# Neural Network Learning: Theoretical Foundations
# Chapter 16-17

## Anthony and Bartlett

Presented by Kuhwan Jeong [1]

[1]Department of Statistics, Seoul National University, South Korea

October, 2017

1. Pattern Classification with Binary-Output Neural Networks (Chap. 2-8)
2. Pattern Classification with Real-Output Neural Networks (Chap. 9-15)
3. Learning Real-Valued Functions (Chap. 16-21)

# Outline


1. 16. Learning Classes of Real Functions
   - 16.1 Introduction
   - 16.2 Learning Framework for Real Estimation
   - 16.3 Learning Finite Classes of Real Functions
   - 16.4 A Substitute for Finiteness
2. 17. Uniform Convergence Results for Real Function Classes
   - 17.1 Uniform Convergence for Real Functions
   - 17.2 Remarks

# 16.1 Introduction

- This part examines supervised learning problems.
- Data are generated by a probability distribution $P$ on $X \times \mathbb{R}$.
- To measure how accurately $f(x)$ approximates $y$, we use the *quadratic loss*.
- We define the *error* of a function $f : X \to \mathbb{R}$ with respect to $P$ as

$$\mathrm{er}_P(f) = \mathbb{E}(f(x) - y)^2.$$

# 16.2 Learning Framework for Real Estimation

- We assume that the random variable $y$ falls in $[0, 1]$.
- $F$ : a set of functions mapping from $X$ into $[0, 1]$
- A learning algorithm $L$ for $F$ is a function

$$L : \bigcup_{m=1}^{\infty} (X \times \mathbb{R})^m \to F$$

with the following property:
for any $\epsilon \in (0, 1),\ \delta \in (0, 1)$, $\exists m_0(\epsilon, \delta)$ s.t.
if $m \geq m_0(\epsilon, \delta)$, then for any probability distribution $P$ on $X \times [0, 1]$,

$$P^m \left\{ \mathrm{er}_P(L(z)) < \inf_{f \in \mathcal{F}} \mathrm{er}_P(f) + \epsilon \right\} \geq 1 - \delta.$$

# 16.2 Learning Framework for Real Estimation

- *sample complexity* $m_L(\epsilon, \delta)$ : the least possible value $m_0(\epsilon, \delta)$
- *estimation error* $\epsilon_L(m, \delta)$ : the least possible value $\epsilon$ s.t.

$$P^m\left\{\mathrm{er}_P(L(z)) < \inf_{f \in \mathcal{F}} \mathrm{er}_P(f) + \epsilon\right\} \geq 1 - \delta$$

# 16.3 Learning Finite Classes of Real Functions

- *sample error* $\hat{\text{er}}_z(f)$ :

$$\hat{\text{er}}_z(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2$$

- *sample error minimization* (SEM) algorithm $L$ for a finite class $F$ : a function from $\bigcup_{m=1}^{\infty} Z^m$ to $F$ s.t.

$$\hat{\text{er}}_z(L(z)) = \min_{f \in \mathcal{F}} \hat{\text{er}}_z(f)$$

# 16.3 Learning Finite Classes of Real Functions

**Theorem 16.2** *SEM* algorithm $L$ is a learning algorithm for finite $F$, whose estimation error satisfies

$$\epsilon_L(m,\delta) \le \left(\frac{2}{m}\log\left(\frac{2|F|}{\delta}\right)\right)^{1/2}$$

*Proof*

$$P^m\left\{|\hat{\mathrm{er}}_z(f) - \mathrm{er}_P(f)| \ge \frac{\epsilon}{2}\right\} \le 2e^{-\epsilon^2 m/2}$$

$$P^m\left\{\exists f \in \mathcal{F} \text{ s.t. } |\hat{\mathrm{er}}_z(f) - \mathrm{er}_P(f)| \ge \frac{\epsilon}{2}\right\} \le 2|F|e^{-\epsilon^2 m/2} =: \delta \qquad (1)$$

Let $f^*$ be a function in $F$ satisfying

$$\mathrm{er}_P(f^*) = \min_{f\in\mathcal{F}} \mathrm{er}_P(f).$$

With probability at least $1-\delta$,

$$\mathrm{er}_P(L(z)) \overset{(1)}{<} \hat{\mathrm{er}}_z(L(z)) + \frac{\epsilon}{2} \overset{SEM}{\le} \hat{\mathrm{er}}_z(f^*) + \frac{\epsilon}{2} \overset{(1)}{<} \left(\mathrm{er}_P(f^*) + \frac{\epsilon}{2}\right) + \frac{\epsilon}{2} = \min_{f\in\mathcal{F}} \mathrm{er}_P(F) + \epsilon. \quad \square$$

# 16.3 Learning Finite Classes of Real Functions

**Theorem 16.3** $N$ is a neural network with arbitrary activation functions whose output takes values in $[0, 1]$. $F$ is the set of functions computable by $N$ when each weight and threshold is represented using $k$ bits. Then SEM algorithm $L$ for $F$ is a learning algorithm with

$$m_L(\epsilon, \delta) \leq \frac{2}{\epsilon^2}\left(kW \log 2 + \log\left(\frac{2}{\delta}\right)\right),$$

where $W$ is the total number of weights and thresholds.

*Proof*
It follows from Theorem 16.2 with $|F| \leq 2^{kW}$.

# 16.4 A Substitute for Finiteness

- $F$ : a set of functions mapping from $X$ into $[0, 1]$
- metric $d_{L_\infty}$ on $F$ :

$$d_{L_\infty}(f, g) = \sup_{x \in X} |f(x) - g(x)|$$

- Approximate-SEM algorithm $\mathcal{A}$ for $F$ :
  a function from $\bigcup_{m=1}^{\infty} Z^m \times \mathbb{R}^+$ to $F$ s.t.

$$\hat{\text{er}}_z(\mathcal{A}(z, \epsilon)) < \inf_{f \in \mathcal{F}} \hat{\text{er}}_z(f) + \epsilon$$

# 16.4 A Substitute for Finiteness

**Theorem 16.5** $F$ is totally bounded w.r.t. $d_{L_\infty}$, and $L$ satisfies $L(z) = \mathcal{A}(z, 1/\sqrt{2m})$ for $z \in Z^m$. Then $L$ is a learning algorithm for $F$ with

$$m_L(\epsilon, \delta) \leq \frac{72}{\epsilon^2} \log\left(\frac{2\mathcal{N}(\epsilon/12, F, d_{L_\infty})}{\delta}\right).$$

*Proof*
*Lemma* If $d_{L_\infty}(f, g) < \alpha$, then $|\mathrm{er}_Q(f) - \mathrm{er}_Q(g)| < 2\alpha$ for any probability distribution $Q$ on $Z$.

Let $\mathcal{C}$ be a $\epsilon/12$-cover for $F$ of cardinality $\mathcal{N}(\epsilon/12, F, d_{L_\infty})$. For any $f \in F$, $\exists \hat{f} \in \mathcal{C}$ s.t. $d_{L_\infty}(f, \hat{f}) < \epsilon/12$, and so

$$|\mathrm{er}_P(f) - \mathrm{er}_P(\hat{f})| < \frac{\epsilon}{6}, \tag{2}$$

$$|\hat{\mathrm{er}}_z(f) - \hat{\mathrm{er}}_z(\hat{f})| < \frac{\epsilon}{6}. \tag{3}$$

Let $f^*$ be s.t.

$$\mathrm{er}_P(f^*) < \inf_{f \in F} \mathrm{er}_P(f) + \frac{\epsilon}{12}. \tag{4}$$

# 16.4 A Substitute for Finiteness

From the proof of Theorem 16.2, we have

$$P^m\left\{\exists f \in \mathcal{C} \text{ s.t. } |\hat{\text{er}}_z(f) - \text{er}_P(f)| \geq \frac{\epsilon}{12}\right\} \leq 2\mathcal{N}(\epsilon/12, F, d_{L_\infty})e^{-\epsilon^2 m/72} =: \delta. \tag{5}$$

For convenience, we denote $L(z)$ by $f_z$. With probability at least $1-\delta$,

$$\begin{aligned}
\text{er}_P(f_z) &\overset{(2)}{<} \text{er}_P(\hat{f}_z) + \frac{\epsilon}{6} \overset{(5)}{<} \left(\hat{\text{er}}_z(\hat{f}_z) + \frac{\epsilon}{12}\right) + \frac{\epsilon}{6} \overset{(3)}{<} \left(\hat{\text{er}}_z(f_z) + \frac{\epsilon}{6}\right) + \frac{3\epsilon}{12} \\
&\overset{\text{aSEM}}{<} \left(\inf_{f\in\mathcal{F}} \hat{\text{er}}_z(f) + \frac{\epsilon}{12}\right) + \frac{5\epsilon}{12} \\
&\leq \hat{\text{er}}_z(f^*) + \frac{\epsilon}{2} \overset{(3)}{<} \left(\hat{\text{er}}_z(\hat{f}^*) + \frac{\epsilon}{6}\right) + \frac{\epsilon}{2} \overset{(5)}{<} \left(\text{er}_P(\hat{f}^*) + \frac{\epsilon}{12}\right) + \frac{2\epsilon}{3} \\
&\overset{(2)}{<} \left(\text{er}_P(f^*) + \frac{\epsilon}{6}\right) + \frac{3\epsilon}{4} \overset{(4)}{<} \left(\inf_{f\in F} \text{er}_P(f) + \frac{\epsilon}{12}\right) + \frac{11\epsilon}{12} \\
&= \inf_{f\in F} \text{er}_P(f) + \epsilon. \ \square
\end{aligned}$$

# 17.1 Uniform Convergence for Real Functions

- Given $x = (x_1, \ldots, x_k) \in X^k$,

$$F_{|x} = \{(f(x_1), \ldots, f(x_k)) : f \in F\} \subset \mathbb{R}^k.$$

- uniform $\epsilon$-covering number $\mathcal{N}_1(\epsilon, F, k)$ :

$$\mathcal{N}_1(\epsilon, F, k) = \max\{\mathcal{N}(\epsilon, F_{|x}, d_1) : x \in X^k\}.$$

**Theorem 17.1**

$$P^m\{\exists f \in F \text{ s.t. } |\mathrm{er}_P(f) - \hat{\mathrm{er}}_z(f)| \geq \epsilon\} \leq 4\mathcal{N}_1(\epsilon/16, F, 2m)e^{-\epsilon^2 m/32}$$

# 17.1 Uniform Convergence for Real Functions

**Lemma 17.2 (Symmetrization)**
Let

$$Q = \{z \in Z^m : \exists f \in \mathcal{F} \text{ s.t. } |\mathrm{er}_P(f) - \hat{\mathrm{er}}_z(f)| \geq \epsilon\},$$
$$R = \left\{(r,s) \in Z^m \times Z^m : \exists f \in \mathcal{F} \text{ s.t. } |\hat{\mathrm{er}}_r(f) - \hat{\mathrm{er}}_s(f)| \geq \frac{\epsilon}{2}\right\}.$$

Then, $P^m(Q) \leq 2P^{2m}(R)$ for $m \geq 4/\epsilon^2$.

*Proof* Since $|\mathrm{er}_P(f) - \hat{\mathrm{er}}_r(f)| \geq \epsilon$ and $|\mathrm{er}_P(f) - \hat{\mathrm{er}}_s(f)| < \epsilon/2$ implies $|\hat{\mathrm{er}}_r(f) - \hat{\mathrm{er}}_s(f)| \geq \epsilon/2$, we have

$$\begin{aligned} P^{2m}(R) &\geq P^{2m}\{\exists f \in \mathcal{F} \text{ s.t. } |\mathrm{er}_P(f) - \hat{\mathrm{er}}_r(f)| \geq \epsilon \;\&\; |\mathrm{er}_P(f) - \hat{\mathrm{er}}_s(f)| < \epsilon/2\} \\ &= \int_Q P^m\{s : \exists f \in \mathcal{F} \text{ s.t. } |\mathrm{er}_P(f) - \hat{\mathrm{er}}_s(f)| < \epsilon/2\} dP^m(r) \\ &\geq P^m(Q)/2 \text{ provided } m \geq 4/\epsilon^2. \square \end{aligned}$$

# 17.1 Uniform Convergence for Real Functions

**Permutation**

- $\Gamma_m$ : the set of permutations on $\{1, \ldots, 2m\}$ in which $i$ can be only switched with $m + i$.
- By Lemma 4.5,

$$P^{2m}(R) \leq \max_{z \in Z^{2m}} \Pr_\sigma(\sigma z \in R),$$

  where $\Pr_\sigma$ is the uniform probability on $\Gamma_m$.

# 17.1 Uniform Convergence for Real Functions

- $l_f : Z \to [0,1]$ is given by $l_f(x,y) = (f(x) - y)^2$.
- $l_F = \{l_f : f \in F\}$

**Lemma 17.4**
For all $z \in Z^m$,

$$\mathcal{N}(\epsilon, (l_F)_{|z}, d_1) \leq \mathcal{N}_1(\epsilon/2, F, m).$$

*Proof* Let $z = ((x_1, y_1), \ldots, (x_m, y_m))$.

$$\begin{aligned} d_1(l_{f_{|z}}, l_{g_{|z}}) &= \frac{1}{m}\sum_{i=1}^{m} |(f(x_i) - y_i)^2 - (g(x_i) - y_i)^2| \\ &= \frac{1}{m}\sum_{i=1}^{m} |(f(x_i) - y_i)(f(x_i) + g(x_i) - 2y_i)| \leq \frac{2}{m}\sum_{i=1}^{m} |f(x_i) - g(x_i)| \end{aligned}$$

$$\mathcal{N}(\epsilon, (l_F)_{|z}, d_1) \leq \mathcal{N}(\epsilon/2, F_{|z}, d_1) \leq \mathcal{N}_1(\epsilon/2, F, m) \ \square$$

# 17.1 Uniform Convergence for Real Functions

**Lemma 17.5**

$$\max_{z\in Z^{2m}} \Pr_\sigma(\sigma z \in R) \le 2\mathcal{N}_1(\epsilon/16, F, 2m)e^{-\epsilon^2 m/32}$$

*Proof* Choose $G \subset F$ s.t. $(l_G)_z$ is a minimal $\epsilon/8$-cover for $(l_F)_{|z}$ w.r.t. $d_1$. By Lemma 17.4,

$$|G| = \mathcal{N}(\epsilon/8, (l_F)_{|z}, d_1) \le \mathcal{N}_1(\epsilon/16, F, 2m).$$

If $\sigma z = (r,s) \in R$, then $\exists f \in G$ s.t. $|\hat{\text{er}}_r(f) - \hat{\text{er}}_s(f)| \ge \epsilon/4$.

$$\begin{aligned}
\Pr_\sigma(\sigma z \in R) &\le \Pr_\sigma\left(\exists f \in G : \left|\frac{1}{m}\sum_{i=1}^{m}\left(l_f(z_{\sigma(i)}) - l_f(z_{\sigma(m+i)})\right)\right| \ge \epsilon/4\right) \\
&\le |G| \max_{f\in G} \Pr_\sigma\left(\left|\frac{1}{m}\sum_{i=1}^{m}\left(l_f(z_{\sigma(i)}) - l_f(z_{\sigma(m+i)})\right)\right| \ge \epsilon/4\right) \\
&= |G| \max_{f\in G} \Pr_\beta\left(\left|\frac{1}{m}\sum_{i=1}^{m}\left|l_f(z_i) - l_f(z_{m+i})\right|\beta_i\right| \ge \epsilon/4\right) \le 2|G|e^{-\epsilon^2 m/32},
\end{aligned}$$

where $\beta_i$ is independently and uniformly drawn from $\{-1,1\}$. $\square$

# 17.2 Remarks

- The results do not depend significantly on the choice of the quadratic loss.
- Let $Y = \mathbb{R}$ and $B \geq 1$.
- Suppose the loss function $l : [0,1] \times Y \to [0,B]$ satisfies

$$l(y_1, y_0) - l(y_2, y_0) \leq L|y_1 - y_2|$$

  for all $y_0, y_1, y_2$.
- Define $l_f(z) = l(f(x), y)$.
- $F$ is a class of $[0,1]$-valued functions.

**Lemma 17.6** (Generalization of Lemma 17.4)

$$\max_{z \in Z^m} \mathcal{N}(\epsilon, (l_F)_{|z}, d_1) \leq \mathcal{N}_1(\epsilon/L, F, m)$$

**Theorem 17.7** (Generalization of Theorem 17.1)

$$P^m \left\{ \exists f \in \mathcal{F} \text{ s.t. } \left| \mathbb{E} l_f - \frac{1}{m} \sum_{i=1}^{m} l_f(z_i) \right| \geq \epsilon \right\} \leq 4\mathcal{N}_1(\epsilon/(8L), F, 2m) e^{-\epsilon^2 m/(32B^4)}$$