Institute of Mathematics, EPFL MATH524

# Chapter 8: Consistency of neural network estimators

*Lecturer: Rajita Chandak* *Spring 2025*

Today, neural networks need no introduction. Neural networks have become synonymous with *machine learning*. However, it is worth understanding the origins of the neural network estimation procedure if we hope to making claims about the accuracy and robustness of these estimators. The idea of a neural network first came about from the goal of using a mathematical formulation (called a *neuron*) to approximate the behaviour of a human brain. The first proposal of this mathematical neuron was published by McCulloch and Pitts (1943) wherein the authors introduce the concept of neuron by virtue of binary thresholding. As can be seen in Figure 1 the neuron takes a linear combination of the inputs and applies a thresholding function to provide an output.

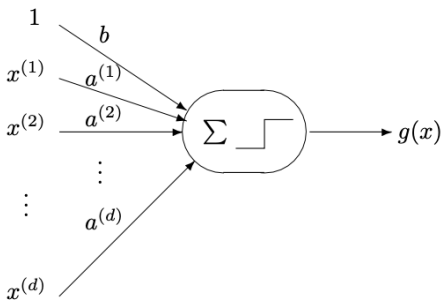


Figure 1: A McCulloch-Pitts neuron

Soon after, Rosenblatt (1958) applied the neuron to pattern recognition and introduced the concept of a *perceptron* (a neuron with a different thresholding function) and proved its convergence (Rosenblatt, 1962). The neural network, a complex structure of multiple neurons was then a natural extension.

## 1 Mathematical formulation

The neurons that most commonly make up the operations of a typical neural network today are usually smoothed versions of the neuron operation proposed by McCulloch and Pitts.

$$g(x) = \sigma(a^{\mathrm{T}}x + b)$$

where $x \in \mathbb{R}^d$, $a \in \mathbb{R}^d$, $b \in \mathbb{R}$ and $\sigma(\cdot) : \mathbb{R} \to [0, 1]$ is known as the sigmoid function. A neural network is a combination of multiple neurons operating simultaneously on the input:

$$f(x) = \sum_{i=1}^{k} c_i \sigma(a_i^{\mathrm{T}}x + b_i) + c_0.$$

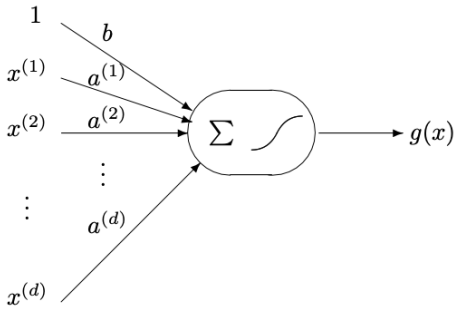


Figure 2: An artificial neuron

The function $f$ is called a *feedforward neural network* with one hidden layer of $k$ neurons. See Figure 3 for a visualization of such an estimator.

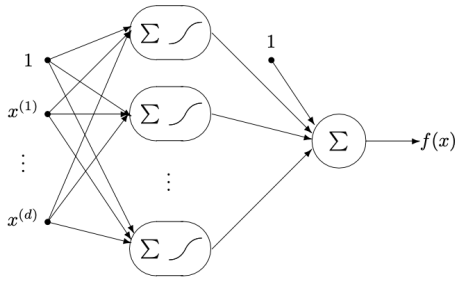


Figure 3: A feedforward neural network

The specific form of the sigmoid $\sigma$ can be quite general but there are certain conditions it must satisfy to provide a contractable output $f$.

**Definition 1** (Squashing functions)**.** *A sigmoid function is called a squashing function if it is nondecreasing,* $\lim_{x\to-\infty}\sigma(x) = 0$ *and* $\lim_{x\to\infty}\sigma(x) = 1$.

The most commonly used squashing functions have at most countable number of discontinuities and as a result are measurable functions. These are the kinds of functions we will study for mathematical convenience. Some examples of such functions that are also commonly used in practice:

- Thresholding: $\sigma(x) = \mathbf{1}(x \in [0, \infty))$
- Ramp: $\sigma(x) = x\mathbf{1}(x \in [0, 1]) + \mathbf{1}(x \in [1, \infty))$
- Cosine: $\sigma(x) = \frac{1}{2}(1 + \cos(x + 3\pi/2))\mathbf{1}(x \in [-\pi/2, \pi/2]) + \mathbf{1}(x \in (\pi/2, \infty))$
- Logistic: $\sigma(x) = (1 + \exp(-x))^{-1}$
- Arctan: $\sigma(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x)$

- Gaussian: $\sigma(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-y^2/2)\mathrm{d}y$.

Our goal for this chapter will be to show consistency of the neural network regression estimator. We will assume i.i.d. data and choose to estimate the parameters of the model $(a, b, c)$ by minimizing the empirical squared error:

$$\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2.$$

In order to make reasonable progress with this, as always, we will need to place some constraints on the class of possible regression functions.

$$\mathcal{F}_{NN} = \left\{ f(x) = \sum_{j=1}^{k_n} c_j \sigma(a_j^{\mathrm{T}} x + b_j) + c_0 : k_n \in \mathbb{N}, a_j \in \mathbb{R}^d, b_j \in \mathbb{R}, \sum_{j=1}^{k_n} |c_j| \leq \beta_n \right\}$$

The NN regression estimator, $m_n \in \mathcal{F}_{NN}$ is then defined as

$$\frac{1}{n}\sum_{i=1}^{n}(m_n(X_i) - Y_i)^2 = \min_{f \in \mathcal{F}_{NN}} \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2.$$

We will assume that the minimizing function exists but will not require it to be unique. Under some additional assumptions on the *width* (number of neurons in a layer, $k_n$) of the network and the magnitude of the parameters $\beta_n$, we can show strong universal consistency of the one hidden layer feedforward neural network.

**Theorem 1 (Consistency of NNs)**
*For the class $\mathcal{F}_{NN}$ let $m_n$ be the empirical risk minimizing estimator. If $k_n, \beta_n \to \infty$ and*

$$\frac{k_n \beta_n^4 \log(k_n \beta_n^2)}{n} \to 0,$$

*then,*

$$\mathbb{E}\left[\int (m_n(x) - m(x))^2 \mu(\mathrm{d}x)\right] \to 0$$

*for all joint distributions of $X, Y$ such that $\mathbb{E}[Y^2] < \infty$. Additionally, if there exists a $\delta > 0$ such that $\beta_n^4 / n^{1-\delta} \to 0$, then*

$$\int (m_n(x) - m(x))^2 \mu(\mathrm{d}x) \to 0.$$

Before we prove this theorem, we will show a powerful approximation result for the class $\mathcal{F}_{NN}$.

**Theorem 2 (Universal approximation)**
*Let $\sigma$ be a squashing function and let $K$ be a compact subsect of $\mathbb{R}^d$. Then, for every continuous function $f : \mathbb{R}^d \to \mathbb{R}$ and every $\varepsilon > 0$, there exists a neural network*

$$h(x) = \sum_{j=1}^{k} c_j \sigma(a_j^{\mathrm{T}} x + b_j) + c_0$$

*such that*

$$\sup_{x \in K} |f(x) - h(x)| < \varepsilon.$$

*Proof of Theorem 2.* The proof will be done in parts.

**Part 1. Cosine network universal approximation** The first step will show that the clas of cosine networks uniformly approximates the space of real continuous functions.

Consider the class of cosine networks:

$$\mathcal{F} = \left\{ f(x) = \sum_{j=1}^{k} c_j \cos(a_j^{\mathrm{T}} x + b_j), k \in \mathbb{N}, a_j \in \mathbb{R}^d, b_j, c_j \in \mathbb{R} \right\}.$$

We note that $\mathcal{F}$ is closed under addition, multiplication and scalar multiplication. Closed under multiplication follows from the cosine identity:

$$\cos a \cos b = \frac{1}{2}(\cos(a + b) + \cos(a - b)).$$

We will say that $\mathcal{F}$ *separates* $K$ if for every $x, y \in K, x \neq y$ there is a function $f \in \mathcal{F}$ such that for ever $x \in K, f(x) \neq f(y)$. $\mathcal{F}$ *vanishes at no point* of $K$ if there exists $f \in \mathcal{F}$ such that for every $x \in K$, $f(x) \neq 0$.

We observe that the cosine network class separates points on $K$ and vanishes nowhere on $K$ by the following argument: For some $x, y \in K, x \neq y$ and some $a \in \mathbb{R}^d, b = 0$, we have $\cos(a^{\mathrm{T}} x + b) \neq \cos(a^{\mathrm{T}} y + b)$ (as long as a is chosen such that $a^{\mathrm{T}} x, a^{\mathrm{T}} y \in (-\pi, \pi)$ and $a^{\mathrm{T}} x \neq a^{\mathrm{T}} y$). Now, select $b$ such that $\cos b \neq 0$ and $a \equiv \mathbf{0}$. Then, for all $x \in K$, we see that $\cos(a^{\mathrm{T}} x + b) \neq 0$. Then, by the Stone-Weierstrass theorem (Rudin, 1964), $\mathcal{F}$ is uniformly dense in the space of real continuous functions on $K$.

**Part 2. Cosine squasher approximation** The next step is to show that the class of cosine squasher functions approximates the class of cosine networks.

Let $c(x) = \frac{1}{2}(1 + \cos(x + 3\pi/2))\mathbf{1}(x \in [-\pi/2, \pi/2]) + \mathbf{1}(x \in (\pi/2, \infty))$. Now, we will show that a collection of the cosine squasher functions can approximate $\cos(\cdot)$ on any compact interval$[-M, M]$. Define

$$f(u) = \sum_{j=1}^{l} \gamma_j c(\alpha_j^{\mathrm{T}} u + \beta_j).$$

We will first show that $f$ can reconstruct $2(\cos(u)+1)$ on any compact interval since changing any of the parameters $\gamma, \alpha, \beta$ allow us to shift and scale the result to then reconstruct $\cos(u)$. Note that

$$(\cos(u)+1)\mathbf{1}(u \in [-\pi, \pi]) = 2(c(u+\pi/2) - c(u-\pi/2)).$$

And so, by adding a finite number of such shifted functions we can reconstruct the cosine on any compact interval. If the neural network is constructed using the cosine squasher, we have proven the result of the theorem and we could end the proof here.

**Part 3. Arbitrary squasher approximation** Now, we show that any other squasher function can approximate the cosine squasher and therefore the theorem holds for neural networks constructed with any reasonable squashing function. That is, for every $\varepsilon > 0$ and arbitrary squashing function $\sigma$, there is a neural network $h$ such that

$$\sup_{u \in \mathbb{R}} |h(u) - c(u)| < \varepsilon.$$

We start by choosing $k$ such that $1/k < \varepsilon/4$. By construction, there exists an $M > 0$ such that $\sigma(-M) < \varepsilon/(2k)$ and $\sigma(M) > 1 - \varepsilon/(2k)$. By continuity and monotonicity of $c$, we can find constants $r_1, \ldots, r_k$ such that $c(r_i) = i/k$ for $i = 1, \ldots, k-1$ and $c(r_k) = 1 - 1/(2k)$. Now, set $a_i = 2M/(r_{i+1} - r_i)$ and $b_i = M(r_i + r_{i+1})/(r_i - r_{i+1})$. This means that $a_i u + b_i$ is a straight line through the points $(r_i, -M)$ and $(r_{i+1}, M)$. Note also that $a_i > 0$. Then, for the network

$$h(u) = \frac{1}{k} \sum_{j=1}^{k-1} \sigma(a_j u + b_j),$$

we have that

$$|c(u) - h(u)| < \varepsilon$$

on each of the subintervals $(-\infty, r_1], (r_1, r_2], \ldots, (r_k, \infty)$. We can show this through the following argument. Let $i = 1, \ldots, k-1$ and $u \in (r_i, r_{i+1}]$, then $i/k \leq c(u) \leq (i+1)/k$. For $j \in \{1, \ldots, i-1\}$,

$$\sigma(a_j u + b_j) \geq \sigma(a_j r_{j+1} + b_j) = \sigma(M) \geq 1 - \frac{\varepsilon}{2k}$$

and for all $j \in \{i+1, \ldots, k-1\}$

$$\sigma(a_j u + b_j) \leq \sigma(a_j r_j + b_j) = \sigma(-M) < \frac{\varepsilon}{2k}.$$

Then,

$$\begin{aligned} &|c(u) - h(u)| \\ &\leq \left| c(u) - \frac{1}{k} \sum_{j=1}^{i-1} \sigma(a_j u + b_j) \right| + \frac{1}{k} \sigma(a_i u + b_i) + \frac{1}{k} \sum_{j=i+1}^{k-1} \sigma(a_j u + b_j) \end{aligned}$$

$$
\begin{aligned}
&\leq \left| c(u) - \frac{i-1}{k} \right| + \left| \frac{i-1}{k} - \frac{1}{k} \sum_{j=1}^{i-1} \sigma(a_j u + b_j) \right| + \frac{1}{k}\sigma(a_i u + b_i) + \frac{1}{k} \sum_{j=i+1}^{k-1} \sigma(a_j u + b_j) \\
&\leq \left( \frac{i+1}{k} - \frac{i-1}{k} \right) + \left( \frac{i-1}{k} - \frac{i-1}{k}(1 - \frac{\varepsilon}{2k}) \right) + \frac{1}{k} + \frac{k-1-i}{k} \frac{\varepsilon}{2k} \\
&= \frac{2}{k} + \frac{i-1}{k} \frac{\varepsilon}{2k} + \frac{1}{k} + \frac{k-1-i}{k} \frac{\varepsilon}{2k} \\
&\leq \frac{3}{k} + \frac{\varepsilon}{2k} \leq \frac{3}{4}\varepsilon + \frac{\varepsilon^2}{8} \leq \varepsilon.
\end{aligned}
$$

For $u \in (-\infty, r_1]$, by $c(r_1) = 1/k$ and the fact that $\sigma(-M) < \varepsilon/(2k)$,

$$
|c(u) - h(u)| \leq \max \left\{ \frac{1}{k}, \frac{k-1}{k} \frac{\varepsilon}{2k} \right\} < \max\{\varepsilon/4, \varepsilon/8\} < \varepsilon.
$$

Similarly, for $u \in (r_k, \infty)$,

$$
\begin{aligned}
|c(u) - h(u)| = |1 - c(u) - (1 - h(u))| &\leq \max\{|1 - c(u)|, |1 - h(u)|\} \\
&= \max \left\{ 1 - (1 - \frac{1}{2k}), 1 - \frac{k-1}{k}(1 - \frac{\varepsilon}{2k}) \right\} \\
&= \max \left\{ \frac{1}{2k}, \frac{1}{k} + \frac{\varepsilon}{2k}(1 - \frac{1}{k}) \right\} \\
&\leq \max \left\{ \frac{1}{2k}, \frac{1}{k} + \frac{\varepsilon}{2k} \right\} \\
&< \max \left\{ \frac{\varepsilon}{8}, \frac{\varepsilon}{4} + \frac{\varepsilon^2}{8} \right\} \\
&< \varepsilon
\end{aligned}
$$

Parts 2 and 3 of the proof together show that for every $\varepsilon > 0$, $M > 0$ and any squashing function, $\sigma$, there exists a neural network that deviates from the cosine function by at most $\varepsilon$ on the interval $[-M, M]$. The only remaining piece of the proof is to extend this to any compact set $K \in \mathbb{R}^d$.

**Part 4. Extension to arbitrary compact sets** We start by defining an abitrary cosine network $g(x) = \sum_{i=1}^{k} \tilde{c}_i \cos(\tilde{a}_i^{\mathrm{T}} x + \tilde{b}_i)$ We want to show that for any squashing function $\sigma$ and compact set $K \in \mathbb{R}^d$, there exists another network $s(x) = \sum_{i=1}^{k} c_i \cos(a_i^{\mathrm{T}} x + b_i)$ such that

$$
\sup_{x \in K} |s(x) - g(x)| < \varepsilon.
$$

Sine $K$ is compact and the inputs to the squashing function $(a_i^{\mathrm{T}} x + b_i)$ are continuous, there is a finite $M$ such that $\sup_{x \in K} |a_i^{\mathrm{T}} x + b_i| \leq M$. Then, by the conclusions of Parts 2 and 3,

$$
\sup_{x \in K} |\sum_{i=1}^{k} \tilde{c}_i \cos(\tilde{a}_i^{\mathrm{T}} x + \tilde{b}_i) - \sum_{i=1}^{k} \tilde{c}_i C_{M,\varepsilon}(\tilde{a}_i^{\mathrm{T}} x + \tilde{b}_i)| \leq \sum_{i=1}^{k} |\tilde{c}_i| \sup_{u \in [-M,M]} |\cos(u) - C_{M,\varepsilon}(u)| < \varepsilon.
$$

where $C_{M,\varepsilon}$ is the appropriately scaled and shifted $\sigma$ neural network. The conclusion of the theorem follows from this inequality combined with the conclusion of Part 1 with the triangle inequality. ■

We have shown through Theorem 2 that neural networks can approximate any real-valued continuous function when appropriately scaled and shifted. This result takes care of the approximation error for the class of continuous functions. In order to prove the statement in Theorem 1, we also need to control the estimation error. As we have seen in previous chapters, the bounding of the estimation error will come from VC arguments. We now look at three key VC and covering number arguments that will help in completing the proof of Theorem 1.

**Lemma 1**

*Let $\mathcal{F}$ be a class of real functions on $\mathbb{R}^m$ and let $g : \mathbb{R} \to \mathbb{R}$ be some nondecreasing function. Let $\mathcal{G} = \{g \circ f : f \in \mathcal{F}\}$. Then,*

$$VC(\mathcal{G}^+) \leq VC(\mathcal{F}^+).$$

*Proof of Lemma 1.* Suppose some collection of points $(s_1, t_1), \ldots, (s_n, t_n)$ are shattered by $\mathcal{G}^+$. Then, there will exist a collection of functions $f_1, \ldots, f_{2^n} \in \mathcal{F}$ such that the vector

$$(\mathbf{1}(g(f_j(s_1)) \geq t_1), \ldots, \mathbf{1}(g(f_j(s_n)) \geq t_n)) \tag{1}$$

takes on all $2^n$ values for $j = 1, \ldots, 2^n$. For all $1 \leq i \leq n$, define

$$u_i = \min_{1 \leq j \leq 2^n} \{f_j(s_i) : g(f_j(s_i)) \geq t_i\}$$
$$l_i = \max_{1 \leq j \leq 2^n} \{f_j(s_i) : g(f_j(s_i)) < t_i\}.$$

By the monotonicity of $g$, $u_i > l_i$ and so $l_i < (l_i + u_i)/2 < u_i$. Furthermore,

$$g(f_j(s_i)) \geq t_i \to f_j(s_i) \geq u_i \to f_j(s_i) > \frac{u_i + l_i}{2}.$$

Similarly,

$$g(f_j(s_i)) < t_i \to f_j(s_i) \leq u_i \to f_j(s_i) < \frac{u_i - l_i}{2}.$$

And so,

$$(\mathbf{1}(f_j(s_1) \geq \frac{u_1 + l_1}{2}), \ldots, \mathbf{1}(f_j(s_n) \geq \frac{u_n + l_n}{2}))$$

take on the same values as (1) for all $j$. Thus, $(s_1, (u_1 + l_1)/2), \ldots, (s_n, (u_n + l_n)/2)$ is shattered by $\mathcal{F}^+$. ■

The next two lemmas show the relationship between composition of functions and covering numbers.

**Lemma 2 (Covering numbers for sums of functions)**

*Let $\mathcal{F}$ and $\mathcal{G}$ be two function classes on $\mathbb{R}^m$ Define $\mathcal{F} \oplus \mathcal{G} = \{f + g : f \in \mathcal{F}, g \in \mathcal{G}\}$. Then, for any data, $z_1^n$ and $\varepsilon, \delta > 0$, we have*

$$N(\varepsilon + \delta, \mathcal{F} \oplus \mathcal{G}, L_1(z_1^n)) \leq N(\varepsilon, \mathcal{F}, L_1(z_1^n)) N(\delta, \mathcal{G}, L_1(z_1^n))$$

*Proof of Lemma 2.* The proof of this theorem will be done in the exercises. ■

**Lemma 3 (Covering numbers for products of functions)**
*Let $\mathcal{F}$ and $\mathcal{G}$ be two function classes on $\mathbb{R}^m$ such that $|f(x)| \leq M_1$ and $|g(x)| \leq M_2$ for all $x \in \mathbb{R}^m$, $f \in \mathcal{F}$ and $g \in \mathcal{G}$. Define $\mathcal{F} \odot \mathcal{G} = \{f \cdot g : f \in \mathcal{F}, g \in \mathcal{G}\}$. Then, for any data, $z_1^n$ and $\varepsilon, \delta > 0$, we have*

$$N(\varepsilon + \delta, \mathcal{F} \odot \mathcal{G}, L_1(z_1^n)) \leq N(\varepsilon/M_2, \mathcal{F}, L_1(z_1^n))N(\delta/M_1, \mathcal{G}, L_1(z_1^n))$$

*Proof of Lemma 3.* Let $\{f_1, \ldots, f_K\}$ and $\{g_1, \ldots, g_L\}$ be minimal $\varepsilon/M_2$ and $\delta/M_1$ covers of $\mathcal{F}$ and $\mathcal{G}$ respectively. Then, for every $f \in \mathcal{F}, g \in \mathcal{G}$, there exists a $k \in [K], l \in [L]$ such that

$$\frac{1}{n}\sum |f(z_i) - f_k(z_i)| < \frac{\varepsilon}{M_2},$$
$$\frac{1}{n}\sum |g(z_i) - g_l(z_i)| < \frac{\delta}{M_1},$$

and $|f_k| \leq M_1$, $|g_l| \leq M_2$. Then, by the triangle inequality,

$$\begin{aligned}
\frac{1}{n}\sum |f(z_i)g(z_i) - f_k(z_i)g_l(z_i)| &= \frac{1}{n}\sum |f(z_i)(g(z_i) + g_l(z_i) - g_l(z_i)) - f_k(z_i)g_l(z_i)| \\
&\leq \frac{1}{n}\sum |g_l(z_i)(f(z_i) - f_k(z_i))| + \frac{1}{n}\sum |f_k(z_i)(g(z_i) - g_l(z_i))| \\
&\leq M_2 \frac{1}{n}\sum f(z_i) - f_k(z_i)| + M_1 \frac{1}{n}\sum |(g(z_i) - g_l(z_i)| \\
&< \varepsilon + \delta
\end{aligned}$$

and so $\{f_k g_l : k \in [K], l \in [L]\}$ is an $(\varepsilon + \delta)$-cover of $\mathcal{F} \odot \mathcal{G}$. ■

We are now ready to prove Theorem 1.

*Proof of Theorem 1.* The approximation error given by

$$\inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mu(\mathrm{d}x)$$

vanishes as $k_n, \beta_n \to \infty$ if $\cup_{k \in \mathbb{N}} \mathcal{F}_k$ is dense in $L_2(\mu)$ for every $\mu$. The proof of denseness is an analytic proof that uses Fourier transforms. For the sake of time and mainting emphasis on statistical concepts we will skip the proof of this part and simply take the result as given.

Now, we are left with controlling the estimation error. Recall Theorem 2 from Chapter 7, which tells us that we can assume $|Y| \leq L$ almost surely and all that remains to be proven is

$$\sup_{f \in \mathcal{F}_n} \left| \frac{1}{n}\sum_{j=1}^{n} |f(\mathbf{X}_j) - Y_j|^2 - \mathbb{E}[|f(\mathbf{X}) - Y|^2] \right| \to 0.$$

We will show this by first defining $Z = (X, Y), Z_i = (X_i, Y_i)$ and

$$\mathcal{H} = \{h : \mathbb{R}^d \to \mathbb{R} : \exists f \in \mathcal{F}_n \text{s.t. } h(x, y) = |f(x) - y|^2\}.$$

WLOG we can assume that $\beta_n \geq L$ (since $L$ is some constant and $\beta_n \to \infty$). In this case, for all $h \in \mathcal{H}$

$$0 \leq h(x, y) \leq 2\beta_n^2 + 2L^2 \leq 4\beta_n^2.$$

Then, by Lemma 1 from Chapter 5, we have for any $\varepsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f \in \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|^2 - \mathbb{E}\left[|f(X) - Y|^2\right] \right| > \varepsilon \right) \\
&\mathbb{P}\left(\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n h(Z_i) - \mathbb{E}\left[h(Z)\right] \right| > \varepsilon \right) \\
&\leq 8\mathbb{E}[N(\frac{\varepsilon}{8}, \mathcal{H}, L_1(Z_1^n))] \exp(-\frac{n\varepsilon^2}{128(4\beta_n)^2}).
\end{aligned}$$

We need to bound the expected covering number. Let $h_i(z) = |f_i(x) - y|^2$ for some $f_i \in \mathcal{F}_n$. Then,

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n |h_1(Z_i) - h_2(Z_i)| &= \frac{1}{n} \sum_{i=1}^n ||f_1(X_i) - Y_i|^2 - |f_2(X_i) - Y_i|^2| \\
&= \frac{1}{n} \sum_{i=1}^n |f_1(X_i) - f_2(X_i)||f_1(X_i) - Y_i + f_2(X_i) - Y_i| \\
&\leq 4\beta_n \frac{1}{n} \sum_{i=1}^n |f_1(X_i) - f_2(X_i)|.
\end{aligned}$$

Thus,

$$N(\frac{\varepsilon}{8}, \mathcal{H}, L_1(Z_1^n)) \leq N(\frac{\varepsilon}{32\beta_n}, \mathcal{F}_n, L_1(X_1^n)).$$

Now, let us define three function classes:

$$\begin{aligned}
\mathcal{G}_1 &= \{a^{\mathrm{T}} x + b : a \in \mathbb{R}^d, b \in \mathbb{R}\} \\
\mathcal{G}_2 &= \{\sigma(a^{\mathrm{T}} x + b) : a \in \mathbb{R}^d, b \in \mathbb{R}\} \\
\mathcal{G}_3 &= \{c\sigma(a^{\mathrm{T}} x + b) : a \in \mathbb{R}^d, b \in \mathbb{R}, c \in [-\beta_n, \beta_n]\}.
\end{aligned}$$

It is clear from the construction that $\mathcal{G}_1$ is a linear vector space of dimension $d + 1$. Then, by Theorem 9 from Chapter 5, $VC(\mathcal{G}_1^+) \leq d + 2$. Furthermore, since $\sigma$ is nondecreasing, Lemma 1 directly gives the bound $VC(\mathcal{G}_2^+) \leq d + 2$. Then, by Theorem 12 from Chapter 5,

$$N(\varepsilon, \mathcal{G}_2, L_1(X_1^n)) \leq 3\left(\frac{2e}{\varepsilon} \log \frac{3e}{\varepsilon}\right)^{d+2}.$$

By Lemma 3, we can see that

$$\begin{aligned} N(\varepsilon, \mathcal{G}_3, L_1(X_1^n)) &\leq N(\varepsilon/2, \{c : |c| \leq \beta_n\}, L_1(X_1^n)) N(\varepsilon/2\beta_n, \mathcal{G}_2, L_1(X_1^n)) \\ &\leq \frac{4\beta_n}{\varepsilon} 3 \left(\frac{6e\beta_n}{\varepsilon}\right)^{2d+4} . \\ &\leq \left(\frac{12e\beta_n}{\varepsilon}\right)^{2d+5} . \end{aligned}$$

Finally, by applying Lemma 2,

$$\begin{aligned} N(\varepsilon, \mathcal{F}_n, L_1(X_1^n)) &\leq N(\frac{\varepsilon}{k_n+1}, \{c_0 : |c_0| \leq \beta_n\}, L_1(X_1^n)) N(\frac{\varepsilon}{k_n+1}, \mathcal{G}_2, L_1(X_1^n))^{k_n} \\ &\leq \frac{2\beta_n(k_n+1)}{\varepsilon} \left(\frac{12e\beta_n(k_n+1)}{\varepsilon}\right)^{k_n(2d+5)} \\ &\leq \left(\frac{12e\beta_n(k_n+1)}{\varepsilon}\right)^{k_n(2d+5)+1} . \end{aligned}$$

Putting all the bound together and scaling the $\varepsilon$'s appropriately,

$$\begin{aligned} &\mathbb{P}\left( \sup_{f \in T_{\beta_n}(\mathcal{F}_n)} \left| \frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|^2 - \mathbb{E}\left[|f(X) - Y|^2\right] \right| > \varepsilon \right) \\ &\leq 8 \left(\frac{384e\beta_n^2(k_n+1)}{\varepsilon}\right)^{k_n(2d+5)+1} \exp\left(-\frac{n\varepsilon^2}{128 \cdot 2^4 \beta_n^4}\right). \end{aligned}$$

Then applying the union bound and Borel-Cantelli, the conclusion of the theorem follows. ■

A final note regarding the results of this chapter: We have shown strong universal consistency of the one-layer feedforward neural network with growing number of neurons ($k_n$). This approach, however, tells us nothing about the rate at which the consistency is achieved. It is not obvious from this proof approach what we would expect the rate of convergence to be. In general, with machine learning type estimators, as we have seen in Chapter 7 and this chapter, consistency proofs either provide a sub-optimal rate of convergence or provide no insight into the true rate of convergence.

Furthermore, we have only shown consistency of one-layer neural networks. In practice, not only are deeper neural networks more commone, but the architecture of these deep neural networks is also often much more complicated than the feedforward structure. Most theoretical generalizations of this chapter remain open questions.

# References

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books.

Rudin, W. (1964). *Principles of mathematical analysis*, volume 3. McGraw-hill New York.