Universal approximation of one-hidden-layer feedforward NN
A simple proof.

Sébastien GERCHINOVITZ

The results and proofs below are combinations of results and proofs from Cybenko (1989) and Barron (1993).
(See also Jones 1992 and Leshno et al. 1993.)

Definition: $\sigma : \mathbb{R} \to \mathbb{R}$ is "universal" iff, for all $a < b$, any $\mathcal{C}^\infty$ function $f : [a,b] \longrightarrow \mathbb{R}$ can be arbitrarily well approximated by a 1-hidden layer NN with activation function σ, i.e.: $\forall \varepsilon > 0$, $\exists N \geq 1$, $\exists v_i, w_i, b_i \in \mathbb{R}$ s.t.

$$\forall x \in [a,b], \quad \left| f(x) - \sum_{i=1}^{N} v_i \, \sigma(w_i x + b_i) \right| \leq \varepsilon .$$

See Eldan and Shamir (2016) for a related definition.

Ex: $\sigma = \mathrm{ReLU}$ or Heaviside (see later for more examples)

Theorem: Let $K \subset \mathbb{R}^d$ compact and $\sigma : \mathbb{R} \to \mathbb{R}$ universal.
Then, any continuous function $f : K \longrightarrow \mathbb{R}$ can be arbitrarily well approximated by a 1-hidden layer NN with activation function σ, i.e.: $\forall \varepsilon > 0$, $\exists N \geq 1$, $\exists v_i \in \mathbb{R}$, $w_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$ s.t.

$$\forall x \in K, \quad \left| f(x) - \sum_{i=1}^{N} v_i \, \sigma(\langle w_i, x \rangle + b_i) \right| \leq \varepsilon .$$

Proof scheme

① Reduce the problem to $f = g_{|K}$, with $g \in C_c^\infty(\mathbb{R}^d, \mathbb{R})$.

② Use the Fourier decomposition

$$g(x) = \int_{\mathbb{R}^d} e^{i\langle w, x\rangle} \hat{g}(w)\, dw \qquad (1)$$

$$= \int_{\mathbb{R}^d} e^{i(\langle w, x\rangle + \theta(w))} |\hat{g}(w)|\, dw \qquad (2)$$

and approximate the first integral (uniformly in $x \in K$) by

$$\tilde{g}(x) = \sum_{k=1}^{K} e^{i\langle w_k, x\rangle} \hat{g}(w_k)\, \lambda(A_k)$$

where the A_k are "small", pairwise disjoint, and s.t.

$$\int_{\mathbb{R}^d \setminus \bigcup_{k=1}^K A_k} |\hat{g}(w)|\, dw \leq \frac{\varepsilon}{2}. \qquad \text{with } \operatorname{diam}\Big(\bigcup_{k=1}^K A_k\Big) < +\infty.$$

This is possible since g and thus $\hat{g}$ belong to the Schwartz space.

Take the real part to see that we can approximate g with

$$\tilde{g} \in \operatorname{span}\left\{\cos(\langle w, \cdot\rangle + \varphi) : w \in \mathbb{R}^d, \varphi \in [0, 2\pi]\right\}$$

③ Approximate $t \mapsto \cos(t)$ by $t \mapsto \sum_{i=1}^{N} v_i \sigma(\nu_i t + b_i)$ uniformly on $[-R_w R_x, R_w R_x + 2\pi]$, where $R_w := \sup\{\|w\| : w \in \bigcup_{k=1}^K A_k\}$

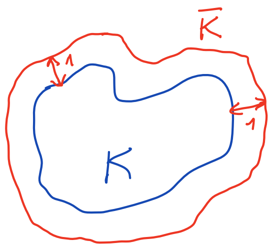
$R_x := \sup\{\|x\| : x \in K\}$

Conclude. ∎

Details : Let $\varepsilon > 0$.

① We show that there exists $g \in \mathcal{C}_c^\infty(\mathbb{R}^d, \mathbb{R})$ such that
$$\forall x \in K, \quad |f(x) - g(x)| \leqslant \varepsilon$$

First, by the Tietze extension theorem, there exists $\bar{f} : \bar{K} \to \mathbb{R}$ continuous on $\bar{K} = \bigcup_{x \in K} \bar{B}(x,1)$ and such that $\bar{f}(x) = f(x)$ for all $x \in K$.



Since $\bar{f}$ is uniformly continuous on the compact $\bar{K}$, let $\delta \in (0,1)$ be s.t.
$$\forall x, x' \in \bar{K}, \quad \|x - x'\| \leqslant \delta \Rightarrow |\bar{f}(x) - \bar{f}(x')| \leqslant \varepsilon$$

Define $\varphi_\sigma(h) = \frac{1}{C_\sigma} \exp\left(-\frac{1}{1 - \left(\frac{\|h\|}{\sigma}\right)^2}\right)$, density function over $\mathbb{R}^d$

Note that $\varphi_\sigma \in \mathcal{C}^\infty(\mathbb{R}^d, \mathbb{R})$ and $\mathrm{supp}(\varphi_\sigma) = B(0, \sigma)$

For $\sigma \in (0, \delta)$ and $x \in \mathbb{R}^d$, we have
$$(\bar{f} * \varphi_\sigma)(x) = \int_{\mathbb{R}^d} \underbrace{\bar{f}(x-h)}_{} \varphi_\sigma(h)\, dh = \mathbb{E}_{Z \sim \varphi_\sigma}\left[\bar{f}(x - Z)\right]$$
with $\bar{f}(x') \overset{\text{conv}}{=} 0$ if $x' \notin \bar{K}$
(never happens if $x \in K$, because $\mathrm{supp}(\varphi_\sigma) = B(0,\sigma)$ and $K + B(0,\sigma) \subset \bar{K}$)

so that $\left|(\bar{f} * \varphi_\sigma)(x) - \bar{f}(x)\right| \leqslant \varepsilon$ for all $x \in K$

and $\bar{f} * \varphi_\sigma \in \mathcal{C}_c^\infty(\mathbb{R}^d; \mathbb{R})$

This proves the result with $g = \bar{f} * \varphi_\sigma$

② We now approximate g with a 1-hidden layer NN.

Since $g \in \mathcal{C}^\infty_c(\mathbb{R}^d, \mathbb{R})$ is "rapidly decreasing" (g lies in the Schwartz space), the inversion formula for the Fourier transform holds: $\forall x \in K$,

$$g(x) = \int_{\mathbb{R}^d} e^{i\langle w, x\rangle} \hat{g}(w)\, dw$$

$$\overset{\pm\varepsilon}{\approx} \int_C e^{i\langle w, x\rangle} \hat{g}(w)\, dw \quad \text{for some hypercube } C \subset \mathbb{R}^d \text{ of sidelength } \rho.$$

(we write $a \overset{\pm\varepsilon}{\approx} b$ to mean that $|a-b| \leq \varepsilon$)

This is because $\int_{\mathbb{R}^d} |\hat{g}(w)|\, dw < +\infty$ ($\hat{g}$ is also rapidly decreasing)

$$\overset{\pm\varepsilon}{\approx} \sum_{k=1}^{m^d} e^{i\langle w_k, x\rangle} \hat{g}(w_k)\, \mathrm{vol}(A_k) \quad \text{for } A_k \text{: subcubes of } C \text{ with sidelength } \frac{\rho}{m}.$$

We can choose $m \geq 1$ large enough and <u>independently of $x \in K$</u> such that the integral-sum ε-approximation above holds, because:

$$\exists \rho' > 0 : \|\;\; - w'\|_\infty \leq \rho' \Rightarrow \forall x \in K, \left| e^{i\langle w, x\rangle} \hat{g}(w) - e^{i\langle w', x\rangle} \hat{g}(w') \right| \leq \frac{\varepsilon}{\mathrm{vol}(C)}$$

(indeed, K is bounded and both $t \mapsto e^{it}$ and $\hat{g}$ are uniformly continuous)

Therefore, for all $x \in K$,

$$\left| g(x) - \underbrace{\underline{\mathrm{Re}} \sum_{k=1}^{m^d} e^{i\langle w_k, x\rangle} \hat{g}(w_k)\, \mathrm{vol}(A_k)}_{} \right| \leq \left| g(x) - \sum_{k=1}^{m^d} e^{i\langle w_k, x\rangle} \hat{g}(w_k)\, \mathrm{vol}(A_k) \right| \leq 2\varepsilon$$

of the form $\sum_{k=1}^{m^d} \alpha_k \cos(\langle w_k, x\rangle + \theta_k)$ with $\alpha_k \in \mathbb{R}$ and $\theta_k \in [0, 2\pi]$

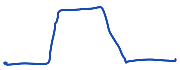
$$\overset{+\varepsilon}{\approx} \sum_{k=1}^{m^d} \alpha_k \left(\sum_{i=1}^{N} a_i \sigma\left(b_i [\langle w_k, x\rangle + \theta_k] + c_i\right) \right)$$

$$= \sum_{k=1}^{m^d} \sum_{i=1}^{N} \alpha_k a_i \sigma\left(\langle b_i w_k, x\rangle + b_i \theta_k + c_i\right). \quad \blacksquare$$

Examples of universal activation functions σ

alternative proof: via linear interpolation

(a) $\sigma(x) = \max\{x, 0\}$

→ approximate indicator function 

+ convex combination of two such bumps at the boundary between two cells in a partition of $[a,b]$.

(b) $\sigma \in \mathcal{C}^\infty(\mathbb{R},\mathbb{R})$ s.t. σ is not a polynomial.

→ Leshno et al. (1993), Step 3

→ further extensions in the same paper.

Dropping the assumption that K is compact?

Under mild assumptions on $f:\mathbb{R}^d \to \mathbb{R}$, the approximation result holds in L^∞_{loc}-norm, with the Lebesgue measure. loc $\Leftrightarrow$ bdd K

→ see Leshno et al. (1993), Theorem 1

→ csq, same paper, Proposition 1: $L^p(\mu)$, $1 \le p < +\infty$, μ proba with compact support, $\mu \ll \lambda$.

In $L^2(\mu)$-norm, there is a quantitative approximation result via an argument due to Maurey ($\overline{\mathrm{cv}(G)}$, $G \subset$ Hilbert space)

→ see Barron (1993), Lemma 1 in particular

→ the integral in (2) above is reinterpreted as a convex combination

NB: Barron assumes $\int \|w\| |\hat{f}(w)| \, dw < +\infty$, but $\int |\hat{f}(w)| \, dw < +\infty$ seems to work too (weaker assumption provided $f \in L^1(\mathbb{R}^d)$).