Mathematics for New Technologies in Finance

Solution sheet 2

Exercise 2.1 (Stone-Weierstrass theorem [1])

(a) Construct a sequence of polynomials converges pointwisely but not uniformly on $[0, 1]$.

(b) Construct a sequence of polynomials converges uniformly to $x \mapsto |x|$ on $[-1, 1]$. (Hint: Corollary 2.3. in [1])

(c) Prove that ReLU can be approximated uniformly by polynomials on $[-1, 1]$.

(d) Use the universal approximation theory of shallow neural networks on $[0, 1]$ to prove the Stone-Weierstrass theorem.

Solution 2.1

(a) Consider the function $f_n(x) = x^n$ for $x \in [0, 1]$.

(b) Consider the following map

$$p_{n+1}(x) = p_n(x) + \frac{1}{2}(x - p_n^2(x)), \tag{1}$$

which is a contraction on $[0, 1)$ and the special case $x = 1$ is obvious.

(c) $g(x) = \frac{1}{2}(x + |x|)$

(d) Since ReLU can be approximated uniformly by polynomials on $[0, 1]$, composition of affine function and ReLU can be uniformly by polynomials on $[0, 1]$. Thus, shallow neural networks can be uniformly by polynomials on $[0, 1]$. Therefore, by UAT, polynomials can uniformly approximate any continuous function on $[0, 1]$.

Exercise 2.2 (Networks on discrete path spaces)

(a) Describe the space of paths $\omega : \{1, \ldots, T\} \to \mathbb{R}^d$ as $\mathbb{R}^{dT}$.

(b) Describe a shallow neural network, which depends on value at time t and on path information up to time t. Formulate a universal approximation theorem in this setting.

Solution 2.2

(a) Maps from $\{1, \ldots, T\}$ to $\mathbb{R}^d$ expressed by $\mathbb{R}^{dT}$.

(b) A neural network with input space $\mathbb{R}^{dt}$ for fixed t, a neural network with input space at least $\mathbb{R}^{dT}$ (might be larger if allow duplicated information in input space). UAT for path space is concerning universal approximation of continuous functional on path spaces e.g. the running max of a discrete path.

Exercise 2.3 (Backpropogation of neural network) Let $\theta = (w, b, a) \in \mathbb{R}^3$ and let σ be the activation function. We consider the shallow neural network $f_\theta\colon \mathbb{R} \to \mathbb{R}$ s.t.

$$f_\theta(x) = a\sigma(wx + b). \tag{2}$$

Then we solve the regression problem with 3 data point $(x_i, y_i) \in \mathbb{R}^2$, $i = 1, 2, 3$ by minimizing the L^2 loss

$$\mathcal{L}_f = \sum_{i=1,2,3} (f_\theta(x_i) - y_i)^2. \tag{3}$$

(a) When solving the regression, do we compute $\nabla_{x_0}\mathcal{L}_f$ or $\nabla_\theta \mathcal{L}_f$?

(b) Compute $\partial_w f$ and $\partial_b f$ by chain rule. Do you find any intermediate value computed twice in both $\partial_w f$ and $\partial_b f$?

(c) Consider regression problem as a constrained optimization problem

$$\begin{aligned} &\min \sum_{i=1,2,3} l_i \\ &l_i = (\tilde{y}_i - y_i)^2 \\ &\tilde{y}_i = a\sigma(z_i), \qquad i = 1,2,3. \\ &z_i = wx_i + b \end{aligned} \tag{4}$$

Solve it by Lagrange multiplier and relate this with backpropagation.

(d) Generalize this idea to deep neural networks.

Solution 2.3

(a) $\nabla_\theta \mathcal{L}_f$

(b) Let $z = wx_0 + b$ then

$$\partial_w \mathcal{L}_f = \partial_z \mathcal{L}_f \cdot x_0 = (a\sigma(w_0 x + b) - y_0)\sigma' a(wx_0 + b)x_0, \tag{5}$$
$$\partial_b \mathcal{L}_f = \partial_z \mathcal{L}_f \cdot 1 = (a\sigma(wx_0 + b) - y_0)a\sigma'(wx_0 + b) \tag{6}$$

(c) Consider the Lagrangian

$$\mathcal{L} = l - \lambda_l(l - (y - y_0)^2) - \lambda_y(y - a\sigma(z)) - \lambda_z(z - (wx_0 + b)) \tag{7}$$

Compute the gradient

$$\begin{aligned} \partial_l \mathcal{L} &= 1 - \lambda_l \\ \partial_y \mathcal{L} &= \lambda_l \frac{\partial (y - y_0)^2}{\partial y} - \lambda_y \\ \partial_z \mathcal{L} &= \lambda_y \frac{\partial a\sigma(z)}{\partial z} - \lambda_z \\ \partial_w \mathcal{L} &= \lambda_z \frac{\partial (wx_0 + b)}{\partial w} \\ \partial_b \mathcal{L} &= \lambda_z \frac{\partial (wx_0 + b)}{\partial b} \end{aligned}$$

Let $\nabla \mathcal{L} = 0$ we get exactly the backpropagation formula.

(d) See [4].

Exercise 2.4 (Functional analysis) Let K be a compact subset of $\mathbb{R}^d$.

(a) Let μ be a finite Borel measure on K. Prove that

$$\mathcal{L}_\mu(f) := \int_K f(x)\mu(dx) \tag{8}$$

for $f \in C(K, \mathbb{R})$ is a bounded linear functional.

(b) Let $\mathcal{L}, C(K, \mathbb{R})$ be a positive linear functional, i.e. $\mathcal{L}(f) \geq 0$ for $f \geq 0$. Then $\mathcal{L}$ is continuous.

(c) Prove that

$$\mathcal{F} := \{f \mapsto \sum_{i=1}^{n} \lambda_i f(x_i) \mid \lambda_i \in \mathbb{R}, n \in \mathbb{N}, x_i \in K, i = 1, 2, ..., n\} \tag{9}$$

is point separating and additive.

Solution 2.4

(a) $\mathcal{L}_\mu$ is linear by the linearity of the integral. We need to show that $\mathcal{L}_\mu$ is bounded. f is bounded, as f is continuous on K and K is compact. In addition, as $\mu(K) < \infty$, there exists $C \in \mathbb{R}$ such that $\mu(K) = C$. Hence

$$\begin{aligned}
\mathcal{L}_\mu(f) &= \int_K f(x)\mu(dx) \\
&\leq \int_K \sup_{x \in K} |f(x)| \mu(dx) \\
&= \int_K ||f(x)||_\infty \mu(dx) \\
&\leq ||f(x)||_\infty \mu(K) \\
&= ||f(x)||_\infty C.
\end{aligned}$$

We have shown that there exists $C \in \mathbb{R}$ such that

$$\mathcal{L}(f) \leq ||f(x)||_\infty C, \forall f \in C(K, \mathbb{R}).$$

So $\mathcal{L}$ is bounded.

(b) We start by giving a reminder of the Riesz-Markov-Kakutani representation theorem.

Theorem 1 *Riesz-Markov-Kakutani representation theorem* Let X be a locally compact Hausdorff space, and $\mathcal{L}$ a positive linear functional on $C_c(X)$. Then there exists a unique positive Borel measure μ on X such that

$$\mathcal{L} = \int_X f(x)\mu(dx)$$

for every $f \in C_c(X)$, and which has the following properties for some M containing the Borel $\delta-$algebra on X:

(1) $\mu(K) < \infty$ for every compact set $K \subset X$

(2) For every $E \in M$, we have $\mu(E) = \inf\{\mu(V) : E \subset V, V \text{open}\}$

(3) The relation $\mu(E) = \sup\{\mu(K) : K \subset E, K \text{compact}\}$ holds for every open set E, and for every $E \in M$ with $\mu(E) < \infty$

(4) If $E \in M$, $A \subset E$, and $\mu(E) = 0$, then $A \in M$.

As $\mathcal{L}$ is positive linear functional, by Riesz-Markov-Kakutani representation theorem, there exists a unique measure μ such that the functional $\mathcal{L}$ on f is defined as $\mathcal{L}(f) := \int_K f(x)\mu(dx)$. Let a sequence of functions f_n in $C(K, \mathbb{R})$ converges uniformly to a function $f \in C(K, \mathbb{R})$, we have for any $\epsilon > 0$, there exists a positive integer N such that for all $n \geq N$ and $x \in K$, $|f_n(x) - f(x)| < \epsilon$. Since K is compact and f is continuous, f is also bounded on K, i.e., there exists a constant M such that $|f(x)| \leq M$ for all $x \in K$. Consequently, for all $n \geq N$,

$|f_n(x)| \le |f_n(x) - f(x)| + |f(x)| < \epsilon + M$. This implies $|f_n(x)|$ is bounded by $\epsilon + M$ for all $n \ge N$ and $x \in K$. Let $g_n(x) = max(|f_n(x)|, |f(x)|)$, we can see g_n is a bounded continuous function on compact set K, hence g_n is integrable. Thus we can apply dominated convergence theorem: If $f_n(x) \ge 0, f_n(x)$ converges to $f(x)$ pointwisely for all $x \in K$, and $|f_n(x)| \le g_n(x)$ for all n and x, where $g_n(x)$ is integrable, then

$$\lim_{n\to\infty} \int_K f_n(x)\mu(dx) = \int_K f(x)\mu(dx)$$

So we have

$$\lim_{n\to\infty} \mathcal{L}(f_n) = \mathcal{L}(f)$$

It proves $\mathcal{L}$ is continuous.

(c) Let p and q be distinct points in K. Since they are distinct, there must exist at least one coordinate where they differ, i.e., $p_i \neq q_i$. Define the function $f(x)$ as follows:

$$f(x) = \begin{cases} 1, & \text{for all } x_j \neq p_j \\ 0, & \text{for } x = p \end{cases}$$

Now consider function $F(f)$:

$$F(f)(p) = \sum_{i=1}^{n} \lambda_i f(p_i) = 0$$

$$F(f)(q) = \sum_{i=1}^{n} \lambda_i f(q_i) = \lambda_i$$

Since λ_i can be non-zero, and consequently, $F(f)(p) \neq F(f)(q)$. The additivity from $\mathcal{F}$ comes from

$$F(f+g) = \sum_{i=1}^{n} \lambda_i (f+g)(x_i) = \sum_{i=1}^{n} \lambda_i f(x_i) + \lambda_i g(x_i) = \sum_{i=1}^{n} \lambda_i f(x_i) + \sum_{i=1}^{n} \lambda_i g(x_i) = F(f) + F(g), \forall F \in \mathcal{F}.$$

References

[1] SAMEER CHAVAN. Problems and notes: Uniform convergence and polynomial approximation.

[2] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[3] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4):917–963, 2019.

[4] Yann LeCun, D Touresky, G Hinton, and T Sejnowski. A theoretical framework for back-propagation. 1:21–28, 1988.