MSRI Summer School — **Precept 1: Approximation in Deep Learning**

Instructor: J. Bruna

TA: S. Jelassi

Problem 1: Universal approximation theorem

The universal approximation theorem informally states that a feed-forward network with a single hidden layer containing an infinite number of neurons can approximate continuous functions on compact subsets of $\mathbb{R}^n$, under mild assumptions on the activation function. In this exercise, we propose to prove the original statement by Cybenko, 1989.

Let I_n be the n-dimensional unit hypercube $[0,1]^n$ and $C(I_n)$ the space of real-valued continuous functions on I_n. We consider S the set of 1-hidden layer networks which correspond to the functions F for which there exist $N > 0$ such that

$$F(x) = \sum_{i=1}^{N} \alpha_i \sigma(w_i^\top x + b_i), \tag{1}$$

where $w \in \mathbb{R}^n$ and $\alpha_i, b_i \in \mathbb{R}$ are the parameters of the network. It is easy to see that S is a linear subspace of $C(I_n)$. We assume that the activation function σ is continuous and discriminatory i.e. if for a signed measure $\mu \in \mathcal{M}(I_n)$

$$\int_{I_n} \sigma(w^\top x + b) d\mu(x) = 0, \tag{2}$$

for all $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$, then $\mu = 0$. The universal approximation theorem states that given any function $f \in C(I_n)$, for all $\epsilon > 0$, there exist $N_\epsilon > 0$ and $F \in S$ such that

$$|F(x) - f(x)| < \epsilon, \quad \forall x \in I_n. \tag{3}$$

Part I: Proof of the theorem

We assume by contradiction that $\bar{S} \neq C(I_n)$, where $\bar{S}$ is the closure of S. In other words, $\bar{S}$ is a closed, proper subspace of $C(I_n)$.

I.1 Application of the Hahn-Banach theorem

1.1) Explain why $\bar{S} \bigoplus \mathbb{R}\varphi$ where $\varphi \in C(I_n) \backslash \bar{S}$ is a direct sum.

1.2) We define the mapping Φ such that

$$\begin{aligned} \Phi \colon \bar{S} \bigoplus \mathbb{R}\varphi &\to \mathbb{R} \\ f = g + \alpha\varphi &\mapsto \alpha, \end{aligned} \tag{4}$$

and p_Φ as

$$\begin{aligned} p_\Phi \colon C(I_n) &\to \mathbb{R} \\ h &\mapsto \|\Phi\| \cdot \|h\|_\infty. \end{aligned} \tag{5}$$

where $\|\Phi\| = \sup_{\|f\|_\infty \leq 1} |\Phi(f)|$ and $\|h\|_\infty = \sup_{x \in I_n} |h(x)|$. Show that by applying Hahn-Banach theorem with these choices of functionals, there exists a bounded linear functional F such that $F|_{\bar{S}} = 0$ and $F \neq 0$.

I.2 Application of the Riesz representation theorem

3) By using the Riesz Representation theorem on F and the discriminatory property (2), prove that there is a contradiction with $\bar{S} \neq C(I_n)$.

I.3 Conclusion

4) In I.2, we proved that $\bar{S} = C(I_n)$. Deduce that (3) holds.

Part II: Discussion

5) Assume that σ is the complex exponential function i.e. $\sigma(x) = e^{ix}$. Is it a discriminatory function? What is the analog of the universal approximation theorem in this case?

6) Give an example of a discriminatory and a non-discriminatory activation function σ.

Hint: For simplicity, you can assume that the parameters are one-dimensional i.e. $w, b \in \mathbb{R}$.

i. Show that the ReLU function is discriminatory by constructing a sigmoid function from the ReLU function and using the fact that sigmoid functions are discriminatory.

ii. Show that a polynomial of degree m is non-discriminatory.

N.B: In this exercise, we showed that a neural network with infinite number of neurons can approximate any continuous function. It is important to notice that the convergence rate for this approximation is very slow since it is exponential in the dimension.

Problem 2: The Power of Depth in feedfoward neural networks

The XOR function (”exclusive or”) is an operation on two binary values x and y. When exactly one of these binary values is equal to 1, the XOR function returns 1. Otherwise, it returns 0. In this exercise, we assume that the XOR function provides the target function $y = f^*(x)$ and we want to learn this target by using a 0-hidden layer and a 1-hidden layer neural network.

1) Write all the possible inputs and outputs of the XOR function.

2) Show that the XOR function cannot be approximated by a 0-hidden layer neural network.

3) Construct a 1- hidden layer neural network that approximates the XOR function.

Hint: Write the XOR function as a composition of simpler binary functions. It is then

possible to express these binary functions using neurons with well-chosen parameters and activation functions.
4) Does this contradict the previous universal approximation result? Justify your answer.

N.B: This exercise underlines the importance of depth in neural networks for approximation. Eldan and Shamir, 2016 and Safran and Shamir, 2017 provide examples of functions that can be approximated with a 2-hidden layers neural network having polynomial width but cannot be approximated by a 1-hidden layer neural networks with exponential width.

Problem 3: Learning with an infinite number of neurons

Analyzing the problem of learning a target function with a 1-hidden layer is in general challenging due to the non-convexity of the problem. Bach, 2017 proposed a method to "convexify" the problem by considering 1-hidden layer neural network having an infinite number of neurons. This can be done by restricting the search space to two possible functional spaces. The goal of this exercise is to understand in a simplified setting the difference between these two spaces. The functional space $\mathcal{G}_1$ is defined as

$$\mathcal{G}_1 := \left\{ g\colon \mathbb{S}^d \to \mathbb{R} \,\middle|\, g(z) = \int_{\mathbb{S}^d} \sigma(v^\top z)\, \mu(dv) \text{ with } \|\mu\|_{\mathrm{TV}} < \infty \right\} \tag{6}$$

where $\mathbb{S}^d \subset \mathbb{R}^{d+1}$ is the unit sphere of the Euclidean norm, $\sigma(u) = \max\{0, u\}$ is the ReLU function, μ is a signed Radon measure on $\mathbb{S}^d$ from which the neurons v are sampled. $\|\mu\|_{\mathrm{TV}}$ is the *total variation* of μ, defined as

$$\|\mu\|_{\mathrm{TV}} := \sup_{|u(v)|\leq 1} \int_{\mathbb{S}^d} u(v)\mu(dv) \ .$$

On the other hand, he defines the functional space $\mathcal{G}_2$

$$\mathcal{G}_2 := \left\{ g\colon \mathbb{S}^d \to \mathbb{R} \,\middle|\, \exists p\colon \mathbb{S}^d \to \mathbb{R},\ g(z) = \int_{\mathbb{S}^d} p(v)\sigma(v^\top z)\, d\tau(v) \right\}, \tag{7}$$

where p is here a $d\tau$-squared-integrable density function $p \in L^2(\mathbb{S}^d, d\tau)$. Remark that $\mathcal{G}_1$ and $\mathcal{G}_2$ just differ by the existence of a density function. $\mathcal{G}_2$ is defined relative to a base measure τ whereas $\mathcal{G}_1$ is 'intrinsic'.

He also associate two norms γ_1 and γ_2 respectively to $\mathcal{G}_1$ and $\mathcal{G}_2$ that are defined as

$$\gamma_1(g) = \inf_{g(z)=\int_{\mathbb{S}^d} \sigma(v^\top z)\mu(dv)} \|\mu\|_{\mathrm{TV}}, \tag{8}$$

$$\gamma_2^2(g) = \inf_{\substack{p \in L^2(\mathbb{S}^d) \\ g(z)=\int_{\mathbb{S}^d} \sigma(v^\top z)p(v)d\tau(v)}} \int_{\mathbb{S}^d} |p(v)|^2 d\tau(v). \tag{9}$$

Part I: Differences between $\mathcal{G}_1$ and $\mathcal{G}_2$

1) Show that $\mathcal{G}_2 \subset \mathcal{G}_1$. *Hint: Consider a function $f \in \mathcal{G}_2$ and show that $\gamma_1(f) \leq \gamma_2(f)$.*

2) Give an example of function that belongs to $\mathcal{G}_1$ but not to $\mathcal{G}_2$.

Hint: Consider the function $g_\epsilon(z) = \int_{I_n} \sigma(v^\top z)p_\epsilon(v)d\tau(v) = \int_{I_n} \epsilon^{-d}\sigma(v^\top z)p(\epsilon^{-1}v)d\tau(v)$ where p is a density with bounded L_2-norm. Show that $\gamma_1(g_\epsilon) < \infty$ and $\gamma_2(g_\epsilon) = \epsilon^{-d/2}\gamma_2(g)$ where $g(z) = \int_{I_n} \sigma(v^\top z)p(v)d\tau(v)$. When $\epsilon \to 0$, this implies that $\gamma_2(g_\epsilon) \to \infty$ meaning that $g_\epsilon \notin \mathcal{G}_2$.

Part II: $\mathcal{G}_2$ is a RKHS with norm γ_2.

II.1 $\mathcal{G}_2$ is a RKHS.

We want to show that $\mathcal{G}_2$ is a RKHS with kernel $k(x, y) = \int_{\mathbb{S}^d} \sigma(v^\top x)\sigma(v^\top y)d\tau(v)$. We consider a linear mapping $T\colon L_2(d\tau) \to \mathcal{G}_2$ defined by $(Tp)(x) = \int_{\mathbb{S}^d} p(v)\sigma(v^\top x)d\tau(v)$ with null space $\mathcal{K}$.

3) Explain why we can define a bijection U from $\mathcal{K}^\perp$ to $\mathcal{G}_2$.

In what follows, we define a dot-product on $\mathcal{G}_2$ as $\langle f, g\rangle = \int_{\mathbb{S}^d}(U^{-1}f)(v)(U^{-1}g)(v)d\tau(v)$.

4) To what functional space does $k(\cdot, y)$ belong for all $y \in \mathbb{S}^d$?

5) For any $y \in \mathbb{S}^d$, we set $p = U^{-1}k(\cdot, y) \in \mathcal{K}^\perp$. Let q defined as $q\colon v \mapsto \sigma(v^\top y) \in L_2(d\tau)$. To what functional space does $p - q$ belong? Justify your answer.

6) Show that for $f \in \mathcal{G}_2$, $\langle f, k(\cdot, y)\rangle = f(y)$. This is called the reproducing property and allows to show that $\mathcal{G}_2$ is a RKHS.

II.2 γ_2 is the norm associated to the RKHS.

Now, we want to show that the RKHS norm of $\mathcal{G}_2$ is γ_2. For any $f \in \mathcal{G}_2$ such that $f = Tp$ for $p \in L_2(d\tau)$ we have $p = U^{-1}f + q$ where $q \in \mathcal{K}$.

7) Show that $\int_{\mathbb{S}^d} p(v)^2 d\tau(v) = \|f\|^2 + \|q\|^2_{L_2(d\tau)}$, where $\|\cdot\|$ is the norm induced by the dot-product defined on $\mathcal{G}_2$. Deduce that $\gamma_2(f) = \|f\|$.

N.B: Learning a target function over the functional space $\mathcal{G}_1$ (resp. $\mathcal{G}_2$) amounts to fit a 1-hidden layer network with infinite width but with a L^1-penalty (resp. L^2-penalty) on the number of neurons. As we have seen on some examples, $\mathcal{G}_2$ is smaller than $\mathcal{G}_1$. However, algorithms for $\mathcal{G}_2$ are significantly more efficient since this functional space is a RKHS. One can then use the usual RKHS representer theorem or the random features technique

(Rahimi and Recht, 2008) to approximate the target function. Algorithms for $\mathcal{F}_1$ are of the type conditional gradient (a.k.a. Frank-Wolfe) and may be exponential in time. An interesting trade-off between approximation and optimization appears here.

Functional analysis tools

Sublinear function: Given a real vector space X, a function $p\colon X \to \mathbb{R}$ is called sublinear if

1. Positive homogeneity: $p(\lambda x) = \lambda p(x)$ for all $\lambda \geq 0$ and $x \in X$.
2. Subadditivity: $p(x+y) \leq p(x) + p(y)$ for all $x, y \in X$.

Hahn-Banach theorem: Let X be a real vector space, $p\colon X \to \mathbb{R}$ a sublinear functional, M a subspace of X, and $\Phi\colon M \to \mathbb{R}$ a linear functional such that $\Phi(x) \leq p(x)$ for all $x \in M$. Then there exists a linear functional $F\colon X \to \mathbb{R}$ such that $F(x) \leq p(x)$ for all $x \in X$ and $F|_M = \Phi$.

Riesz representation theorem: If I is a linear functional on $C(X)$ with compact support, there is a unique signed measure μ on X such that $I(f) = \int f d\mu$ for all $f \in C(X)$ with compact support.

Sigmoid function: A function $\sigma\colon \mathbb{R} \to \mathbb{R}$ is called a sigmoid if it satisfies the two following properties: $\lim_{t\to\infty} \sigma(t) = 1$ and $\lim_{t\to-\infty} \sigma(t) = 0$.

Sigmoid functions are discriminatory: Any bounded measurable sigmoid function is discriminatory.

Reproducing Kernel Hilbert Space (RKHS): Let X be an arbitrary set and $(H, \langle\cdot,\cdot\rangle)$ a Hilbert space of real-valued functions on X. H is a RKHS if for all $x \in X$, there exist a unique $K_x \in H$ such that $f(x) = \langle f, K_x \rangle$ for all $f \in H$.

References

Bach, Francis (2017). "Breaking the curse of dimensionality with convex neural networks." In: *The Journal of Machine Learning Research* 18.1, pp. 629–681.

Cybenko, George (1989). "Approximation by superpositions of a sigmoidal function." In: *Mathematics of control, signals and systems* 2.4, pp. 303–314.

Eldan, Ronen and Ohad Shamir (2016). "The power of depth for feedforward neural networks." In: *Conference on learning theory*, pp. 907–940.

Rahimi, Ali and Benjamin Recht (2008). “Random features for large-scale kernel machines.” In: *Advances in neural information processing systems*, pp. 1177–1184.

Safran, Itay and Ohad Shamir (2017). “Depth-width tradeoffs in approximating natural functions with neural networks.” In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, pp. 2979–2987.