Neural networks from the point of view of function approximation theory

Neural net as a composition of simple functions

- A complicated mapping can be expressed as composition of simple mappings.
- The idea of approximation of complicated functions using simple functions was studied years ago by many mathematicians such as Hilbert or Kolmogorov.
- For neural nets, the most important employed simple functions are:
  - Logistic activation function on $\mathcal{R}$, $f(x) = \frac{1}{1+e^{-x}}$
  - Radial basis function (RBF) on $\mathcal{R}^{|I|}$, $f_v(x) = \exp\big(-\frac{1}{2} x^T \Sigma_v x \big)$
  - RBF with Σ_v as identity matrix, $f_v(x) = e^{-\frac{1}{2}||x||^2}$

Hilbert's 13th problem (1900)

- Introduced at the 2nd world mathematical congress in Paris as one of 23 most important open problems of mathematics.
- Hilbert considered the seventh-degree equation:

$$x^7 + ax^3 + bx^2 + cx + 1 = 0$$

  and asked whether its solution, x, considered as a function of the three variables a, b and c, can be expressed as the composition of a finite number of arbitrary finite sums and, apart from them, only at most two-variable functions.
- His conjecture was that the answer is negative.

Kolmogorov-Arnold representation theorem (1957)

- Showed that Hilbert's conjecture was wrong and proved that the composed functions, apart from sums, can be even of only 1 variable.
- Let $k \in \mathcal{N}, k \geq 2$, and $C(\langle 0,1\rangle^k)$ denotes class of continuous functions on the $k-$dimensional unit cube $\langle 0,1\rangle^k$. Then, there exist $k(2k+1)$ continuous functions on $\langle 0,1\rangle$, $h_{1,1},...,h_{1,2k+1},h_{2,1},....h_{k,2k+1}$ such that

$$(\forall f \in \mathcal{C}(\langle 0,1\rangle^k))(\exists g_1,...,g_{2k+1} - \text{functions continuous}$$

$$\text{on a suitable subset of } \mathcal{R})(\forall x \in \langle 0,1\rangle^k)$$

$$f(x) = \sum_{j=1}^{2k+1} g_j(\sum_{i=1}^{k} h_{i,j}(x_i))$$

Vitushkin theorem (1954)

- However, Kolmogorov theorem can not be generalized to continuously differentiable functions.
- This would contradict the Vitushkin theorem:
- Let $r, k \in \mathcal{N}, k \geq 2$. Then there exist r-times continuously differentiable functions of k variables, that can not be expressed as the composition of a finite number of arbitrary finite sums and, apart from them, only of function of at most k-1 variables.

Multilayer perceptron - function approximation

- We will discuss MLP with the activations

$$z_v = f\left(\sum_{u \in i(v)} w_{(u,v)} z_u + \theta_v\right)$$

- For further analysis, we need the following notation:
- Set of all linear functionals on $\mathcal{R}^k$

$$\mathcal{L}_k = \{\varphi : \mathcal{R}^k \to \mathcal{R} \& (\exists a \in \mathcal{R}^k)(\exists b \in \mathcal{R})(\forall x \in \mathcal{R}^k)\varphi(x) = a^T x + b\}$$

- Linear span of a tuple of vectors $(\xi_1, ..., \xi_n)$.

$$[\xi_1, ... \xi_n] = \{\xi : (\exists \alpha_1, ..., \alpha_n \in \mathcal{R})\xi = \sum_{k=1}^{n} \alpha_k \xi_k\}$$

Important sets of functions

- For each $k, n \in \mathcal{N}$ and each function $f : \mathcal{R} \to \mathcal{R}$

$$\Lambda_k^{(n)}(f) = \bigcup_{\xi_1 \in \mathcal{L}_k} \dots \bigcup_{\xi_n \in \mathcal{L}_k} [f \circ \xi_1, \dots, f \circ \xi_n]_\lambda$$

$$\Lambda_k(f) = \bigcup_{n=1}^{\infty} \Lambda_k^{(n)}(f)$$

- For each set of functions Φ on set X and for each subset $Y \subset X$ the symbol $\Phi|Y$ represents restriction of Φ to Y.

$$\Phi|Y = \{\psi : (\exists \varphi \in \Phi)\psi = \varphi|Y\}$$

Important Banach spaces I

- Banach space $L_p(\mu)$, $p \geq 1, \mu$ is a finite measure on $\mathcal{R}^k$

$$L_p(\mu) = \{\varphi : \mathcal{R}^k \to \mathcal{R} \& \int_{\mathcal{R}^k} |\varphi|^p d\mu < +\infty\}$$

- Banach space $C(X)$, where $X \subset \mathcal{R}^k$ is bounded closed (i.e., compact)

$$C(X) = \{\varphi : X \to \mathcal{R} \& \varphi \text{ is a continuous function on } X\}$$

Important Banach spaces II

- Let $k \in \mathcal{N}, f : \mathcal{R}^k \to \mathcal{R}, x \in \mathcal{R}^k$ and $\alpha = (\alpha_1, ..., \alpha_k) \in \mathcal{N}_0^k$. If the partial derivative $\frac{\partial^{\alpha_1 + ... + \alpha_k} f}{\partial^{\alpha_1} x_1 ... \partial^{\alpha_k} x_k}$ exists, we will denote it as

$$D^\alpha f = \frac{\partial^{\alpha_1 + ... + \alpha_k} f}{\partial^{\alpha_1} x_1 ... \partial^{\alpha_k} x_k}$$

- Let $k \in \mathcal{N}, \mu$ be a non-negative measure on $\mathcal{R}^k$ and a set $S \subset R^k$ fulfills $\mu(\mathcal{R}^k \setminus S) = 0$, then S is called support of the measure μ.
- Let $k \in \mathcal{N}, p \geq 1, m \in \mathcal{N}_0$. Define

$$C^{m,p}(\mu) = \{\varphi : \mathcal{R}^k \to \mathcal{R} \& (\forall \alpha \in \mathcal{N}_0^k) ||\alpha|| \leq m \Rightarrow \int_{\mathcal{R}^k} |D^\alpha \varphi|^p d\mu < +\infty\}$$

- This space, more precisely the space of disjoint classes of functions that are equal almost surely with respect to μ, is called Sobolev space.

Important Banach spaces III

- We can see that $L_p(\mu)$ is a special case of $C^{m,p}(\mu)$ for $m = 0$.
- Let $X \subset \mathcal{R}^k$ be a compact set, then:

  $$C^m(X) = \{\varphi : X \to \mathcal{R} \& (\forall \alpha \in \mathcal{N}_0^k) ||\alpha|| \leq m \Rightarrow D^\alpha \varphi \text{ is continuous on } X\}$$

  is a Banach space.
- We can see that $C(X)$ is a special case of $C^m(X)$ for $m = 0$.

Corresponding networks

- From the definition of $\Lambda_k^{(n)}(f)$ follows that $\Lambda_k^{(n)}(f)$ is a set of all mappings that can be computed by a MLP with k input neurons, one hidden layer with n neurons and one output neuron.
- We assume that the output neuron is linearly dependent on the neurons in the hidden layer, i.e., the activation function is identity.

Let a function $f : \mathcal{R} \to \mathcal{R}$ be Borel measurable, non-constant and bounded. Let $k \in \mathcal{N}, p \in \langle 1, \infty), X \subset R^k$ be a compact set and μ be a finite Borel measure defined on $\mathcal{R}^k$. Then:

1. $\Lambda_k(f)$ is dense in $L_p(\mu)$,
2. if f is continuous, $\Lambda_k(f)|X$ is dense in $C(X)$.

Differentiability vs. approximation

- We would like to see whether the differentiability of a function f can be refelcted in its approximation by $\Lambda_k(f)$.
- We can show that $L_p(\mu)$ and $C(X)$ can be replaced with analogous spaces of differentiable functions.

Let $m \in \mathcal{N}$ and a function $f \in C^m(R)$ be non-constant and bounded. Let $k \in \mathcal{N}, p \in \langle 1, \infty), X \subset R^k$ be a compact set and μ be a finite Borel measure defined on R^k. Then:

1. $\Lambda_k(f)|X$ is dense in $C^m(X)$,
2. if all partial derivations are bounded up to a degree m, then $\Lambda_k(f)$ is dense in $C^{m,p}(\mu)$,
3. if μ has a compact support, then $\Lambda_k(f)$ is dense in $C^{m,p}(\mu)$.

Approximation with sigmoid activation functions I

- Commonly, as sigmoid function is known any function f such that:

$$f : \mathcal{R} \rightarrow < L, U > \ \& \ f \text{ is non-decreasing Borel measureable } \ \&$$
$$L < U \ \& \lim_{t \rightarrow -\infty} f(t) = L \ \& \lim_{t \rightarrow +\infty} f(t) = U$$

- logistic function
- arctan function

$$f(x) = \frac{1}{\pi} \text{arctan}(x) + \frac{1}{2}$$

- Usually, it is also required that a sigmoid function is non-decreasing.
- Any sigmoid function is borel measurable, non-constant and bounded. Therefore, the theorems from the previous slides can be applied. However, it allows and additional kind of aproximations, more similar to Kolmogorov theorem.

Approximation with sigmoid activation functions II

Let $k \in \mathcal{N}, k \geq 2$ and $f : R \to \langle 0, 1\rangle$ be a sigmoid function. Let

$$\Sigma(f) = \Big\{ s : \langle 0,1\rangle^k \to \mathcal{R} \& (\exists g, h_1, \ldots, h_k \in \Lambda_1(f))(\forall x \in \langle 0,1\rangle^k)$$

$$s(x) = g\Big(\sum_{i=1}^{k} h_i(x_i) \Big) \Big\}$$

Then:

$$\bigcup_{n=1}^{\infty} \bigcup_{\xi_1,\ldots,\xi_n \in \Sigma(f)} [\xi_1, \ldots, \xi_n]_\lambda \text{ is dense in } C(\langle 0,1\rangle^k).$$

Corresponding networks

- We get a set of all mappings that can be computed by incompletely connected MLPs with the following properties:
  - k input neurons,
  - 1 output neuron,
  - each hidden neuron is connected with exactly one input neuron,
  - activation function f is assigned to hidden neurons.
- As to $\Sigma(f)$:
  - 1 layer of k hidden neurons,
- As to $\bigcup_{n=1}^{\infty} \bigcup_{\xi_1,\dots,\xi_n \in \Sigma(f)} [\xi_1, \dots, \xi_n]_\lambda$:
  - 2 layers of hidden neurons,
  - the 1st layer of hidden neurons contains k-times as many hidden neurons as the 2nd layer.