Donsker class).

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

The following theorem is proven in Rynkiewicz [5].

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

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

Even when the set of possible regression functions $\mathcal{F}$ may be heavily over-parametrized or not differentiable, this theorem proves the tightness of the SSE, if the set $\mathcal{S}$ is a Donsker class. Note that assumption **B-1** is true for MLP with ReLu hidden units even if the functions f_θ are not differentiable everywhere because it involves only differentiability in quadratic mean as in Le Cam [3]. Now, using the same reparametrization technique and following the same ideas that in Rynkiewicz [5], we can prove assumption **B-2** and give the general description of the asymptotic behavior of the SSE:

Reparameterization. If k_0 is the minimal number of hidden units to get the best function f_0, then the writing of f_0 with a neural network with k_0 hidden units is unique, up to some permutations:

$$f_0=\beta^0+\sum_{i=1}^{k_0}a_i^0\phi\left({w_i^0}^T x+b_i^0\right). \tag{6}$$

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

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

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

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

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

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

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

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

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

We get then the following expansion for the numerator of generalized derivative functions:

Lemma 2.3 *For a fixed t, in the neighborhood of the identifiable parameter Φ_t^0:*

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

with

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

where $\mathbb{I}_{\mathbb{R}^+}$ is the indicator function of $\mathbb{R}^+$: $\mathbb{I}_{\mathbb{R}^+}(z) = 0$ if $z < 0$ and $\mathbb{I}_{\mathbb{R}^+}(z) = 1$ if $z \geq 0$

Now, with this reparameterization, the proposition 1 of Rynkiewicz [5] shows that the assumption (**B-2**) is true for our model.

Finally, we can give the asymptotic behavior of the SSE:

Theorem 2.4 *Let the map $\Omega : \mathcal{L}^2(P) \to \mathcal{L}^2(P)$ be defined as $\Omega(f) = \frac{f}{\|f\|_2}$. Under the assumptions* **B-1** *and* **B-2**, *a centered Gaussian process $\{W(d), d \in \mathcal{D}\}$ with continuous sample paths and a covariance kernel $P(W(d_1)W(d_2)) = P(d_1 d_2)$ exists so that*

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

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

$$\begin{aligned}&\mathcal{D}_t = \Big\{\Omega\Big(\gamma + \textstyle\sum_{i=0}^{k^0} \epsilon_i \phi({w_i^0}^T X + b_i^0) + \sum_{i=0}^{k^0} \mathbb{I}_{\mathbb{R}^+}({w_i^0}^T X + b_i^0)(\zeta_i^T X + \alpha_i)\\ &+ \textstyle\sum_{i=t_{k^0+1}+1}^{k} \mu_i \phi({w_i}^T X + b_i)\Big),\\ &\gamma, \epsilon_1, \cdots, \epsilon_{k^0}, \alpha_1, \cdots, \alpha_{k^0} \in \mathbb{R}, \mu_{t_{k^0+1}+1}, \cdots, \mu_k \in \mathbb{R}^+; \zeta_1, \cdots, \zeta_{k^0} \in \mathbb{R}^d,\\ &(w_{k^0+1+1}, b_{k^0+1+1}), \cdots, (w_k, b_k) \in \Theta \backslash \left\{(w_1^0, b_1^0), \cdots, (w_{k^0}^0, b_{k^0}^0)\right\}\Big\}.\end{aligned}$$

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

This theorem shows that the degree of over-fitting is bounded in probability, but depends on the size of the asymptotic set $\mathcal{D}$. Hence, the over-fitting that occurs in the over-realizable case is for the extreme values of the input weights as in an MLP with sigmoidale activation functions (see Hagiwara and Fukumizu [2]). In order to reduce the over-fitting we need to control the size of the limit functions in $\mathcal{D}$, this can be done by reducing the size of the inputs weights $(w_i, b_i)_{1 \leq i \leq k}$ either by L_2 penalization (weight decay method) or L_1 penalization (Lasso method). However, note that the size of the weights has to be large enough so that Θ contains some parameters of Θ_0, so we need to find a trade-off for this penalization.

3 Conclusion

MLP models have been used for many years, but have evolved dramatically these last years. The ReLu activation function is now one of the most popular even if the statistical properties of models using these functions are not well known. This paper is an attempt to fill this gap. By using modern theory which deals with over-parameterized models we can give the asymptotic behavior of the SSE for MLP with one hidden layer using ReLu functions. It is clear that the networks used in practice are deeper and the layer after the hidden units is often a pooling function like the mean, the maximum or a norm, but pooling functions can also be seen as constraints over parameters and so over asymptotic set $\mathcal{D}$. Finally, our methodology seems to be promising to give some statistical understanding of models involved in deep learning.

References

[1] Fukumizu, K., Likelihood ratio of unidentifiable models and multilayer neural networks, Ann. Statist. 31 (2003) 833-851.

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

[3] Le Cam, L., On the assumptions used to prove asymptotic normality of maximum likelihood estimators, nnals of Mathematical Statistics 41 (1970) 802-828.

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

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

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