Foundations of Deep Learning

Lecture 02

Approximation Theory

Aurelien Lucchi

Fall 2024

Section 1

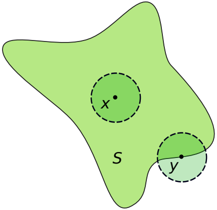
Universality

Review - Topology

Let S be a subset of a topological space X.

Limit point. $p \in X$ is a limit point of S if every open neighborhood of p contains one point in S (other than p).

Closure. The closure of a set E is the union of all its limit points. It is usually denoted by $\bar{E}$ or $cl(E) = E \cup E'$, where E' is the set of all limit points.



Review - Topology

Dense set. A subset A of a topological space X is called dense (in X) if every point $x \in X$ either belongs to A or is a limit point of A. Alternatively, A is dense if it has **non-empty intersection** with an arbitrary non empty open subset $B \subset X$.

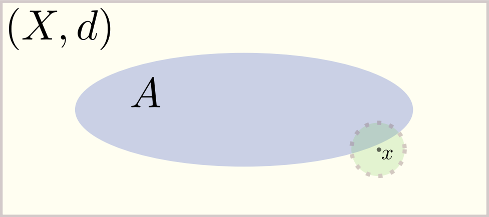


Figure: The set A shown in blue is dense in X if every $x \in X$ is a limit point of A.

Review - Topology

A well-known example is the fact that the rationals are dense in the set of reals, which we formalize in the next theorem.

Theorem 1 (Density theorem, $\mathbb{Q}$ is dense on $\mathbb{R}$)

$$\forall a < b \in \mathbb{R}, \exists x \in \mathbb{Q} \textit{ s.t. } x \in (a, b), \textit{ i.e. } a < x < b$$

Review - Topology

Compact set. There are typically two characterizations of compact spaces, one in terms of open sets and another one in terms of convergent sequences. We start with the definition in terms of open sets.

Definition 2 (Compact set, definition 1)

A topological space X is called compact if each of its open covers [a] has a finite subcover.

[a] A cover of a set X is a collection of sets whose union includes X as a subset.

Review - Topology

Explicitly, this means that for every arbitrary collection $\{U_\alpha\}_{\alpha \in A}$ of open subsets of X such that $X = \bigcup_{\alpha \in A} U_\alpha$, there is a **finite** subset J of A such that $X = \bigcup_{i \in J} U_i$.

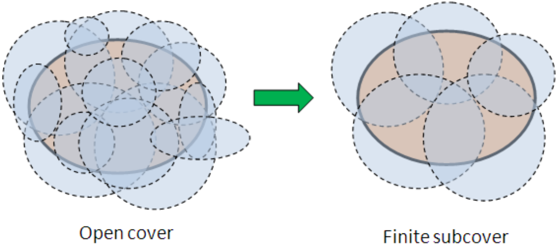


Figure: Source: `https://mathstrek.blog/`

Review - Topology

Compact set.

Another definition in terms of convergent sequences is given below.

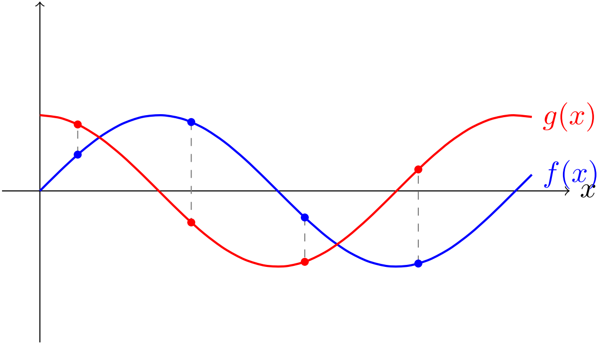
Definition 3 (Compact set, definition 2)

We call X a compact set if all sequences $(f_n)_{n\geq 1} \subset X$ have a convergent subsequence $(f_{n(k)})$ with limit point in X.

Finally, we note that the following characterization of compact sets is often used in the literature: a subset of $\mathbb{R}^d$ is compact if it is closed and bounded (Heine-Borel theorem).

Sup-Norm

How do we measure the quality of approximation?



Sup-Norm

How do we measure the quality of approximation?

Supremum norm

$$\boxed{\|f\|_\infty := \sup_{x \in S} |f(x)|\,, \quad \text{approx-err}(f, g) := \|f - g\|_\infty\,.}$$

Extension to approximating function classes $\mathcal{G}$

$$\boxed{\text{approx-err}(f, \mathcal{G}) := \inf \left\{ g \in \mathcal{G} : \text{approx-err}(f, g) \right\}\ .}$$

Approximability

Consider family of models $\mathcal{G}_m$, $m = 1, 2, \ldots$; $\mathcal{G} = \bigcup_m \mathcal{G}_m$.

E.g. MLPs with m hidden units.

$$\boxed{f \simeq \mathcal{G} : \quad \text{approx-err}(f, \mathcal{G}) = 0}$$

Clearly:

$$\Rightarrow \boxed{f \in \mathcal{G} \quad \Longrightarrow \quad f \simeq \mathcal{G}\,.}$$

But also:

$$\Rightarrow \boxed{\mathcal{G} \ni g_m \xrightarrow{\text{unif.}} f \Longrightarrow f \simeq \mathcal{G}}$$

where **uniform convergence** is defined as

$$\boxed{(g_m) \xrightarrow{\text{unif.}} f \iff \forall \epsilon > 0 : \ \exists m \geq 1 : \ \|g_m - f\|_\infty < \epsilon\,.}$$

Denseness

Generalizing to function classes: denseness

$$\boxed{\begin{array}{l} \mathcal{G} \subseteq \mathcal{F} \text{ is dense in } \mathcal{F} \iff \mathcal{F} \simeq \mathcal{G} \\ \qquad\quad\iff \\ \forall f \in \mathcal{F} : f \simeq \mathcal{G} \end{array}}$$

The largest class of approximated functions is the closure $\mathrm{cl}(\mathcal{G})$.

$$\Rightarrow \boxed{\mathrm{cl}(\mathcal{G}) \simeq \mathcal{G} \quad \text{and} \quad \mathcal{F} \simeq \mathcal{G} \Longrightarrow \mathcal{F} \subseteq \mathrm{cl}(\mathcal{G})}$$

Universal Approximator

Continuous functions $C(\mathbb{R}^n)$:

$\mathcal{G}$ is a universal approximator $\iff$

$C(S) \simeq \mathcal{G}(S)$ for any compact $S \subset \mathbb{R}^n$

$\mathcal{G}_S$: restriction of functions in $\mathcal{G}$ to S.

One also says: $\mathcal{G}$ is dense in $C(\mathbb{R}^n)$ in the topology of uniform convergence on compacta.

Section 2

Elementary folklore construction

Our plan

We will start with some **constructive results**, where we explicitly choose the form of the function used to approximate a given target function.

1. First discuss the unitary case where $f : [0, 1] \to \mathbb{R}$
2. Then discuss the extension to the multivariate case where $f : \mathbb{R}^d \to \mathbb{R}$

Univariate case

Theorem 4 ([Telgarsky(2021)])

Let $\epsilon > 0$ and assume $g : [0,1] \to \mathbb{R}$ is ρ-Lipschitz and f is a 2-layer neural network with $\lceil \frac{\rho}{\epsilon} \rceil$ threshold nodes. Then we have

$$\sup_{x \in [0,1]} |f(x) - g(x)| \leq \epsilon.$$

Proof

Multivariate case

First present an auxiliary lemma that allows us to approximate continuous functions with piecewize constant functions.

Lemma 5

Let g, δ, ϵ be given as in Theorem 4. Assume we are given $U \subseteq \mathbb{R}^d$ and a partition $\mathcal{P}$ of U into rectangles of side lengths at most δ, i.e. $\mathcal{P} = (R_1, \dots, R_n)$. Then $\exists(\alpha_1, \dots \alpha_n)$ such that $\sup_{\mathbf{x} \in U} |g(\mathbf{x}) - h(\mathbf{x})| \leq \epsilon$ where

$$h(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \mathbf{1}_{R_i}(\mathbf{x}) \text{ and } \mathbf{1}_{R_i}(\mathbf{x}) = \begin{cases} 1 \textit{ if } \mathbf{x} \in R_i \\ 0 \textit{ else.} \end{cases}$$

Proof

Multivariate case

Theorem 6 ([Telgarsky(2021)])

Consider a continuous function $g : \mathbb{R}^{d} \to \mathbb{R}$. Assume that, given $\epsilon > 0$, there exists $\delta > 0$ such that for $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ with $\|\mathbf{x} - \mathbf{x}'\| \leq \delta$, then $|g(\mathbf{x}) - g(\mathbf{x}')| \leq \epsilon$. Then there exists a 3-layer neural network $f : \mathbb{R}^d \to \mathbb{R}$ with $\Omega(\delta^{-d})$ ReLU activation functions such that

$$\int_{[0,1]^d} |f(\mathbf{x}) - g(\mathbf{x})| dx \leq 2\epsilon.$$

Proof

Section 3

Weierstrass Theorem

Weierstrass Theorem

Theorem 7 (Weierstrass Theorem)

Polynomials $\mathcal{P}$ are dense in $C(I)$, where $I = [a; b]$ for any $a < b$.

Without loss of generality we can focus on $I = [0; 1]$. Define :

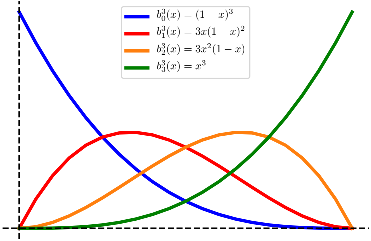
$$\phi : C([0; 1]) \to C([a; b]),\ \phi(f) = f \circ \phi,\ \phi = (1 - t)a + tb.$$

Proof of Weierstrass Theorem

Bernstein basis poynomials (degree m)

$$\boxed{b_k^m(x) = \binom{m}{k} x^k (1-x)^{m-k}\,.}$$

$$\Rightarrow \boxed{\sum_{k=0}^{m} b_k^m(x) = 1 \quad \forall x \in [0;1]}$$



Proof of Weierstrass Theorem

Define **approximation on lattice** with spacing $1/m$

$$\boxed{q_m(x) = \sum_{k=0}^{m} f\left(\frac{k}{m}\right) b_k^m(x), \quad b_k^m(x) = \binom{m}{k} x^k (1-x)^{m-k}\,.}$$

Consider residuals

$$|f(x) - q_m(x)| = \left|\sum_{k=0}^{m} r_k^m(x)\right|$$

$$r_k^m(x) := \left[f(x) - f\left(\frac{k}{m}\right)\right] b_k^m(x)$$

Proof of Weierstrass Theorem

Upper bound by splitting

$$\mathcal{I} := \{k : |x - \tfrac{k}{m}| \leq \delta\}, \quad \text{and} \quad \mathcal{I}^c,$$

where δ is chosen such that $|f(x) - f(y)| \leq \epsilon/2$ and thus

$$\sum_{k \in \mathcal{I}} r_k^m(x) \leq \frac{\epsilon}{2} \sum_{k \in \mathcal{I}} b_k^m(x) \leq \frac{\epsilon}{2} \sum_{k=1}^{m} b_k^m(x) = \frac{\epsilon}{2}$$

Note that by concentration

$$\sum_{k \notin \mathcal{I}} r_k^m(x) \leq R \sum_{k \notin \mathcal{I}} b_k^m(x) \xrightarrow{m \to \infty} 0$$

and we can choose m to make it $< \epsilon/2$. (See lecture notes)

Section 4

Approximation with smooth functions

Approximation with smooth functions (1d Case)

Let us now consider an **arbitrary smooth function** $\sigma \in C^\infty(\mathbb{R})$ and the span of its composition with affine functions

$$\boxed{\begin{aligned} \mathcal{G}_\sigma^1 &= \{g : g(x) = \sigma(ax+b) \text{ for some } a, b \in \mathbb{R}\} \\ \mathcal{H}_\sigma^1 &= \mathsf{span}(\mathcal{G}_\sigma^1) \end{aligned}}$$

Then the following holds:

Theorem 8 (Leshno. Lin, Pinkus, Schocken 1993)

For any $\sigma \in C^\infty(\mathbb{R})$ not a polynomial. $\mathcal{H}_\sigma^1$ is a universal approximator.

Approximation with smooth functions (1d Case)

1. Approximate derivative of σ: for all $h \neq 0$,

$$\frac{\sigma((a+h)x+b) - \sigma(ax+b)}{h} \in \mathcal{H}^1_\sigma$$

2. It follows that (generalizing to all k-th derivatives)

$$\frac{d^k}{da^k}\sigma(ax+b)\Big|_{a=0} = x^k \sigma^{(k)}(b) \in \mathrm{cl}(\mathcal{H}^1_\sigma)$$

3. If there is always a b_0 such that $\sigma^{(k)}(b_0) \neq 0$ then we are guaranteed that $x^k \in \mathrm{cl}(\mathcal{H}^1_\sigma)$ and hence all polynomials. (known theorem)
4. By the **Weierstrass theorem** this implies the result.

1d Case

What we have shown now is that for the case of $n = 1$ (one-dimensional inputs), an MLP with smooth activation function σ is a universal approximator, unless σ is a polynomial.

This is because the linear output layer is picking an element in the span of the hidden units, each one of which computes a function in $\mathcal{G}^1_\sigma$.

So as far as universal function approximation is concerned, there is nothing special about the logistic function or the hyperbolic tangent as choices of activation functions.

Sketch: 1d Case

Ridge Function Theorem (generalization to $\mathbf{x} \in \mathbb{R}^n$)

Define

$$\boxed{\begin{aligned} &\mathcal{G}_\sigma^n = \{g : g(\mathbf{x}) = \sigma(\boldsymbol{\theta} \cdot \mathbf{x}),\ \boldsymbol{\theta} \in \mathbb{R}^n\}, \quad \mathcal{H}_\sigma^n = \mathsf{span}(\mathcal{G}_\sigma^n) \\ &\mathcal{G}^n = \bigcup_{\sigma \in C(\mathbb{R})} \mathcal{G}_\sigma^n, \quad \mathcal{H}^n = \mathsf{span}(\mathcal{G}^n) \end{aligned}}$$

Theorem 9 (Vostrecov and Kreines, 1961)

$\mathcal{H}^n$ is a universal function approximator.

3-layer neural networks with adaptive activation functions "could be" universal approximators. However, this in itself is not practical. $\rightsquigarrow$ We will soon see how to remove this limitation using the concept of **dimension lifting**.

Sketch: Adaptive Activations

Dimension Lifting

Recall we showed earlier that for any $\sigma \in C^\infty(\mathbb{R})$ *not* a polynomial, $\mathcal{H}_\sigma^1$ is a universal approximator.

The following theorem allows us to extend the universality result we have for $n = 1$ to $\mathbb{R}^n$ for a single fixed $\sigma \in C^\infty(\mathbb{R})$.

Theorem 10 (Pinkus 1999)

For a **fixed** $\sigma \in C^\infty$,

$$\mathcal{H}_\sigma^1 \text{ universal for } C(\mathbb{R})$$

$$\implies \mathcal{H}_\sigma^n \text{ universal for } C(\mathbb{R}^n) \text{ for any } n \geq 1$$

Dimension Lifting

1. Fix f and compact $K \subset \mathbb{R}^n$. By Theorem 9, we can find ridge functions g_k s.t.

$$\left| f(\mathbf{x}) - \sum_{k=1}^{m} g_k(\boldsymbol{\theta}^k \cdot \mathbf{x}) \right| < \frac{\epsilon}{2} \quad (\forall \mathbf{x} \in K).$$

2. Since K is compact, $\boldsymbol{\theta}^k \cdot \mathbf{x} \in [\alpha_k, \beta_k]$ for $\mathbf{x} \in K$.
3. Because $\mathcal{H}_\sigma^1$ is dense in each $C([\alpha_k, \beta_k])$, Theorem 8 implies that we can find constants s.t.

$$\left| g_k(z) - \sum_{j=1}^{m_k} c_{kj} \sigma(a_{kj} z + b_{kj}) \right| \leq \frac{\epsilon}{2m} \quad (\forall k = 1, \ldots, m)$$

4. Plugging things together yields the result.

Summary

(1) For $n = 1$, an MLP with any continuous, non-polynomial activation function is a universal approximator.

(2) Spans of ridge functions are universal approximators for $C(\mathbb{R}^n)$.

(3) The non-linear part of any/each ridge function can be approximated according to (1).

(4) Hence MLPs are universal function approximators.

Matus Telgarsky.
Deep learning theory lecture notes.
https://mjt.cs.illinois.edu/dlt/, 2021.
Version: 2021-10-27 v0.0-e7150f2d (alpha).