Universal Approximation Theorems

Josef Teichmann
(joint work with Christa Cuchiero and Philipp Schmocker)

ETH Zürich

April 2021

Bernstein polynomials

A simple and beautiful application of the law of large numbers (LLN) is Sergey Bernstein's proof of Weierstrass approximation theorem:

A Bernstein polynomial of type (n, k) is defined by

$$B_{n,k}(x) = \binom{n}{k} x^k (1-x)^{n-k} \quad (k = 0, 1, \ldots, n)\,. \tag{1}$$

Then every continuous function f on $[0, 1]$ can be uniformly approximated by the following polynomial

$$B_n^f(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) B_{n,k}(x)\,,$$

where a quantitative estimate is given below.

Bernstein polynomials

Let (X_n) be a sequence of independent, identically distributed Bernoulli random variables with success parameter $x \in [0,1]$, then by LLN

$$\frac{X_1 + \ldots + X_n}{n} \to x$$

almost surely. We furthermore have

$$P\Big[X_1 + \ldots + X_n = k\Big] = B_{n,k}(x)\,.$$

Denote by S_n the sum $X_1 + \ldots + X_n$.

Bernstein polynomials

Whence

$$\begin{aligned} |B_n^f(x) - f(x)| &= \left| E\left[f\left(\frac{S_n}{n}\right) - f(x) \right] \right| \leq E\left[\left| f\left(\frac{S_n}{n}\right) - f(x) \right| \right] \\ &\leq 2 \sup_u |f(u)| \; P\left[\left| \frac{S_n}{n} - x \right| > \delta \right] \\ &+ \sup_{|u-v| \leq \delta} |f(u) - f(v)| \; P\left[\left| \frac{S_n}{n} - x \right| \leq \delta \right]. \end{aligned}$$

Since f is uniformly continuous we can bound the second term on the right hand side by ϵ for small enough δ. Due to Chebychev's inequality the first term is bounded by

$$2 \sup_u |f(u)| \frac{x(1-x)}{n\delta^2} \leq \frac{1}{2n\delta^2} \sup_u |f(u)| \leq \epsilon,$$

for n large enough. Therefore

$$\|B_n^f(x) - f(x)\|_\infty \longrightarrow 0 \text{ for } n \to \infty.$$

Weierstrass approximation theorem

This proves in particular the following theorem:

The polynomials are dense in $C([0,1]) = C([0,1],\mathbb{R})$ (Weierstrass approximation theorem).

A substantial generalization of this result tells that on compact topological Hausdorff spaces K every point separating subalgebra of the algebra of continuous functions $C(K) := C(K;\mathbb{R})$ is actually dense, too (Stone-Weierstrass approximation theorem). Point separating just means that for every two points $x \neq y$ there is a function $f \in A$ such that $f(x) \neq f(y)$.

There is an order theoretic version of this theorem and Bernstein's proof also paves the path towards a probabilistic version of this theorem.

Proof of the Stone Weierstrass approximation theorem

Let K be a compact topological Hausdorff space and let $A \subset C(K)$ be a point separating subalgebra ((sub-)algebras here always contain the 1). Let $f \in C(K)$ and $\epsilon > 0$ be fixed. Then we can proceed as follows:

- With $g \in A$, we have that $|g| \in \overline{A}$. Indeed $g(K) \subset [a, b]$ for some a, b, and take a polynomial p which approximates $x \mapsto |x|$ on $[a, b]$ up to accuracy ϵ. Then $\| |g| - p(g) \|_\infty \leq \epsilon$, however $p(g) \in A$.
- With $g, h \in A$ we have that $\max(g, h) = \frac{|g+h|}{2} + \frac{|g-h|}{2} \in \overline{A}$.
- With $g, g \in \overline{A}$ we have that $\max(g, h) \in \overline{A}$.

Proof of the Stone Weierstrass approximation theorem

- For every $x \in K$ we construct $f_x \in \overline{A}$ such that $f_x \leq f + \epsilon$ and $f_x(x) = f(x)$. Indeed we can find (point separation) for every $z \in K$ a function $g_{x,z} \in A$ with $g_{x,z}(x) = f(x)$ and $g_{x,z}(z) = f(z)$. Then there exists an open neighborhood $V_z \ni z$ such that $g_{x,z}|_{V_z} \leq f|_{V_z} + \epsilon$. Due to compactness there is a finite subcover of (V_z) indexed by $z_1, \ldots, z_n \in K$. Define now $f_x = \min(g_{x,z_1}, \ldots, g_{x,z_n}) \in \overline{A}$.
- With an analogue argument we can construct an open cover (U_x) such that $f_x \geq f - \epsilon$ on $U_x \ni x$, which has again a finite subcover indexed by $x_1, \ldots, x_m$. Define now $g = \max(f_{x_1}, \ldots, f_{x_m}) \in \overline{A}$, then $f - \epsilon \leq g \leq f + \epsilon$ globally.

Remarks

- We could equally take a point separating, linear subspace A such that with $f, g \in \overline{A}$ also $\max(f, g) \in \overline{A}$ (order theoretic version of the Stone Weierstrass approximation theorem).
- A probabilistic version could look as follows: let ν be a measure with full support on K and let $\mu_{n,x} = g_{n,x}\nu$ be a family of probability measures converging weakly to δ_x as $n \to \infty$, for $x \in K$. Assume that $x \mapsto g_{n,x}(y)$ is continuous for every y in the support of ν. Then the span of $x \mapsto g_{n,x}(y)$ is dense in $C(K)$.

Vector valued Stone Weierstrass approximation theorem

Let Y be a Banach space. Let $B \subset C(K;Y)$ be an A-submodule, where A a point separating subalgebra of $C(K)$. Assume furthermore that $(g(x))_{g\in B}$ is a dense family in Y for every $x \in K$. Then B is dense in $C(K;Y)$ (this is related to Nachbin's theorem).

The proof is simple: without restriction we can assume that $A = C(K)$ and that B is closed. Take $f \in C(K;Y)$ and choose $\epsilon > 0$. For every $x \in K$ choose $g_x \in B$ such that $g_x(x) = f(x)$. Then $(\{y \in K \mid \|f(y) - g_x(y)\| < \epsilon\})$ is an open cover of K which has a finite subcover indexed by $x_1, \ldots, x_n \in X$. Choose a partition of unity $\sum_i \psi_i = 1$ for this finite subcover, then $g := \sum_i \psi_i g_{x_i} \in B$ is approximating f up to accuracy ϵ.

Weighted Spaces

For several applications it is necessary to go beyond compact spaces. We therefore consider weighted spaces (E, ρ), i.e. topological Hausdorff spaces with $\rho : E \to \mathbb{R}_{\geq 1}$ such that $\{\rho \leq R\}$ is compact for all R, where a similar analysis as on compact spaces is possible.

We consider the closure $B^\rho(E)$ of bounded continuous functions $C_b(E; \mathbb{R}) = C_b(E)$ with respect to the ρ-norm

$$\|f\|_\rho := \sup_x \frac{|f(x)|}{\rho(x)} .$$

In a similar manner we can define $B^\rho(E; Y)$ for vector valued functions.

Stone Weierstrass approximation theorem for weighted spaces E

Let A a point separating subalgebra of $B^\rho(E)$ of bounded functions, then A is dense in $B^\rho(E)$.

The proof follows directly from the compact case: it is sufficient to show that $f \in C_b(E) \subset B^\rho(E)$ can be approximated by elements from A. Choose $R > 0$, then we can find $g \in A$, such that g is close to f on $\{\rho \leq R\}$ with distance less than $1 > \epsilon > 0$. Assume f has range bounded by M, whence there is a polynomial p which closely approximates on $[-M - \|g\|_\infty - 1, M + \|g\|_\infty + 1]$ a function being $x \mapsto x$ on $[-M - 1, M + 1]$ and bounded by $M + 1$ otherwise. Consequently $p(g) \in A$ is close to f with distance less than $\epsilon + \frac{M+1}{R}$, but now globally in ρ-norm (if R is chosen big enough such that M/R is small).

Vector valued Stone Weierstrass approximation theorem for weighted spaces E

Let Y be a Banach space. Let $B \subset B^\rho(E;Y)$ be an A-submodule, where A a point separating subalgebra of $B^\rho(E)$ of bounded continuous functions. Assume furthermore that $(g(x))_{g\in B}$ is a dense family in Y for every $x \in E$. Then B is dense in $B^\rho(E;Y)$.

Again without restriction we can assume that $A = B^\rho(E;Y)$ and again it is sufficient to show that $f \in C_b(E;Y) \subset B^\rho(E;Y)$ can be approximated by elements from B. Choose $R > 0$, then we can choose $g \in B$, such that g is close to f on $\{\rho \leq R\}$ with distance less than $1/3 > \epsilon > 0$. Assume without restriction that f has range bounded by $1/3$. The function $h = 1 \wedge \frac{1}{1/3+\|g\|}$ is bounded continuous on E, therefore it lies in A. $hg \in B$ is still close to f with distance less than $\epsilon + \frac{1}{R}$ but now globally (if R is chosen large enough as above).

Remark

- We can replace the Banach space Y by any locally convex vector space and obtain analogue results for the locally convex spaces of vector valued continuous functions on K or E, respectively.
- In the real valued case an order theoretic version is possible, too.
- In both cases we can generalize the assumptions to subsets A or B, respectively, of bounded functions, whose restrictions on compacts of the form $\{\rho \leq R\}$ contain a point separating subalgebra. The proof is analogous.

Universal approximation theorems

Universal approximation theorems aim for easy constructions of subalgebras or submodules on weighted spaces in order to apply Stone Weierstrass type approximation theorems.

We shall introduce the notion of *activating families* and *additive families* for this purpose.

Additive point separating families

Let E be a weighted space. A set of bounded continuous functions $\mathcal{L} \subset B^{\rho}(E)$ is called additive point separating family if it is closed under addition, contains the 1 and point separating.

We remark that this definition also makes sense for vector valued functions.

Activating families

Let Y be a Banach space. A family Φ of continuous functions $\varphi : \mathbb{R} \to \mathbb{R}$ is called activating if the space

$$A_\Phi := \Big\{ \sum_i \alpha_i \varphi_i(\beta_i . + \gamma_i) | \text{ for } \alpha_i \in Y,\ \beta_i, \gamma_i \in \mathbb{N}, \varphi_i \in \Phi \text{ and } n \in \mathbb{N} \Big\}$$

is dense in $C([0,1]; Y)$.

Typically Φ is a singleton ('an activation function'). Notice that it is sufficient that this property holds for $Y = \mathbb{R}$, since then it holds for all finite dimensional spaces, whence for all finite dimensional subspaces of Y, wherefrom the general assertion follows by vector valued Stone-Weierstrass on $[0,1]$.

UAT

Let Y be a Banach space, E a weighted space, Φ and activating family of functions and $\mathcal{L}$ an additive family, then

$$\mathrm{NN}_\Phi = \left\{ \sum_i \alpha_i \varphi_i(l_i) | \text{ for } \alpha_i \in Y,\ l_i \in \mathcal{L} \text{ and } n \in \mathbb{N} \right\}$$

is dense in $B^\rho(E; Y)$.

Proof of UAT

For the proof we have to show that the closure B of NN_Φ is a $B^\rho(E)$ submodule which satisfies the condition that $(g(x))_{g\in B}$ is dense for every $x \in E$.

Assume first that $Y = \mathbb{R}$, then the algebra A generated by $\mathcal{L}$ is point separating and therefore dense. This algebra, however, lies in the closure of NN_Φ. Indeed consider $l \in \mathcal{L}$, then $\sin(l)$ as well as $\cos(l)$ lie in the closure since we can approximate *sin* and *cos* by functions from A_Φ uniformly (notice that l has bounded range). Therefore $\sin(k_1 l_1 + \cdots + k_n l_n)$ and $\cos(k_1 l_1 + \cdots + k_n l_n)$ lie in the closure, for $l_i \in \mathcal{L}$ and $k_i \in \mathbb{N}$ (addivity!). By uniform trigonometric approximation we obtain therefore that all polynomials of elements from $\mathcal{L}$ lie in the closure.

For the general case it is sufficient to show it for finite dimensional subspaces of Y, where it clearly holds.

Remark

- We could replace $[0, 1]$ in the definition of activating families above by compacts in a weighted topological vector space Z. Of course members of the additive family have to map into compacts of Z then. A completely analogous result holds true then.
- We can also consider activating families of functions $\varphi : Z \to Z$, α_i should then be linear maps from Z to Y. The important property of the dense subspace of $B^\rho(K \subset Z; Y)$ is its invariance under affine maps.
- Elements of NN_Φ are called neural networks with activating family Φ initialized by $\mathcal{L}$.
- The space of real valued neural networks NN_Φ is again an additive family. Whence deeper networks are dense, too.

Activating families

- If $\varphi : \mathbb{R} \to \mathbb{R}$ is a discriminatory function, i.e. a Borel measure is vanishing if and only if

$$\int phi(\beta x + \gamma)\mu(dx) = 0$$

  for integers $\beta, \gamma \in \mathbb{N}$, then $\Phi = \{\varphi\}$ is an activating family.

- If the Fouriertransform of $\varphi : \mathbb{R} \to \mathbb{R}$ only vanishes at 0, then $\Phi = \{\varphi\}$ is an activating family.

- If $\varphi(x) = \max(0, x)$, then $\Phi = \{\varphi\}$ is an activating family.

Different topologies

From the previous results we can also conclude densities in L^p spaces or C^k or Besov spaces $B^{\alpha,p}$.