Weinan and Wang *Res Math Sci(2017)4:2*
DOI 10.1186/s40687-016-0093-6

Research in the Mathematical Sciences

RESEARCH Open Access

Optimal convergence rate of the universal estimation error

E. Weinan[1,2,3] and Yao Wang[2*]

*Correspondence: yw10@princeton.edu
[2]Department of Mathematics, Princeton University, Princeton, NJ, USA
Full list of author information is available at the end of the article

Abstract

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 $\text{fat}_\epsilon(\mathcal{F}) = O(\epsilon^{-p})$, 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 minimax optimal rate of $O(n^{-1/2})$. We also show that if the hypothesis space is the compact supported convex Lipschitz continuous functions in $\mathbb{R}^d$ with $d > 4$, then the rate is approximately $O(n^{-2/d})$.

1 Background

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 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. In this paper, we will focus on the estimation error.

To begin with, 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 μ, here X_i is the i-th input and Y_i is the corresponding output. L is the loss function and $\mathcal{H}$ is the hypothesis space which contains functions from X to Y. Let h^* be the minimizer of the risk associated with $\mathcal{H}$ and $\hat{h}$ be the minimizer of the empirical risk:

$$h^* := \operatorname*{argmin}_{h\in H} \mathbb{E}_\mu[L(h)],$$
$$\hat{h} := \operatorname*{argmin}_{h\in H} \mathbb{E}_{\mu_n}[L(h)].$$

Here for simplification, we use $L(h)$ in place of $L(h(X), Y)$, $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$ for the empirical measure and $\mathbb{E}_{\mu_n}[L(h)] = \frac{1}{n}\sum_{i=1}^n (L(h(X_i), Y_i))$ to denote the empirical risk. The estimation error is defined to be $\mathbb{E}_\mu[L(\hat{h}) - L(h^*)]$.

Springer Open

To estimate the estimation error, instead of looking at the space $\mathcal{H}$, we will look at the excess loss class associated with $\mathcal{H}$, denoted as $\mathcal{F}$, see [4]

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

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 $\mathbb{E}_\mu[\hat{f}]$. Since f^* is the minimizer of $\mathbb{E}_\mu[f]$ and $\mathbb{E}_\mu[f^*] = 0$, we know that $\mathbb{E}_\mu[\hat{f}] \geq 0$. Similarly, we know that $\hat{f}$ is the minimizer of $\mathbb{E}_{\mu_n}[f]$ and because $\mathbb{E}_{\mu_n}[f^*] = 0$, we have $\mathbb{E}_{\mu_n}[\hat{f}] \leq 0$. Therefore, we have

$$0 \leq \mathbb{E}_\mu[\hat{f}] \leq \mathbb{E}_\mu[\hat{f}] - \mathbb{E}_{\mu_n}[\hat{f}]. \tag{1.1}$$

To bound the $\mathbb{E}_\mu[\hat{f}]$, it is enough to bound $\mathbb{E}_\mu(\hat{f}) - \mathbb{E}_{\mu_n}[\hat{f}]$. Intuitively, for any fixed function f, if we blindly apply the Law of Large Number and the Central Limit Theorem, we get

$$\mathbb{E}_\mu[f] - \mathbb{E}_{\mu_n}[f] \to 0 \text{ almost surely,}$$
$$\sqrt{n}(\mathbb{E}_\mu[f] - \mathbb{E}_{\mu_n}[f]) \to N(0, \mathbb{E}_\mu(f - \mathbb{E}_\mu[f])^2) \text{ in distribution.}$$

However, we cannot use the Law of Large Number or the Central Limit Theorem for $\hat{f}$ since $\hat{f}$ is the empirical minimizer, the iid assumption does not hold.

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 $\mathbb{E}_{\mu_n}\hat{f} = 0$. But there is no guarantee that $\mathbb{E}_\mu\hat{f} = 0$ and hence no guarantee that $\mathbb{E}_\mu[\hat{f}] - \mathbb{E}_{\mu_n}[\hat{f}]$ 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 rather than focusing only on $\hat{f}$. Thus we define the empirical process $\{(\mathbb{E}_{\mu_n} - \mathbb{E}_\mu)(f) : f \in \mathcal{F}\}$ as the family of the random variables indexed by $f \in \mathcal{F}$. Instead of bounding $\mathbb{E}_\mu[\hat{f}] - \mathbb{E}_{\mu_n}[\hat{f}]$, it is better to bound the supremum of the empirical process. Define $||Q||_\mathcal{F} = \sup\{|Qf| : f \in \mathcal{F}\}$. The quantity $||\mathbb{E}_{\mu_n} - \mathbb{E}_\mu||_\mathcal{F}$ will be called the *empirical process supremum* and its expectation $\mathbb{E}_\mu||\mathbb{E}_{\mu_n} - \mathbb{E}_\mu||_\mathcal{F}$ will be called the *μ-estimation error*, and it naturally provides a good bound for the estimation error.

Next we define a $\mathcal{F}$-indexed empirical process G_n by

$$f \mapsto G_n f = \sqrt{n}(\mathbb{E}_{\mu_n} - \mathbb{E}_\mu)[f] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(Z_i) - \mathbb{E}_\mu(f)). \tag{1.2}$$

We now make the assumption that

$$\sup_{f \in \mathcal{F}} |f(Z) - \mathbb{E}_\mu(f)| < \infty \tag{1.3}$$

for all Z. Under this condition, the empirical process $\{G_n : f \in \mathcal{F}\}$ can be viewed as a map in $l^\infty(\mathcal{F})$. Consequently, it makes sense to investigate conditions under which

$$G_n = \sqrt{n}(\mathbb{E}_{\mu_n} - \mathbb{E}_\mu) \to G \text{ in distribution,} \tag{1.4}$$

where G is a tight process in $l^\infty(\mathcal{F})$. This is actually the $\mathcal{F}$-version Central Limit Theorem. Function spaces that satisfy this property are called *Donsker class* [10]. Moreover, a class $\mathcal{F}$ is called a *Glivenko–Cantelli class (GC)* [10] if the $\mathcal{F}$-version Law of Large Numbers

$$||\mathbb{E}_{\mu_n} - \mathbb{E}_\mu||_\mathcal{F} \to 0 \text{ almost surely}$$

holds.

We now simplify the assumption (1.3). If we let $g = f - E_\mu[f]$, we have

$$\mathbb{E}_\mu[g] - \mathbb{E}_{\mu_n}[g] = \mathbb{E}_\mu[f] - \mathbb{E}_{\mu_n}[f]. \tag{1.5}$$

Thus we can assume $\mathbb{E}_\mu f = 0$ for any $f \in \mathcal{F}$. Then (1.3) can be simplified to be

$$\sup_{f\in\mathcal{F}} |f(Z))| < \infty. \tag{1.6}$$

Without loss of generality, we further assume that

$$\sup_{f\in\mathcal{F}} |f(Z))| \le 1.$$

Equivalently, we are interested in the following class of distributions

$$\mathcal{P} = \left\{\mu : |f(Z)| \le 1 \text{ for any } f \in \mathcal{F} \text{ and any } Z \text{ generated from } \mu\right\}.$$

Since μ is actually unknown (otherwise we have achieved our goal for learning), we study the worst case of μ-estimation error, so we define the

$$\sup_{\mu\in\mathcal{P}} \mathbb{E}_\mu ||\mathbb{E}_{\mu_n} - \mathbb{E}_\mu||_{\mathcal{F}}$$

to be the *universal estimation error*.

2 Preliminaries

There are many classical approaches to describe the complexity of a class of functions. For instance, growing number and VC dimension can be used to describe the binary classification hypothesis space. In more general settings, one can also use the Rademacher complexity. However, it seems that this quantity is not very intuitive. When using these terms, one cannot tell how fast the empirical loss minimizer comes close to the loss minimizer as the data size increases. In this paper, we will use the entropy and the Fat-shattering dimension to describe the complexity.

2.1 Rademacher average

The first step to study the μ-estimation error is to study the Rademacher average: for fixed empirical measure μ_n, we define the *Rademacher average* [3,14] by

$$R(\mathcal{F}/\mu_n) = \mathbb{E}_r \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n r_i f(Z_i)\right| \tag{2.1}$$

where $r_i, \ldots, r_n$ are iid 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. Also, we define the *Rademacher process* associated with the empirical measure μ_n as

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

It is known that the Rademacher averages control the μ estimation error:

Theorem 2.1 [10] *If $\mathcal{F}$ is a class of functions map into $[-M, M]$, then for every integer n, we have*

$$\begin{aligned}\mathbb{E}_\mu \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n (f(Z_i) - \mathbb{E}_\mu f)\right| &\le 2\mathbb{E}_{\mu\times r} \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n r_i f(Z_i)\right| \\ &\le M\mathbb{E}_\mu \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n (f(Z_i) - \mathbb{E}_\mu f)\right| + \mathbb{E}_r \left|\frac{1}{n}\sum_{i=1}^n r_i\right|.\end{aligned} \tag{2.3}$$

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

2.2 Covering number and fat-shatter dimension

To get more explicit bounds, we need two more concepts. In what follows, the logarithm always takes 2 as base and $L_p(\mu_n)$ norm of f is defined as $\left(1/n\sum_{i=1}^n |f(Z_i)|^p\right)^{1/p}$.

Definition 2.2 For an arbitrary semi-metric space (T, d), the *covering number* $\mathbb{N}(\epsilon, T, d)$ is the minimal number of the closed d-balls of radius ϵ required to cover T. See [8,10]. The associated *entropy* $\log\mathbb{N}(\epsilon, T, d)$ is the logarithm of the covering number.

We also define another concept which is always easy to calculate: the Fat-shattering dimension.

Definition 2.3 For every $\epsilon > 0$, a set $A = \{Z_1, \ldots, Z_n\}$ is said to be ϵ- shattered by $\mathcal{F}$ if there exists some real function $s : A \to \mathbb{R}$ such that for every $I \in \{1, \ldots, n\}$ 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 $i \notin I$.

$$\mathrm{fat}_\epsilon(\mathcal{F}) := \sup\left\{|A|\middle|A \in \Omega, A \text{ is } \epsilon\text{-shattered by } \mathcal{F}\right\}$$

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

Note that both the Fat-shattering dimension and the covering number are non-decreasing as ϵ decreases and since $||f||_{L_1(\mu_n)} \leq ||f||_{L_2(\mu_n)} \leq ||f||_{L_\infty(\mu_n)}$, we know that

$$\mathbb{N}(\epsilon, \mathcal{F}, L_1(\mu_n)) \leq \mathbb{N}(\epsilon, \mathcal{F}, L_2(\mu_n)) \leq \mathbb{N}(\epsilon, \mathcal{F}, L_\infty(\mu_n)). \tag{2.4}$$

The Fat-shattering dimension is actually linear with respect to the entropy up to a logarithm factor of the Fat-shattering dimension[14], on page 253 and page 252:

Lemma 2.4 *If $|f| \leq 1$ for any $f \in \mathcal{F}$, then*

$$\sup_{\mu_n} \log \mathbb{N}(\epsilon, \mathcal{F}, L_1(\mu_n)) \geq fat_{16\epsilon}(\mathcal{F})/8. \tag{2.5}$$

Lemma 2.5 *for every empirical measure μ_n and $p \geq 1$, there is some constant c_p such that*

$$\log \mathbb{N}(\epsilon, \mathcal{F}, L_p(\mu_n)) \leq c_p fat_{\frac{\epsilon}{8}}(\mathcal{F}) \log^2\left(\frac{2fat_{\frac{\epsilon}{8}}(\mathcal{F})}{\epsilon}\right). \tag{2.6}$$

2.3 Maximal inequality

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

$$||\max_{1\leq i\leq m} X_i||_p \leq (\mathbb{E}\max_{1\leq i\leq m}|X_i|^p)^{1/p} \leq m^{1/p}\max_{1\leq i\leq m}||X_i||_p. \tag{2.7}$$

As m increase, this type bound increases very fast, so we cannot get satisfied result. To overcome this, we introduce the following Orlicz 2-norm and the corresponding maximal inequality:

Definition 2.6 Let $\psi_2(x) = e^{x^2} - 1$, and *Orlicz norm* for random variables $||\cdot||_{\psi^2}$ is defined by (see [10] for more details)

$$||X||_{\psi_2} := \inf\left\{c > 0 : \mathbb{E}\psi_2\left(\frac{|X|}{c}\right) \leq 1\right\}. \tag{2.8}$$

Note that $||X||_{\psi_2} \geq ||X||_{L_1}$ since $\psi_2(x) \geq x$. The Orlicz norm is more sensitive to the behavior of in the tail of X, which makes it possible to have a better bound if we bound the maxima of many variables with a light tails. The following lemma gives a better bound [11], in chapter 8:

Lemma 2.7 *Let $X_1, X_2, \ldots, X_m$ be random variables,*

$$|| \sup_{1\leq i\leq m} X_i||_{\psi_2} \leq 4\sqrt{\log(m+1)} \sup_{1\leq i\leq m} ||X_i||_{\psi_2}. \tag{2.9}$$

Random variables from Rademacher process actually have a nice property that their tails decrease very fast. The following result was proved by Kosorok in [11], in chapter 8:

Lemma 2.8 *Define*

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

where $r_i, \ldots, r_n$ are i.i.d. Rademacher random variables satisfying $P(r = -1) = P(r = 1) = 1/2$. Let $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$, Then we have

$$P\left(\left|\sum_{i=1}^{n} r_i a_i\right| > x\right) \leq 2e^{-\frac{1}{2}x^2/||a||^2} \tag{2.10}$$

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

Our main technique comes from Mendelson [14], who 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 $1/n$. 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. In [13,17], it was that the Gaussian and the Rademacher averages are closely related and have the following connection:

Theorem 2.9 *There are absolute constants c and C such that for every n and $\mathcal{F}$*

$$c(1+\log n)^{\frac{1}{2}}\mathbb{E}_g \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n} g_i f(Z_i)\right| \leq \mathbb{E}_r \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n} r_i f(Z_i)\right| \leq C\mathbb{E}_g \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n} g_i f(Z_i)\right|. \tag{2.11}$$

Using the above theorem and the result in [14], the upper bound was given for expectation of the Rademacher average. But we cannot say whether the bound is tight. In the following 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 section 4.

3 Upper bound

To bound the empirical Rademacher average, we use the following theorem, this follows from the standard "chaining" method, see [11], chapter 8.

Theorem 3.1 *Let μ_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 to 0 with $\epsilon_0 = 1$. Then, there exists an absolute constant C such that for any integer N,*

$$R(\mathcal{F}/\mu_n) \leq Cn^{-\frac{1}{2}} \sum_{k=1}^{N} \epsilon_{k-1} \log^{\frac{1}{2}} \mathbb{N}(\epsilon_k, \mathcal{F}, L_2(\mu_n)) + \epsilon_N. \tag{3.1}$$

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

Construct a sequence of finite covering sets $\mathcal{F}_0, \mathcal{F}_1, \ldots, \mathcal{F}_N$ such that $\mathcal{F}_i \subset \mathcal{F}$ and $\mathcal{F}_i$ is minimal ϵ_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 μ_n and study the associated Rademacher process $X_{rad}(f) = \frac{1}{n}\sum_{i=1}^{n} r_i f(Z_i)$. Applying the triangle inequality to the Rademacher average, we get

$$R(\mathcal{F}/\mu_n) = \mathbb{E}_r \sup_{f\in\mathcal{F}} |X_{rad}(f)| \leq \mathbb{E}_r \sup_{f\in\mathcal{F}} |X_{rad}(f - f_N)| + \mathbb{E}_r \sup_{f_N\in\mathcal{F}_N} |X_{rad}(f_N)|. \tag{3.2}$$

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

$$\begin{aligned} \mathbb{E}_r \sup_{f\in\mathcal{F}} \frac{1}{n} \Big| \sum_{i=1}^{n} r_i (f - f_N)(Z_i) \Big| &\leq \sup_{f\in\mathcal{F}} \sqrt{\frac{1}{n}\sum_{i=1}^{n} (f - f_N)^2(Z_i)} \\ &= \sup_{f\in\mathcal{F}} ||f - f_N||_{L_2(\mu_n)} \leq \epsilon_N. \end{aligned} \tag{3.3}$$

The magnitude of the second term $\mathbb{E}_r \sup_{f_N\in\mathcal{F}_N} |X_{rad}(f_N)|$ 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 ϵ_{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 $f_{k-1} \to f_k$. Using the triangle inequality, we have

$$\sup_{f_N\in\mathcal{F}_N} |X_{rad}(f_N)| \leq \sum_{k=1}^{N} \sup_{f_{k-1}\to f_k} |X_{rad}(f_k) - X_{rad}(f_{k-1})| + |X_{rad}(f_0)| \tag{3.4}$$

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 $|X_{rad}(f_0)|$ vanishes. Taking the ψ_2 norm on both sides and using the triangle inequality again for the ψ_2 norm, we obtain

$$\Big|\Big| \sup_{f_N\in\mathcal{F}_N} |X_{rad}(f_N)| \Big|\Big|_{\psi_2} \leq \sum_{k=1}^{N} \Big|\Big| \sup_{f_{k-1}\to f_k} |X_{rad}(f_k - f_{k-1})| \Big|\Big|_{\psi_2}. \tag{3.5}$$

Since $\mathbb{N}(\epsilon_k, \mathcal{F}, L_2(\mu_n)) \geq \mathbb{N}(\epsilon_{k-1}, \mathcal{F}, L_2(\mu_n))$, the number of choices of the chaining pair $(f_{k-1} \to f_k)$ is bounded by $\mathbb{N}^2(\epsilon_k, \mathcal{F}, L_2(\mu_n))$. Applying Lemma 2.7, for the maximal inequality on each term on the right-hand side of (3.5), we have

$$\Big|\Big| \sup_{f_{k-1}\to f_k} |X_{rad}(f_k - f_{k-1})| \Big|\Big|_{\psi_2} \leq 4\log^{\frac{1}{2}}(\mathbb{N}^2(\epsilon_k, \mathcal{F}, L_2(\mu_n)) + 1)\Big(\sup_{f_{k-1}\to f_k} ||X_{rad}(f_k - f_{k-1})||_{\psi_2}\Big). \tag{3.6}$$

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

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

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

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

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

$$\begin{aligned}\mathbb{E}_r \sup_{f_N\in\mathcal{F}_N} |X_{rad}(f_N)| &\le \Big|\Big| \sup_{f_N\in\mathcal{F}_N} |X_{rad}(f_N)|\Big|\Big|_{\psi_2} \\ &\le Cn^{-1/2}\sum_{k=1}^{N}\epsilon_{k-1}\log^{\frac{1}{2}}\mathbb{N}(\epsilon_k, \mathcal{F}, L_2(\mu_n)).\end{aligned} \tag{3.7}$$

□

In [14], Mendelson found a similar upper bound for the Gaussian average, the details of this chaining technique also can be found in [15].

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

Theorem 3.2 *Assume that for some $\gamma > 1$, $fat_\epsilon(\mathcal{F}) \le \gamma\epsilon^{-p}$ holds for any $\epsilon > 0$, then there exists a constant C_p, which depends only on p, such that for any empirical measure μ_n*

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

Proof Let μ_n be an empirical measure. When $p < 2$, we know the sum on the right-hand side of inequality (3.1) can be bounded using Lemma 2.5 as follows:

$$\begin{aligned} n^{-1/2}\sum_{k=1}^{N}\epsilon_{k-1}\log^{\frac{1}{2}}\mathbb{N}(\epsilon_k, \mathcal{F}, L_2(\mu_n)) &\le n^{-1/2}\int_0^\infty \log^{\frac{1}{2}}\mathbb{N}(\epsilon, \mathcal{F}, L_2(\mu_n))d\epsilon \\ &\le C_p\gamma^{\frac{1}{2}}\log\gamma\ n^{-1/2}. \end{aligned} \tag{3.9}$$

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

$$\begin{aligned} R(\mathcal{F}/\mu_n) &\le C_p n^{-1/2}\log\gamma\sum_{k=1}^{N}\epsilon^{1-\frac{p}{2}}\log\Big(\frac{2}{\epsilon_k}\Big) + 2\epsilon_N \\ &\le C_p n^{-1/2}\gamma^{\frac{1}{2}}\log\gamma\sum_{k=1}^{N}k2^{k(\frac{p}{2}-1)} + 2n^{-\frac{1}{p}}. \end{aligned} \tag{3.10}$$

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

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

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

$$R(\mathcal{F}/\mu_n) \leq \begin{cases} C_p \gamma^{\frac{1}{2}} n^{-1/2} & \text{if } 0 < p < 2, \\ C_2 \gamma^{\frac{1}{2}} n^{-1/2} \log n & \text{if } p = 2, \\ C_p \gamma^{\frac{1}{2}} n^{-1/p} & \text{if } p > 2. \end{cases} \tag{3.11}$$

By taking the expectation of $R(\mathcal{F}/\mu_n)$ in Theorem 3.2 and Theorem 3.3, then apply the Theorem 2.1, we can also get the upper bounds for corresponding μ-estimation error and universal estimation error.

4 Lower bound

In this section, we prove that for some proper underlying distribution μ, 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 [14].

Theorem 4.1 *If* $fat_\epsilon(\mathcal{F}) \geq \gamma \epsilon^{-p}$ *for some* γ, *then there exists a measure* $\mu \in \mathcal{P}$ *and constant c such that*

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

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

$$R(\mathcal{F}/\mu_n) = \frac{1}{n} \mathbb{E}_r \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \sum_{k=1}^{n_i} r_{i,k} f(Z_i) \right| \tag{4.1}$$

$$\geq \frac{1}{2n} \mathbb{E}_r \sup_{f, f' \in \mathcal{F}} \sum_{i=1}^{n} \sum_{k=1}^{n_i} r_{i,k} (f(Z_i) - f'(Z_i)) \tag{4.2}$$

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,

$$\sup_{f, f' \in \mathcal{F}} \sum_{i=1}^{n} \sum_{k=1}^{n_i} r_{i,k} (f(Z_i) - f'(Z_i)) \geq \sum_{i=1}^{n} \sum_{k=1}^{n_i} (r_{i,k} (f_A(Z_i) - f_{A^c}(Z_i))). \tag{4.3}$$

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

$$R(\mathcal{F}/\mu_n) \geq \frac{1}{2n} \mathbb{E}_r \sum_{i=1}^{n} \sum_{k=1}^{n_i} (r_{i,k} (f_A(Z_i) - f'_{A^c}(Z_i))) \geq \frac{1}{2n} \epsilon n^*. \tag{4.4}$$

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

$$\mathbb{E}_\mu R(\mathcal{F}/\mu_n) \geq \mathbb{E}_\mu \left(\frac{1}{2n} \epsilon n^* \right), \tag{4.5}$$

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

$$\mathbb{E}_\mu(n^*) = n \left(1 - \left(\frac{n-1}{n} \right)^n \right) > \left(1 - \frac{1}{e} \right) n. \tag{4.6}$$

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

$$\mathbb{E}_\mu R(\mathcal{F}/\mu_n) \geq \left(\frac{1}{2} - \frac{1}{2e} \right) \gamma^{1/p} n^{-\frac{1}{p}}. \tag{4.7}$$

Addition with Theorem 2.1, For $p \geq 2$, we know there also exists a constant c_1 such that

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

□

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 2.1, 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 $\mathcal{F}$ is the Donsker class [10]. As long as the limit of the empirical process is non-trivial, the rate $O(n^{-1/2})$ is optimal.

5 Excess loss class or hypothesis class

It seems a little bit obscure to study the excess loss class $\mathcal{F}$ 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

$$|L(\hat{y}_1, y) - L(\hat{y}_2, y)| \leq K|\hat{y}_1 - \hat{y}_2|. \tag{5.1}$$

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:

$$R(\mathcal{F}/\mu_n) \leq KR(\mathcal{H}/\mu_n). \tag{5.2}$$

Thus we know that the Rademacher average of $\mathcal{H}$ can bound 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 [14].

Lemma 5.1 *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 μ, we have*

$$\log \mathbb{N}(\epsilon, |\mathcal{H} - g|^q, L_2(\mu)) \leq \log \mathbb{N}(C_q \epsilon, \mathcal{H}, L_2(\mu)). \tag{5.3}$$

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

Lemma 5.2 *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

$$\log \mathbb{N}(c\epsilon, \mathcal{H}, L_2(\mu)) \leq \log \mathbb{N}(\epsilon, (\mathcal{H} - g)^2, L_2(\mu)). \tag{5.4}$$

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

$$\mathbb{N}(\epsilon, \mathcal{H}, L_2(\mu_n)) = \mathbb{N}(\epsilon, \mathcal{H} - g, L_2(\mu_n)). \tag{5.5}$$

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 $\mathcal{H}^*$:

$$\mathbb{N}(4\epsilon, \mathcal{H}, L_2(\mu_n)) \leq \mathbb{N}(\epsilon, \mathcal{H}^*, L_2(\mu_n)). \tag{5.6}$$

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.

6 Application

6.1 VC classes for classification

We consider the binary classification problem. Assume $\mathcal{F}$ has finite VC dimension V. Then 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 [7] for more details.

From the definition of VC dimension, we know that $\text{fat}_\epsilon(\mathcal{F}) = V$ for $\epsilon < 1$. In this case, we can set γ to be V and p to be 1. Under this setting, from Theorem 3.2, 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 and only a logarithm factor of V worse than the best bound.

Remark 6.1 Faster rates can be achieved under some margin assumptions for the distribution of μ, see [12].

6.2 Regularized linear class

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 $1 \leq p \leq 2$. Consider the following linear function hypothesis space $\mathcal{H}_p$ containing all the functions in the form of $W \cdot X$. In [19], Zhang derived the following bound:

$$\log \mathbb{N}(\epsilon, \mathcal{H}_p, L_2(\mu_n)) \leq \left\lceil \frac{a^2 b^2}{\epsilon^2} \right\rceil \log(2d + 1). \tag{6.1}$$

He then obtained a bound for the estimation error for classification error. Now we can use his result (6.1) for more general setting, for example, real value problems.

Fix the regularization condition $||W||_p \leq b$ and let $\mathcal{H}_1$ is the hypothesis space for lasso regression and $\mathcal{H}_2$ for ridge regression as following:

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

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

From the Holder inequality, we have $|W \cdot X| \leq 1$ for $W \cdot X \in \mathcal{H}_1, \mathcal{H}_2$. The bound of the entropy together with Theorem 3.3 gives the upper bound of the Rademacher average:

$$R(\mathcal{H}_p/\mu_n) \leq C_2\sqrt{\frac{\log(2d+1)}{n}}\log n. \tag{6.2}$$

where C_2 is the constant from Theorem 3.3. This bound provides a convergence rate bound for regression estimation error.

6.3 Non-decreasing class and bounded variation class

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 non-decreasing 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 [5], Bartlett proved the following theorem:

Theorem 6.2 *For all* $\epsilon \leq V/12$

$$\log \mathbb{N}(\epsilon, \mathcal{H}_2, L_1(\mu)) \leq \frac{13V}{\epsilon}. \tag{6.3}$$

From Lemma 2.4, we know that the Fat-shattering dimension has the bound:

$$fat_\epsilon(\mathcal{H}_2) \leq \frac{128V}{\epsilon}. \tag{6.4}$$

From Theorem 3.2, we know the convergence rate of Rademacher average of $\mathcal{H}_2$ can achieve $O(n^{-1/2})$ and so does $\mathcal{H}_1$.

6.4 Multiple layer neural nets

We will present some evidence to why deep learning works. We make the assumption that the input magnitude of each neuron is bounded and consider the following architecture for the neural net:

$$\Omega = \left\{x \in \mathbb{R}^d : ||X||_\infty \leq B\right\}.$$

Let $\mathcal{H}_0$ be the class of functions on Ω defined by

$$\mathcal{H}_0 = \left\{X = (X^1, X^2 \ldots, X^d) \to X^i : 1 \leq i \leq d\right\}.$$

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

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

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

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

By the definition of Rademacher average, one can show

$$CR(\mathcal{H}/\mu_n) = R(\mathrm{conv}_C(\mathcal{H})/\mu_n). \tag{6.5}$$

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

$$R((\sigma \circ \mathcal{H})/\mu_n) \leq LR(\mathcal{H}/\mu_n). \tag{6.6}$$

Since the number of functions in the space $\mathcal{H}_0$ is d, which is finite, the ϵ-covering number can be bounded by d for any ϵ. 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

$$R(\mathcal{H}_l/\mu_n) \leq C_1 C^l \sqrt{\frac{\log d}{n}}. \tag{6.7}$$

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 [2] of Bartlett:

$$\log \mathbb{N}(\epsilon, \mathcal{H}_l, L_2(\mu_n)) \leq a\left(\frac{b}{\epsilon}\right)^{2l}. \tag{6.8}$$

Here a, b are factors independent of ϵ. From this bound, we can only get the universal estimation error bound in the form of $O(n^{-1/2l})$, which means that the learning 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.

6.5 Boosting

Using simple function class 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. Boosting [9] has proven to be an attractive strategy in their regard both in theory and in practice. In each step t, based on the error, the current function h_{t-1} made, boosting greedily choose a function g_t from the base function space $\mathcal{B}$, multiplied by the learning rate γ_t and added to the current function h_{t-1} to reduce the error h_{t-1} made. We denote by T the total number of steps. Let us consider the following hypothesis space:

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

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

In [16], Schapire et al. have shown that for AdaBoosting, the margin error on training data decreases exponentially fast in T. They also 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 $\mathcal{B}$, defined in the last section:

$$R(\mathcal{H}/\mu_n) = R(\mathrm{conv}_C(\mathcal{B})/\mu_n) = CR(\mathcal{B}/\mu_n). \tag{6.9}$$

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 $C\mathbb{E}_\mu R(\mathcal{B}/\mu_n)$. Since the base function space $\mathcal{B}$ 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.

6.6 Convex functions

This example illustrates the fact that if $\mathcal{H}$ is rich enough, the rate of $O(n^{-1/2})$ cannot be achieved. Consider the hypothesis space $\mathcal{H}$ containing all the real-valued convex functions defined on $[a, b]^d \subset \mathbb{R}^d$, which are uniformly bounded by B and uniformly L-Lipschitz.

In Bronshtein's paper [6], it was proved that for ϵ sufficiently small, the logarithm of the covering number $\mathbb{N}(\epsilon, \mathcal{H}, L_\infty(\mu))$ can be bounded from above and below by a positive constant times $\epsilon^{-d/2}$, here μ is the ordinary Lesbegue measure.

We use both Fat-shattering dimension and entropy in this case. By Lemma 2.4, we have

$$\log\mathbb{N}(\epsilon, \mathcal{H}, L_\infty(\mu)) \geq \sup_{\mu_n} \log\mathbb{N}(\epsilon, \mathcal{H}, L_2(\mu_n)) \geq \mathrm{fat}_{16\epsilon}(\mathcal{H})/8. \tag{6.10}$$

From Theorem 3.3, we conclude that $R(\mathcal{H}/\mu_n)$ is bounded above by $Cn^{-2/d}$ for some constant C.

To bound the associated Rademacher average from below, we use the inequality from lemma 2.5:

$$\begin{aligned} &\mathrm{fat}_\epsilon(\mathcal{F})\log^2\left(\frac{2\mathrm{fat}_{\frac{\epsilon}{8}}(\mathcal{F})}{\epsilon}\right) \geq \sup_{\mu_n} \log \mathbb{N}(\epsilon, \mathcal{H}, L_\infty(\mu_n)) \\ &= \log \mathbb{N}(\epsilon, \mathcal{H}, L_\infty(\mu)) \geq c\epsilon^{-d/2}. \end{aligned} \tag{6.11}$$

By solving this inequality for $\mathrm{fat}_\epsilon(\mathcal{F})$, we conclude that there exists a function $\delta(\epsilon)$ which decreases to 0 as ϵ goes to 0 such that

$$fat_\epsilon(\mathcal{F}) \geq c\epsilon^{-d/2-\delta(\epsilon)}. \tag{6.12}$$

Now apply Theorem 4.1, we can conclude that there exists $\gamma(n)$ which goes to 0 as n goes to 0 such that the Rademacher average is bounded below by $O(n^{-(2/d-\gamma(n))})$.

Note that $\mathcal{H}$ also satisfies the requirement in Lemma 5.2, if we use L_2 norm for the loss function, we know that the universal estimation error has a rate between $O(n^{-(2/d-\gamma(n))})$ and $O(n^{-2/d})$. This shows that the general convex function space in high dimension can be very complex for learning problems.

Author details
[1]Beijing Institute of Big Data Research, Peking University, Beijing, China, [2]Department of Mathematics, Princeton University, Princeton, NJ, USA, [3]PACM, Princeton University, Princeton, NJ, USA.

Acknowledgements
The work presented here is supported in part by the Major Program of NNSFC under grant 91130005, DOE grant DE-SC0009248, and ONR grant N00014-13-1-0338. Dedicated to Professor Bjorn Engquist on occasion of his 70th birthday.

Received: 19 July 2016 Accepted: 28 December 2016
Published online: 10 February 2017

References

1. Alon, N., Ben-David, S., Cesa-Bianchi, N., Haussler, D.: Scale sensitive dimensions, uniform convergence and learnability. J. Assoc. Comput. Math **44**(4), 615–631 (1997)
2. Bartlett, L.: The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. IEEE Trans. Inf. Theory **44**(2), 525–536 (1998)
3. Barlett, L., Mendelson, S.: Rademacher and Gaussian complexities: risk bounds and structural results. J. Mach. Learn. Res. **3**(2002), 463–482 (2011)
4. Bartlett, L., Bousquet, O., Mendelson, S.: Local Rademacher complexities. Ann. Stat. **33**(4), 1497–1537 (2005)
5. Bartlett, L., Kulkarni, R., Posner, E.: Covering number for real valued Function classes. IEEE Trans. Inf. Theory **43**(5) 1721–1724 (1997)
6. Bronshtein, E.M.: ϵ-Entropy for classes of convex functions. Sib. Math. J. **17**, 393–398 (1976)
7. Devorye, L., Lugosi, G.: Lower bounds in pattern recognition and learning. Pattern Recogn. **28**, 1011–1018 (1995)
8. Dudley, R.M., Giné, E., Zinn, J.: Uniform and universal Glivenko-Cantelli classes. J. Theor. Probab. **4**, 485–510 (1991)
9. Freund, Y., Schapire, R.: A short introduction to boosting. J. Jpn. Soc. Artif. Intell. **14**(5), 771–780 (1999)
10. Gine, E.: Empirical processes and applications: an overview. Bernoulli **12**(4), 929–989 (1984)
11. Kosorok, Micheal R.: Introduction to Empirical Processes and Semiparametric Inference. Springer, Berlin (2008)

12. Lugosi, G.: Principles of Nonparametric Learning. In: CISM International Centre for Mechanical Sciences, vol. 434. Springer, Verlag, pp. 1–56 (2002)
13. Lindenstrauss, J., Milman, V.D.: The local theory of normed spaces and its application to convexity. In: Proceedings of 14th Annual Conference Computational Learning Theory, pp. 256–272 (2001)
14. Mendelson, S.: Rademahcer averages and phase transitions in Glivenko–Cantelli classes. IEEE Trans. Inf. Theory **48**(1), 251–263 (2002)
15. The Volume of Convex Bodies and Banach Space Geometry. Cambridge University Press, Cambridge (1989)
16. Schapire, R., Freund, Y., Bartlett, P., Lee, W.: Boosting the margin: a new explanation for the effectiveness of voting methods. In: Proceedings of the Fourteenth International Conference on Machine Learning (1997)
17. Tomczak-Jaegermann, N.: Banach Mazur distance and finite dimensional operator ideals. Pitman monographs and surveys in pure and applied mathematics. Pure and Applied Mathematics, vol. 38, p 395 (1989)
18. Vapnik, V., Chervonenkis, A.: Necessary and sufficient conditions for uniform convergence of the means to mathematical expectations. Theory Prob. Appl. **26**, 532–553 (1971)
19. Zhang, T.: Covering number bounds of certain regularized linear function classes. J. Mach. Learn. Res. **2**(2002), 527–550 (2002)