Neural Network Learning: Theoretical Foundations Chapter 20,21

IDEA Seminar Speaker : Dongha Kim

Department of Statistics, Seoul National University, South Korea

November 29, 2017

Lower and Upper bounds on sample complexity

- In chapter 19, there is a considerable gap between our lower and upper bounds on sample complexity.
- The sample complexity of any learning algorithm for a class F satisfies

$$m(\epsilon, \delta, B) = \Omega\left(\frac{1}{\epsilon} + \mathrm{fat}_F(4\epsilon)\right).$$

- And there is a learning algorithm(approximate-SEM) with sample complexity

$$m(\epsilon, \delta, B) = O\left(\frac{1}{\epsilon^2}\left(\mathrm{fat}_F(\epsilon/256)\log^2\left(\frac{1}{\epsilon}\right)\right)\right).$$

✓ There are function classes demonstrating that both rates are possible.

Main results of this chapter

1. If a function class F is *almost convex*, the sample complexity of this class is of order $1/\epsilon$.
2. And if F is not *almost convex*, the sample complexity in this case is of order at least $1/\epsilon^2$

What is *almost convex*?

Definition 20.1 *For a probability distribution P_X on X, define the* norm induced by P_X *on the set of functions $f : X \to \mathbb{R}$ as*

$$\|f\| = \left(\int_X f^2(x)\, dP_X(x)\right)^{1/2}.$$

For a class F of real-valued functions defined on a set X and a probability distribution P_X on X, let $\bar{F}$ denote the closure of F with respect to this norm. We say that such a class F is closure convex *if, for all probability distributions P_X on X, $\bar{F}$ is convex.*

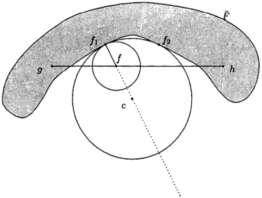
Lower bounds for non-convex classes

Main theorem 1.

Theorem 20.2 *For every class F that is not closure convex, there is a positive constant k and a bound B' such that for all $0 < \delta < 1$, all sufficiently small $\epsilon > 0$, all $B \geq B'$, and all learning algorithms L for F, the sample complexity satisfies*

$$m_L(\epsilon, \delta, B) \geq \frac{k \ln(1/\delta)}{\epsilon^2}.$$

Proof)



- This is enough to show that, by positioning $E(y|x)$ inside the ball approximately equidistant from f_1 and f_2, we can make the learning problem as the problem of estimating the probability of a Bernoulli random variable.

Lower bounds for non-convex classes

Lemma 5.1 *Suppose that α is a random variable uniformly distributed on $\{\alpha_-, \alpha_+\}$, where $\alpha_- = 1/2 - \epsilon/2$ and $\alpha_+ = 1/2 + \epsilon/2$, with $0 < \epsilon < 1$. Suppose that $\xi_1, \ldots, \xi_m$ are i.i.d. (independent and identically distributed) $\{0,1\}$-valued random variables with $\Pr(\xi_i = 1) = \alpha$ for all*

i. Let f be a function from $\{0,1\}^m$ to $\{\alpha_-, \alpha_+\}$. Then

$$\Pr(f(\xi_1, \ldots, \xi_m) \neq \alpha) > \frac{1}{4}\left(1 - \sqrt{1 - \exp\left(\frac{-2\lceil m/2 \rceil \epsilon^2}{1 - \epsilon^2}\right)}\right). \quad (5.1)$$

Hence, if this probability is no more than δ, where $0 < \delta < 1/4$, then

$$m \geq 2\left\lfloor \frac{1-\epsilon^2}{2\epsilon^2} \ln\left(\frac{1}{8\delta(1-2\delta)}\right) \right\rfloor. \quad (5.2)$$

2-layered networks class is not convex

- Consider the class F_k of two-layer networks, with a linear output unit and k first-layer computation units, each with the standard sigmoid activation function, $\sigma(\alpha) = 1/(1 + e^{-\alpha})$.

Theorem 20.5 *For any $k \in \mathbb{N}$, the class F_k is not convex, even if the input space is $X = \mathbb{R}$.*

- As a result, if the parameters are restricted to any compact set, the sample complexity of this class grows as $\log(1/\delta)/\epsilon^2$.

Upper Bounds for Convex Classes

Main theorem 2.

Theorem 20.7 *Suppose F is a closure convex class of functions that map to the interval $[0,1]$, $\mathcal{A}$ is an approximate-SEM algorithm for F, and $L(z) = \mathcal{A}(z, 1/m)$ for $z \in Z^m$. Suppose that the distribution P on $X \times \mathbb{R}$ is such that $|f(x) - y| \leq B$ almost surely. Then*

$$\begin{aligned} P^m & \left\{ \mathrm{er}_P(L(z)) \geq \inf_{f \in F} \mathrm{er}_P(f) + \epsilon \right\} \\ & \leq \ 6\mathcal{N}_1\left(\frac{\epsilon}{96B^3}, F, 2m\right) \exp\left(-\frac{\epsilon m}{5216 B^4}\right). \end{aligned}$$

Hence, if F has finite fat-shattering dimension, then L is a learning algorithm with

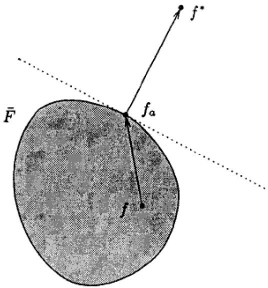
$$m_L(\epsilon, \delta) = O\left(\frac{B^4}{\epsilon}\left(d \ln^2\left(\frac{B}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right),$$

where $d = \mathrm{fat}_F\left(\epsilon/(768B^3)\right)$. Furthermore, if $d = \mathrm{Pdim}(F)$ is finite, L is a learning algorithm, and

$$m_L(\epsilon, \delta) = O\left(\frac{B^4}{\epsilon}\left(d \ln\left(\frac{B}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right).$$

Upper Bounds for Convex Classes

Proof)



- Set $g = l_f - l_{f_a}$ where $l_f = (y - f(x))^2$ and apply the following lemma.

Upper Bounds for Convex Classes

Proof(cont.))

Lemma 20.8 *Fix constants* $K_1 > 0$ *and* $K_2 \geq 1$. *Consider a class* G *of real functions defined on a set* Z, *and suppose that for every* $g \in G$ *and every* $z \in Z$, $|g(z)| \leq K_1$. *Let* P *be a probability distribution on* Z *for which* $\mathbf{E}g(z) \geq 0$ *and* $\mathbf{E}(g(z))^2 \leq K_2 \mathbf{E}g(z)$ *for all* g *in* G. *Then for* $\epsilon > 0$, $0 < \alpha \leq 1/2$ *and* $m \geq \max\left\{4(K_1 + K_2)/(\alpha^2\epsilon), K_1^2/(\alpha^2\epsilon)\right\}$,

$$P^m\left\{\exists g \in G, \frac{\mathbf{E}g - \hat{\mathbf{E}}_z g}{\mathbf{E}g + \epsilon} \geq \alpha\right\}$$
$$\leq 2\mathcal{N}_1\left(\frac{\alpha\epsilon}{8}, G, 2m\right)\exp\left(-\frac{3\alpha^2\epsilon m}{8K_1 + 324K_2}\right) +$$
$$4\mathcal{N}_1\left(\frac{\alpha\epsilon}{8K_1}, G, 2m\right)\exp\left(-\frac{\alpha^2\epsilon m}{4K_1^2}\right),$$

where $\hat{\mathbf{E}}_z g = \frac{1}{m}\sum_{i=1}^m g(z_i)$ *for* $z = (z_1, \ldots, z_m)$.

Restricted model

- If the conditional expectation $E(y|x) \in F$, the rate of uniform convergence is the same as the fast rate achieved by convex classes.

Theorem 20.10 *Suppose that F is a class of functions that map to the interval $[0,1]$, $\mathcal{A}$ is an approximate-SEM algorithm for F, $L(z) = \mathcal{A}(z, 1/m)$ for $z \in Z^m$, and the distribution P on $X \times \mathbb{R}$ is such that $|f(x) - y| \leq B$ almost surely and $\mathbf{E}(y|x)$ is in F. Then*

$$\begin{aligned} \mathcal{P}^m \left\{ \mathrm{er}_P(L(z)) \geq \inf_{f \in F} \mathrm{er}_P(f) + \epsilon \right\} \\ \leq \quad 6\,\mathcal{N}_1\left(\frac{\epsilon}{96B^3}, F, 2m\right) \exp\left(-\frac{\epsilon m}{5216 B^4}\right). \end{aligned}$$

Loss Functions in General

- We shall assume that the loss function l maps to the interval $[0, 1]$. (ex: $Y \in [0, 1]$)
- Given a particular loss function l, we define, for $f \in F$, the function $l_f : X \times Y \to [0, 1]$ by

$$l_f(x, y) = l(f(x), y),$$

and we let $l_F = \{l_f : f \in F\}$ be the corresponding loss class.

- The l-error of $f \in F$ with repect to a distribution P on $Z = X \times Y$ is the expected value of l_f with respect to P,

$$er_P^l(f) = El_f = El(f(x), y),$$

and, for $z \in Z^m$, the l-sample error $\hat{er}_z^l(f)$ is

$$\hat{er}_z^l(f) = \frac{1}{m}\sum_{i=1}^{m} l_f(x_i, y_i) = \frac{1}{m}\sum_{i=1}^{m} l(f(x_i), y_i).$$

Convergence for General Loss Functions

Theorem 17.1 *Suppose that F is a set of functions defined on a domain X and mapping into the real interval $[0,1]$. Let P be any probability distribution on $Z = X \times [0,1]$, ϵ any real number between 0 and 1, and m any positive integer. Then*

$$\begin{aligned} & P^m \{\textit{some } f \textit{ in } F \textit{ has } |\mathrm{er}_P(f) - \hat{\mathrm{er}}_z(f)| \geq \epsilon\} \\ \leq \; & 4\mathcal{N}_1(\epsilon/16, F, 2m) \exp\left(-\epsilon^2 m/32\right). \end{aligned}$$

Theorem 21.1 *Suppose that F is a class of functions mapping into the interval $[0,1]$, and that $\ell : [0,1] \times Y \to [0,1]$ is a loss function. Let P be any probability distribution on $Z = X \times Y$, $0 < \epsilon < 1$, and m any positive integer. Then*

$$\begin{aligned} & P^m \left\{|\mathrm{er}_P^\ell(h) - \hat{\mathrm{er}}_z^\ell(h)| \geq \epsilon \text{ for some } h \in F\right\} \\ \leq \; & 4\mathcal{N}_1\left(\frac{\epsilon}{8}, \ell_F, 2m\right) \exp\left(-\frac{\epsilon^2 m}{32}\right). \end{aligned}$$

Corollary 21.2 *Let ℓ denote the absolute loss function. Then, for all positive integers k and for all positive numbers ϵ,*

$$\mathcal{N}_1(\epsilon, \ell_F, k) \leq \mathcal{N}_1(\epsilon, F, k).$$

Learning in Multiple-Output Networks

- Suppose that F maps from a set X into $\mathbb{R}^s$ where $s > 1$.
- It would seem appropriate to use the loss function $l^s : \mathbb{R}^s \times \mathbb{R}^s \to [0, 1]$, as follows:

$$l^s(y, y') = \frac{1}{s}\sum_{i=1}^{s} l(y_i, y_i').$$

- For instance, l^s measures the loss as the average quadratic loss over the outputs,

$$l^s(y, y') = \frac{1}{s}\sum_{i=1}^{s}(y_i - y_i')^2.$$

- For $1 \le s \le s$ and $f \in F$, let $f_i(x) = (f(x))_i$, the ith entry of $f(x) \in \mathbb{R}^s$, and let $F_i = \{f_i : f \in F\}$.
- For $f \in F$, we define $l_{f_i} : \mathbb{R}^s \times \mathbb{R}^s \to [0, 1]$ by $l_{f_i}(x, y) = l(f_i(x), y)$ and we let $l_{F_i} = \{l_{f_i}, f \in F\}$.

Learning in Multiple-Output Networks

Theorem 21.3 *With the above notations,*

$$\begin{aligned}\mathcal{N}_1(\epsilon, \ell_F^s, k) &\leq \mathcal{N}_1(\epsilon, \ell_{F_1}, k)\,\mathcal{N}_1(\epsilon, \ell_{F_2}, k)\cdots\mathcal{N}_1(\epsilon, \ell_{F_s}, k)\\ &= \prod_{i=1}^{s} \mathcal{N}_1(\epsilon, \ell_{F_i}, k),\end{aligned}$$

for all positive integers k and all $\epsilon > 0$.

Corollary 21.4 *If ℓ is the quadratic loss function then*

$$\begin{aligned}\mathcal{N}_1(\epsilon, \ell_F^s, k) &\leq \mathcal{N}_1\left(\frac{\epsilon}{2}, F_1, k\right)\mathcal{N}_1\left(\frac{\epsilon}{2}, F_2, k\right)\cdots\mathcal{N}_1\left(\frac{\epsilon}{2}, F_s, k\right)\\ &= \prod_{i=1}^{s} \mathcal{N}_1\left(\frac{\epsilon}{2}, F_i, k\right),\end{aligned}$$

for all positive integers k and all $\epsilon > 0$. If ℓ is the absolute loss function, then

$$\mathcal{N}_1(\epsilon, \ell_F^s, k) \leq \prod_{i=1}^{s} \mathcal{N}_1(\epsilon, F_i, k)$$

for all ϵ and k.

Learning in Multiple-Output Networks

Theorem 21.5 *Suppose that a feed-forward network N has W weights and k computation units arranged in L layers, where s of these computation units are output units. Suppose that each computation unit has a fixed piecewise-polynomial activation function with p pieces and degree no more than l. Let F be the class of functions computed by N. Then any approximate-SEM algorithm for F can be used to define a learning algorithm for F, and for fixed p and l, the sample complexity of this algorithm is*

$$O\left(\frac{1}{\epsilon^2}\left(s\left(WL\ln W + WL^2\right)\ln\left(\frac{1}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right).$$

Theorem 21.6 *Consider the class of two-layer networks defined in Corollary 14.16, but with s output units. These networks have inputs in $[-A, A]^n$, and each computation unit has a bound V on the sum of the magnitudes of the associated parameters, and an activation function that is bounded and satisfies a Lipschitz constraint. Let F be the class of vector-valued functions computed by this network. Any approximate-SEM algorithm can be used to define a learning algorithm L for F that has sample complexity satisfying*

$$m_L(\epsilon, \delta) = O\left(\frac{1}{\epsilon^2}\left(\frac{sV^6A^2}{\epsilon^4}\ln n + \ln\left(\frac{1}{\delta}\right)\right)\right).$$

Interpolation Models

- In this section, we take a fresh approach to the question of how to extend a basic learning model of Part I for binary classification to models of learning applicable to real-valued function classes.

Theorem 4.8 *Suppose that H is a set of functions from a set X to $\{0,1\}$ and that H has finite Vapnik-Chervonenkis dimension $d \geq 1$. Let L be a consistent algorithm; that is, for any m and for any $t \in H$, if $x \in X^m$ and z is the training sample corresponding to x and t, then the hypothesis $h = L(z)$ satisfies $h(x_i) = t(x_i)$ for $i = 1, 2, \ldots, m$. Then L is a learning algorithm for H in the restricted model, with sample complexity*

$$m_L(\epsilon, \delta) \leq \frac{4}{\epsilon}\left(d \ln\left(\frac{12}{\epsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right)$$

Interpolation Models

- Therefore, there is $m(\epsilon, \delta)$ such that for $m \geq m(\epsilon, \delta)$, for any probability distribution μ in X and any function $t \in H$, the following holds:

$$P^m \text{ (for any function } f \text{ such that } f(x_i) = t(x_i) \text{ for } i = 1, ..., m,$$
$$\mu\{f(x) = t(x)\} > 1 - \epsilon) > 1 - \delta.$$

Real-valued problem

- We extend in two different ways the restricted model of learning for $\{0, 1\}$-classes.
- Suppose t is any function from X to $[0, 1]$ (not necessarily in the class F), and μ is a probability distribution on X.
- In real-valued problem, we replace the condition $f(x) = t(x)$ to $|f(x) - t(x)| < \eta$.

Interpolation Models

Definition 21.7 *Suppose that F is a class of functions mapping from a set X to the interval* $[0, 1]$. *Then F* strongly generalizes from approximate interpolation *if for any $\epsilon, \delta, \eta \in (0, 1)$, there is $m_0(\epsilon, \delta, \eta)$ such that for $m \geq m_0(\epsilon, \delta, \eta)$, for any probability distribution μ in X and any function $t : X \to [0, 1]$, the following holds: with probability at least $1 - \delta$, if $x = (x_1, x_2, \ldots, x_m) \in X^m$, then for any $f \in F$ satisfying $|f(x_i) - t(x_i)| < \eta$ for $i = 1, 2, \ldots, m$, we have*

$$\mu\{x : |f(x) - t(x)| < \eta\} > 1 - \epsilon.$$

Definition 21.8 *Suppose that F is a class of functions mapping from a set X to the interval* $[0, 1]$. *Then F* generalizes from approximate interpolation *if for any $\epsilon, \delta, \eta, \gamma \in (0, 1)$, there is $m_0(\epsilon, \delta, \eta, \gamma)$ such that for $m \geq m_0(\epsilon, \delta, \eta, \gamma)$, for any probability distribution μ in X and any function $t : X \to [0, 1]$, the following holds: with probability at least $1 - \delta$, if $x = (x_1, x_2, \ldots, x_m) \in X^m$, then for any $f \in F$ satisfying $|f(x_i) - t(x_i)| < \eta$ for $i = 1, 2, \ldots, m$, we have*

$$\mu\{x : |f(x) - t(x)| < \eta + \gamma\} > 1 - \epsilon.$$

Interpolation Models

Strong generalization from interpolation

Theorem 21.12
Suppose that F is a set of functions from a set X to $[0,1]$. Then F strongly generalizes from approximate interpolation if and only if F has finite pseudo-dimension. Furthermore, if F has finite pseudo-dimension $\mathrm{Pdim}(F)$ *then a sufficient sample length function for generalization from approximate interpolation is*

$$m_0(\epsilon,\delta,\eta) = \frac{4}{\epsilon}\left(15\,\mathrm{Pdim}(F)\ln\left(\frac{12}{\epsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right),$$

and any suitable sample length function must satisfy

$$m_0(\epsilon,\delta,\eta) \geq \frac{1}{24\epsilon}\left(\frac{\mathrm{Pdim}(F)}{2\ln(2/\eta)} - 1 + 6\ln\left(\frac{1}{\delta}\right)\right)$$

for all $\eta > 0$, $\epsilon \in (0,1/2)$ and $\delta \in (0,1)$.

Interpolation Models

Generalization from interpolation

Theorem 21.14 *Suppose that F is a class of functions mapping into $[0,1]$. Then F generalizes from approximate interpolation if and only if F has finite fat-shattering dimension. Furthermore, there is a constant c such that if F has finite fat-shattering dimension, then a sufficient sample length for generalization from approximate interpolation is*

$$m_0(\epsilon,\delta,\gamma,\eta) = \frac{c}{\epsilon}\left(\ln\left(\frac{1}{\delta}\right) + \mathrm{fat}_F\left(\frac{\gamma}{8}\right)\ln^2\left(\frac{\mathrm{fat}_F(\gamma/8)}{\gamma\epsilon}\right)\right).$$

A result on large margin classification

- It is possible to use our results on generalization from approximate interpolation to derive a result useful for a restricted form of the classification learning model of Part 2.
- Recall that in this framework, for a probability distribution P on $X \times \{0,1\}$, a positive number γ, and $f \in F$, we define

$$er_P^\gamma(f) = P\{\, margin(f(x), y) < \gamma\}.$$

- In Chapter 10, we proved the following convergence result:

$$P^m\{\text{some } f \text{ in } F \text{ has } er_P(f) \geq \hat{er}_z^\gamma(f) + \epsilon\}$$
$$\leq 2\mathcal{N}_\infty\left(\frac{\gamma}{2}, F, 2m\right) \exp\left(-\frac{\epsilon^2 m}{8}\right).$$

A result on large margin classification

Theorem 21.15 *Suppose that F is a set of functions mapping from a set X to $[0,1]$, that $t : X \to \{0,1\}$, and that μ is a probability distribution on X. Let $\gamma \in (0, 1/2]$ and $\epsilon \in (0,1)$. For $f \in F$, define $\mathrm{er}_\mu(f,t)$ to be $\mu\{x : \mathrm{sgn}(f(x) - 1/2) \neq t(x)\}$, the error incurred in using the function f for binary classification. Let P_{bad} be the probability of $x \in X^m$ for which some $f \in F$ has $\mathrm{margin}(f(x_i), t(x_i)) > \gamma$ for $i = 1, \ldots, m$, but $\mathrm{er}_\mu(f,t) \geq \epsilon$. Then $P_{\mathrm{bad}} \leq 2\mathcal{N}_\infty(\gamma/2, F, 2m)\, 2^{-\epsilon m/2}$.*

- the above result is similar to this, but is more specialized in two ways:
  - restricted model.
  - $er_\mu(f) \geq \epsilon$ and $\hat{er}^\gamma_\mu(f) = 0$, not $er_P(f) \geq \hat{er}^\gamma_z(f) + \epsilon$.