Estimation Error for Regression and Optimal Convergence Rate

Yao Wang

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Mathematics
Adviser: Professor Weinan E

June 2018
Abstract

In this thesis, we study the optimal convergence rate for the universal estimation error. Let $\mathcal{F}$ be the excess loss class associated with the hypothesis space and $n$ be the size of the data set, we prove that if the Fat-shattering dimension satisfies $\text{fat}_r(\mathcal{F}) = O(\epsilon^{-p})$, then the universal estimation error is of $O(n^{-1/2})$ for $p < 2$ and $O(n^{-1/p})$ for $p > 2$. Among other things, this result gives a criterion for a hypothesis class to achieve the minmax optimal rate of $O(n^{-1/2})$. Examples are also provided for optimal rates not equal to $O(n^{-1/p})$, such as compact supported convex Lipschitz continuous functions in $\mathbb{R}^d$ with $d > 4$ with optimal rate approximately about $O(n^{-2/d})$.

Training in practice may only explore a certain subspace in $\mathcal{F}$. It is useful to bound the complexity of the subspace explored instead of the whole $\mathcal{F}$. This is done for the gradient descent method.
Acknowledgements

I would like to thank my advisor Professor Weinan E for always encouraging me and providing his invaluable guidance. Professor E always asked me to step out my comfortable zone, dive into the unknown for deep understanding and always look for the big picture. His hard work and focus is a testament to his dedication to his students. Inspired by him, I built my confidence and ambition, which played a very important role in my academic progress and will certainly guide me in professional career. I consider it a tremendous honor to have the privilege to work with Professor E at Princeton and I will remain forever in his debt.

I would also like to thank all the faculty and staff of the mathematics department of Princeton University. It is through their support that my time at Princeton proved enjoyable.

I likewise express my sincere gratitude to Jianqing Fan, Qiang Du and Zuowei Shen for agreeing to serve on my committee and provide feedback to this dissertation.

I thoroughly enjoyed my time at Princeton and truly appreciate the opportunity to learn so much, though hard work, determination and focus, both in and out of the classroom. I thank all the members of Professor Weinan E’s group: Jiequn Han, Lei Wu, Steven Chen, Qingcan Wang, Ruying Bao, Chao Ma, Lingfeng Zhang, Yucheng Yang and Fan Xiao.

Of course, I must express my sincere thanks and appreciation to my friends, who always believed in me, stayed with me through my challenges and helped make my graduate experience delightful: Hongyang Jia, Xinlei Sheng, XiaoJie Shi, Chunhung Liu, Inliang Tseng, Justen Pinto and Runpu Zong.

Above all, I would like to thank my parents, Juncai Wang and Wenjie Chen, for their enduring love and support. I dedicate this dissertation to them.
To my parents.
Contents

Abstract ................................................................. iii
Acknowledgements ..................................................... iv

1 Introduction ............................................................. 1

2 Preliminary Results ................................................... 4
   2.1 Empirical Process .................................................. 4
   2.2 Generalization Error and Estimation Error ....................... 7
   2.3 Function Space Complexity ....................................... 9

3 Convergence Rate for Estimation Error ............................. 14
   3.1 Maximal Inequality ................................................ 14
   3.2 Upper Bound ..................................................... 16
   3.3 Lower Bound ................................................... 20

4 Applications to Common Function Classes ......................... 23
   4.1 Regularized Linear Class .......................................... 23
   4.2 Bounded Variation Class .......................................... 25
   4.3 Smooth Function Class ........................................... 28
   4.4 Piecewise Smooth Function Space ............................... 29
   4.5 Convex Functions ................................................ 31

5 Applications to Function Classes in Statistical Learning ......... 33
   5.1 Excess Loss Class or Hypothesis Class ......................... 33
Chapter 1

Introduction

Analysis of the generalization error for a learning model is one of the main topics regarding the mathematical foundations of learning theory. It is important to understand which factors influence the performance of a model and how to improve the model. Although the first step is to consider the asymptotic analysis of the behavior of the generalization error when the number of data points goes to infinity, it is more valuable to study a finite data generalization error bound since it is helpful for model selection and parameters tuning. The generalization bounds consist of two parts: one term to bound the empirical error on the training data and the other is the penalty term related to the number of data and the complexity of the function space that model can explore. We call this penalty term the estimation error, and it can be studied through the empirical process measuring the difference between the true error and the error on the data. The empirical process is indexed by the function space explored, and it embodies the difficulty associated with finding the best fitting function in a certain hypothesis space using the data at hand. However, there is no straightforward way of analyzing this object since the data distribution is always unknown. The classical approach is to introduce distribution-free complexity measure to bound the penalty term.

Vapnik [7] pioneered this research and introduce the concept of VC dimension to describe the complexity of \{0,1\}-valued function space. Other measures include Fat-
shattering dimension which is a scale sensitive version of VC dimension, covering number which measures the number of balls required to cover the function space with respect to some norm and is used to describe the 'size' of real value functions on a finer scale. Another quantity is Rademacher average with an idea comes from concentration inequalities and empirical process theory.

In Chapter 2, preliminaries for this thesis are provided. In Section 2.1, we introduce the concept of empirical process. The Donsker class and Glivenko-Cantelli class are defined in this section. Roughly speaking, Glivenko-Cantelli class is the class of functions that uniformly satisfies the Law of Large Numbers. Donsker class is the class of functions that satisfies the Central Limit Theorem. Section 2.2 provides definitions of the generalization error and estimation error, which are the main topics of this thesis. Section 3 introduces the measures for the richness of function space, these measures provide the main methodologies for later chapters.

In Chapter 3, we first discuss the maximal inequality in Section 3.1. Knowing that the Rademacher random variable \(X(a) = \sum_{i=1}^{n} r_i a_i\) has an even lighter tail than the Gaussian variable, bounding a large number of such random variables uniformly becomes possible, and can be done by using \(\psi_2\)-Orlicz norm. We provide the main theorem in this thesis about the optimal rate in Section 3.2 and Section 3.3. In the statistics community, the rate of \(1/\sqrt{n}\) for estimation error can be guaranteed, in most cases. These two sections provide criteria about when this rate can be achieved. Moreover, they also provide the optimal rate in the case when the underlying function space is too rich to achieve a convergence rate of \(1/\sqrt{n}\).

In Chapter 4, we focus on the function spaces which are fundamental and basic to approximation theory. Classical results about their approximation error have been thoroughly studied for a long time in various literature. In this thesis, we will give some results regarding their estimation error, hence completing the whole picture for their generalization error. In Section 4.1, we give results about the universal estimation error for regularized linear class. We also discuss the space of bounded variation functions on \(\mathbb{R}\)
and the space of monotonic functions on \( \mathbb{R}^d \) in Section 4.2. Convergence rate results related to smooth functions and convex functions are given from Section 4.3 to Section 4.5.

In Chapter 5, we show how to calculate the estimation error bound in practice. In Section 5.1, we study the relationship between function space and excess loss function space. In general, the complexities of the original function space and the excess loss function space can control each other. We give the estimation bound for VC class. In machine learning, function space with finite VC dimension can yield a good bound for estimation error in classification problems. However, through the work of Massart and et al. [6], a convergence rate faster than \( 1/\sqrt{n} \) can be achieved if the distribution has some additional nice properties. Various results regarding the covering number of neural networks, boosting, and support vector machines have been developed recently. We apply our estimation bound calculations to these results and get the corresponding results from section 5.3 to section 5.5.

In all the chapters we discussed so far, we assume that there is a fixed function space \( \mathcal{F} \) for the model. To give bound for the generalization error, we are always trying to bound the empirical process index by \( \mathcal{F} \). Essentially, we always consider the worst case scenario occurred in our hypothesis space. However, in practice, the training process is unlikely to explore the worst functions in \( \mathcal{F} \). In this Chapter 6, we prove that for neural networks, given fixed initial values, the function space that we need to deal with is not as large as we think. The training algorithm, such as gradient descent method, actually explores a subspace in \( \mathcal{F} \) instead of the whole \( \mathcal{F} \). This result sheds light on the fact that we need to consider the algorithm used when trying to bound the generalization error, instead of only the model. In section 6.2, we present both the training dynamics and the generalization dynamics when training linear model with gradient descent.
Chapter 2

Preliminary Results

2.1 Empirical Process

In this section, we define and sketch the main features and asymptotic behavior of the empirical process and we introduce two main classes: the Glivenko-Cantelli class and the Donsker class. Empirical process has become increasingly important recently. On one hand, it provides a powerful tool for proving many of the main limit theorems in statistics. On the other hand, it has a variety of applications in many fields, such as semiparametric estimation, functional delta-method and bootstrap.

A stochastic process is a collection of random variables \( \{X(t), t \in T\} \) on the same probability space \((\Omega, \mu)\), indexed by an arbitrary index set \(T\). The corresponding empirical process is defined as a stochastic process based on a random sample; given random samples \(X_1, \ldots, X_n\) independently generated from a probability measure \(\mu\), on a sample space \(X\). We define the empirical measure \(\mu_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}\), where \(\delta_x\) is the Dirac delta function.

For a measurable function \(f : X \to \mathbb{R}\), we denote \(\mathbb{E}_{\mu_n} f = n^{-1} \sum_{i=1}^{n} f(X_i)\). The empirical measure then induces a map from \(\mathcal{F}\) to \(\mathbb{R}\) given by

\[
f \mapsto \mathbb{E}_{\mu_n} f\]
Here, we use abbreviation $E_\mu f = \int fd\mu$ for a given measurable function $f$ and measure $\mu$. Then we can define a $\mathcal{F}$-indexed empirical process $\{G_n f : f \in \mathcal{F}\}$, where $G_n f$ is given by

$$f \mapsto G_n f = \sqrt{n}(E_{\mu_n} - E_{\mu})f$$

Then

$$G_n = \sqrt{n} \sum_{i=1}^{n}(\delta_{X_i} - E_{\mu})$$

For a given function $f$, if $E_{\mu}f$ exists and $E_{\mu}f^2 < \infty$, we have the Law of Large Numbers and the Central Limit Theorem that:

$$E_{\mu_n} f \xrightarrow{a.s.} E_{\mu} f,$$

$$G_n f \rightsquigarrow N(0, E_{\mu}(f - E_{\mu}f)^2),$$

where $\rightsquigarrow$ denotes convergence in distribution.

In order to use the Law of Large Numbers and the Central Limit Theorem for a function space, we introduce the notation $\|Q\|_{\mathcal{F}} = \sup\{|Qf| : f \in \mathcal{F}\}$, also we assume that $\{f : f \in \mathcal{F}\}$ is uniformly bounded. Then the uniform version of the Law of Large Numbers becomes:

$$\|E_{\mu_n} - E_{\mu}\|_{\mathcal{F}} \to 0$$

A class $\mathcal{F}$ for which this holds, is called a Glivenko-Cantelli class.

The definition for function space version Central Limit Theorem requires the following Lemma [1]:

**Lemma 2.1.1.** The space $l^\infty(\mathcal{F})$ is defined as the collection of all uniformly bounded, real functions on $\mathcal{F}$. Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be processes indexed by $F$ and induce tight Borel measurable maps from probability space $(\Omega, \mu)$ to $l^\infty(\mathcal{F})$. Then $\mathcal{G}_1$ and $\mathcal{G}_2$ are equal in distribution if and only if all corresponding marginals of $\mathcal{G}_1$ and $\mathcal{G}_2$ are equal in distribution.
For the function space version of the Central Limit Theorem, it makes sense to study the conditions under which

\[ G_n = \sqrt{n}(E_{\mu_n} - E_{\mu}) \Rightarrow G, \text{ in } l^\infty(F) \]

where the limit \( G \) is a tight Borel measurable map into \( l^\infty(F) \) and we need to think of the convergence of distribution in terms of sample path in this setting. To get an intuition about what \( G \) should be, consider the marginal distribution of \( G_n \). For \( f_1, \ldots, f_k \) where \( E_{\mu} f_i^2 < \infty \), the multivariate Central Limit Theorem gives:

\[ (G_n f_1, \ldots, G_n f_k) \Rightarrow N_k(0, \Sigma), \]

where the \((i,j)\)-th element of \( \Sigma \) is given by

\[ E_{\mu} (f_i - E_{\mu} f_i)(f_j - E_{\mu} f_j) = E_{\mu} f_i f_j - E_{\mu} f_i E_{\mu} f_j \]

Since convergence in \( l^\infty(F) \) means marginal convergence, the limit process \( G \) should be a mean zero Gaussian process indexed by \( F \), with covariance \( E_{\mu}(fg) - E_{\mu} f E_{\mu} g \) for all \( f, g \in F \). By applying the above lemma, we know the marginal distribution determines the distribution of \( G \) in \( l^\infty(F) \). A class that satisfies the above condition is then called a Donsker class.

Whether a given function space \( F \) is a Glivenko-Cantelli or a Donsker class depends on its complexity. For example, a function space containing only finite number of square integrable functions is always Donsker, but the function space containing all bounded functions on \( \mathbb{R} \) can never be Donsker class. We will give criteria for a function space to be Glivenko-Cantelli or Donsker in later chapters.
2.2 Generalization Error and Estimation Error

In this section, we will introduce the concept of generalization error and estimation error in learning problems.

Intuitively, given some data independently generated by the same underlying distribution and some model class, we are interested in how close the model trained with the data is to the best possible model for the underlying distribution. The gap is known as the generalization error in the context of supervised learning. The model class is called the hypothesis space. We can decompose the generalization error into two parts. One is the difference between the best possible model and the best model in the hypothesis space. This is known as the approximation error. The second part is called the estimation error, which is the difference between the best model from the hypothesis space and the model trained with the data.

We will use the following notations: We denote the data set by $\{Z_i = (X_i, Y_i)\}_1^n$, which is generated independently from the same underlying distribution $\mu$, here $X_i$ is the $i$-th input and $Y_i$ is the corresponding output. $L$ is the loss function and $H$ is the hypothesis space which contains functions from $X$ to $Y$. The risk is the expectation of the loss function with respect to $\mu$, and the empirical loss is the expectation of the loss function with respect to $\mu_n$. Let $h_{\text{best}}$ be the minimizer of the risk, $h^*$ be the minimizer of the risk associated with $H$ and $\hat{h}$ be the minimizer of the empirical risk:

\[
\begin{align*}
    h_{\text{best}} &:= \text{argmin}_{h} \mathbb{E}_{\mu}(L(h)) \\
    h^* &:= \text{argmin}_{h \in H} \mathbb{E}_{\mu}(L(h)) \\
    \hat{h} &:= \text{argmin}_{h \in H} \mathbb{E}_{\mu_n}(L(h)).
\end{align*}
\]

Here, for simplification, we use $L(h)$ in place of the function $L(h(X), Y)$ and $\mu_n$ is the empirical measure from last section. Then the generalization error is defined as
$L(\hat{h}) - L(h_{\text{best}})$, which can be rewritten as follows:

$$E_\mu(L(\hat{h}) - L(h_{\text{best}})) = E_\mu((L(\hat{h}) - L(h^*)) + E_\mu((L(h^*) - L(h_{\text{best}})))$$

We will call the first term on the right hand side the estimation error and second term the approximation error.

To estimate the estimation error, instead of looking at the space $\mathcal{H}$, we will instead look at the excess loss class associated with $\mathcal{H}$, denoted as $\mathcal{F}$, see [11].

$$\mathcal{F} := \{Z = (X,Y) \rightarrow L(h) - L(h^*) : h \in \mathcal{H}\}.$$ 

Every function $h \in \mathcal{H}$ corresponds to an element in $\mathcal{F}$. Let $\hat{f}$ and $f^*$ in $\mathcal{F}$ be the corresponding elements of $\hat{h}$ and $h^*$ in $\mathcal{H}$, respectively. Obviously $f^* \equiv 0$. Now the estimation error can be written as $E_\mu \hat{f}$. Since $f^*$ is the minimizer of $E_\mu f$ and $E_\mu f^* = 0$, we know that $E_\mu \hat{f} \geq 0$. Similarly, we know that $\hat{f}$ is the minimizer of $\mathbb{P}_n f$ and because $E_\mu f^* = 0$, we have $E_{\mu_n} \hat{f} \leq 0$. Therefore, we have

$$0 \leq E_\mu \hat{f} \leq E_{\mu_n} \hat{f} - E_{\mu_n} \hat{f}$$

To bound the $E_\mu \hat{f}$, it is enough to bound $E_\mu \hat{f} - E_{\mu_n} \hat{f}$. However, we can not directly use the Law of Large Number or the Central Limit Theorem for $\hat{f}$ since $\hat{f}$ is the empirical minimizer, the independent and identically distributed assumption no longer holds.

The following example is informative. Suppose $\mathcal{F}$ contains all continuous functions with range bounded below by 0. Then the the empirical minimizer $\hat{f}$ can be any function interpolating the data set with value 0. This implies that $E_{\mu_n} \hat{f} = 0$. But there is no guarantee that $E_\mu \hat{f} = 0$, and hence no guarantee that $E_\mu \hat{f} - E_{\mu_n} \hat{f}$ converges as $n$ goes to infinity.

The solution to this dilemma is to study the differences between the true and empirical expectation of all functions in the whole excess loss class $\mathcal{F}$ rather than focusing only on
Thus it makes sense to study the empirical process \( \{ (\mathbb{E}_{\mu_n} - \mathbb{E}_\mu)f : f \in \mathcal{F} \} \) and its asymptotic behavior as \( n \) goes to infinity. We call \( \| \mathbb{E}_{\mu_n} - \mathbb{E}_\mu \|_F \) the empirical process supremum and its expectation, \( \mathbb{E}_\mu \| \mathbb{E}_{\mu_n} - \mathbb{E}_\mu \|_F \), will be called the \( \mu \)-estimation error. It naturally provides a good bound for the estimation error.

Also, to simplify our following analysis, we make the assumption that the function space \( \mathcal{F} \) is uniformly bounded and \( \sup_{f \in \mathcal{F}} \| f \| \leq 1 \).

### 2.3 Function Space Complexity

The main challenge to obtain Glivenko-Cantelli class or Donsker class is to extend point-wise convergence of one function to uniform convergence of a function space. Clearly the complexity of the function space plays an essential role. In this section, we will introduce some common complexity measures.

**Rademacher Average**

Rademacher average measures the complexity of a real value function space. It was first introduced into empirical process theory by Koltchinskii [28] and Pollard [29], and serves as a very useful tool for proving uniform convergence bounds for real value function class and provides natural bound for estimation error.

**Definition 2.3.1.** For empirical measure \( \mu_n \), we define the Rademacher average [18][8] by

\[
R(\mathcal{F}/\mu_n) = \mathbb{E}_r \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} r_i f(Z_i) \right|
\]

where \( r_1, \ldots, r_n \) are independent and identically distributed Rademacher random variables satisfying \( P(r = -1) = P(r = 1) = 1/2 \) and \( \mathbb{E}_r \) is the expectation with respect to the Rademacher variables.

This sample dependent quantity measures how correlated the most-correlated hypothesis is to a random labeling of data. Also, we define the Rademacher process associated
with the empirical measure $\mu_n$ as

$$X_{\text{rad}}(f) = \frac{1}{n} \sum_{i=1}^{n} r_i f(Z_i). \quad (2.2)$$

It is an accepted fact that the Rademacher averages control the $\mu$ estimation error [1]:

**Theorem 2.3.2.** If $\mathcal{F}$ is a class of functions that map into $[-M, M]$, then for every integer $n$, we have

$$E_{\mu} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} (f(Z_i) - E_{\mu} f) \right| \leq 2E_{\mu} R(\mathcal{F}/\mu_n)$$

$$\leq M E_{\mu} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} (f(Z_i) - E_{\mu} f) \right| + E_{\sigma} \left| \frac{1}{n} \sum_{i=1}^{n} r_i \right| \quad (2.3)$$

**Proof.** To understand the intuition of Rademacher average, we sketch the proof of the first part of Theorem 2.3.1 here. Let $Z'_1, \ldots, Z'_n$ be an independent copy of $Z_1, \ldots, Z_n$ and set $\mu'$ to be the corresponding empirical measure for $Z'_1, \ldots, Z'_n$, then

$$E_{\mu} \sup_{f \in \mathcal{F}} (E_{\mu} f - E_{\mu_n} f) = E_{\mu} \sup_{f \in \mathcal{F}} (E_{\mu} \mu'_n f - E_{\mu_n} f)$$

$$\leq E_{\mu} \sup_{f \in \mathcal{F}} (E_{\mu'_n} f - E_{\mu_n} f)$$

$$= \frac{1}{n} E_{\mu} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} r_i f(Z'_i) - r_i f(Z_i) \quad (2.1)$$

$$\leq \frac{1}{n} E_{\mu} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} r_i f(Z'_i) - \frac{1}{n} E_{\mu} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} r_i f(Z_i)$$

$$= 2E_{\mu} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} r_i f(Z_i) \right|$$

Using the same argument, it also holds when the left hand side is replaced by $E_{\mu} \sup_{f \in \mathcal{F}} (E_{\mu_n} f - E_{\mu})$. Hence complete the proof.

From Theorem 2.3.1, we see that the term $E_{\mu} \|E_{\mu_n} - E_{\mu}\|_{\mathcal{F}}$ is comparable to the expectation of the Rademacher average up to a term of $O(n^{-1/2})$. 

10
Covering Number

Although the Rademacher average provide a sharp bound for empirical process, it is
difficult to compute sometimes. We introduce covering number here that can be easier to
compute than Rademacher average, in most cases. In the following, the logarithm always
takes 2 as base and the $p$-norm $L_p(\mu_n)$ of $f \in \mathcal{F}$ on empirical measure $\mu_n$ is defined as
\[
\left(\frac{1}{n}\sum_{i=1}^{n}|f(Z_i)|^p\right)^{1/p}.
\]

**Definition 2.3.3.** For an arbitrary semi-metric space $(T, d)$ the covering number
$N(\epsilon, T, d)$ with $d$ being the metric is the minimal number of closed $d$-balls of radius $\epsilon$
required to cover $T$, see [16, 1]. The associated entropy $\log N(\epsilon, T, d)$ is the logarithm of
the covering number.

Intuitively, if the function space is compressed such that the distortion ( the difference
between the original and the compressed function in norm $d$) is less than $\epsilon$, then the
entropy is the least bit needed to code the function space. Note that both the covering
number is non-decreasing as $\epsilon$ decreases, and since $\|f\|_{L_1(\mu_n)} \leq \|f\|_{L_2(\mu_n)} \leq \|f\|_{L_{\infty}(\mu_n)}$, we know that
\[
N(\epsilon, \mathcal{F}, L_1(\mu_n)) \leq N(\epsilon, \mathcal{F}, L_2(\mu_n)) \leq N(\epsilon, \mathcal{F}, L_{\infty}(\mu_n))
\]

There are several results which connect the Clivenko-Cantelli class and Donsker class
to the corresponding covering number:

**Theorem 2.3.4.** [16] Let $\mathcal{F}$ be a set of functions which is uniformly bounded, then $\mathcal{F}$ is
Clivenko-Cantelli if and only if for every $1 \leq p \leq \infty$ and every $\epsilon > 0$
\[
\lim_{n \to \infty} \sup_{\mu_n} \log \frac{N(\epsilon, \mathcal{F}, L_p(\mu_n))}{n} = 0
\]
where the supreme is taken with respect to all finite empirical measures.

**Theorem 2.3.5.** [17] Let $\mathcal{F}$ be a set of functions which is uniformly bounded, if
\[
\int_0^{\infty} \sup_{\mu_n} \log^{1/2} N(\epsilon, \mathcal{F}, L_2(\mu_n))d\epsilon \leq \infty
\]
then \( \mathcal{F} \) is Donsker class. On the other hand, if \( \mathcal{F} \) is Donsker class, then there exists a constant \( C \), such that for any \( \epsilon > 0 \)

\[
\sup_{\mu_n} \log N(\epsilon, \mathcal{F}, L_2(\mu_n)) \leq \frac{C}{\epsilon^2}
\]

**Fat-shattering Dimension**

We also define another concept that is easy to calculate: Fat-shattering dimension. Fat-shattering dimension can be viewed as a scale sensitive version of VC dimension, and it is always used for real value problem.

**Definition 2.3.6.** For every \( \epsilon > 0 \), a set \( A = \{Z_1, \cdots, Z_n\} \) is said to be \( \epsilon \)-shattered by \( \mathcal{F} \) if there exists some real function \( s : A \to \mathbb{R} \) such that for every \( I \in \{1, \cdots, n\} \) there exists some \( f_I \in F \) such that \( f_I(Z_i) \geq s(Z_i) + \epsilon \) if \( i \in I \), and \( f_I(Z_i) \leq s(Z_i) - \epsilon \) if \( i \notin I \).

\[
fat_{\epsilon}(\mathcal{F}) := \sup \{ |A| \mid A \in \Omega, A \text{ is } \epsilon\text{-shattered by } \mathcal{F} \}
\]

is called the Fat-shattering dimension, \( f_I \) is called the shattering function of the set \( I \) and the set \( \{s(Z_i) \mid Z_i \in A\} \) is called a witness to the \( \epsilon \)-shatter.

The Fat-shattering dimension is actually linear with respect to the entropy up to a logarithm factor of the Fat-shattering dimension\(\text{[18]}\):

**Lemma 2.3.7.** If \( |f| \leq 1 \) for any \( f \in \mathcal{F} \), then

\[
\sup_{\mu_n} \log N(\epsilon, \mathcal{F}, L_1(\mu_n)) \geq fat_{16\epsilon}(\mathcal{F})/8.
\]

**Lemma 2.3.8.** For every empirical measure \( \mu_n \) and \( p \geq 1 \), there is some constant \( c_p \) such that

\[
\log N(\epsilon, \mathcal{F}, L_p(\mu_n)) \leq c_p fat_{\frac{2}{4\epsilon}}(\mathcal{F}) \log^2 \left( \frac{2 \text{fat}_{\frac{2}{4\epsilon}}(\mathcal{F})}{\epsilon} \right).
\]
Using the above lemma together with Theorem 2.3.4, we can conclude that a function space $\mathcal{F}$ is Glivenko-Cantelli if and only if it has finite Fat-shattering dimension for every $\epsilon > 0$. 
Chapter 3

Convergence Rate for Estimation Error

In this chapter we will prove that if the Fat-shattering dimension satisfies $\text{fat}_\epsilon(F) = O(\epsilon^{-p})$, then the universal estimation error rate is $O(n^{-1/2})$ for $p < 2$ and $O(n^{-1/p})$ for $p > 2$. The phase transition occurs when $p = 2$. Similar results are also given for the entropy. In later chapters, we will also give the applications for the results in this chapter.

3.1 Maximal Inequality

In order to study the supreme of a class of random variables, we begin with the simple case when the function class is finite. In this case, we have

$$||\max_{1 \leq i \leq m} X_i||_p \leq (\mathbb{E}\max_{1 \leq i \leq m}|X_i|^p)^{1/p} \leq m^{1/p}\max_{1 \leq i \leq m}||X_i||_p$$

As $m$ increases, bounds like this type would increase in the polynomial order with respect to the number of variables $m$, so we can not get satisfying results for a large number of variables. To overcome this, we introduce the following Orlicz norm[1] and the corresponding maximal inequality:
**Definition 3.1.1.** If $\psi$ is a non-decreasing, convex function with $\psi(0) = 0$, the Orlicz norm $|| \cdot ||_\psi$ for random variables is defined by

$$||X||_\psi := \inf\{c > 0 : \mathbb{E}\psi\left(\frac{|X|}{c}\right) \leq 1\}$$

when $\psi_2(x) = e^{x^2} - 1$, the norm $|| \cdot ||_{\psi_2}$ is called $\psi_2$-Orlicz norm.

Note that for $\psi(x) = x^p$, the corresponding Orlicz norm is exactly the $L_p$ norm. Also $||X||_{\psi_2} \geq ||X||_{L_1}$ since $\psi_2(x) \geq x$. The Orlicz norm is more sensitive to the behavior of the tail of $X$, which makes it possible to have a better bound, if we bound the maxima of many variables with light tails. The following lemma gives a better bound [13], in Chapter 8:

**Lemma 3.1.2.** Let $X_1, X_2, \cdots, X_m$ be random variables, then we have

$$|| \sup_{1 \leq i \leq m} X_i ||_{\psi_2} \leq 4\sqrt{\log(m+1)} \sup_{1 \leq i \leq m} ||X_i||_{\psi_2}$$

Random variables from Rademacher process actually have a nice property that their tails decrease very fast. The following result was proved by Kosorok in [13]:

**Lemma 3.1.3.** Define

$$X(a) = \sum_{i=1}^{n} r_i a_i, \ a \in \mathbb{R}^n$$

where $r_1, \cdots, r_n$ are independent and identically distributed Rademacher random variables satisfying $P(r = -1) = P(r = 1) = 1/2$. Let $a = (a_1, \cdots, a_n) \in \mathbb{R}^n$, Then we have

$$P\left(\left|\sum_{i=1}^{n} r_i a_i\right| > x\right) \leq 2e^{-\frac{1}{2}x^2/||a||^2}$$

for the Euclidean norm $|| \cdot ||$. Hence $|| \sum r a ||_{\psi_2} \leq \sqrt{6} ||a||$. 

15
### 3.2 Upper Bound

In [18], Mendelson studied the Gaussian average rather than Rademacher average, which is defined by:

$$l(\mathcal{F}/\mu_n) = \frac{1}{\sqrt{n}} \mathbb{E}_{g} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} g_i f(Z_i) \right|,$$

where $g_i$ are independent standard Gaussian random variables and $\mathbb{E}_g$ means taking expectation of these Gaussian random variables. Note that the factor $1/\sqrt{n}$ was used in his result instead of $1/n$. Assume there exists some $\gamma > 1$, such that $\text{fat}_\epsilon(\mathcal{F}) \leq \gamma \epsilon^{-p}$ holds for any $\epsilon > 0$. Mendelson proved that if $p < 2$, the Gaussian averages are uniformly bounded; if $p > 2$, they may grow at the rate of $n^{\frac{1}{2} - \frac{1}{p}}$, and this bound is tight for Gaussian averages.

Also it is known that the Gaussian and the Rademacher averages are closely related and have the following connection:

**Theorem 3.2.1.** [15, 5] There are absolute constants $c$ and $C$ such that for every $n$ and $\mathcal{F}$

$$C(1 + \log n)^\frac{1}{2} \mathbb{E}_g \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} g_i f(Z_i) \right| \geq \mathbb{E}_r \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} r_i f(Z_i) \right| \geq c \mathbb{E}_g \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} g_i f(Z_i) \right|$$

Using the above theorem and the result in [18], the upper bound was given for expectation of the Rademacher average. But we can not say whether the bound is tight or not.

In this section, we will give a direct proof of the upper bound for the expectation of the Rademacher average and we will make the argument that the bound is tight in a later section.

To bound the empirical Rademacher average, we use the following theorem, this follows from the standard chaining method, see [18], Chapter 8.

**Theorem 3.2.2.** Let $\mu_n$ be the empirical measure and $|f| \leq 1$ for all $f \in \mathcal{F}$ and $f_0 \equiv 0 \in \mathcal{F}$. Let $(\epsilon_k)_{k=0}^\infty$ be a decreasing monotone sequence converging to 0 with $\epsilon_0 = 1$. Then, there exists an absolute constant $C$ such that for any integer $N$,
\[ R(\mathcal{F}/\mu_n) \leq Cn^{-\frac{1}{2}} \sum_{k=1}^{N} \epsilon_{k-1} \log^{\frac{1}{2}} N(\epsilon_k, \mathcal{F}, L_2(\mu_n)) + \epsilon_N. \]

**Proof.** Note that if for any \( \epsilon_i \), if \( N(\epsilon_i, \mathcal{F}, L_2(\mu_n)) \) is infinity, the inequality trivially holds. Hence we can, without loss of generality, assume the covering numbers that appear in the inequality are all finite.

Construct a sequence of finite covering sets \( \mathcal{F}_0, \mathcal{F}_1, \cdots, \mathcal{F}_N \) such that \( \mathcal{F}_i \subset \mathcal{F} \) and \( \mathcal{F}_i \) is minimal \( \epsilon_i \)-cover for the semi-metric space \( (\mathcal{F}, L_2(\mu_n)) \). For each \( f \in \mathcal{F} \) we could find \( f_N \in \mathcal{F}_N \), such that \( ||f - f_N||_{L_2(\mu_n)} \leq \epsilon_N \). Now we fix the empirical measure \( \mu_n \) and study the associated Rademacher process \( X_{\text{rad}}(f) = \frac{1}{n} \sum_{i=1}^{n} r_i f(Z_i) \). Applying the triangle inequality to the Rademacher average, we get

\[ R(\mathcal{F}/\mu_n) = \mathbb{E}_r \sup_{f \in \mathcal{F}}|X_{\text{rad}}(f)| \leq \mathbb{E}_r \sup_{f \in \mathcal{F}}|X_{\text{rad}}(f - f_N)| + \mathbb{E}_r \sup_{f_N \in \mathcal{F}_N} |X_{\text{rad}}(f_N)|. \]

The first term on the right hand side can be bounded as follows

\[ \mathbb{E}_r \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} r_i (f - f_N)(Z_i) \leq \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f - f_N)^2(Z_i) \]

\[ = \sup_{f \in \mathcal{F}} ||f - f_N||_{L_2(\mu_n)} \leq \epsilon_N. \]

The magnitude of the second term \( \mathbb{E}_r \sup_{f_N \in \mathcal{F}_N} |X_{\text{rad}}(f_N)| \) is determined by the size of \( \mathcal{F}_N \). Now we use the following chaining method: For any \( f_k \in \mathcal{F}_k \), there is a \( f_{k-1} \in \mathcal{F}_{k-1} \) such that \( f_k \) is in the \( \epsilon_{k-1} \) ball centered at \( f_{k-1} \) in the semi-metric space \( (\mathcal{F}, L_2(\mu_n)) \). We say that \( f_{k-1} \) is chaining with \( f_k \), denote as \( f_{k-1} \rightarrow f_k \). Using the triangle inequality, we have

\[ \sup_{f_N \in \mathcal{F}_N} |X_{\text{rad}}(f_N)| \leq \sum_{k=1}^{N} \sup_{f_{k-1} \rightarrow f_k} |X_{\text{rad}}(f_k) - X_{\text{rad}}(f_{k-1})| + |X_{\text{rad}}(f_0)| \]

Since for any \( f \in \mathcal{F} \), \( ||f - f_0||_{L_2(\mu_n)} \leq 1 \), and \( \mathcal{F}_0 = \{f_0 \equiv 0\} \), the term \( |X_{\text{rad}}(f_0)| \) vanishes. Taking the \( \psi_2 \) norm on both sides and using the triangle inequality again for
the $\psi_2$ norm, we obtain

$$\left\| \sup_{f_N \in \mathcal{F}_N} |X_{rad}(f_N)| \right\|_{\psi_2} \leq \sum_{k=1}^{N} \left\| \sup_{f_{k-1} \rightarrow f_k} |X_{rad}(f_k - f_{k-1})| \right\|_{\psi_2}$$

Since $N(\epsilon_k, \mathcal{F}, L_2(\mu_n)) \geq N(\epsilon_{k-1}, \mathcal{F}, L_2(\mu_n))$, the number of choices of the chaining pair $(f_{k-1} \rightarrow f_k)$ is bounded by $N^2(\epsilon_k, \mathcal{F}, L_2(\mu_n))$. Applying Lemma 2.7, for the maximal inequality on each term on the right hand side, we have

$$\left\| \sup_{f_{k-1} \rightarrow f_k} |X_{rad}(f_k - f_{k-1})| \right\|_{\psi_2} \leq 4\log^2(N^2(\epsilon_k, \mathcal{F}, L_2(\mu_n)) + 1)(\sup_{f_{k-1} \rightarrow f_k} \left\| X_{rad}(f_k - f_{k-1}) \right\|_{\psi_2}).$$

As long as the covering number is bigger than 1, the factor

$$4\log^{1/2}(N^2(\epsilon_k, \mathcal{F}, L_2(\mu_n)) + 1)$$

is bounded by $9\log^{1/2}(N(\epsilon_k, \mathcal{F}, L_2(\mu_n)))$. Moreover, by Lemma 2.8, we have

$$\sup_{f_{k-1} \rightarrow f_k} \left\| X_{rad}(f_k - f_{k-1}) \right\|_{\psi_2} \leq \sup_{f_{k-1} \rightarrow f_k} \sqrt{6n^{-1/2}}|f_k - f_{k-1}|_{L_2(\mu_n)}$$

By construction, it is bounded by $\sqrt{6n^{-1/2}}\epsilon_{k-1}$. So we have

$$\mathbb{E}_n \sup_{f_N \in \mathcal{F}_N} |X_{rad}(f_N)| \leq \left\| \sup_{f_N \in \mathcal{F}_N} |X_{rad}(f_N)| \right\|_{\psi_2} \leq Cn^{-1/2} \sum_{k=1}^{N} \epsilon_{k-1} \log^{1/2}N(\epsilon_k, \mathcal{F}, L_2(\mu_n)).$$

In [18], Mendelson found a similar upper bound for the Gaussian average, the details of this chaining technique also can be found in [3].

We now present the bound for Radmacher average using Fat-shattering dimension:

$$\quad$$
Theorem 3.2.3. Assume that for some $\gamma > 1$, $\text{fat}_\epsilon(\mathcal{F}) \leq \gamma \epsilon^{-p}$ holds for any $\epsilon > 0$, then there exists a constant $C_p$, which depends only on $p$, such that for any empirical measure $\mu_n$

$$R(\mathcal{F}/\mu_n) \leq \begin{cases} C_p \gamma^{\frac{1}{2}} \log \gamma n^{-1/2} & \text{if } 0 < p < 2 \\ C_2 \gamma^{\frac{1}{2}} \log \gamma n^{-1/2} \log^2 n & \text{if } p = 2 \\ C_p \gamma^{\frac{1}{2}} \log \gamma n^{-1/p} & \text{if } p > 2. \end{cases}$$ (3.8)

Proof. Let $\mu_n$ be an empirical measure. When $p < 2$, we know the sum on the right hand side of inequality from theorem 3.2.2 can be bounded using Lemma 2.5 as follows:

$$n^{-1/2} \sum_{k=1}^N \epsilon_k \log \frac{1}{2} N(\epsilon_k, \mathcal{F}, L_2(\mu_n)) \leq n^{-1/2} \int_0^\infty \log \frac{1}{2} N(\epsilon, \mathcal{F}, L_2(\mu_n)) d\epsilon \leq C_p \gamma^{\frac{1}{2}} \log \gamma n^{-1/2}$$

Assume that $p \geq 2$. Let $\epsilon_k = 2^{-k}$ and $N = p^{-1}\log n$. Using Theorem 3.1 and Lemma 2.5, we have

$$R(\mathcal{F}/\mu_n) \leq C_p n^{-1/2} \log \gamma \sum_{k=1}^N \epsilon_k^{1/2} \log \left( \frac{2}{\epsilon_k} \right) + 2\epsilon_N$$

$$\leq C_p n^{-1/2} \log \gamma \sum_{k=1}^N k^{2^{(\frac{p}{2})-1}} + 2n^{-\frac{1}{p}}$$

If $p = 2$, the geometric sum is bounded by:

$$C_p n^{-1/2}(\gamma^{\frac{1}{2}} \log \gamma)N^2 \leq C_p(\gamma^{\frac{1}{2}} \log \gamma)n^{-1/2} \log^2 n$$

If $p > 2$, it is bounded by

$$C_p(\gamma^{\frac{1}{2}} \log \gamma)n^{-1/p}$$
We also present the entropy version upper bound, the proof follows from the same argument.

**Theorem 3.2.4.** Assume that for some $\gamma > 1$, $\log N(\epsilon, \mathcal{F}, L_2(\mu_n)) \leq \gamma \epsilon^{-p}$ holds for all $\epsilon > 0$. Then there exists a constant $C_p$, which depends only on $p$, such that for any empirical measure $\mu_n$

$$R(\mathcal{F}/\mu_n) \leq \begin{cases} C_p \gamma^{1/2} n^{-1/2} & \text{if } 0 < p < 2 \\ C_2 \gamma^{1/2} n^{-1/2} \log n & \text{if } p = 2 \\ C_p \gamma^{1/2} n^{-1/p} & \text{if } p > 2. \end{cases}$$

By taking the expectation of $R(\mathcal{F}/\mu_n)$ in Theorem 3.2.3 and Theorem 3.2.4, then apply the Theorem 2.3.2, we can also get the upper bounds for corresponding $\mu$-estimation error and universal estimation error.

### 3.3 Lower Bound

In this section, we prove that for some proper underlying distribution $\mu$, the Fat-shattering dimension provides a lower bound for the Rademacher average (hence for the universal estimation error), and this bound is tight. A similar lower bounds for the Gaussian average can be found in [18].

**Theorem 3.3.1.** If $\text{fat}_\epsilon(\mathcal{F}) \geq \gamma \epsilon^{-p}$ for some $\gamma$, then there exists a measure $\mu \in \mathcal{P}$ and constant $c$ such that

$$\mathbb{E}_\mu R(\mathcal{F}/\mu_n) \geq cn^{-1/p}$$

**Proof.** By the definition of Fat-shattering dimension, for every integer $n$, let $\epsilon = (\gamma/n)^{1/p}$, there exists a set $\{Z_1, Z_2, \cdots, Z_n\}$ which is $\epsilon$ shattered by $\mathcal{F}$ and all $Z_i$ are distinct. Let $\mu$ be the measure uniformly distributed on $\{Z_1, Z_2, \cdots, Z_n\}$. By the definition of shattering, we know all $Z_i$ are distinct.
Let $Z_1^*, \ldots, Z_n^*$ be the data generated uniformly and independently from $\mu$ and let $\mu_n$ be the corresponding empirical measure. Assume that $Z_i$ appears $n_i$ times in the support of $\mu_n$. Then we have:

$$R(\mathcal{F}/\mu_n) = \frac{1}{n} \mathbb{E}_r \sup_{f \in \mathcal{F}} |\sum_{i=1}^{n} \sum_{k=1}^{n_i} r_{i,k} f(Z_i)|$$

(4.2)

$$\geq \frac{1}{2n} \mathbb{E}_r \sup_{f,f' \in \mathcal{F}} \sum_{i=1}^{n} \sum_{k=1}^{n_i} r_{i,k}(f(Z_i) - f'(Z_i))$$

(4.2)

where the $\{r_{i,k}\}$’s are independently Rademacher random variables.

As we know for those $i$ where $n_i > 0$, the probability of $P(\sum_{k=1}^{n_i} r_{i,k} = 0) \leq \frac{1}{2}$. For a realization of $r_{i,k}$, set $A = \{i : \sum_{k=1}^{n_i} r_{i,k} > 0\}$. Let $f_A$ to be the Fat-shattering function of the set $A$, and $f_{A^c}$ be the shattering function of its complement $A^c$. Also, denote by $n^*$ the number of $i$’s for which $n_i > 0$. Then we have,

$$\sup_{f,f' \in \mathcal{F}} \sum_{i=1}^{n} \sum_{k=1}^{n_i} r_{i,k}(f(Z_i) - f'(Z_i)) \geq \sum_{i=1}^{n} \sum_{k=1}^{n_i} (r_{i,k}(f_A(Z_i) - f_{A^c}(Z_i)))$$

(4.3)

As long as $\sum_{k}^{n_i} r_{i,k} \neq 0$, for each $i$, $\sum_{k=1}^{n_i} (r_{i,k}(f_A(Z_i) - f_{A^c}(Z_i))) \geq 2\epsilon$. So we know

$$R(\mathcal{F}/\mu_n) \geq \frac{1}{2n} \mathbb{E}_r \sum_{i=1}^{n} \sum_{k=1}^{n_i} (r_{i,k}(f_A(Z_i) - f_{A^c}(Z_i))) \geq \frac{1}{2n} \epsilon n^*.$$  

(4.4)

The last inequality holds because for each $i$ with $n_i > 0$, the probability of $\sum_{k=1}^{n_i} r_{i,k} = 0$ is no more than $1/2$.

Now take the expectation for inequality (4.4), we have

$$\mathbb{E}_\mu R(\mathcal{F}/\mu_n) \geq \mathbb{E}_\mu (\frac{1}{2n} \epsilon n^*)$$

(4.5)

$n^*$ here is the number of $Z_i$’s that appear in $Z_1^* \cdot \cdot \cdot, Z_n^*$. We know

$$\mathbb{E}_\mu (n^*) = n - \left(\frac{n-1}{n}\right)^n > (1 - \frac{1}{e})n$$

(4.6)
For $\epsilon = (\gamma/n)^{1/p}$, we obtain the following lower bound

$$\mathbb{E}_{\mu} R(F/\mu_n) \geq \left(\frac{1}{2} - \frac{1}{2e}\right)\gamma^{1/p} n^{-\frac{1}{p}}$$

(4.7)

Addition with Theorem 2.3.2, For $p \geq 2$, we know there also exists a constant $c_1$ such that

$$\mathbb{E}_{\mu} \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} f(Z_i) - \mathbb{E}_{\mu} f \right| > c_1 n^{-\frac{1}{p}}$$

\[\square\]

In the previous section and this section, we have proved that for $p > 2$, the expectation of the Rademacher average is bounded above and below by $O(n^{-1/p})$. Since $O(n^{-1/2})$ is negligible comparing $O(n^{-1/p})$, from Theorem 2.3.2, we know that the universal estimation error is bounded by $n^{-1/p}$ and this bound is tight.

For $p < 2$, the upper bound gives us convergence rate as $O(n^{-1/2})$ and in this case $F$ is the Donsker class [1]. As long as the limit of the empirical process is non-trivial, the rate $O(n^{-1/2})$ is optimal. Note this is only for Rademacher average, if the underlying distribution satisfies some additional property, the rate can be better, we will show some examples in Chapter 4.
Chapter 4

Applications to Common Function Classes

In this chapter, we will mainly focus on the estimation error for some common function classes, such as the linear function class, the smooth function class and the convex class. It is interesting to see how the dimension of the input data, the regularities of the model, and most importantly, the size of the data influences the behavior of the estimation error. Based on a variety of results from the topics of covering number, Fat-shattering dimension and Rademacher average developed recently, we can apply our results from the last chapter and get interesting results on the estimation error bound.

4.1 Regularized Linear Class

In this section, we will study the estimation error for linear class. The results in this section can also be helpful for deriving estimation error bound for model involving linear functions such as neural networks and support vector machines.

Consider the function space containing linear functions of the form $W \cdot X$. In order to make its covering number finite, we impose the following restrictions: Assume that the input $X \in \mathbb{R}^d$, $\|X\|_q \leq a$ and linear weight vector satisfies the regularization condition $\|W\|_p \leq b$, where $1/p + 1/q = 1$ and $1 \leq p \leq 2$. In [21], Zhang derived the following
bound:
\[
\log N(\epsilon, \mathcal{H}_p, L_2(\mu_n)) \leq \left\lceil \frac{a^2b^2}{\epsilon^2} \right\rceil \log(2d + 1)
\]

here \(\left\lceil \frac{a^2b^2}{\epsilon^2} \right\rceil\) is the smallest integer that is not less than \(\frac{a^2b^2}{\epsilon^2}\). By assuming \(\epsilon < 1\) in the following analysis, we can replace \(\left\lceil \frac{a^2b^2}{\epsilon^2} \right\rceil\) with \(\frac{a^2b^2+1}{\epsilon^2}\) for simplicity.

It is well known that the VC dimension for linear classifier in \(\mathbb{R}^d\) is \(O(d)\), but because of the constraints imposed, the above bound on entropy depends logarithmically on \(d\), which is weak already. However, in [21], it is also possible to remove the dimensional dependency. For example, when \(p = 2\)

\[
\log N(\epsilon, \mathcal{H}_2, L_2(\mu_n)) \leq \left\lceil \frac{a^2b^2}{\epsilon^2} \right\rceil \log(2n + 1).
\]

Intuitively, the reason for getting rid of the dimension dependency is that the effective dimension of \(w\), acting on \(n\) data is at most \(n\). This result can be very useful in the high dimensional regime when the number of data points is much less than the dimension. In [21], Zhang then obtained a bound for the estimation error for classification problems.

Now we can apply the covering number results for more general setting, for example, real value problems.

**Lemma 4.1.1.** Let the \(\mathcal{H}_p\) be the functions space defined above. Then we have

\[
\mathbb{E}_\mu \sup_{f \in \mathcal{H}_p} \left| \frac{1}{n} \sum_{i=1}^{n} (f(Z_i) - \mathbb{E}_\mu f) \right| \leq 2C_2 \sqrt{(a^2b^2 + 1) \log(2d + 1) \frac{\log n}{\sqrt{n}}}
\]

Here the constant \(C_2\) is the constant from Theorem 3.2.4.

The proof is straightforward, just apply covering number results to Theorem 3.2.4 and Theorem 2.3.2.

When \(p = 1\) and \(p = 2\), the function space \(\mathcal{H}_p\) is the lasso regression style hypothesis space and ridge regression style hypothesis space, which we have the most interest in. For
some constant $b$, consider the following spaces:

$$\mathcal{H}_1 = \{W \cdot X : \|W\|_1 \leq b \text{ and } \|X\|_\infty \leq 1/b\},$$

$$\mathcal{H}_2 = \{W \cdot X : \|W\|_2 \leq b \text{ and } \|X\|_2 \leq 1/b\}.$$  

From the Holder inequality, we have $|W \cdot X| \leq 1$ for $W \cdot X \in \mathcal{H}_1, \mathcal{H}_2$. We can apply the dimension-free covering number result of $\mathcal{H}_2$ to Lemma 4.1.1, and get the following result for the estimation error

$$E_{\mu} \sup_{f \in \mathcal{H}_2} \left| \frac{1}{n} \sum_{i=1}^{n} (f(Z_i) - E_{\mu}f) \right| \leq 2C_2 \sqrt{2 \log(2d+1) \log n \over n}.$$  

### 4.2 Bounded Variation Class

In this section, we will give results for both monotonic class and bounded variation class. We first consider one dimensional case, let $\mathcal{H}_1$ and $\mathcal{H}_2$ be the set of all functions on $[0, T]$ taking values in $[-1, 1]$ with the requirements that $h_1$ is nondecreasing for any $h_1 \in \mathcal{H}_1$ and the total variation of $h_2$ is bounded by $V$ for any $h_2 \in \mathcal{H}_2$. If $V \geq 2$, we have $\mathcal{H}_1 \subset \mathcal{H}_2$.

The Rademacher average of $\mathcal{H}_2$ provides an upper bound for Rademacher average of $\mathcal{H}_1$. In [10], Bartlett proved the following theorem:

**Theorem 4.2.1.** For all $\epsilon \leq V/12$

$$(\log e) \frac{V}{54 \epsilon} \leq \log N(\epsilon, \mathcal{H}_2, L_1(\mu)) \leq \frac{13V}{\epsilon}.$$  

We give the sketch of the proof of the upper bound here [10].

**Proof.** The first step of the proof is to show that because any bounded variation function can be expressed as the difference of two monotonic functions, their corresponding covering numbers has the relationship
\[ \mathbb{N}(\epsilon, \mathcal{H}_2, L_1) \leq (\mathbb{N}(\epsilon/2, \mathcal{H}_1, L_1))^2 \]

When it comes to the entropy, \( \mathcal{H}_1 \) and \( \mathcal{H}_2 \) are only different by a constant factor. We only need to find the covering number for monotonic functions. Bartlett then partition \([0, T]\) into sub intervals of length less than \( \epsilon/2V \) and partition on the \( y \) axis \([-1, 1]\) into sub intervals of length than \( \epsilon/T \). These two partitions make a grid on the domain \([0, T] \times [-1, 1]\). Then all non decreasing functions which are constant on these sub intervals form a cover for \( \mathcal{H}_2 \), and then count the number of all these functions, we can get the upper bound.

We can also directly derive the Fat-shattering results

**Lemma 4.2.2.** Let \( \mathcal{H} \) be a function space containing functions with variation bound \( V \), then

\[ \text{fat}_\epsilon \mathcal{H} \leq \frac{V}{\epsilon} \]

*Proof.* We prove by contradiction. Assume that for function space \( \mathcal{H} \), the Fat-shattering dimension \( \text{fat}_\epsilon(\mathcal{F}) = K > V/\epsilon \), Since for any \( f_1, f_2 \in \mathcal{H} \), the variation is bounded by \( V \), the variation of their difference \( f_1 - f_2 \) is bounded by \( 2V \). By definition of shattering we know there exists \( A = \{X_1, \cdots X_K\} \) and a function \( s \), such that for any set \( I \in 1, \cdots K \), there exists some \( f_I \in \mathcal{H} \) that satisfies \( f_I(X_i) \geq s(X_i) + \epsilon \) for \( i \in I \) and \( f_I(X_i) \leq s(X_i) - \epsilon \) for \( i \notin I \). Then we consider the difference between \( f_I \) and \( f_{A/I} \), the variation of this difference is at least \( 2K\epsilon > 2V \), a contradiction to its variation bound.

Using above theorem and Lemma 2.3.7, we know that the Fat-shattering dimension has the bound:

\[ \text{fat}_\epsilon(\mathcal{H}_2) \leq \frac{128V}{\epsilon} \]
From Theorem 3.2.4, we conclude that the convergence rate of Rademacher average of $\mathcal{H}_2$ can achieve $O(n^{-1/2})$ and so does $\mathcal{H}_1$ since $\mathcal{H}_1 \subset \mathcal{H}_2$. For estimation error, we have

$$\mathbb{E}_\mu \sup_{f \in \mathcal{H}_2} \left| \frac{1}{n} \sum_{i=1}^{n} (f(Z_i) - \mathbb{E}_\mu f) \right| \leq 2C_1 \sqrt{\frac{13V}{n}}$$

where the constant $C_1$ is from Theorem 3.2.3 when $p = 1$. This bound also can be applied to many other classes such as the Lipschitz class.

The grid method can also be used to bound the entropy of the class of monotonic functions on $[0, 1]^d \subset \mathbb{R}^d$ which have uniform bound of 1. Denote this function space by $\mathcal{H}_d$, in [38], Gao and Wellner extended the grid method to bound the entropy of $\mathcal{H}_d$

**Theorem 4.2.3.** [38] For $1 \leq p < \infty$ and $d \geq 2$, there exists constants $c_1$ and $c_2$ depending only on $p$ and $d$, such that if $(d-1)p \neq d$, then

$$c_1 \epsilon^{-\alpha} \leq \log N(\epsilon, \mathcal{H}_d, L_p) \leq c_2 \epsilon^{-\alpha},$$

where $\alpha = \max\{d, (d-1)p\}$. If $(d-1)p$, then

$$c_1 \epsilon^{-d} \leq \log N(\epsilon, \mathcal{H}_d, L_p) \leq c_2 \epsilon^{-d} (\log 1/\epsilon)^d,$$

From this theorem, for $p = 2$, we know that as long as $d \geq 2$, the smallest entropy for the class of monotonic functions is approximately $\gamma \epsilon^{-2}$ only when $d = 2$. The convergence rate of the estimation error for $\mathcal{H}_d$ is at most $O(n^{-1/2} \log n)$. When $d > 2$, we have $(d-1)p > 2$, the universal estimation error for $\mathcal{H}_d$ would have a convergence rate of $O(n^{-1/((d-1)2)})$. The class $\mathcal{H}_d$ is no longer a Donsker class. Also we can also claim that this rate is optimal.
4.3 Smooth Function Class

In this section, we consider the smooth functions. For \( k > 0 \), we study the class of all functions on a compact set \( I^d \subset \mathbb{R}^d \) that have uniformly bounded partial derivatives up to order \( k \). The following results have been established in [33 34 35]

**Theorem 4.3.1.** Let \( C^k(d, K) \subset C_0(I^d) \) be a (relative compact) subset of \( C_0(I^d) \) consisting of all \( C^k \)-smooth functions on \( I^d \) with all the derivatives up to order \( k \) bounded by \( K \), then there exists a constant \( A \) depending on \( k, K \) and \( d \) such that

\[
\log \mathbb{N}(\epsilon, C^k(d, K), L_\infty(\mu)) \leq A \lambda(I^d)(\frac{1}{\epsilon})^{\frac{d}{k}}
\]

where for every \( \epsilon > 0 \), where \( \lambda(I^d) \) is the Lebesgue measure of the set \( \{ X : ||X - I^d|| \leq 1 \} \).

Since the regularity of the function space, the \( L_\infty \) result can natural extend to \( L_p \) where \( p > 1 \) as long as the domain is smooth. Kolmogorov [33] also proved that the bound of this form is sharp, we have

**Lemma 4.3.2.** Let \( C^k(d, K) \) as defined above, let \( I^d \) be a bounded, convex subset in \( \mathbb{R}^d \) with nonempty interior. Then there exists a constant \( A_1, A_2 \) with \( A_1 > A_2 \) depending on \( k, K, \) diameter of \( I^d, d, \) and constant \( p \) with \( p > 1 \), such that

\[
A_2(\frac{1}{\epsilon})^{\frac{d}{k}} < \log \mathbb{N}(\epsilon, C^k(d, K), L_p(\mu)) < A_1(\frac{1}{\epsilon})^{\frac{d}{k}}
\]

Using this result we can give the sharp estimate of the Rademacher average and universal estimation error for smooth function space. We only give the results for the Rademacher average since bounds for universal estimation error are essentially the same.

**Lemma 4.3.3.** Define \( C^k(d, K) \) and \( I^d \) in the same way as in Lemma 4.3.2. Then there exists some constant \( A \) such that
\[ R((C^k(d, K))/\mu_n) \leq \begin{cases} An^{-1/2} & \text{if } d < 2k \\ An^{-1/2} \log n & \text{if } d = 2k \\ An^{-d/k} & \text{if } d > 2k. \end{cases} \]

We conclude that where \( d \leq 2k \), the class is Donsker class and as the dimension outgrows the regularity of the function space. When \( d > 2k \), the complexity increases and the class is no longer a Donsker class anymore.

Moreover, results about analytic functions are given in [33].

**Lemma 4.3.4.** Let \( C(\eta, K) \subset C_0([-\pi, \pi]) \) consist of periodic real analytic functions on \([-\pi, \pi]\), extendable into a complex \( \eta \)-neighborhood of \([-\pi, \pi]\) for some \( \eta > 0 \) and uniformly bounded there by \( K \). Then

\[
\log N(\epsilon, C(\eta, K), L_2(\mu)) \asymp \log^2 \left( \frac{1}{\epsilon} \right)
\]

The entropy for the analytic function space grows very slowly, and can reach the universal estimation convergence rate of \( n^{-1/2} \) easily. Moreover, if we assume the underlying data distribution \( \mu \in \mathcal{P}(h) \), defined later in (5.2), and the associated loss function, is binary loss function defined later in (5.2), one can conclude that for \( C^k(d, K) \), the convergence rate of estimation error can be as fast as \( O(n^{-k/(k+d)}) \). When \( d < k \), the estimation rate can achieve faster rate than \( n^{-1/2} \) under this condition. For the analytic function space, the convergence rate under this distribution can be as fast as \( n^{-1} \).

### 4.4 Piecewise Smooth Function Space

In this section we study the convergence rate of estimation error for piecewise smooth functions. One should think that piecewise smooth functions is natural extension of smooth functions. But be careful here: the complexity of the boundary can significantly increase the entropy for the function space.
In [36], Batenkov et al considered piecewise $C^k$ smooth functions $f(x_1, \cdots, x_d)$ defined on the torus $T^d$ which can be thought of as the $d$-dimensional box $I^d = [-\pi, \pi]^d$ with the opposite faces glued together. Moreover, assume that the jumps of $f$ occur at certain $C^k$ smooth, pairwise disjoint, compact hyper surface $\Sigma_j \subset T^d$, $j = 1, \cdots, K$. Then the complement in $T^n$ of $\Sigma = \bigcup_{j=1}^K \Sigma_j$ is the union of certain $n$ dimensional manifolds $D_s$ whose boundaries $\partial D_s$ are unions of some of $\Sigma_j$. So we assume that $f = f_s$ on each $D_s$, with $f_s$ being $C^k$-smooth functions on $D_s$.

As a smooth function space, we assume that all derivatives of the smooth pieces $f_s$ of $f$ up to order $k$ are uniformly bounded, the curvature for the jump manifold $\Sigma$ can be bounded. Denote the function space defined as $PC^k(d)$. Then Batnkov et al [36] proved that

$$\log N(\epsilon, PC^k(d), L_2(\mu)) \asymp \left(\frac{1}{\epsilon}\right)^m$$

where $m = 1/k$ for $d = 1$ and $m = 2(d - 1)/k$ for $d \geq 2$. By this result, one can conclude that the complexities of the piecewise smooth function class in one or two dimension is still the same for their smooth counterparts. When $d > 2$, the entropy of $PC^k(d)$ become strictly larger. Actually, we have

**Theorem 4.4.1.** For the function space $PC^k(d)$ defined above, then there exists a constant such that

$$R((PC^k)/\mu_n) \leq \begin{cases} Cn^{-1/2} & \text{if } d - 1 < k \\ Cn^{-1/2} \log n & \text{if } d - 1 = k \\ Cn^{-d/k} & \text{if } d - 1 > k. \end{cases}$$

Like the smooth counterparts, similar faster convergence rate can be achieved if we make additional assumptions about the distribution in $\mathcal{P}(h)$, defined later in (5.2), the results are omitted here. Note that when discontinuity points appear, the convergence rate drops when $d > 2$. 

30
4.5 Convex Functions

In this section, we discuss the covering numbers for both the class of uniformly bounded convex sets and convex function space. They are related since a convex function in $\mathbb{R}^d$ can be regarded as an convex set in $\mathbb{R}^{d+1}$.

We first give results for convex set in $\mathbb{R}^d$. For subsets $C$ and $D$ in a metric space $(M, d)$, the Hausdorff distance is defined as

$$h(D, C) = \min\{\sup_{x \in C} d(x, D), \sup_{x \in D} d(x, C)\}$$

Bronshtein [2] first proved the following lemma for convex sets

Lemma 4.5.1. [2] Let $C$ be the set of all close convex subsets in a fixed, bounded subset $U$ of $\mathbb{R}^d$ with respect to Hausdorff metric. If $d \geq 2$, then there exists two constants $K_1$ and $K_2$ only depending on $U$ such that

$$K_1(\frac{1}{\epsilon})^{(d-1)/2} \leq \log N(\epsilon, C, h) \leq K_2(\frac{1}{\epsilon})^{(d-1)/2}$$

Next we consider convex function space $\mathcal{H}$, defined on a bounded convex domain $U \subset \mathbb{R}^d$. We also impose two restrictions to the functions we are interested in: The functions are uniformly bounded such that $\sup_{f \in \mathcal{H}} ||f||_{\infty} \leq B_1$. In addition, the function space is uniformly Lipschitz such that $|f(X) - f(Y)| \leq B_2||X - Y||$ for every $f \in \mathcal{H}$ and $X, Y \in U$, we denote this function space as $\mathcal{H}(B_1, B_2)$.

The entropy of the set of convex sets can be used to derive the entropy of convex functions space. Let $C_f$ be the supergraph $\{(X, t) : f(X) \leq t\}$ of a function $f$, one can easily prove that there exists a constant $A$, such that for every pair functions $f$ and $g$, we have

$$||f - g||_{\infty} \leq AB_1(1 + B_2)h(C_f, C_g)$$

This bound, together with Lemma 4.5.1, gives upper bound for the entropy of $\mathcal{H}(B_1, B_2)$ with respect to the $L_{\infty}$ norm.
Lemma 4.5.2. Let $\mathcal{H}(B_1, B_2)$ be defined above, there exists two constants $K_1$ and $K_2$, such that for any $\epsilon \in (0, 1)$

$$K_1 B_1 (1 + B_2) \left(\frac{1}{\epsilon}\right)^{d/2} \leq \log N(\epsilon, \mathcal{H}(B_1, B_2), L_\infty) \leq K_2 B_1 (1 + B_2) \left(\frac{1}{\epsilon}\right)^{d/2}$$

Moreover, in [22], Guntuboyina et al proved in more general setting that the above lemma holds after replacing $L_\infty$ with $L_p$ for $1 \leq p \leq \infty$.

In conclusion, we have

$$R((\mathcal{H}(B_1, B_2))/\mu_n) \leq \begin{cases} 
Cn^{-1/2} & \text{if } d < 4 \\
Cn^{-1/2} \log n & \text{if } d = 4 \\
Cn^{-2/d} & \text{if } d > 4.
\end{cases}$$

The regime switching happens when $d = 4$. This result shows surprisingly that when $d > 4$, the convex function space is not Donsker any more. The complexity increases with the dimension and it is much harder to fit high dimensional convex function.
Chapter 5

Applications to Function Classes in Statistical Learning

5.1 Excess Loss Class or Hypothesis Class

It seems obscure to study the excess loss class $\mathcal{F}$, defined in Section 2.2, rather than $\mathcal{H}$ itself. However, when it comes to the most common loss functions $L$, the complexity of excess loss class $\mathcal{F}$ can be controlled by the complexity of the hypothesis space $\mathcal{H}$. For example, assuming that the loss function $L$ is $K$-Lipschitz in its first argument, i.e. for all $\hat{y}_1, \hat{y}_2, y$, we have

$$|L(\hat{y}_1, y) - L(\hat{y}_2, y)| \leq K|\hat{y}_1 - \hat{y}_2|$$

(5.1)

Since we also have $f^* \equiv 0 \in \mathcal{F}$, it is not hard to prove that the Rademacher average of the excess loss class can be bounded in terms of the average of the hypothesis space:

$$R(\mathcal{F}/\mu_n) \leq KR(\mathcal{H}/\mu_n).$$

(5.2)

Thus we know that the Rademacher average of $\mathcal{H}$ can provide bound for the Rademacher average of $\mathcal{F}$. We also have the following lemma to characterize how to
bound the entropy of $\mathcal{F}$ by the entropy of $\mathcal{H}$ when using $q$-loss function. The proof can be found in [18].

**Lemma 5.1.1.** If $\mathcal{H}$ has uniform bound of 1, then for every $1 \leq q \leq \infty$ there is a constant $C_q$ such that for every $\epsilon > 0$, $g$ bounded by 1, and probability $\mu$, we have

$$
\log N(\epsilon, |\mathcal{H} - g|^q, L_2(\mu)) \leq \log N(C_q \epsilon, \mathcal{H}, L_2(\mu))
$$

In the following case, we can further claim that the complexity of the excess loss class controls hypothesis space.

**Lemma 5.1.2.** Assume $\mathcal{H}$ has a uniform bound of 1. Let $\mathcal{H}^* = \{(h/4 + 3/4) : h \in \mathcal{H}\}$ and if $\mathcal{H}^* \subset \mathcal{H}$, then there exists constant $c$ such that for any $g$ bounded by 1, we have

$$
\log N(c \epsilon, \mathcal{H}, L_2(\mu)) \leq \log N(\epsilon, (\mathcal{H} - g)^2, L_2(\mu))
$$

**Proof.** It is easily seen from the definition that the covering number is translation invariant:

$$
N(\epsilon, \mathcal{H}, L_2(\mu_n)) = N(\epsilon, \mathcal{H} - g, L_2(\mu_n)).
$$

Also by the property that $\mathcal{H}^* \subset \mathcal{H}$, one can prove that by enlarging the radius of the covering balls, the covering number of $\mathcal{H}$ can be bounded by $\mathcal{H}^*$:

$$
N(4 \epsilon, \mathcal{H}, L_2(\mu_n)) \leq N(\epsilon, \mathcal{H}^*, L_2(\mu_n)).
$$

Moreover, since $\mathcal{H}^*$ is bounded below by $1/2$, we have $|h_1^2 - h_2^2| \geq |h_1 - h_2|$, therefore the covering number of $\mathcal{H}^*$ can be bounded by the covering number of $(\mathcal{H}^*)^2$. And because $\mathcal{H}^* \subset \mathcal{H}$, the covering number of $(\mathcal{H})^2$ can bound the covering number of $(\mathcal{H}^*)^2$, and hence
the covering number of $\mathcal{H}^*$ and $\mathcal{H}$. Together with the translation invariant property, the result follows.

We will see in later applications that the condition $\mathcal{H}^* \subset \mathcal{H}$ can actually be achieved in many scenarios.

5.2 VC Classes for Classification

In this part, we study function space for a classification problem which has bounded VC dimension. VC dimension, introduced by Vladimir Vapnik and Alexey Chervonenkis [27], is a measure to describe the capacity of a function space, and often used for analyzing classification problems. By definition, it is the maximum number of points that can be shattered by this function space with no error.

Assume $\mathcal{F}$ has finite VC dimension $V$. It is well known that there exists a constant $C$ such that the estimation error is bounded by $C\sqrt{V/n}$, which is optimal in the minimax sense, see [12] for more details.

Since we have Theorem 3.2, we can come up with similar results as well known ones, from the definition of VC dimension, we know that $\text{fat}_\epsilon(\mathcal{F}) = V$ for $\epsilon < 1$. In this case, we can set $\gamma$ to be $V$ and $p$ to be $1$. Under this setting, the associated Rademacher average is bounded above by $C_1 \log V \sqrt{V/n}$. It is clearly optimal in terms of the data size by only a logarithm factor of $V$.

Here we give a classification example when faster convergence rate can be achieved if additional assumptions for the distribution are made. Without loss of generality, we assume that $Y \in \{0, 1\}$. Let $\eta_P(x) = P[Y = 1|X = x]$ denotes the regression function of $Y$ given $X = x$. Naturally the best classifier in this setting is $f^*(x) = 1_{\eta_P(x) \geq 1/2}$. Denote the empirical minimizer as $\hat{f}$ and $l(f) = P(Y \neq f(x)) - P(Y \neq f^*(x))$. In addition, for constant $h > 0$, define

$$\mathcal{P}(h) = \{P : |2\eta_P(x) - 1| \geq h\}$$
We have the theorem \cite{20} that, for some positive constant $k$, either
\[
\sup_{P \in \mathcal{P}(h)} \mathbb{E}[l(\hat{f})] \leq k \sqrt{\frac{V}{n}} \text{ if } h \leq \sqrt{\frac{V}{n}}
\]
or
\[
\sup_{P \in \mathcal{P}(h)} \mathbb{E}[l(\hat{f})] \leq k \frac{V}{nh} (1 + \log(\frac{nh^2}{V})) \text{ if } h \geq \sqrt{\frac{V}{n}}
\]
It turns out that this upper bound is also optimal in the minimax sense up to a logarithmic factor.

The paper \cite{20} actually gives a more subtle bound for the general classification problems (without explicit VC dimension bound) with distribution $P \in \mathcal{P}(h)$.

**Lemma 5.2.1.** If the entropy of the function space $\mathcal{F}$ satisfies:
\[
\log \mathcal{N}(\epsilon, \mathcal{F}, L_1(\mu)) \leq \gamma \epsilon^{-p} \text{ for every } \epsilon \in (0, 1)
\]
then, for some constant $C$ depending only on $\gamma$ and $p$, we have
\[
\sup_{P \in \mathcal{P}(h)} \mathbb{E}[l(\hat{f})] \leq C((nh^{1-p})^{-1/(p+1)} \wedge n^{-1/2})
\]
This bound has also proven to be optimal in the minimax sense. It result shows that for certain function space under margin type distribution, the convergence rate can be $n^{-\alpha}$ with $1/2 \leq \alpha < 1$.

### 5.3 Multiple Layer Neural Nets

We will present some evidence on why deep learning works in this section. We make the assumption that the input magnitude of each neuron is bounded and consider the following architecture for the neural net:
\[
\Omega = \{ x \in \mathbb{R}^d : ||x||_\infty \leq B \}
\]
Let $\mathcal{H}_0$ be the class of functions on $\Omega$ defined by

$$
\mathcal{H}_0 = \{ X = (X^1, X^2, \cdots, X^d) \rightarrow X^i : 1 \leq i \leq d \}
$$

Let $\sigma$ be the standard logistic sigmoid function, which is 1-Lipschitz. Define the hypothesis space recursively by:

$$
\mathcal{H}_l = \left\{ \sigma \left( \sum_{i=1}^{N} w_i h_i \right) : N \in \mathbb{N}, h_i \in \mathcal{H}_{l-1}, \sum_{i=1}^{N} |w_i| \leq C \right\}
$$

Define the $C$-convex hull of $\mathcal{H}$ as

$$
\text{conv}_C(\mathcal{H}) = \left\{ \sum c_i h_i : h_i \in \mathcal{H}, \sum |c_i| \leq C \right\}
$$

By the definition of Rademacher average, one can show

$$
CR(\mathcal{H}/\mu_n) = R(\text{conv}_C(\mathcal{H})/\mu_n)
$$

One can also check by compositing $\mathcal{H}$ with a $L$-Lipschitz function $\sigma$, we have

$$
R((\sigma \circ \mathcal{H})/\mu_n) \leq LR(\mathcal{H}/\mu_n)
$$

Since the number of functions in the space $\mathcal{H}_0$ is $d$, which is finite, the $\epsilon$-covering number can be bounded by $d$ for any $\epsilon$. Then by applying Theorem 3.3 and setting $\gamma = \log d$ and $p = 1$, we can bound $R(\mathcal{H}_0/\mu_n)$ by $C_1 \sqrt{\log d/n}$ for a positive constant $C_1$. 

Do induction on the number of layers, in each layer, we use (6.5) and (6.6) alternatively and get

$$
R(\mathcal{H}_l/\mu_n) \leq C_1 C^l \sqrt{\frac{\log d}{n}}.
$$
Note that $\mathcal{H}_l$ satisfies the requirement in Lemma 5.2. Hence for $L_2$ loss function, the Rademacher average of $\mathcal{F}$ has a similar upper bound which differs by a constant factor and so does the universal estimation error.

Our result can be compared with the result in [6] of Bartlett:

$$\log N(\epsilon, \mathcal{H}_l, L_2(\mu_n)) \leq a\left(\frac{b}{\epsilon}\right)^{2l},$$

Here $a, b$ are factors independent of $\epsilon$. From this bound, we can only get the universal estimation error bound in the form of $O(n^{-1/2l})$, which means that the convergence rate decays very fast when more layers are used.

Deep neural nets often use hundreds of layers. One might think that this may lead to large estimation error and overfitting. However, our result shows that as long as we control the magnitude of the weights, overfitting is not a problem.

### 5.4 Boosting

Boosting is a general method to improve the accuracy for a learning algorithm and has proven to be an attractive strategy both in theory and practice. The main idea is to create a highly accurate predictor by combining many weak and inaccurate learners. It has applications in statistics, game theory, convex optimization, and information geometry.

In [19], Schapire et al gave the result that boosting does not over-fitting easily. They find empirical evidence that boosting tends to increase the margins associated with data and converge to a margin distribution in which data have large margins. The AdaBoost algorithm, introduced by Freud and Schapire even has the property that under certain assumptions, its training error can be exponentially decreased with the number of steps in boosting [37]. Also, they found that there is positive correlation between a reduction in test error when margin on training data is small. We will study the estimation error for boosting in our framework in this section.
Using simple function classes such as decision stumps as hypothesis space usually leads to low estimation error but high approximation error. In order to reduce the approximation error, we can enrich the hypothesis space by boosting. In each step $t$, based on the error the current function $h_{t-1}$ made, boosting greedily chooses a function $g_t$ from the base function space $B$, multiplied by the learning rate $\gamma_t$ and added to the current function $h_{t-1}$ to reduce the error $h_{t-1}$ made on data. We denote by $T$ the total number of steps. Let us consider the following hypothesis space:

$$\mathcal{H} = \left\{ \sum_{t=1}^{T} \gamma_t g_t \mid \sum_{t=1}^{T} |\gamma_t| \leq C, g_t \in B \right\}$$

which contains all possible functions produced by boosting with constraint on its learning rate.

In [19], Schapire, Freund, Bartlett and Lee provided a bound on generalization error by assuming that the VC dimension is finite.

In the following we will derive a bound for boosting in more general setting. Note that the hypothesis space $\mathcal{H}$ we considered can also be regarded as a $C$-convex hull of $B$, defined in the last section:

$$R(\mathcal{H}/\mu_n) = R(\text{conv}_C(B)/\mu_n) = C R(B/\mu_n)$$  \hfill (6.9)

As we argued previously, the Rademacher average can bound the estimation error. This result essentially tells us that the estimation error of boosting can be bounded by $C E_n R(B/\mu_n)$. Since the base function space $B$ is fixed in boosting, the bound is actually determined by $C$, the $L_1$ norm of the learning rate.

$C$ here controls the complexity of $\mathcal{H}$. When one uses too many steps and the corresponding learning rate does not decay fast enough, $C$ becomes too large and overfitting becomes a problem.

39
5.5 Support Vector Machines

In this section, we will introduce the covering number for function space generated by support vector machines. Support vector machines are learning algorithms try to find the maximum hyperplanes by using an implicit mapping into feature space and are largely used for classification problem. Of course, one can calculate the VC dimension of functions of support vector machine and get corresponding estimation for the producing function space and estimation error. However, this calculation ignores the effect of the kernel. Since one should expect that by using a smoother kernel, support vector machines can generate function space of less complexity and has smaller estimation error. In this section, the methodology for support vector machines can also be applied to other kernel methods in general.

Let us first introduce the function space generated by kernels. Given an abstract space $\mathcal{X}$, a kernel function $K : X_1 \times X_2 \mapsto \mathbb{R}$ is a nonnegative and symmetric map from a pair of object $X_1$ and $X_2$ in $\mathcal{X}$ to $\mathbb{R}$. It measures the similarity between $X_1$ and $X_2$ in some feature space we will define later. Let $\{X_1, \cdots, X_n\}$ be any $n$ data from $\mathcal{X}$, if the $n \times n$ matrix $(K(X_i, X_j))_{1 \leq i,j \leq n}$ is positive definite, we call this kernel Mercer Kernel. For Mercer Kernel, $K$ has eigen expansions ensured by the following theorem

**Theorem 5.5.1.** [25] Let $K$ be a Mercer Kernel and the associated integral operator $T_K : L_2(\mathcal{X}) \mapsto L_2(\mathcal{X})$ defined as

$$T_K f(\cdot) := \int_{\mathcal{X}} K(\cdot, X)f(X)dX$$

Let $\psi_i \in L_2(\mathcal{X})$ where $i \in \mathbb{N}$ be the continuous eigenfunction of $T_K$ with $\|\psi_i\|_{L^2(\mathcal{X})} = 1$ and $\lambda_i \neq 0$ is the associated eigenvalue. Then

1. $\sum_{i=1}^{\infty} |\lambda_i| \leq \infty$.

2. $\psi_i \in L_\infty(\mathcal{X})$ and $\sup_i \|\psi_i\| \leq \infty$. 

40
3. \( K(X_1, X_2) = \sum_{i \in \mathbb{N}} \lambda_i \psi_i(X_1) \psi_i(X_2) \) holds for all \( (X_1, X_2) \in \mathcal{X} \times \mathcal{X} \), and the series converges absolutely and uniformly.

The corresponding function space \( \mathcal{H}_K \) for \( K \) is called reproducing kernel Hilbert space, it contains the linear span of \( \{ K(\cdot, X), X \in \mathcal{X} \} \). Then function \( f \in \mathcal{H}_K \) has the representation

\[
f(X) = \sum_{i=1}^{\infty} c_i \psi_i(X)
\]

Moreover, we impose the inner product on \( \mathcal{H}_K \) to satisfy \( \langle K(\cdot, X), f(\cdot) \rangle_{\mathcal{H}_K} = f(X) \). Then one can conclude that for \( f = \sum_{i=1}^{\infty} c_i \psi_i \), we have

\[
\|f\|_{\mathcal{H}_K} = \sum_{i=1}^{n} c_i^2 / \lambda_i < \infty.
\]

In [31], Guo and et al get the covering number bound for the following function space

\( \mathcal{H}_K(B) = \{ f : f \in \mathcal{H}_K \text{ and } ||f||_{\mathcal{H}_K} \leq B \} \).

**Theorem 5.5.2.** [31] Let the Mercer kernel \( K \), \( \mathcal{H}_K(B) \), eigen functions \( \psi_i \) and eigenvalues \( (\lambda_i) \) are defined above. Let \( \{X_1, \ldots, X_n\} \) be any \( n \) data from \( \mathcal{X} \) and

\[
\sup_i ||\psi_i||_{L_{\infty}} = C_k < \infty
\]

For any \( m \in \mathbb{N} \), set

\[
\varepsilon_m^* = 6BC_k \sqrt{\frac{j^*(\frac{\lambda_1 \cdots \lambda_{j^*}}{m^2})^\frac{2}{j^*}} + \sum_{i=j^*}^{\infty} \lambda_i}
\]

with

\[
 j^* = \min\{j : \lambda_{j+1} < \left(\frac{\lambda_1 \cdots \lambda_j}{m^2}\right)^\frac{1}{j}\}
\]

Then

\[
\mathbb{N}(\varepsilon_m^*, \mathcal{H}_K(B), L_{\infty}(\mu_n)) \leq m
\]
We can think of the number $j^*$ as the effective dimension of the function space, and only depends on the decay rate of the eigenvalues. Note that this theorem gives a obscure covering number bound for space $\mathcal{H}_K(B)$. In [31], an explicit covering number for the Gaussian radial basis function kernel $K(X_1, X_2) = e^{-(X_1 - X_2)^2/\sigma^2}$ is estimated. And we have

$$N(\epsilon^*_m, \mathcal{H}_K(B), L_\infty(\mu_n)) = O(\log \frac{3}{\epsilon}).$$

Since $L_\infty$ norm can bound $L_2$ norm, we conclude that the optimal rate $\sqrt{n}$ of the universal estimation error for $\mathcal{H}_K(B)$ can be achieved. Although $H_K(B)$ has infinite dimension representation as $\sum_{i=1}^{\infty} c_i \psi_i$, the $B$-ball in $\mathcal{H}_K(B)$ is not that complicated as we thought.
Chapter 6

Generalization Dynamics

Previous work on estimating generalization loss or generalization error has focused on the
stability of the training algorithm, such as recent work of algorithmic stability \[39, 40\]. An
time algorithm is stable if single observation change in the training set does not affect
the output function significantly. For stable algorithms, one can use leave-one-out style
estimates for the generalization error. However, the arguments of stability still rely heavily
on traditional notions such as VC dimension.

In order to include the consideration of the training algorithm into the generalization
error, we provide a general framework of the dynamics of training process and general-
ization process when training with stochastic gradient descent or gradient descent. These
dynamics depend on the data generation distribution, the randomness of stochastic pro-
cess, but not the loose complexity measure of functional space.

By denote $L(X_i, \theta_t)$ the loss function on data $X_i$ and let $\theta_t$ be the training parameters
at time $t$ in training process. We have

$$L_n(\theta_t) = \frac{1}{n} \sum_{i=1}^{n} L(X_i, \theta_t) \quad \text{and} \quad L(\theta_t) = \mathbb{E}L(X, \theta_t)$$
where $L_n$ and $L$ are the training loss and corresponding generalization loss. $\theta_t$ is trained with stochastic gradient descent. The continuous version of its dynamics is

$$d\theta_t = \nabla L_n(\theta_t) dt + (\eta \Sigma(\theta_t)^{1/2} dW_t)$$

where $\eta$ is the scaling factor associated with learning rate, $W_t$ is the Wiener process and $\Sigma(\theta_t)$ is the sample covariance given by

$$\Sigma(\theta_t) = \frac{1}{n} \sum_{i=1}^{n} \left[ \nabla L(X_i, \theta_t) - \nabla L_n(\theta_t) \right] \left[ \nabla L(X_i, \theta_t) - \nabla L_n(\theta_t) \right]^T$$

Then the dynamics for training loss is:

$$dL_n(\theta_t) = -||\nabla L_n(\theta_t)||^2 dt + \left[ \eta \nabla^T L_n(\theta_t) \Sigma(\theta_t) \nabla L_n(\theta_t) \right]^{1/2} dW_t$$

and the generalization loss is:

$$dL(\theta_t) = -\nabla^T L_n(\theta_t) \nabla L(\theta_t) dt + \left[ \eta \nabla^T L_n(\theta_t) \Sigma(\theta_t) \nabla L_n(\theta_t) \right]^{1/2} dW_t$$

If we use gradient descent instead of stochastic gradient descent, we do not have the stochastic terms above. In the training loss equation, the first term $-||\nabla L_n(\theta_t)||^2 dt$ is always negative thus is negatively correlated with $L_n$. While in the generalization loss equation, the first term depends on the correlation between the gradients of $L$ and $L_n$. The intuition is, if the magnitude of gradient of $L$ is the same as the gradient of $L_n$, the training error is always decreasing no slower than generalization error.

### 6.1 Invariants for the Gradient Descent Dynamics

In this section, we focus on the fact that gradient descent only explores a subspace in $\mathcal{F}$ instead of the whole $\mathcal{F}$, given fixed initial value for neural networks. In particular, we show some time invariant equations in parameter space must hold during the training process.
It implies that the complexity of \( \mathcal{F} \) may not provide sharp bound for generalization error in practice, since it can only bound generalization error uniformly over the whole \( \mathcal{F} \). The results in this section may also be interesting for other theoretical analysis.

For training linear neural networks and deep ReLU (or leaky ReLU) neural networks with stochastic gradient descent (SGD) or gradient descent (GD), we assume that the network \( f(x) \) has \( H \) hidden layers, where the \( k \)-th layer is

\[
y_k = W_k z_{k-1} + b_k, \quad z_k = \sigma(y_k), \quad y_k, z_k, b_k \in \mathbb{R}^{d_k}, W_k \in \mathbb{R}^{d_k \times d_{k-1}}
\]

and \( z_0 = x \in \mathbb{R}^d, f(x) = y_{H+1} \in \mathbb{R} \). Here \( \sigma \) is element-wise nonlinearity: for \( y_k = [y_{k,[1]}, y_{k,[2]}, \ldots, y_{k,[d_k]}]^T \in \mathbb{R}^{d_k} \), let

\[
\sigma(y_k) = [\sigma(y_{k,[1]}), \ldots, \sigma(y_{k,[d_k]})]^T \in \mathbb{R}^{d_k}, \\
\sigma'(y_k) = \text{diag} \left( \sigma'(y_{k,[1]}), \ldots, \sigma'(y_{k,[d_k]}) \right) \in \mathbb{R}^{d_k \times d_k}.
\]

For simplicity, let \( \tilde{W}_k = [W_k, b_k], \tilde{z}_k = [z_k]^T \), then \( y_k = \tilde{W}_k \tilde{z}_{k-1} \).

Let \( l(f(x), y) \) be the loss function, we have

\[
\frac{\partial l}{\partial \tilde{W}_k} = \frac{\partial l}{\partial y_k} \tilde{z}_{k-1}^T, \quad \frac{\partial l}{\partial \tilde{W}_{k+1}} = \frac{\partial l}{\partial y_{k+1}} z_{k}^T.
\]

In the forward and backward propagation,

\[
\frac{\partial l}{\partial y_k} = \sigma'(y_k) W_{k+1}^T \frac{\partial l}{\partial y_{k+1}}, \quad z_k = \sigma(\tilde{W}_k \tilde{z}_{k-1}).
\]

For ReLU or leaky ReLU activation function, \( \sigma(y) = \sigma'(y)y \), thus

\[
\sigma'(y_k) W_{k+1}^T \frac{\partial l}{\partial \tilde{W}_{k+1}} = \frac{\partial l}{\partial \tilde{W}_k} \tilde{W}_k^T \sigma'(y_k).
\]
Consider the $j$-th neural in the $k$-th layer, we have

$$\sigma'(y_{k,j}) W_{k+1,[:],j}^T \frac{\partial l}{\partial W_{k+1,[:],j}} = \frac{\partial l}{\partial W_{k,[:],j}} \tilde{W}_{k,[:],j}^T \sigma'(y_{k,j}),$$

$$W_{k+1,[:],j}^T \frac{\partial l}{\partial W_{k+1,[:],j}} = \frac{\partial l}{\partial W_{k,[:],j}} \tilde{W}_{k,[:],j}^T.$$

Notice that the second equation holds even for $\sigma'(y_{k,j}) = 0$.

Suppose the training data set is $\{(x^{(i)}, y^{(i)})\}_{i=1}^N$. Let $l^{(i)} = l(f(x^{(i)}), y^{(i)})$. The parameter update rules during the training are

$$\Delta W_k = -\eta \frac{\partial l^{(\gamma)}}{\partial W_k}, \quad \text{(SGD)}$$

$$\Delta W_k = -\frac{\eta}{N} \sum_{i=1}^N \frac{\partial l^{(i)}}{\partial W_k}, \quad \text{(GD)}$$

where $\eta$ is the learning rate, $\gamma$ is randomly chosen from $\{1, \ldots, N\}$ in each step. We can see for both SGD and GD,

$$W_{k+1,[:],j}^T \Delta W_{k+1,[:],j} = \Delta \tilde{W}_{k,[:],j} \tilde{W}_{k+1,[:],j}^T.$$

Therefore,

$$\Delta \| W_{k+1,[:],j} \|^2 = \Delta \| \tilde{W}_{k,[:],j} \|^2 + o(\eta),$$

thus

$$\| W_{k+1,[:],j} \|^2 - \| \tilde{W}_{k,[:],j} \|^2 = C + o(\eta T).$$

For each neuron in layer $k$, the square of $L_2$ norms difference between weights connecting it in layer $k$ and in layer $k + 1$ remains approximately constant in the training process. Since there are $\sum_{k=1}^H d_k$ nodes in the neural networks, this result implies a subspace of dimension $\sum_{k=1}^H d_k$ in the parameter space is effectively frozen, and the subspace that gradient descent can explore is of dimension $\sum_{k=1}^H d_k d_{k+1} - \sum_{k=1}^H d_k$. Intuitively, the number of $\epsilon$-balls required to cover this subspace should be less that the number required
to cover the whole $\mathcal{F}$. The corresponding generalization bound should be smaller if we take it into consideration.

The results for linear neural networks seem to be more surprising, the frozen dimension is as large as $\sum_{k=1}^{H} d_k^2$. Let us consider the model

$$y = W_H W_{H-1} \cdots W_1 x$$

where $W_k \in \mathbb{R}^{d_{k+1} \times d_k}$ with $d_{H+1} = 1$ and $d_1 = d$. Like the neural network case, we will train this linear neural network with gradient descent and the loss function is square loss. We also assume all $W_k$ update simultaneously in the training process. We give the following lemma about the invariant of linear neural networks during the training process

**Lemma 6.1.1.** There exists a $(k + 1) \times (k + 1)$ matrix $C$, such that for GD and SGD

$$(W_{k+1})^T W_{k+1} - (W_k)^T W_k = C$$

**Proof.** Let $f(x) = W_H W_{H-1} \cdots W_1 x$ and $h_k = W_k W_{k-1} \cdots W_1 x$ for $k \leq H$. Then for the square loss $L$, we have

$$\frac{\partial L}{\partial W_k} = \frac{\partial L}{\partial h_k} h_{k-1}^T$$

and

$$\frac{\partial L}{\partial W_{k+1}} = \frac{\partial L}{\partial h_{k+1}} h_k^T$$

Also, because $h_{k+1} = W_{k+1}h_k$, $h_{k+1}$ and $h_k$ satisfy

$$\frac{\partial L}{\partial h_{k+1}} = \frac{\partial L}{\partial h_k} W_{k+1}$$

Thus we have

$$(W_{k+1})^T \frac{\partial L}{\partial W_{k+1}} = \frac{\partial L}{\partial W_k} (W_k)^T$$

Assume that the learning rate for gradient descent is 1, we have $dW_k = \partial L/\partial W_k$. Now we can conclude that $(W_{k+1})^T dW_{k+1} = dW_k (W_k)^T$ and hence the term $(W_{k+1})^T W_{k+1} - (W_k)^T W_k$ is invariant with respect to time $t$. \hfill \square
6.2 High Dimensional Linear Regression

The convergence rate we discussed in previous chapters only deals with the asymptotic behavior when the data size is large enough. In practice, this may not be the case. Sometimes, we have more parameters than the number of data we can use to train the model. In this case, the framework we have discussed can not provide a good bound for the generalization error since the functional space complexity term dominates data size. This contradicts the fact that models with more parameters may be generalized better. Moreover, the underlying distribution may play an import role to bound generalization error. Zhang et al [32] showed that while deep networks with large parameters can achieve zero training error on a data set that has been randomly labeled, they still can generalize well when training on the true labels.

In previous chapters, we propose very loose assumptions about the underlying data distribution and intentionally ignore the process of how we use specific algorithm to find the empirical minimizer. In this section, we will analyze the training error dynamics and generalization error dynamics of linear models trained with gradient descent. The high dimension case attracts more interests as we just discussed, so we will study the case when the the number of parameters is larger than the data size.

We begin with a standard linear model with input $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$:

$$y = W^* x + e \quad \text{where} \quad x \sim \mathcal{N}(0, \Sigma_{d \times d}), e \sim \mathcal{N}(0, \sigma^2)$$

where $W^*$ is a 1 by $d$ parameters vector to be estimated. Assume we have $n$ samples draw from this model

$$Y = W^* X + E$$

where $X = (X_1, \cdots, X_n) \in \mathbb{R}^{d \times n}$, $Y = (Y_1, \cdots, Y_n) \in \mathbb{R}^{1 \times n}$ and $E = (E_1, \cdots, E_n) \in \mathbb{R}^{1 \times n}$. We focus on high dimensional model with $d > n$. We also use gradient descent method and square loss function.
For any estimator $W$ of $W^*$, the training error can be written as

$$L_n(W) = \frac{1}{2n} \| Y - WX \|_2^2$$

and the associated generalization error is

$$L(W) = \frac{1}{2} \mathbb{E} \| y - Wx \|_2^2$$

The gradient descent begins with initial value $W_0$. During training, the parameters always satisfy

$$\Delta W = -r \nabla L_n(W)$$

where $r$ is the learning rate and will be set to 1 for simplicity. The corresponding continuous version for $W_t$ should satisfy

$$dW_t = -\nabla L_n(W_t)dt = \frac{1}{n}X^T(Y - W_tX)dt$$

solving the ODE leads to the process for $\beta_t$

$$W_t = W_0e^{-Bt} + YX^\dagger(1 - e^{-Bt})$$

where $B = \frac{1}{n}XX^T \in \mathbb{R}^{d \times d}$ is the covariance matrix and $X^\dagger = (X^TX)^{-1}X^T \in \mathbb{R}^{n \times d}$ is the pseudo inverse of $X$. The long term behavior of this process is clear. $W_t$ will converge to $YX^\dagger$. Rewrite the process by incorporating $Y = W^*X + E$ and decomposing $W_t - W_0$ into the time invariant and time variant parts for the training process, we have

$$W_t - W^* = (W_0 - W^*)^\perp + (W_0 - W^*)^\parallel e^{-Bt} + EX^\dagger(1 - e^{-Bt})$$

where $(W_0 - W^*)^\perp = (W_0 - W^*)(1 - XX^\dagger)$ is a constant vector with respect to $t$ and is orthogonal to the subspace spanned by data (columns of $X$), and we have $(1 - XX^\dagger)X = 0$. 

49
The term \((W_0 - W^*)/ = (W_0 - W^*)X X^\dagger\) is parallel to the subspace spanned by data and decreasing to 0 in an exponential rate with respect to time \(t\). The third term is the noise term and converges to \(EX^\dagger\) when \(t\) goes to infinity. Then we can calculate the process of the training error

\[
L_n(t) = \frac{1}{2n} ||Y - W_t X||^2
= \frac{1}{2n} |((W^* - W_t)X + E)||^2
= \frac{1}{2n} |((W^* - W_0)e^{-Bt}X + EX^\dagger e^{-Bt}X)||^2
\]

Let \(X = USV^T\) be the singular value decomposition for \(X\) where \(U \in \mathbb{R}^{d \times n}\) and \(S, V \in \mathbb{R}^{n \times n}\). Also, by taking the expectation of the noise term \(E\), the cross term from \((W_0 - W^*)e^{-Bt}X\) and \(EX^\dagger e^{-Bt}X\) is 0. We can conclude that

\[
\mathbb{E}_E L_n(t) = \frac{1}{2n} [(W_0 - W^*)US^2 e^{-\frac{1}{2}S^2 t}U^T(W_0 - W^*)^T + \sigma^2 \text{tr}(e^{-\frac{2}{2}S^2 t})]
\]

It is obvious that with \(t \to \infty\), \(\mathbb{E}_E L_n(t)\) goes to 0 exponentially, and the decay rate of the training error is controlled by the spectral of \(S\). Another interesting observation is that as both terms on the right hand side exponentially decay, their ratio stays unchanged during the training. Note that the model is \(y = W^* x + \epsilon\), the first error term \((W_0 - W^*)US^2 e^{-\frac{1}{2}S^2 t}U^T(W_0 - W^*)^T\) can be interpreted as the error from \(W^* x\) and the second term \(\sigma^2 \text{tr}(e^{-\frac{2}{2}S^2 t})\) can be thought as the error from noise \(\epsilon\).

Now we switch to study the process of generalization error, we have

\[
L(t) = \mathbb{E} ||W_t x - y||^2
= \frac{1}{2} [(W_t - W^*)\Sigma(W_t - W^*)^T + \sigma^2]
\]
where $\Sigma$ is covariance matrix for $x$. Without loss of generality, we assume $\Sigma = I$. Again, by taking expectation of the noise term $E$ in $W_t - W^*$, the generalization error is

$$E_E L(t) = \frac{1}{2} \left[ \| (W_0 - W^*)^\perp \|^2 + (W_0 - W^*)/e^{-2Bt}(W_0 - W^*)// + \sigma^2 \text{tr}(S^{-2}(1 - e^{-\frac{1}{n}S^2t})^2) + \sigma^2 \right]$$

Moreover, we have

$$E_E L(t) \rightarrow \frac{1}{2} \left[ \| (W_0 - W^*)^\perp \|^2 + \sigma^2 \text{tr}(S^{-2}) + \sigma^2 \right] \text{ when } t \rightarrow \infty$$

The first term $\| (W_0 - W^*)^\perp \|^2$ is the discrepancy between the initial value $W_0$ and $W^*$ that can not be learned by the data. Like the decomposition of training error process, the term $(W_0 - W^*)/e^{-2Bt}(W_0 - W^*)//$ can be interpreted as the error from $W^*x$ and converges to 0 when $t$ goes to infinity. The term $\sigma^2 \text{tr}(S^{-2}(1 - e^{-\frac{1}{n}S^2t})^2)$ corresponds to the noise $\epsilon$. Since $S$ is a diagonal matrix, we denote the $ii$-th entry as $\sigma_i$ with $\sigma_1 \geq \cdots \geq \sigma_n$.

We can rewrite this term as

$$\sigma^2 \text{tr}(S^{-2}(1 - e^{-\frac{1}{n}S^2t})^2) = \sigma^2 \sum_{i=1}^{n} \left(1 - e^{-\frac{1}{n}\sigma_i^2t} \right)^2.$$  

This term monotonically increases with respect to time $t$. It implies that the limit of $E_E L(t)$ is not guaranteed to be the best generalization error one can achieve, and there should be an optimal time for the generalization process. Before the optimal time, the training improves the generalization error and the dominate term is $\| (W_0 - W^*)^\perp \|^2$. After that time, the training process switches to the over fitting regime and is dominated by $\sigma^2 \text{tr}(S^{-2}(1 - e^{-\frac{1}{n}S^2t})^2)$. Taking the derivative with respect to $t$ for both term, we can conclude that the optimal time $T^*$ satisfies the following equation:

$$(W_0 - W^*)/e^{-2Bt}(W_0 - W^*)//= \frac{\sigma^2}{n} \sum_{i=1}^{n} e^{-\frac{1}{n}\sigma_i^2T^*}(1 - e^{-\frac{1}{n}\sigma_i^2T^*})$$

51
If $n$ goes to infinity (with the constraint $n < d$), $B$ converges to $I$ and the distribution of singular values $\{\sigma_1^2, \ldots, \sigma_n^2\}$ should converge to some spectral distribution of a covariance matrix of Gaussian variable. The noise term would converge to a constant.

From the examples of the generalization error dynamics in this section, we know it is important to consider the underlying data distribution when analyzing the generalization error. We also know $\sigma$ measures the noise of the output. Though the training error can always converge to 0 exponentially in high dimensional case but the generalization error can not converge to 0. When $\sigma$ becomes larger, the generalization error increases and it may takes more time to learn.

In conclusion, unlike the empirical error whose expectation is decreasing with respect to time, the generalization error has U-shaped behavior during training. There exits an optimal time for the generalization error and early stopping is necessary.
Bibliography


