ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 315-320

Functional approximation by perceptrons: a new approach

Jean-Gabriel ATTALI, Gilles PAGÈS

Samos, Univ. Paris I, 90, rue de Tolbiac Cedex 13 & Univ. Paris 12, 61, av. du Gal de Gaulle, F-94010 Créteil Cedex, mail: attali@univ-paris1.fr,
Lab. de Proba., URA 224, Univ. Paris 6, 4 pl. Jussieu, F-75252 Cedex 05 & Univ. Paris 12, 61, av. du Gal de Gaulle, F-94010 Créteil Cedex, mail: gpa@ccr.jussieu.fr

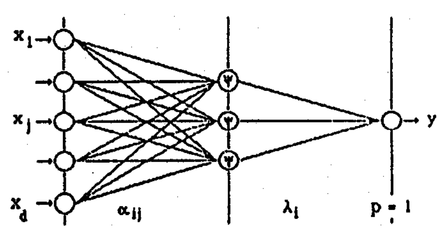
Abstract. We provide a radically elementary proof of the universal approximation property of the 1-hidden layer perceptron based on the Taylor Young formula and the Vandermonde determinant. It works for both L^p and uniform approximation on a compact set. This method naturally yields some bounds for the design of the hidden layer and some convergence results for the derivatives.

1. Introduction

In 1993, Hornik established in [1], using the Riesz representation theorem, that a 1-hidden layer perceptron can uniformly approximate continuous functions on compact sets. First, we show that any C^p-function on $\mathbb{R}^d$ can be locally uniformly approximated with all its (partial) derivatives using a 1-hidden layer perceptron. Some bounds for the design of the hidden layer are also proposed. Our results differ from Barron one's (who deals with mean square approximation see [2]), namely our bounds for the design of the hidden layer are dimension dependent. This is no surprise as the uniform convergence on compact sets is far more stringent.

Notations: • $C(K,\mathbb{R})$ will denote the set of continuous real-valued functions defined on the compact set K of $\mathbb{R}$ or $\mathbb{R}^d$, and for $f \in C(K,\mathbb{R})$ we set $\|f\|_K := \sup_{x\in K} |f(x)|$.

• for f_k and g in $C(K,\mathbb{R})$, $f_k \xrightarrow{U_K} g$ will denote the uniform convergence of f_k to g on a compact set K.



1-hidden layer perceptron

• $C^n(K,\mathbb{R})$ will denote the set of all real-valued functions defined on K, n times continuously differentiable, and for f_k and g in $C^n(K,\mathbb{R})$ $n \geq 1$, we will

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 315-320

write $f_k \xrightarrow{U_K^{(n)}} g$ if $f_k^{(\ell)} \xrightarrow{U_K} g^{(\ell)}$, for all $\ell \in \{0, \cdots, n\}$, where $g^{(\ell)}$ denotes the ℓ-th derivative of g.

Definition: We call "$(n+1,\psi)$-perceptron" any function of the form: $x \mapsto \sum_{i=0}^{n} \lambda_i \psi(\alpha_i . x)$,where $x \in \mathbb{R}^d$ and . denotes the canonical inner product on $\mathbb{R}^d$.

2. Approximation on a compact set of $\mathbb{R}$

Assume first that ψ is C^n. For any polynomial P of degree $p \leq n$, we exhibit a sequence of $(p+1,\psi)$- perceptrons that $U_K^{(n)}$-converges to P.

Proposition 1: *Let $\psi \in C^n(\mathbb{R},\mathbb{R})$ such that $\forall k,\ \ 0 \leq k \leq n,\ \psi^{(k)}(0) \neq 0$. Let $p \in \{0, \cdots, n\}$ and $(c_i)_{0 \leq i \leq n}$ nonzero pairwise distinct real numbers, then for every polynomial P such that $d^0 P = p$, there exist $p+1$ rational functions $\lambda_i(h) := Q_i(\frac{1}{h})$ where Q_i are some polynomials of degree p, such that:*

$$\forall K \textit{ compact set of } \mathbb{R}, \quad \sum_{i=0}^{p} \lambda_i(h) \psi(c_i h x) \xrightarrow{U_K^{(n)}} P(x) \textit{ when } h \to 0.$$

Proof: 1) Convergence of the perceptron: Let $P(x) = \sum_{i=0}^{p} a_i x^i,\ \ p \leq n$, be the polynomial we want to approximate. Let $(\alpha_i, \lambda_i)_{i \in \{0,\cdots,p\}}$ be $2(p+1)$ arbitrary real numbers. The Taylor-Young formula applied to ψ at the p-th order yields:

$$\psi(\alpha_i x) - \psi(0) - \alpha_i x \psi'(0) - \cdots - \frac{\alpha_i^p x^p}{p!} \psi^{(p)}(0) = \frac{(\alpha_i x)^p}{p!} \epsilon(\alpha_i x), \text{ for } 0 \leq i \leq p, \quad (1)$$

with $\lim_{x \to 0} \epsilon(x) = 0$. Hence, setting $A_K := \sup_{x \in K} \frac{|x^p|}{p!}$ and summing (1) over i:

$$\left| \sum_{i=0}^{p} \lambda_i \psi(\alpha_i x) - \psi(0) \sum_{i=0}^{p} \lambda_i - \cdots - x^p \frac{\psi^{(p)}(0)}{p!} \sum_{i=0}^{p} \lambda_i \alpha_i^p \right| \leq A_K \sum_{i=0}^{p} |\lambda_i \alpha_i^p \epsilon(\alpha_i x)|. \quad (2)$$

So we solve the system in $\lambda_0, \cdots, \lambda_p$:
$$(S_p) \equiv \begin{cases} \lambda_0 + & \cdots & + \lambda_p = \frac{a_0}{\psi(0)} \\ \vdots & & \vdots \\ \lambda_0 \alpha_0^p + & \cdots & + \lambda_p \alpha_p^p = \frac{a_p . p!}{\psi^{(p)}(0)} \end{cases}$$

The solution of (S_p) is given by:

$$\lambda_i(\alpha_0, \cdots, \alpha_p) = \begin{vmatrix} 1 & \cdots & 1 & \frac{a_0}{\psi(0)} & 1 & \cdots & 1 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \alpha_0^p & \cdots & \alpha_{i-1}^p & \frac{a_p . p!}{\psi^{(p)}(0)} & \alpha_{i+1}^p & \cdots & \alpha_p^p \end{vmatrix}_{(p+1)\times(p+1)} \frac{1}{\prod_{i>j} (\alpha_i - \alpha_j)}. \quad (3)$$

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 315-320

The key of the proof is to show that $\lambda_i(\alpha_0,\cdots,\alpha_p)\alpha_i^p$ has a finite limit as $\alpha := (\alpha_0,\cdots,\alpha_p) \rightarrow 0$, as least for some subclass α, then inequality (2) yields the announced result. Namely, we set $\alpha_i := c_i h$, where $c_i > 0$, $c_i \neq c_j$ if $i \neq j$ and $h > 0$. Then we develop the determinant of the numerator in (3) with respect to the column i, which gives, setting $a_j^{'} := \dfrac{a_j . j!}{\psi^{(j)}(0)}$:

$$\lambda_i(h) = h^{-\frac{p(p+1)}{2}} \left(\prod_{i>j}(c_i - c_j)\right)^{-1} \sum_{j=0}^{p}(-1)^{i+j} a_j^{'} \Delta_j(c_0,\cdots,c_p) h^{\frac{p(p+1)}{2}-j}.$$

Then we can see that $\lambda_i(h) = Q_i(\frac{1}{h})$ where Q_i is a polynomial function, $d^0 Q_i = p$. Hence $\lim_{h\rightarrow 0} \lambda_i(h)(c_i h)^p$ is finite for $\lambda_i(h)(c_i h)^p$ is a polynomial in h.

On the other hand, $\lim_{h\rightarrow 0} \sup_{x\in K} |\epsilon(c_i h x)| = 0$, so $A_K \sum_{i=0}^{p} |\lambda_i(h)(c_i h)^p||\epsilon(c_i h x)| \xrightarrow{U_K} 0$.

But (2) implies: $\left|\sum_{i=0}^{p} \lambda_i(h)\psi(c_i h x) - P(x)\right| \leq A_K \sum_{i=0}^{p} |\lambda_i(h)(c_i h)^p||\epsilon(c_i h x)|$.

Finally, $\forall K$ compact set, $\sum_{i=0}^{p} \lambda_i(h)\psi(c_i h x) \xrightarrow{U_K} P(x)$.

2) U_K-convergence of the derivatives with order $k \in \{0,\cdots,n\}$: for $k \leq p$, the Taylor-young formula with order $p-k$ applied to $\psi^{(k)}$ yields:

$$\left|\psi^{(k)}(c_i h x) - \psi^{(k)}(0) - \cdots - \frac{(c_i h)^{p-k}\psi^{(p)}(0)}{(p-k)!} x^{p-k}\right| \leq A_K^k |(c_i h)^{p-k} \epsilon_k(c_i h x)|, \quad (4)$$

$\lim_{y\rightarrow 0} \epsilon_k(y) = 0$, and $A_K^k := \sup_{x\in K} \dfrac{|x^{p-k}|}{(p-k)!}$. It is straightforward to check that:

$$\sum_{\ell=k}^{p}\left(\sum_{i=0}^{p} \lambda_i(h)(c_i h)^{\ell}\right) \psi^{(\ell)}(0) \frac{x^{\ell-k}}{(\ell-k)!} = \frac{p!}{(p-k)!} a_p x^{p-k} + \cdots + k! a_k = P^{(k)}(x).$$

Thus, multiplying each equation (4) by $\lambda_i(h)(c_i h)^k$ and summing over i, it gives :

$$\left|\sum_{i=0}^{p} \lambda_i(h)(c_i h)^k \psi^{(k)}(c_i h x) - P^{(k)}(x)\right| \leq A_K^k \sum_{i=0}^{p} |\lambda_i(h)(c_i h)^p||\epsilon_k(c_i h x)|,$$

which gives the result for $k \leq p$ as the right member goes to 0 as $h \rightarrow 0$.
If $p = n$ it is over, if $p \leq n-1$, the result holds for $0 \leq k \leq p$.
Considering now $k \in \{p+1,\cdots,n\}$, and h satisfying $h \leq \min_i \dfrac{1}{c_i}$.

$$\sup_{x\in K}\left|\sum_{i=0}^{p} \lambda_i(h)(c_i h)^k \psi^{(k)}(c_i h x)\right| \leq \left\|\psi^{(k)}\right\|_K \sum_{i=0}^{p} \left|\lambda_i(h)(c_i h)^k\right|. \quad (5)$$

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 315-320

$\lambda_i(h)(c_i h)^k$ is polynomial with valuation $\geq k-p>1$, thus $\lim_{h\to 0}\lambda_i(h)(c_i h)^k=0$.

Hence: $\sum_{i=0}^{p}\lambda_i(h)(c_i h)^k\psi^{(k)}(c_i hx)\xrightarrow{U_K}0=P^{(k)}(x)$ when $h\to 0$. □

It is now possible to give the approximation theorem on $\mathbb{R}$.

Theorem 1: *Let $\psi\in C^\infty(\mathbb{R},\mathbb{R})$ such that $\forall k\in\mathbb{N}$, $\psi^{(k)}(0)\neq 0$. Then for every $\eta>0$, for every compact set K of $\mathbb{R}$ and for every $n\in\mathbb{N}$, the space* $\left\{x\mapsto\sum_{i=0}^{m}\lambda_i\psi(\alpha_i x),\ \alpha_i\in]0,\eta[,\ \ m\in\mathbb{N},\ \lambda_i\in\mathbb{R}\right\}$ *is dense in* $(C^n(K,\mathbb{R}),U_K^{(n)})$.

Proof: Following proposition 1, it amounts to show that any function $f\in C^n[0,1]$ is a $U_{[0,1]}^{(n)}$-limit of polynomials. Now, this result simply follows by considering the Bernstein polynomials: $B_j(f):=\sum_{k=0}^{j}C_j^k f(\frac{k}{j})x^k(1-x)^{j-k}$ see [3]. □

Remarks: • The Bernstein polynomials are not an optimal choice as far as rate of convergence is concerning (see part 4).
• If ψ is analytic and nonpolynomial, $D:=\{\theta/\ \exists k\in\mathbb{N},\ \psi^{(k)}(\theta)=0\}$ is at most countable. So we can apply theorem 1 with any $\phi_\theta(x):=\psi(x-\theta)$, for $\theta\in D^c$.

3. Approximation on a compact set of $\mathbb{R}^d$

Proposition 2: *Let $\psi\in C^n(\mathbb{R},\mathbb{R})$ such that $\forall k,\ \ 0\leq k\leq n,\ \psi^{(k)}(0)\neq 0$. Let $p\in\{0,\cdots,n\}$ and $P\in\mathbb{R}_p[X_1,\cdots,X_d]$. Let denote $N_p^d:=dim_{\mathbb{R}}\,\mathbb{R}_p[X_1,\cdots,X_d]$. Then there exist N_p^d $\mathbb{R}^d$-valued vectors $(c_i)_{1\leq i\leq N_p^d}$ and N_p^d rational functions $\lambda_i(h):=Q_i(\frac{1}{h})$, where $Q_1,\cdots,Q_{N_p^d}\in\mathbb{R}_p[X]$ s.t.:*

$\forall K$ compact set of $\mathbb{R}^d$, $\sum_{1\leq i\leq N_p^d}\lambda_i(h)\psi(hc_i.x)\xrightarrow{U_K^{(n)}}P(x_1,\cdots,x_d)$ *when $h\to 0$.*

Proof: see [AP1] for a detailed proof. □

Theorem 2: *Let $\psi\in C^\infty(\mathbb{R},\mathbb{R})$ such that $\forall k\in\mathbb{N}$ $\psi^{(k)}(0)\neq 0$.
Then for every $\eta>0$ and for every compact set K of $\mathbb{R}^d$, the space*
$$\left\{x\mapsto\sum_{i=0}^{m}\lambda_i\psi(\alpha_i.x),\ \ x\in\mathbb{R}^d,\ \alpha_i:=(\alpha_i^1,\cdots,\alpha_i^d)\in(]0,\eta[)^d,\ \ m\in\mathbb{N},\ \lambda_i\in\mathbb{R}\right\}$$
is dense in $(C^n(K,\mathbb{R}),U_K^{(n)})$ for every $n\in\mathbb{N}$.

Proof: As in theorem 1 using the d-dim. Berstein polynomials on $[0,1]^d$:
$$B_j(f):=\sum_{0\leq k_1\cdots+k_d\leq j}\frac{j!f(\frac{k_1}{j},\cdots,\frac{k_d}{j})x_1^{k_1}\cdots x_d^{k_d}(1-x_1-\cdots-x_d)^{j-k_1\cdots-k_d}}{k_1!\cdots k_d!(j-k_1\cdots-k_d)!}\quad\text{(see [3]).}\ \square$$

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 315-320

4. Design of the hidden layer

4.1. 1-dimensional case

We give here two error bounds, depending if we want only to approximate the function or if we want to approximate together the function and its derivatives. If K is a compact set of $\mathbb{R}^d$, we denote $M_K = \sup_{x \in K} \|x\|$ and $\delta_K = \sup_{(x,y) \in K^2} \|x-y\|$.

Theorem 3: *Let $\psi \in C^\infty(\mathbb{R},\mathbb{R})$ such that $\forall k \in \mathbb{N}$ $\psi^{(k)}(0) \neq 0$. Let $f \in C^p(K,\mathbb{R})$ such that $f^{(p)}$ is ρ-Lipschitz. Let ε_n be a sequence of positive real numbers such that $\lim_{n\to\infty} \varepsilon_n = 0$. Then:*
i) There exists a sequence $(\phi_n)_{n\geq 0}$ of $(n+1,\psi)$-perceptrons functions such that:

$$\|f - \phi_n\|_K \leq \rho A_p M_K^{p+1} \frac{(1+\varepsilon_n)}{n^{p+1}}.$$

ii) There exists a sequence $(\Phi_n)_{n\geq 0}$ of $(n+1,\psi)$-perceptrons functions such that: $\forall k \in \{0,\cdots,p\}$, $\left\|f^{(k)} - \Phi_n^{(k)}\right\|_K \leq \rho A_0 M_K \max(1,(\delta_K)^p) \frac{(1+\varepsilon_n)}{n}$.

Proof: a) For every $n \in \mathbb{N}$ the polynomial of best approximation of degree n, $P_n(f)$ satisfies:

$\|f - P_n(f)\|_K \leq \rho A_p M_K^{p+1} \frac{1}{n^{p+1}}$, where A_p depends only on p (see [4] p. 75).

The result follows from proposition 1 as we can choose ϕ_n such that:

$$\forall n \in \mathbb{N}, \quad \|\phi_n - P_n(f)\|_K \leq \rho A_p M_K^{p+1} \frac{\varepsilon_n}{n^{p+1}}.$$

b) There exists a sequence of polynomials $Q_n(f)$ such that:

$$\forall k \in \{0,\cdots,p\}, \quad \left\|f^{(k)} - Q_n^{(k)}(f)\right\|_K \leq \rho A_0 M_K \max(1,(\delta_K)^p) \frac{1}{n}.$$

So we have the result using again proposition 1. □

Remarks: • The $P_n(f)$ are generally not explicit. But the Tchebychev ones $T_n(f)$ are and satisfy: $\|T_n(f) - f\|_K \leq (3+\ln(n))\,\|P_n(f) - f\|_K$ (see [4]).
So we can explicitly construct a sequence ϕ_n of $(n+1,\psi)$-perceptrons with:

$$\|f - \phi_n\|_K \leq \rho A_p M_K^{p+1} \frac{(3+\ln(n))(1+\varepsilon_n)}{n^{p+1}}.$$

• The $Q_n(f)$ are not explicit but there exist some explicit polynomials $R_n(f)$ with $d^0 n$ s.t.:

$$\forall k \in \{0,\cdots,p\}, \quad \left\|f^{(k)} - R_n^{(k)}(f)\right\|_K \leq \rho A_0 M_K \max(1,(\delta_K)^p) \frac{(3+\ln(n))}{n},$$

So, it is possible to construct a sequence Φ_n such that:

$$\forall k \in \{0,\cdots,p\}, \quad \left\|f^{(k)} - \Phi_n^{(k)}\right\|_K \leq \rho A_0 M_K \max(1,(\delta_K)^p) \frac{(3+\ln(n))(1+\varepsilon_n)}{n}.$$

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 315-320

4.2. Multidimensional case

Theorem 4: *Let $\psi \in C^\infty(\mathbb{R},\mathbb{R})$ such that $\forall k \in \mathbb{N}$ $\psi^{(k)}(0) \neq 0$. Let $f \in C^p(K,\mathbb{R})$ such that for all i, $1 \leq i \leq d$, $\frac{\partial^p f}{\partial x_i^p}$ is ρ-Lipschitz. Let ε_n be a sequence of positive real numbers such that $\lim_{n\to\infty} \varepsilon_n = 0$. Then:*
i) there exists a sequence $(\phi_n)_{n\in\mathbb{N}}$ of $(n+1,\psi)$-perceptrons functions such that:

$$\|f - \phi_n(f)\|_K \leq \frac{\rho A_{d,p} M_K^{p+1} d^{p+1}}{d!^{\frac{p+1}{d}}} \frac{(1+\varepsilon_n)}{n^{\frac{p+1}{d}}}.$$

ii) there exists a sequence $(\Phi_n)_{n\in\mathbb{N}}$ of $(n+1,\psi)$-perceptrons functions such that:

$$\forall k = k_1 + \cdots + k_d \leq p, \quad \left\| \frac{\partial^k f}{\partial x_1{}^{k_1} \cdots \partial x_d{}^{k_d}} - \frac{\partial^k \Phi_n(f)}{\partial x_1{}^{k_1} \cdots \partial x_d{}^{k_d}} \right\|_K \leq \rho a_{d,p} M_K^{p+1} \frac{1}{n^{\frac{1}{2d}}}.$$

Proof: a) The result is given by considering the polynomials of best approximation (see [4] page 89).
b) The result is given by considering the Bernstein polynomials (see [3]). □

5. Conclusion

Our results contain the L^p-approximations results as the polynomial functions are also L^p-dense. However, our bounds for the design of the hidden layer strongly depend on the convergence mode. So they cannot be compared with results obtained in L^p-settings in [2].

References

[1] K. Hornik, Some New Results on Neural Network Approximation *Neural Networks*, 6, 1993, p.1069-1072.

[2] A.R. Barron, Universal Approximation Bounds for Superpositions of a Sigmoidal Function, *Information Theory*, vol. 39 n. 3, 1993, p.930-945.

[3] G.G. Lorentz, *Approximation of Functions*, Chelsea Publishing Company, New-York, 1966, 188p.

[4] G.G. Lorentz, *Berstein Polynomials*, Chelsea Publishing Company, New-York, 1986, 134p.

[AP1] J.G. Attali, G. Pagès, Approximation of functions by perceptrons: a new approach, preprint of SAMOS (Paris, France).