The Generalization Performance of Learning Machine Based on Phi-mixing Sequence

Bin Zou
Faculty of Mathematics and Computer Science, Hubei University
Wuhan 430062, P. R. China
zoubin0502@hubu.edu.cn

Luoqing Li
Faculty of Mathematics and Computer Science, Hubei University
Wuhan 430062, P. R. China
lilq@hubu.edu.cn

Abstract

The generalization performance is the important property of learning machines. It has been shown previously by Vapnik, Cucker and Smale that, the empirical risks of learning machine based on i.i.d. sequence must uniformly converge to their expected risks as the number of samples approaches infinity. This paper extends the results to the case where the i.i.d. sequence is replaced by phi-mixing sequence. We establish the rate of uniform convergence of learning machine by using Bernstein’s inequality for phi-mixing sequence, and estimate the sample error of learning machine. In the end, we compare these bounds with known results.

1. Introduction

The key property of learning machines is generalization performance. The empirical risks (or empirical errors) must converge to their expected risks (or expected errors) when the number of examples increases. The generalization performance of learning machines has been the topic of ongoing research in recent years [4]. The important theoretical tools for studying the generalization performance of learning machines are the principle of empirical risk minimization (ERM) [12], the stability of learning machines [3] and the leave-one-out error (or cross-validation error) [8]. Vapnik [12] applied Chernoff’s inequalities to obtain exponential bounds on the rate of uniform convergence and relative uniform convergence for i.i.d. sequence. Cucker and Smale [5] considered the least squares error and obtained the bound of the empirical errors based on i.i.d. sequence uniform converge to their expected errors over the compact subset of hypothesis space.

However, independence is a very restrictive concept in several ways [13]. First, it is often an assumption, rather than a deduction on the basis of observations. Second, it is an all or nothing property, in the sense that two sense that two random variables are either independent or they are not— the definition does not permit an intermediate notion of being nearly independent. As a result, many of the proofs based on the assumption that the underlying stochastic sequence is i.i.d. are rather “fragile”. Therefore, Vidyasagar [13] considered the notions of mixing and proved that most of the desirable properties (e.g. PAC property or UCEMUP property) of i.i.d. random sequence are preserved when the underlying sequence is mixing. Nobel and Dembo [11] proved that, if a family of functions has the property that empirical means based on i.i.d. sequence converge uniformly to their values as the number of samples approaches infinity, then the family of functions continues to have the same property if the i.i.d. sequence is replaced by β-mixing sequence. Karandikar and Vidyasagar [10] extended this result to the case where the underlying probability is itself not fixed, but varies over a family of measures. Vidyasagar [13] obtained the rate of uniform convergence of empirical means with α-mixing sequence for a finite family of measurable functions taking values in \([0, F]\) (Theorem 3.5).

In this paper, we obtain the bounds of generalization performance of learning machine based on φ-mixing sequences: The bound on the rate of uniform convergence for learning machine based on φ-mixing sequence and the bound on the sample error of learning machine. The rest of the paper is organized as follows. In section 2, we introduce some notations and main tools. In section 3 we obtain the bound on the rate of uniform convergence for learning machine based on φ-mixing sequence by using Bernstein’s inequality for φ-mixing sequence. We obtain the bound of the sample error of learning machine based on φ-mixing sequence in section 4. In section 5, we compare these results with previous results.

* Supported in part by NSFC under grant 60403011.
2. Preliminaries

We introduce some notations and do some preparations in this section. Let $Z = (z_i)_{i \geq 1}$ be a stationary real-valued sequence with unknown distribution $P$, which implies $z_i, i \geq 1$, all have the same distribution $P$. For the sequence $Z$, let $\sigma_i = \sigma(z_1, z_2, \ldots, z_i)$ denote the $\sigma$-algebra generated by the random variables $z_i, 1 \leq i \leq l$, and similarly let $\sigma_{i+k}^l = \sigma(z_{i+k}, z_{i+k+1}, \ldots)$ denote the $\sigma$-algebra generated by the random variables $z_i, i \geq l + k$. With these definitions, there are several definitions of mixing, but we shall be concerned with only one, namely $\phi$-mixing in this literature [9].

Definition 1. The $\phi$-mixing coefficients of the sequence $Z$ is defined as

$$\phi(k) = \sup_{A \in \sigma_1, B \in \sigma_{i+k}} \{|P(B/A) - P(B)|, l \geq 1 \text{ (1)}$$

and the sequence $Z$ is said to be $\phi$-mixing, or uniformly regular if $\phi(k) \to 0$ as $k \to \infty$.

Let a sample set $S = \{z_1, z_2, \cdots, z_n\}$ drawn from the first $n$ observations of the $\phi$-mixing sequence $Z$. The goal of machine learning from random sampling is to find a function $f^*_S$ that assigns values to objects such that if new objects are given, the function $f^*_S$ will forecast them correctly. Here $\kappa$ is a parameter from the set $\Lambda$. Let

$$R(f^*_S) = \int (f^*_S(z) - B/A) dP, \quad \kappa \in \Lambda \quad \text{ (2)}$$

be the expected error (or expected risk) of function $f^*_S, \kappa \in \Lambda$, where the function $\ell(f^*_S, z)$, which is integrable for any $f^*_S, \kappa \in \Lambda$ and depends on $z$ and $f^*_S$, is called loss function. We will not focus on special cases only, instead, we would like to establish a general framework which includes classification and regression estimation. Throughout the article, we require that $0 \leq \ell(f^*_S, z) \leq M, \kappa \in \Lambda$. Let $Q = \{\ell(f^*_S, z) : \kappa \in \Lambda\}$ be the set of totally bounded functions $\ell(f^*_S, z), \kappa \in \Lambda$ with respect to the sample set $S$. For convenience, we use the notation $\kappa \in \Lambda$ to mean $\ell(f^*_S, z) \in Q$. According to the idea that the quality of the chosen function can be evaluated by the expected error (2), the choice of required function from the set $Q$ is to minimize the expected error (2) based on the sample set $S$ [12]. We can not minimize the expected error (2) directly since the distribution $P$ is unknown. By the principle of ERM, we minimize, instead of the expected error (2), the so-called empirical error (or empirical risk) $R_{\text{emp}}(f^*_S) = \frac{1}{n} \sum_{i=1}^n \ell(f^*_S, z_i), \kappa \in \Lambda$. Let $f^*_S$ be a function minimizing the empirical error $R_{\text{emp}}(f^*_S)$ for all $\kappa \in \Lambda$, and we define $f^*_S$ to be a function minimizing the empirical error $R_{\text{emp}}(f^*_S)$ over $\kappa \in \Lambda$. Let the error in $Q$ of a function $\ell(f^*_S, z) \in Q$ be $R_Q(f^*_S) = R(f^*_S) - R(f^*_S)$. It follows from equality above that $R(f^*_S) = R_Q(f^*_S) + R(f^*_S)$.

The first term $R_Q(f^*_S)$ is called the sample error. The second term in this sum depends on the choice of $Q$, but is independent of sampling, we call it the approximation error. The approximation error should be estimated by the knowledge from approximation theory [5], so we be concerned with only the sample error in the sequel. By the definition of $\phi$-mixing sequence, Collomb [6] obtained the Bernstein’s inequality for $\phi$-mixing process.

Lemma 1. Let $\{\xi_i\}_{i=1}^\infty$ be a sequence of $\phi$-mixing random variables satisfying $E(\xi_i) = 0, \xi_i \leq d, E(\xi_i^2) \leq \eta$, and $E(\xi_i^2) \leq D$. Denote by $\tilde{\phi}(k) = \sum_{i=1}^n \phi(k)$ for each $k \in \mathbb{N}$. Then for any $\epsilon > 0$ and $n \in \mathbb{N}$

$$P\{|\sum_{i=1}^n \xi_i^2 > \epsilon \} \leq 2 \exp\left(-\epsilon n \tilde{\phi}(k) + \frac{3\sqrt{\eta} n \phi(k)}{k} + 6\sqrt{\eta} \psi\right).$$

Here $\psi = D + 4\eta d \tilde{\phi}(k)$, $t$ is a positive real number and $k$ a positive integer less than or equal to $n$ satisfying $tkd \leq \frac{1}{4}$.

3. The Bounds of Uniform Convergence

The study we describe in this section intends to bound the difference between the empirical risks and their expected risks on the set $Q$ based on the $\phi$-mixing sequence $Z$. For any $\epsilon > 0$, our goal is to bound the term

$$P\left\{|\sup_{\kappa \in \Lambda} |R(f^*_S) - R_{\text{emp}}(f^*_S)| > \epsilon \right\}. \quad \text{ (3)}$$

To bound (3), intuition suggests that we might have to regulate the size of $Q$. One measure of the size of a collection of random variables is the covering number and packing numbers [1], entropy numbers, VC-dimension for indicator functions [7], $V_{\text{dim}}$-dimension (or $P_{\text{dim}}$-dimension) for real-valued functions [1, 2]. It has been shown in [2] that VC-dimension is not suitable for real-valued function classes. As for the $V_{\text{dim}}$-dimension or $P_{\text{dim}}$-dimension, though their finiteness is sufficient and necessary for a function class to be a uniform Glivenko-Cantelli set [1], no satisfactory relationship has been found between them and the covering numbers in order to derive sharp estimates. So we introduce the covering numbers of function set.

Let $\{M, d\}$ be a pseudo-metric space and $A \subset M$ a subset. For every $\epsilon > 0$, the covering number of $A$ by balls of radius $\epsilon$ with respect to $d$ is defined as the minimal number of balls of radius $\epsilon$ whose union covers $A$, that is $N(A, \epsilon, d) = \min \{k \in \mathbb{N} : \exists \{s_j\}_{j=1}^k \subset M$ such that $A \subset \bigcup_{j=1}^k B(s_j, \epsilon)\}$, where $B(s_j, \epsilon) = \{s \in M : d(s, s_j) \leq \epsilon\}$ is the ball in $M$. Let $d_p$ denote the normalized $l^p$-metric on the Euclidean space $\mathbb{R}^n$ given by

$$d_p(a, b) = \left(\frac{1}{n} \sum_{i=1}^n |a_i - b_i|^p\right)^{\frac{1}{p}}, \quad a = (a_i)_{i=1}^n, b = (b_i)_{i=1}^n.
and let $Q(x) = \{l(f^x_S, z_i) \leq \ell(f^x_S, z) \in Q\}$. For $1 \leq p \leq \infty$, we denote $N_p(Q, \varepsilon) = \sup_{z \in \mathbb{E}} \sup_{x \in X} N(Q, \varepsilon, d_p)$. Because the function set $Q$ is totally bounded, the covering number $N_p(Q, \varepsilon)$ is finite for a fixed $\varepsilon > 0$, then we have the following lemma and theorem.

**Theorem 1.** Let $L_S(f^x_S) = R(f^x_S) - R_{emp}(f^x_S)$, $\kappa \in \Lambda$. Assume that for all $i \in \{1, 2, \ldots, n\}$ and any $\kappa \in \Lambda$, $|\ell(f^x_S, z_i) - E(\ell(f^x_S, z_i))| \leq B$. Then for any $\varepsilon > 0$,

$$P(\{|L_S(f^x_S)| > \varepsilon\} \leq 2 \exp\left(-\frac{-n\varepsilon^2 C_1}{(6B^2 + 4kM\varepsilon)^2}\right),$$

where $\tilde{\phi}(k) = \sum_{i=1}^{n} \phi(k)$, and

$$C_1 = 4kM\varepsilon - \frac{3\sqrt{v(\phi(k))}}{k\varepsilon^2}(6B^2 + 4kM\varepsilon)^2 - 24M^2 \tilde{\phi}(k).$$

Proof. Let $X_i = |\ell(f^x_S, z_i) - E(\ell(f^x_S, z_i))$, then we have $L_S(f^x_S) = \frac{1}{n} \sum_{i=1}^{n} X_i$. We can get that $|X_i| = |\ell(f^x_S, z_i) - E(\ell(f^x_S, z_i))| \leq M$, and $\sup_{i \in \Lambda} |X_i| \leq M$.

By using Lemma 1, we obtain

$$P(\{|L_S(f^x_S)| > \varepsilon\} \leq 2 \exp\left(-\frac{-n\varepsilon^2 C_1}{(6B^2 + 4kM\varepsilon)^2}\right),$$

where $\nu = B^2 + 4M^2 \tilde{\phi}(k)$. The statement now follows from inequality above by replacing $t$ by $\frac{\varepsilon}{6B^2 + 4kM\varepsilon}$.

**Lemma 2.** Let $Q = S_1 \cup S_2 \cup \cdots \cup S_m$ and let $\Lambda = \Lambda_1 \cup \Lambda_2 \cup \cdots \Lambda_m$. For any $\varepsilon > 0$,

$$P\left(\sum_{\kappa \in \Lambda} |L_S(f^x_S)| \geq \varepsilon\right) \leq \sum_{j=1}^{m} P\left(\sup_{\kappa \in \Lambda_j} |L_S(f^x_S)| \geq \varepsilon\right).$$

Proof. For any $1 \leq i \leq m$, we denote $\ell(f^x_S, z) \in S_i$ by $\kappa \in \Lambda_i$. If $\sup_{\kappa \in \Lambda_i} |L_S(f^x_S)| \geq \varepsilon$, then there exists $j, 1 \leq j \leq m$, such that $\sup_{\kappa \in \Lambda_j} |L_S(f^x_S)| \geq \varepsilon$. Lemma 2 follows from the inequality above and the fact that the probability of a union of events is bounded by the sum of the probabilities of these events.

**Theorem 2.** With all notation as in Theorem 1, for any $\varepsilon > 0$, we have

$$P\left(\sup_{\kappa \in \Lambda} |L_S(f^x_S)| \geq \varepsilon\right) \leq 2N_p(Q, \varepsilon) \exp\left(-\frac{-n\varepsilon^2 C_2}{4\tau^2}\right),$$

where $C_2 = 2kM\varepsilon - \frac{12\sqrt{v(\phi(k))}}{k\varepsilon^2}(6B^2 + 2kM\varepsilon)^2 - 24M^2 \tilde{\phi}(k)$, $\tilde{\phi}(k) = \sum_{i=1}^{n} \phi(k), \tau = B^2 + 2kM\varepsilon$.

Proof. Define $m = N_p(Q, \varepsilon/4)$ and consider $\ell(f^x_S, z), \ell(f^x_S, z), \cdots, \ell(f^x_S, z)$ such that the disks $D_j$ centered at $\ell(f^x_S, z)$, $j \in \{1, 2, \ldots, m\}$ and with radius $\frac{\varepsilon}{4}$ cover $Q$. For any $z \in Z$ and all $\ell(f^x_S, z) \in D_j$,

$$R(f^x_S) - R(f^x_S) \leq d_p(\ell(f^x_S, z), \ell(f^x_S, z)) \leq d_p(\ell(f^x_S, z), \ell(f^x_S, z)) \leq d_p(\ell(f^x_S, z), \ell(f^x_S, z)).$$

4. **The bounds of Sample Error**

According to the principle of ERM, we shall consider the function $f^x_S$ as an approximation to the function $f^x_S$. However, how good can we expect $f^x_S$ to be as an approximation of $f^x_S$? In this section Theorem 3 gives an answer.

**Lemma 3.** With all notation in Theorem 1. Let $\varepsilon > 0$ and $0 < \delta < 1$ such that $P\left(\sup_{\kappa \in \Lambda} |L_S(f^x_S)| \geq \varepsilon\right) \geq 1 - \delta$. Then $P\left(R_Q(f^x_S) \leq 2e\right) \geq 1 - \delta$.

**Proof.** By the hypothesis of Lemma 3 we have, with probability at least $1 - \delta$, $R(f^x_S) \leq R_{emp}(f^x_S) + \varepsilon$, and $R_{emp}(f^x_S) \leq R(f^x_S) + \varepsilon$. Moreover, since $f^x_S$ minimizes $R_{emp}(f^x_S)$ on functions set $Q$, we have $R_{emp}(f^x_S) \leq R(f^x_S)$. Then with probability at least $1 - \delta$

$$R(f^x_S) \leq R_{emp}(f^x_S) + \varepsilon \leq R(f^x_S) + 2\varepsilon.$$.  


Thus $R_Q(f_S^{n*}) = R(f_S^{n*}) - R(f_S^{no}) \leq 2\epsilon$. The statement now follows from inequality above.

Replacing $\epsilon$ by $2\epsilon$ in Lemma 3, and using Theorem 2, we obtain the following Theorem on the sample error.

**Theorem 3.** Let $L_S(f_S^{n*}) = R(f_S^{n*}) - R_{emp}(f_S^{n*})$, $\kappa \in \Lambda$, assume that for all $i \in \{1, 2, \ldots, n\}$ and any $\kappa \in \Lambda$, $|L(f_S^{n*}, z_i) - E(\ell(f_S^{n*}, z_i))| \leq B$. Then for any $\epsilon > 0$,

$$P\{R_Q(f_S^{n*}) > \epsilon\} < 2N_p(Q, \frac{\epsilon}{8}) \exp\left(\frac{-n\epsilon^2C_3}{16(6B^2 + kM\epsilon)^2}\right).$$

Here $\tilde{\phi}(k) = \sum_{i=1}^{n} \phi(k), C_3 = kM\epsilon - \frac{48\sqrt{\epsilon} \tilde{\phi}(k)}{k \epsilon^2} (6B^2 + kM\epsilon)^2 - 24M^2 \tilde{\phi}(k)$.

**Remark 3.** Using Theorem 3, given $\epsilon, \delta > 0$, to have $P\{R_Q(f_S^{n*}) \leq \epsilon\} \geq 1 - \delta$, it is sufficient that the number $n$ of samples satisfies $n \geq \frac{16(6B^2 + kM\epsilon)^2}{\epsilon^2 C_3^2} \ln \left(\frac{2N_p(Q, \frac{\epsilon}{8})}{\delta}\right)$.

To prove this, take $\delta = 2N_p(Q, \frac{\epsilon}{8}) \exp\left(\frac{-n\epsilon^2C_3}{16(6B^2 + kM\epsilon)^2}\right)$ and solve for $n$. In particular, if $Z$ is an i.i.d. sequence, we have the following corollary.

**Corollary 2.** With all notation as in Theorem 3, and assume that $Z$ is an i.i.d. sequence. Then for any $\epsilon > 0$,

$$P\{R_Q(f_S^{n*}) > \epsilon\} < 2N_p(Q, \frac{\epsilon}{8}) \exp\left(\frac{-nkM\epsilon^3}{16(6B^2 + kM\epsilon)^2}\right).$$

**5. Final Remarks**

**Remark 4.** The bounds in Theorems 2 and 3 describe the generalization ability of learning machine based on $\phi$-mixing sequence. The bound in Theorem 2 evaluates the risk for the chosen function, and the bound in Theorem 3 evaluates how close $f_S^{n*}$ to be as an approximation of $f_S^{no}$ for a given set of functions $Q$.

**Remark 5.** Theorems 2 and 3 differ from those of [12] and [5]. In [12] and [5], the difference in (3) is bounded based on i.i.d. random sequence. In the paper, the results of [12] and [5] are extended to the case where the sequence is not i.i.d., but $\phi$-mixing sequence. On the other hand, Vapnik’s results depend on the capacity of a set of loss functions, the VC-dimension, while Theorems 2 and 3 depend on the covering number of the set of loss functions. Theorem B and Theorem C in [5] are dependent on the covering number of hypothesis space $\mathcal{H}$.

**Remark 6.** If loss function $\ell(f_S^{n*}, z)$, $\kappa \in \Lambda$, is the least squares error and $p = \infty$ in the definition $N_p(Q, \epsilon)$, Theorems 2 and 3 can be regarded to be the extension of Theorem B and Theorem C in [5], interested readers are referred to that paper for details.

**References**


