Mach Learn (2009) 75: 275–295
DOI 10.1007/s10994-009-5104-z

The generalization performance of ERM algorithm with strongly mixing observations

Bin Zou · Luoqing Li · Zongben Xu

Received: 18 May 2008 / Revised: 23 December 2008 / Accepted: 8 January 2009 /
Published online: 7 February 2009

Abstract The generalization performance is the main concern of machine learning theoretical research. The previous main bounds describing the generalization ability of the Empirical Risk Minimization (ERM) algorithm are based on independent and identically distributed (i.i.d.) samples. In order to study the generalization performance of the ERM algorithm with dependent observations, we first establish the exponential bound on the rate of relative uniform convergence of the ERM algorithm with exponentially strongly mixing observations, and then we obtain the generalization bounds and prove that the ERM algorithm with exponentially strongly mixing observations is consistent. The main results obtained in this paper not only extend the previously known results for i.i.d. observations to the case of exponentially strongly mixing observations, but also improve the previous results for strongly mixing samples. Because the ERM algorithm is usually very time-consuming and overfitting may happen when the complexity of the hypothesis space is high, as an application of our main results we also explore a new strategy to implement the ERM algorithm in high complexity hypothesis space.

Keywords Generalization performance · ERM principle · Relative uniform convergence · Exponentially strongly mixing

Editor: Nicolo Cesa-Bianchi.

Supported by National 973 project (2007CB311002), NSFC key project (70501030), NSFC project (10771053) and Foundation of Hubei Educational Committee (Q200710001).

B. Zou · Z. Xu (✉)
Institute for Information and System Science, Faculty of Science, Xi'an Jiaotong University, Xi'an, 710049, People's Republic of China
e-mail: zbxu@mail.xjtu.edu.cn

B. Zou
e-mail: zoubin0502@hubu.edu.cn

B. Zou · L. Li
Faculty of Mathematics and Computer Science, Hubei University, Wuhan, 430062, People's Republic of China

L. Li
e-mail: lilq@hubu.edu.cn

1 Introduction

Recently there has been a great increase in the interest for theoretical issues in the machine learning community, which is mainly due to the fact that statistical learning theory has demonstrated its usefulness by providing the ground for developing successful and well-founded learning algorithms such as Support Vector Machines (SVMs) (Vapnik 1998). This renewed interest for theory naturally boosted the development of performance bounds for learning machines (see e.g. Bartlett and Long 1998; Bousquet 2003; Cesa-Bianchi et al. 2004; Cucker and Zhou 2007; Lugosi and Pawlak 1994; Smale and Zhou 2003, 2004; Wu and Zhou 2005; Zhou 2003 and references therein). In order to measure the generalization ability of the empirical risk minimization algorithm with i.i.d. observations, Vapnik (1998) first established the bound on the rate of uniform convergence and that on the rate of relative uniform convergence for i.i.d. observations respectively. Bousquet (2003) obtained a generalization of Vapnik and Chervonenkis' bounds by using a new measure of the size of function classes, local Rademacher average (Bartlett and Mendelson 2002). Cucker and Smale (2002a) considered the least squares error and decomposed the error (or generalization error) into two parts: the sample error and the approximation error, and then they bounded the sample error and the approximation error based on i.i.d. observations respectively for a compact hypothesis space. Chen et al. (2004) obtained the bound on the excess expected risk for pattern recognition with i.i.d. observations by introducing a projection operator.

However, independence is a very restrictive concept in several ways (Vidyasagar 2002). 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 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". The notion of mixing allows one to put the notion of "near independence" on a firm mathematical foundation, and moreover, permits one to derive a robust rather than a "fragile" theory. In addition, the i.i.d. assumption can not be strictly justified in real-world problems, for example, many machine learning applications such as market prediction, system diagnosis, and speech recognition are inherently temporal in nature, and consequently not i.i.d. processes (Steinwart et al. 2006). Therefore, relaxations of such i.i.d. assumption have been considered for quite a while in both machine learning and statistics literatures. For example, Yu (1994) established the rates on the uniform convergence of the empirical means to their means for stationary mixing sequences. White (1989) considered cross-validated regression estimators for strongly mixing processes and established convergence, without rates, of their estimators. Modha and Masry (1996) established the minimum complexity regression estimation with m-dependent observations and strongly mixing observations respectively. Vidyasagar (2002) considered several notions of mixing (e.g. α-mixing, β-mixing and φ-mixing) and proved that most of the desirable properties (e.g. PAC property or UCEMUP property) of i.i.d. sequences are preserved when the underlying sequence is mixing sequence. Nobel and Dembo (1993) proved that, if a family of functions has the property that the empirical means, based on i.i.d. sequences, converge uniformly to their expected 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 (2002) extended this result to the case where the underlying probability is itself not fixed, but varies over a family of measures. Vidyasagar (2002) obtained the bound on the rate of uniform convergence of the empirical means to their means for mixing sequences. Steinwart et al. (2006) proved that the SVMs

for both classification and regression are consistent if the data-generating process (e.g. mixing process, Markov process) satisfies a certain type of law of large numbers (e.g. WLLNE, SLLNE). Zou and Li (2007) established the bound on the rate of uniform convergence of learning machines with exponentially strongly mixing observations.

To extend the previous bounds in Bousquet (2003), Cucker and Smale (2002a), Vapnik (1998) on the rate of relative uniform convergence to the case where the i.i.d. observations are replaced by exponentially strongly mixing observations, and to improve the results in Vidyasagar (2002), Zou and Li (2007) based on strongly mixing sequences, in this paper we first establish the bound on the rate of relative uniform convergence of the ERM algorithm with exponentially strongly mixing samples, and then we obtain the generalization bounds of the ERM algorithm with exponentially strongly mixing samples. Because when the complexity of the given function set is high, the problem of solving ERM algorithm is usually very time-consuming and overfitting may happen, as an application of our main results we also explore a new method to solve the problem of ERM learning with exponentially strongly mixing samples.

The rest of this paper is organized as follows: In Sect. 2, we introduce some notions and notations. In Sect. 3, we present the main results of this paper. In Sect. 4 we establish the bound on the rate of relative uniform convergence of the ERM algorithm with exponentially strongly mixing observations. We prove the generalization bounds of the ERM algorithm with exponentially strongly mixing sequences in Sect. 5. Finally, we conclude the paper with some useful remarks in Sect. 6.

2 Preliminaries

In this section we introduce the definitions and notations used throughout the paper.

Let $\mathcal{Z} = \{z_i = (x_i, y_i)\}_{i=-\infty}^{\infty}$ be a stationary real-valued sequence on a probability space $(\Omega, \mathcal{B}, P)$. For $-\infty < i < \infty$, let σ_i^{∞} and $\sigma_{-\infty}^{i}$ denote the σ-algebra events generated by the random variables z_j, $j \geq i$ and z_j, $j \leq i$ respectively. With these notations, there are several definitions of mixing, but we shall be concerned with only one, namely, α-mixing in this literature (see Ibragimov and Linnik 1971; Modha and Masry 1996; Rosenblatt 1956; Vidyasagar 2002; Yu 1994).

Definition 1 (Vidyasagar 2002) The sequence $\mathcal{Z}$ is called α-mixing, or strongly mixing (or strongly regular), if

$$\sup_{A \in \sigma_{-\infty}^{0}, B \in \sigma_{k}^{\infty}} \{|P(A \cap B) - P(A)P(B)|\} = \alpha(k) \to 0 \quad \text{as } k \to \infty.$$

Here $\alpha(k)$ is called the α-mixing coefficient.

Assumption 1 (Exponentially strongly mixing) (Modha and Masry 1996) *Assume that the α-mixing coefficient of the sequence $\mathcal{Z}$ satisfies*

$$\alpha(k) \leq \overline{\alpha} \exp(-ck^{\beta}), \quad k \geq 1,$$

for some $\overline{\alpha} > 0$, $\beta > 0$, and $c > 0$, where the constants β and c are assumed to be known.

Remark 1 (Modha and Masry 1996) Assumption 1 is satisfied by a large class of processes, for example, certain linear processes (which includes certain ARMA processes) satisfy

the assumption with $\beta = 1$ (Withers 1981), and certain aperiodic, Harris-recurrent Markov processes (which includes certain bilinear processes, nonlinear ARX processes, and ARH processes) satisfy the assumption (Davydov 1973). As a trivial example, i.i.d. random variables satisfy the assumption with $\beta = \infty$.

Denote by $\mathbf{z}$ the sample set of size n observations

$$\mathbf{z} = \{z_1, z_2, \ldots, z_n\}$$

drawn from the exponentially strongly mixing sequence $\mathcal{Z}$. Set

$$n^{(\alpha)} = \lfloor n \lceil \{8n/c\}^{1/(\beta+1)} \rceil^{-1} \rfloor,$$

where n denotes the number of observations drawn from $\mathcal{Z}$ and $\lfloor u \rfloor (\lceil u \rceil)$ denotes the greatest (least) integer less (greater) than or equal to u.

The goal of machine learning from random sampling is to find a function f that assigns values to objects such that if new objects are given, the function f will forecast them correctly. Let

$$\mathcal{E}(f) = \mathrm{E}[\ell(f, z)] = \int \ell(f, z) dP$$

be the expected risk (or expected error) of function f, where the function $\ell(f, z)$, which is integrable for any f and depends on f and z, called loss function. In this paper, we would like to establish a general framework which includes pattern recognition and regression estimation, so we consider the loss function of general form $\ell(f, z)$. The important feature of the regression estimation problem is that the loss function $\ell(f, z)$ can take arbitrary nonnegative values whereas in pattern recognition problem it can take only two values.

A learning task is to find the minimizer of the expected risk $\mathcal{E}(f)$ over a given hypothesis space $\mathcal{H}$. Since one knows only the set $\mathbf{z}$ of random samples instead of the distribution P, the minimizer of the expected risk $\mathcal{E}(f)$ can not be computed directly. According to the principle of Empirical Risk Minimizing (ERM) (Vapnik 1998), we minimize, instead of the expected risk $\mathcal{E}(f)$, the so called empirical risk (or empirical error)

$$\mathcal{E}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f, z_i).$$

Let $f_{\mathcal{H}}$ be a function minimizing the expected risk $\mathcal{E}(f)$ over $f \in \mathcal{H}$, i.e.,

$$f_{\mathcal{H}} = \arg\min_{f \in \mathcal{H}} \mathcal{E}(f) = \arg\min_{f \in \mathcal{H}} \int \ell(f, z) dP.$$

We define the empirical target function $f_{\mathbf{z}}$ to be a function minimizing the empirical risk $\mathcal{E}_n(f)$ over $f \in \mathcal{H}$, i.e.,

$$f_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}} \mathcal{E}_n(f) = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(f, z_i). \tag{1}$$

According to the principle of ERM, we shall consider the function $f_{\mathbf{z}}$ as an approximation function of the target function $f_{\mathcal{H}}$. Thus a central question of ERM learning is how well

$f_{\mathbf{z}}$ really approximate $f_{\mathcal{H}}$. If this approximation is good, then the ERM algorithm is said to generalize well. A ERM algorithm with generalization capability implies that although it is found via minimizing the empirical risk $\mathcal{E}_n(f)$, it can eventually predict as well as the optimal predictor $f_{\mathcal{H}}$. To characterize the generalization capability of a learning algorithm requires in essence to decipher how close $f_{\mathbf{z}}$ is from $f_{\mathcal{H}}$. This is a very difficult issue in general (Vapnik 1998). In the framework of statistical learning, however, this is then relaxed to considering how close the expected risk $\mathcal{E}(f_{\mathbf{z}})$ is from $\mathcal{E}(f_{\mathcal{H}})$, or equivalently, how small can we expect the difference $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}})$ to be. We call the refined upper bound estimations on $\mathcal{E}(f_{\mathbf{z}})$ or on the deviation between $\mathcal{E}(f_{\mathbf{z}})$ and $\mathcal{E}(f_{\mathcal{H}})$ the generalization bounds of the ERM algorithm.

Since $f_{\mathbf{z}}$ is dependent on the sample set $\mathbf{z}$, in other words, the minimization (1) is taken over the discrete quantity $\mathcal{E}_n(f)$, intuitively, we have to estimate the capacity of the function set $\mathcal{H}$. It has been shown that VC-dimension is not suitable for real-valued function classes (Evgeniou and Pontil 1999). As for the V_γ-dimension or P_γ-dimension, though their finiteness is sufficient and necessary for a function class to be a uniform Glivenko-Cantelli (Alon et al. 1997), no satisfactory relationship has been found between them and the covering numbers in order to derive sharp estimates. So the capacity of the function set $\mathcal{H}$ is measured by the covering number in this paper.

Definition 2 (Cucker and Smale 2002a) For a subset $\mathcal{F}$ of a metric space and $\varepsilon > 0$, the covering number $\mathcal{N}(\mathcal{F}, \varepsilon)$ of the function set $\mathcal{F}$ is the minimal integer $b \in \mathbf{N}$ such that there exist b disks with radius ε covering $\mathcal{F}$.

To estimate the generalization ability of the ERM algorithm (1) with exponentially strongly mixing samples, we give some basic assumptions on the hypothesis space $\mathcal{H}$ and the loss function $\ell(f, z)$:

(i) Assumption on the hypothesis space: We suppose that $\mathcal{H}$ is contained in a ball of a Hölder space $\mathcal{C}^p$ on a compact subset of a Euclidean space $\mathbf{R}^d$ for some $p > 0$, that is,

$$\mathcal{H} = \{f \in B_R(\mathcal{C}^p) : r < \mathcal{E}(f) \leq s\},$$

where R is the radius of ball $B_R(\mathcal{C}^p)$.

(ii) Assumption on the loss function: We define

$$M = \sup_{f \in \mathcal{H}} \max_{z \in \mathcal{Z}} |\ell(f, z)|$$

and

$$L = \sup_{g_1, g_2 \in \mathcal{H}, g_1 \neq g_2} \max_{z \in \mathcal{Z}} \frac{|\ell(g_1, z) - \ell(g_2, z)|}{|g_1 - g_2|}.$$

We assume that M and L are finite in this paper.

Because the function set $\mathcal{H}$ is assumed to be compact, the covering number $\mathcal{N}(\mathcal{H}, \varepsilon)$ is finite for a fixed $\varepsilon > 0$. Then there exists constant $C_0 > 0$ such that (Zhou 2003)

$$\mathcal{N}(\mathcal{H}, \varepsilon) \leq \exp\{C_0 \varepsilon^{\frac{-2d}{p}}\}. \tag{2}$$

3 Main results

To measure the generalization performance of a learning machine, Bousquet (2003), Cucker and Smale (2002a), Vapnik (1998) obtained the bound on the rate of the empirical risks

uniform convergence to their expected risks in a given set $\mathcal{H}$ (or Q) based on i.i.d. sequences, that is, for any $\varepsilon > 0$, they bounded the term

$$\text{Prob}\left\{\sup_{f\in\mathcal{H}} |\mathcal{E}(f) - \mathcal{E}_n(f)| > \varepsilon\right\}. \tag{3}$$

Vidyasagar (2002) also established the bounds on the term (3) based on β-mixing sequences and α-mixing sequences respectively. Yu (1994) obtained the convergence rates of the term (3) for mixing sequences. Zou and Li (2007) established the bound on the term (3) based on exponentially strongly mixing observations. The interested reader can consult (Zou and Li 2007; Vidyasagar 2002; Yu 1994) for the details. For more inequalities on probabilities of uniform deviations, see, for example, Alexander (1984), Bartlett and Lugosi (1999), Devroye (1982), Pollard (1984), Talagrand (1994).

However, the term (3) fails to capture the phenomenon that for those functions $f \in \mathcal{H}$ for which the expected risk $\mathcal{E}(f)$ is small, the deviation $\mathcal{E}(f) - \mathcal{E}_n(f)$ is also small with large probability (see Bartlett and Lugosi 1999; Bousquet 2003; Vapnik 1998). In order to extend these results in Bousquet (2003), Cucker and Smale (2002a), Vapnik (1998) to the case where the i.i.d. sequence is replaced by α-mixing sequence, and to improve these estimations in Zou and Li (2007), Vidyasagar (2002), our purpose in this paper is to bound the term (for any $\varepsilon > 0$)

$$\text{Prob}\left\{\sup_{f\in\mathcal{H}} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} > \varepsilon\right\} \tag{4}$$

for the ERM algorithm (1) with exponentially strongly mixing samples. Our main results are stated as follows.

Theorem 1 *Let $\mathcal{Z}$ be a stationary α-mixing sequence with the mixing coefficient satisfying Assumption 1, that is*

$$\alpha(k) \le \overline{\alpha}\exp(-ck^{\beta}), \quad k \ge 1,\ \overline{\alpha} > 0,\ \beta > 0,\ c > 0.$$

Set $n^{(\alpha)} = \lfloor n\lceil\{8n/c\}^{1/(\beta+1)}\rceil^{-1}\rfloor$, and assume that the variance $D[\ell(f,z)] \le \sigma^2$ for all $z \in \mathcal{Z}$ and for all functions in $\mathcal{H}$. Then for any ε, $0 < \varepsilon \le 2r$, the inequality

$$\text{Prob}\left\{\sup_{f\in\mathcal{H}} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \ge \varepsilon\right\} \le C\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon r\sqrt{r}}{(s+4r)L}\right)\exp\left\{\frac{-r\tau^2 n^{(\alpha)}}{2(\sigma^2 + \tau\sqrt{s}M/3)}\right\} \tag{5}$$

holds, where $C = 1 + 4e^{-2}\overline{\alpha}$, and $\tau = \frac{r\sqrt{r}}{\sqrt{s}(s+4r)}\varepsilon$.

In particular, if $\mathcal{Z}$ is an i.i.d. sequence, according to Remark 1, we take $\beta = \infty$ in Theorem 1 and ignore the multiplicative constant $1 + 4e^{-2}\overline{\alpha}$. The following bound follows from Theorem 1 immediately.

Corollary 1 *Let $\mathcal{Z}$ be an i.i.d. sequence, and assume that the variance $\mathrm{D}[\ell(f,z)] \le \sigma^2$ for all $z \in \mathcal{Z}$ and for all functions in $\mathcal{H}$. Then for any ε, $0 < \varepsilon \le 2r$, the inequality*

$$\text{Prob}\left\{\sup_{f\in\mathcal{H}} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \ge \varepsilon\right\} \le \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon r\sqrt{r}}{(s+4r)L}\right)\exp\left\{\frac{-r\tau^2 n}{2(\sigma^2 + \tau\sqrt{s}M/3)}\right\}$$

holds.

Remark 2 (i) $n^{(\alpha)}$ arises from the Bernstein inequality (Theorem 4.3) for strongly mixing processes in Modha and Masry (1996) and is called the "effective number of observations" for strongly mixing processes. From Theorem 1 and Corollary 1, we can find that $n^{(\alpha)}$ plays the same role in our analysis as that played by the number n of observations in the i.i.d. case.

(ii) Since $n^{(\alpha)} \to \infty$ as $n \to \infty$, by Theorem 1, we then have that for any $0 < \varepsilon \leq 2r$

$$\text{Prob}\left\{ \sup_{f \in \mathcal{H}} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \geq \varepsilon \right\} \to 0, \quad \text{as } n \to \infty.$$

This shows that as long as the covering number of the hypothesis space $\mathcal{H}$ is finite, the empirical risks $\mathcal{E}_n(f)$ can uniformly converge to their expected risks $\mathcal{E}(f)$, and the convergence speed may be exponential. This assertion is well known for the ERM algorithm with i.i.d. samples (see e.g. Bousquet 2003; Cucker and Smale 2002a; Vapnik 1998). We have generalized these classical results in Bousquet (2003), Cucker and Smale (2002a), Vapnik (1998) to the exponentially strongly mixing sequences.

(iii) Theorem 1 is on the rate of relative uniform convergence of the ERM algorithm (1) with exponentially strongly mixing sequences. As far as we know, this is the first result on this topic. The bound in Theorem 1 usually has smaller confidence interval than that bound on the rate of uniform convergence (this is the reason why Bousquet 2003; Cucker and Smale 2002a; Vapnik 1998 bounded the term (4)).

Theorem 1 will be proven in the next section. Before going into the technical proofs, we first deduce the generalization bounds of the ERM algorithm (1) with exponentially strongly mixing samples.

Proposition 1 *Let $\mathcal{Z}$ be a stationary α-mixing sequence with the mixing coefficient satisfying Assumption 1. Assume that the variance $D[\ell(f,z)] \leq \sigma^2$ for all $z \in \mathcal{Z}$ and for all functions in $\mathcal{H}$. Then for any $\eta \in (0,1]$, the following inequalities hold true provided that*

$$n^{(\alpha)} \geq \max\left\{ \frac{\ln(C/\eta)}{2C_1 r^2}, \frac{C_0[(s+4r)L]^{\frac{2d}{p}}}{C_1 2^{\frac{2d}{p}} r^{\frac{5d+2p}{p}}}, \frac{\ln(C/\eta(\sigma^2 + sM/3))}{2r^2} \right\}.$$

(i) *With probability at least $1-\eta$,*

$$\mathcal{E}(f_{\mathbf{z}}) \leq \mathcal{E}_n(f_{\mathbf{z}}) + \frac{\varepsilon^2(n,\eta)}{2}\left(1 + \sqrt{1 + \frac{4\mathcal{E}_n(f_{\mathbf{z}})}{\varepsilon^2(n,\eta)}}\right). \tag{6}$$

(ii) *With probability at least $1-2\eta$,*

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \leq \varepsilon'(n,\eta) + \frac{\varepsilon^2(n,\eta)}{2}\left(1 + \sqrt{1 + \frac{4\mathcal{E}_n(f_{\mathbf{z}})}{\varepsilon^2(n,\eta)}}\right), \tag{7}$$

where

$$\varepsilon(n,\eta) \leq \max\left\{ \left[\frac{2\ln(C/\eta)}{C_1 n^{(\alpha)}}\right]^{\frac{1}{2}}, \left[\frac{2C_0 r^{\frac{-3d}{p}}[(s+4r)L]^{\frac{2d}{p}}}{C_1 n^{(\alpha)}}\right]^{\frac{p}{2p+2d}} \right\},$$

$$\varepsilon'(n,\eta) = \frac{M\ln(C/\eta)}{3n^{(\alpha)}}\left(1 + \sqrt{1 + \frac{18n^{(\alpha)}\sigma^2}{M^2\ln(C/\eta)}}\right),$$

$$C_1 = \frac{3r^4}{2s(s+4r)[3(s+4r)\sigma^2 + 2r^{\frac{5}{2}}M]}.$$

Remark 3 (i) Since when $n \to \infty$, $n^{(\alpha)} \to \infty$, we have

$$\varepsilon(n,\eta) \to 0, \quad \varepsilon'(n,\eta) \to 0, \quad \text{as } n \to \infty.$$

By inequality (7), we then have

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \to 0, \quad \text{as } n \to \infty.$$

This shows that the ERM algorithm (1) with exponentially strongly mixing observations is consistent whenever the covering number of the target function set $\mathcal{H}$ is finite.

(ii) Bounds (6) and (7) describe the generalization performance of the ERM algorithm (1) with exponentially strongly mixing observations in the given function set $\mathcal{H}$: Bound (6) evaluates the risk for the chosen function in the target function set $\mathcal{H}$, and bound (7) evaluates how close this risk is to the smallest possible risk for the target functions set $\mathcal{H}$.

(iii) Strongly mixing samples usually contain less information than i.i.d. samples, and they therefore might lead to worse learning rates. This property of dependent samples is just what we can expect as reflected in our results.

In addition, from Proposition 1, we can find that if $p \gg d$, the learning rates of the ERM algorithm with exponentially strongly mixing samples are close to or as same as those for learning rate with i.i.d. samples.

The ERM algorithm is known to be a classical learning algorithm in statistical learning theory (Vapnik 1998). However, when the complexity of the given function set $\mathcal{H}$ is high, the ERM algorithm (1) is usually very time-consuming and overfitting may happen (see Wu and Zhou 2005). Thus, regularization techniques are frequently adopted. Two kinds of regularization methods are the most interesting: the Tikhonov regularization and the Ivanov regularization. The interested reader can consult Wu and Zhou (2005) for the details. As an application of Proposition 1, in this paper we also explore a new method to solve the time-consuming problem of the ERM algorithm (1) by following the enlightening idea of Giné and Koltchinski (2006). We simply state our ideas as follows: first, we decompose the given target function set $\mathcal{H}$ into different disjoint compact subsets such that the complexities of all subsets are small. To be more precise, for the given $r, s, r < s$, we take $q > 1$ and $a \in \mathcal{N}$ such that $s = rq^a$, and

$$a = \log_q\left(\frac{s}{r}\right).$$

Let $\rho_i = rq^i, i = 0, 1, \ldots, a$ (with $\rho_0 = r, \rho_a = s$). We set

$$\mathcal{H}(\rho_{i-1}) = \{f \in \mathcal{H} : \mathcal{E}(f) \le \rho_{i-1}\},$$

and

$$\mathcal{H}(\rho_{i-1}, \rho_i] = \mathcal{H}(\rho_i) \backslash \mathcal{H}(\rho_{i-1}).$$

Then we have

$$\mathcal{H} = \bigcup_{i=1}^{a} \mathcal{H}(\rho_{i-1}, \rho_i],$$

where a is finite because $\mathcal{H}$ is assumed to be compact. Second, for a given function subset $\mathcal{H}(\rho_{i-1}, \rho_i]$, $i \in \{1, 2, \ldots, a\}$, by the ERM algorithm (1), we can obtain the corresponding empirical target function $f_{\mathbf{z}}^i$, and then we can obtain the upper bound of their risk $\mathcal{E}(f_{\mathbf{z}}^i)$ by the same argument conducted in Proposition 1. Thus we choose the minimizer of these upper bounds of the risks $\mathcal{E}(f_{\mathbf{z}}^i)$, $i \in \{1, 2, \ldots, a\}$ as the risk of the chosen function in the hypothesis space $\mathcal{H}$. We can obtain the following proposition.

Proposition 2 *With all notations as in Proposition 1, let $f_{\mathbf{z}}^i$, $i \in \{1, 2, \ldots, a\}$ be the function minimizing the empirical risk $\mathcal{E}_n(f)$ over $f \in \mathcal{H}(\rho_{i-1}, \rho_i]$. Then for any $\eta \in (0, 1]$, the following inequalities hold true provided that*

$$n^{(\alpha)} \geq \max\left\{\frac{\ln(C/\eta)}{2C_1 r^2}, \frac{C_0[(s+4r)L]^{\frac{2d}{p}}}{C_1 2^{\frac{2d}{p}} r^{\frac{5d+2p}{p}}}, \frac{\ln(C/\eta(\sigma^2 + sM/3))}{2r^2}\right\}.$$

(i) *With probability at least $1-\eta$,*

$$\mathcal{E}(f_{\mathbf{z}}) \leq \min_{1\leq i\leq a}\left\{\mathcal{E}_n(f_{\mathbf{z}}^i) + \frac{\varepsilon_i^2(n,\eta)}{2}\left(1 + \sqrt{1 + \frac{4\mathcal{E}_n(f_{\mathbf{z}}^i)}{\varepsilon_i^2(n,\eta)}}\right)\right\}.$$

(ii) *With probability at least $1-2\eta$,*

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \leq \varepsilon'(n,\eta) + \min_{1\leq i\leq a}\left\{\frac{\varepsilon_i^2(n,\eta)}{2}\left(1 + \sqrt{1 + \frac{4\mathcal{E}_n(f_{\mathbf{z}}^i)}{\varepsilon_i^2(n,\eta)}}\right)\right\},$$

where C_1 and $\varepsilon'(n,\eta)$ are defined as in Proposition 1, C_i is defined as at the end of Sect. 5, and

$$\varepsilon_i(n,\eta) \leq \max\left\{\left[\frac{2\ln(C/\eta)}{C_i n^{(\alpha)}}\right]^{\frac{1}{2}}, \left[\frac{2C_0(\rho_{i-1})^{\frac{-3d}{p}}[(\rho_i + 4\rho_{i-1})L]^{\frac{2d}{p}}}{C_i n^{(\alpha)}}\right]^{\frac{p}{2p+2d}}\right\}.$$

Propositions 1 and 2 will be proven in the next section. Before going into the technical proofs, in order to have a better understanding of the significance and value of the established results in this paper, we compare our results with the previously known results in Vidyasagar (2002). Therefore, we first give the equivalent form of inequality (7) in Proposition 1 as follows.

By Theorem 1, we have that for any $0 < \varepsilon < 2r$, and for the function $f_{\mathbf{z}}$ that minimizes the empirical risk $\mathcal{E}_n(f)$ over $\mathcal{H}$, the inequality

$$\text{Prob}\left\{\frac{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_n(f_{\mathbf{z}})}{\sqrt{\mathcal{E}(f_{\mathbf{z}})}} \geq \varepsilon\right\} \leq C\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon r\sqrt{r}}{(s+4r)L}\right)\exp\left\{\frac{-r\tau^2 n^{(\alpha)}}{2(\sigma^2 + \tau\sqrt{s}M/3)}\right\}$$

is valid, where $C = 1 + 4e^{-2}\overline{\alpha}$, and $\tau = \frac{r\sqrt{r}}{\sqrt{s}(s+4r)}\varepsilon$.

It follows that for any $0 < \varepsilon < 2r$,

$$\text{Prob}\left\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_n(f_{\mathbf{z}}) \geq \varepsilon\sqrt{s}\right\} \leq C\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon r\sqrt{r}}{(s+4r)L}\right)\exp\left\{\frac{-r\tau^2 n^{(\alpha)}}{2(\sigma^2 + \tau\sqrt{s}M/3)}\right\}. \quad (8)$$

By Theorem 2 in the next section, we also have that for any $\varepsilon > 0$, and for the function $f_{\mathcal{H}}$ that minimizes the expected risk $\mathcal{E}(f)$ over $\mathcal{H}$, the inequality

$$\text{Prob}\left\{\mathcal{E}_n(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}) \geq \varepsilon\sqrt{s}\right\} \leq C\exp\left\{\frac{-r\varepsilon^2 n^{(\alpha)}}{2(\sigma^2 + \varepsilon\sqrt{s}M/3)}\right\} \tag{9}$$

holds true. Note that

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \leq \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_n(f_{\mathbf{z}}) + \mathcal{E}_n(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}). \tag{10}$$

Taking inequality (8) into account from (9) and (10), and replacing ε by $\frac{\varepsilon}{2\sqrt{s}}$, we conclude that for any $0 < \varepsilon < 2r$, the inequality

$$\text{Prob}\left\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \geq \varepsilon\right\} \leq 2C\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon r\sqrt{r}}{2\sqrt{s}(s+4r)L}\right)\exp\left\{\frac{-r\tau'^2 n^{(\alpha)}}{2(\sigma^2 + \tau'\sqrt{s}M/3)}\right\} \tag{11}$$

holds, where $C = 1 + 4e^{-2}\overline{\alpha}$, and $\tau' = \frac{r\sqrt{r}}{2s(s+4r)}\varepsilon$.

Thus we have the following remarks.

Remark 4 Comparing bound (11) with the bound in Theorem 6.12 obtained by Vidyasagar (2002), we can find that although we adopt the same measure of the complexity of function set, the covering number, and our proof techniques have many steps similar to that of Theorem 3.5 in Vidyasagar (2002). The differences between bound (11) and the bound in Theorem 6.12 are obvious: First, the key proof technique and method are different. Vidyasagar (2002) first established the bound (Theorem 3.5) on the empirical means uniform convergence to their true values, then he proved that the minimal empirical risk algorithm based on a function family of finite metric entropy is PAC (see Theorem 6.12 in Vidyasagar 2002). In this paper, we first adopted the sign $n^{(\alpha)}$ introduced by Modha and Masry (1996) to establish a new bound on the relative uniform convergence of the ERM algorithm with exponentially strongly mixing samples, which consists of only one exponential term, and then we obtained the generalization bounds of the ERM algorithm and proved that the ERM algorithm with exponentially strongly mixing samples is consistent.

Second, in Theorem 6.12, Vidyasagar (2002) merely established the bound in the case of $n = kl$, that is, Theorem 6.12 in Vidyasagar (2002) is on the case of the number n of samples can be exactly divisible by integer k. In the case of $n = kl$, comparing the bound in Theorem 6.12 with bound (11), we can find that if k and l satisfy

$$k \leq \left[\frac{\ln(4\overline{\alpha}l)}{c} + \frac{l\varepsilon^2}{8c} + \frac{4l\varepsilon}{c}\right]^{\frac{1}{\beta}},$$

bound (11) has better convergence rate than the bound in Theorem 6.12, otherwise bound (11) has the same convergence rate as that bound in Theorem 6.12.

In addition, since $n^{(\alpha)} = O(n^{\frac{\beta}{\beta+1}})$, the convergence rate of bound (11) is close to or as same as those for convergence rate with i.i.d. samples in Bousquet (2003), Cucker and Smale (2002a), Vapnik (1998).

4 Proof of relative uniform convergence bound

To prove the main results presented in the last section, we first establish a new bound on the relative difference between the empirical risks and their expected risks by using an argument

similar to that used by Modha and Masry (1996) and by Vidyasagar (2002). Our approach is however based on the Bernstein moment condition (see Craig 1933; Modha and Masry 1996) and the covariance inequality for α-mixing sequences in Vidyasagar (2002).

Lemma 1 (Craig 1933) *Let W be a random variable such that $\mathrm{E}(W) = 0$, and W satisfies the Berstein moment condition, that is , for some $K_1 > 0$,*

$$\mathrm{E}|W|^k \leq \frac{\mathrm{Var}(W)}{2} k! K_1^{k-2}$$

for all $k \geq 2$. Then, for all $0 < \xi < 1/K_1$,

$$\mathrm{E}[\exp(\xi W)] \leq \exp\left[\frac{\xi^2 \mathrm{E}|W|^2}{2(1-\xi K_1)}\right].$$

Lemma 2 (Vidyasagar 2002) *Suppose $\mathcal{Z}$ is an α-mixing stochastic process. Suppose $g_0, g_1, \ldots, g_l$ are essentially bounded functions, where g_i depends only on z_{ik}. Then*

$$\left|\mathrm{E}\left[\prod_{i=0}^{l} g_i\right] - \prod_{i=0}^{l} \mathrm{E}(g_i)\right| \leq 4l\alpha(k) \prod_{i=0}^{l} \|g_i\|_\infty.$$

To exploit the α-mixing property, we decompose the index set $I = \{1, 2, \ldots, n\}$ into different parts as follows: Given an integer n, choose any integer $k_n \leq n$ and define $l_n = \lfloor n/k_n \rfloor$ to be the integer part of n/k_n. For the time being, k_n and l_n are denoted respectively by k and l so as to reduce notational clutter. The dependence of k and l on n is restored near the end of the paper.

Let $p = n - kl$ and define the index sets I_i, $i = 1, 2, \ldots, k$ as follows

$$I_i = \begin{cases} \{i, i+k, \ldots, i+lk\} & 1 \leq i \leq p, \\ \{i, i+k, \ldots, i+(l-1)k\} & p+1 \leq i \leq k. \end{cases}$$

Note that $\bigcup_i I_i$ equals the index set $I = \{1, 2, \ldots, n\}$ and that within each set I_i the elements are pairwise separated by at least k. Then we have the following theorem.

Theorem 2 *Let $\mathcal{Z}$ be a stationary α-mixing sequence with the mixing coefficient satisfying Assumption 1. Assume that the variance $D[\ell(f, z)] \leq \sigma^2$ for all $z \in \mathcal{Z}$ and for all functions in $\mathcal{H}$. Then for all $\varepsilon > 0$, the inequality*

$$\mathrm{Prob}\left\{\frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} > \varepsilon\right\} \leq (1 + 4e^{-2}\overline{\alpha}) \exp\left\{\frac{-n^{(\alpha)} r \varepsilon^2}{2(\sigma^2 + \varepsilon\sqrt{s} M/3)}\right\} \tag{12}$$

holds.

Proof For any i, $1 \leq i \leq n$, let

$$X_i = \mathrm{E}[\ell(f, z_1)] - \ell(f, z_i), \qquad S_n = \sum_{i=1}^{n} X_i,$$

then we have

$$\mathcal{E}(f) - \mathcal{E}_n(f) = \frac{1}{n} S_n.$$

Let $p_i = \frac{|I_i|}{n}$ for $i = 1, 2, \ldots, k$, where $|I_i|$ is the number of terms in the i-th part, it then follows that

$$\sum_{i=1}^{k} p_i = \frac{1}{n} \sum_{i=1}^{n} |I_i| = 1.$$

Then

$$S_n = \sum_{i=1}^{k} \left[\sum_{m \in I_i} X_m \right] = \sum_{i=1}^{k} T(i), \tag{13}$$

where $T(i) = \sum_{m \in I_i} X_m$.

Now we can apply $\frac{1}{n} S_n$ to the exponential $\exp(\frac{\gamma S_n}{n})$ for all $\gamma > 0$

$$\mathrm{E}\left[\exp\left(\gamma \frac{S_n}{n} \right) \right] \leq \sum_{i=1}^{k} p_i \mathrm{E}\left[\exp\left(\gamma \frac{T(i)}{|I_i|} \right) \right]. \tag{14}$$

We now bound the second term on the right-hand side of inequality (14) which is denoted henceforth by ϕ. For any $i \in \{1, 2, \ldots, k\}$, we have

$$\begin{aligned} \phi &= \mathrm{E}\left[\exp\left(\sum_{m \in I_i} \frac{\gamma X_m}{|I_i|} \right) \right] = \mathrm{E}\left[\prod_{m=1}^{|I_i|} \exp\left(\frac{\gamma X_m}{|I_i|} \right) \right] \\ &\leq \prod_{m=1}^{|I_i|} \mathrm{E}\left[\exp\left(\frac{\gamma X_m}{|I_i|} \right) \right] + \left| \mathrm{E}\left[\prod_{m=1}^{|I_i|} \exp\left(\frac{\gamma X_m}{|I_i|} \right) \right] - \prod_{m=1}^{|I_i|} \mathrm{E}\left[\exp\left(\frac{\gamma X_m}{|I_i|} \right) \right] \right|. \end{aligned} \tag{15}$$

For simplicity, we denote the first term in inequality (15) by S_1, and denote the second term in inequality (15) by S_2. Now we proceed through the following two steps.

Step 1 Estimate S_1. By the stationary property of the α-mixing sequence $\mathcal{Z}$, we have

$$S_1 = \prod_{m=1}^{|I_i|} \mathrm{E}\left[\exp\left(\frac{\gamma X_m}{|I_i|} \right) \right] = \left\{ \mathrm{E}\left[\exp\left(\frac{\gamma X_1}{|I_i|} \right) \right] \right\}^{|I_i|}.$$

Since $\frac{X_1}{|I_i|}$ satisfies the Bernstein moment condition with $K_1 = \frac{M}{3|I_i|}$ (Modha and Masry 1996) in Lemma 1 and

$$\mathrm{E}[X_1] = \mathrm{E}\big\{ \mathrm{E}[\ell(f, z_1)] - \ell(f, z_1) \big\} = 0.$$

Hence for all $0 < \gamma \leq \frac{3|I_i|}{M}$, we have

$$\prod_{m=1}^{|I_i|} \mathrm{E}\left[\exp\left(\frac{\gamma X_m}{|I_i|} \right) \right] \leq \exp\left[\frac{\gamma^2 |I_i| \mathrm{E}\left| \frac{X_1}{|I_i|} \right|^2}{2(1 - \gamma M/(3|I_i|))} \right] \leq \exp\left[\frac{\gamma^2 \mathrm{E}\left| X_1 \right|^2}{2|I_i|(1 - \gamma M/(3|I_i|))} \right].$$

For all $i = 1, 2, \ldots, k$, $|I_i| \geq l$, thus we have

$$1 - \frac{\gamma M}{3|I_i|} \geq 1 - \frac{\gamma M}{3l},$$

and furthermore

$$\exp\left[\frac{\gamma^2 \mathrm{E}|X_1|^2}{2|I_i|(1-\gamma M/(3|I_i|))}\right] \le \exp\left[\frac{\gamma^2 \mathrm{E}|X_1|^2}{2l(1-\gamma M/(3l))}\right].$$

Thus we obtain

$$S_1 \le \exp\left[\frac{\gamma^2 \mathrm{E}|X_1|^2}{2l(1-\gamma M/(3l))}\right].$$

Step 2 Estimate S_2. With the same method in Modha and Masry (1996), by Lemma 2 and Assumption 1, we can get

$$\begin{aligned} S_2 &= \left|\mathrm{E}\left[\prod_{m=1}^{|I_i|}\exp\left(\frac{\gamma X_m}{|I_i|}\right)\right] - \prod_{m=1}^{|I_i|}\mathrm{E}\left[\exp\left(\frac{\gamma X_m}{|I_i|}\right)\right]\right| \\ &\le 4\alpha(k)(|I_i|-1)\prod_{m=1}^{|I_i|}\left\|\exp\left[\frac{\gamma X_m}{|I_i|}\right]\right\|_\infty \\ &\le 4(|I_i|-1)\alpha(k)e^{\gamma M} \\ &\le e^{|I_i|}e^{-2}4\overline{\alpha}\cdot e^{-ck^\beta}\cdot e^{\gamma M} \\ &\le 4e^{-2}\overline{\alpha}\exp\{|I_i|+\gamma M - ck^\beta\}. \end{aligned}$$

The final inequality follows from the fact that $|I_i|-1 \le e^{|I_i|-2}$ (this is deduced from $|I_i| \ge 2$ and Assumption 1).

Returning to inequality (15) and since $\gamma M \le 3|I_i|$, we obtain

$$\mathrm{E}\left[\exp\left(\gamma\frac{T(i)}{|I_i|}\right)\right] \le \exp\left[\frac{\gamma^2 \mathrm{E}|X_1|^2}{2l(1-\gamma M/(3l))}\right] + 4e^{-2}\overline{\alpha}\exp(4|I_i| - ck^\beta).$$

We require $\exp(4|I_i| - ck^\beta) \le 1$, which holds if $4|I_i| \le ck^\beta$. But $|I_i| \le (\frac{n}{k}+1)$, thus the bound holds if $4((n/k)+1) \le ck^\beta$, or if $4(n+k) < ck^{\beta+1}$. Since $n+k \le 2n$ the bound holds if $8n \le ck^{\beta+1}$, or if $\{8n/c\}^{\frac{1}{\beta+1}} \le k$. Let

$$k = \lceil \{8n/c\}^{\frac{1}{\beta+1}} \rceil.$$

Since $l = l_n = \lfloor n/k \rfloor$, we have

$$\mathrm{E}\left[\exp\left(\gamma\frac{T(i)}{|I_i|}\right)\right] \le \exp\left[\frac{\gamma^2 \mathrm{E}|X_1|^2}{2l(1-\gamma M/(3l))}\right] + 4e^{-2}\overline{\alpha}. \tag{16}$$

Since inequality (16) is true for all γ, $0 < \gamma < \frac{3|I_i|}{M}$, to make the constraint uniform over all i, we then require that γ satisfies

$$0 < \gamma < \frac{3l}{M} < \frac{3|I_i|}{M}.$$

Since

$$\frac{\gamma^2 \mathrm{E}|X_1|^2}{2l(1-\frac{\gamma M}{3l})} > 0,$$

we have

$$\mathrm{E}\left[\exp\left(\gamma\frac{T(i)}{|I_i|}\right)\right] \le (1+4e^{-2}\overline{\alpha})\exp\left[\frac{\gamma^2\mathrm{E}|X_1|^2}{2l(1-\gamma M/3l)}\right].$$

Returning to inequality (14), we have

$$\mathrm{E}\left[\exp\left(\gamma\frac{S_n}{n}\right)\right] \le (1+4e^{-2}\overline{\alpha})\exp\left[\frac{\gamma^2\mathrm{E}|X_1|^2}{2l(1-\gamma M/3l)}\right].$$

By Markov's inequality, we have that for any $\delta > 0$,

$$\begin{aligned}\mathrm{Prob}\{\mathcal{E}(f)-\mathcal{E}_n(f)>\delta\mathcal{E}(f)\} &= \mathrm{Prob}\left\{e^{\gamma(\mathcal{E}(f)-\mathcal{E}_n(f))} > e^{\gamma\delta\mathcal{E}(f)}\right\}\\ &\le \frac{\mathrm{E}[e^{\gamma(\mathcal{E}(f)-\mathcal{E}_n(f))}]}{e^{\gamma\delta\mathcal{E}(f)}}\\ &\le C\exp\left\{-\gamma\delta\mathcal{E}(f)+\frac{\gamma^2\mathrm{E}|X_1|^2}{2l(1-\gamma M/3l)}\right\},\end{aligned}$$

where $C = 1+4e^{-2}\overline{\alpha}$. Now by substituting

$$\gamma = \frac{\delta l\mu}{\mathrm{E}|X_1|^2 + M\delta\mu/3},$$

where $\mu = \mathcal{E}(f)$, and noting that γ satisfies $\gamma < \frac{3l}{M}$, we obtain

$$\mathrm{Prob}\{\mathcal{E}(f)-\mathcal{E}_n(f)>\delta\mathcal{E}(f)\} \le (1+4e^{-2}\overline{\alpha})\exp\left\{\frac{-\delta^2 l\mu^2}{2(\mathrm{E}|X_1|^2+\delta\mu M/3)}\right\}.$$

Replacing δ by $\frac{\varepsilon}{\sqrt{\mu}}$, we then have

$$\mathrm{Prob}\left\{\frac{\mathcal{E}(f)-\mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}}>\varepsilon\right\} \le (1+4e^{-2}\overline{\alpha})\exp\left\{\frac{-\varepsilon^2 l\mu}{2(\mathrm{E}|X_1|^2+\varepsilon\sqrt{\mu}M/3)}\right\}.$$

Since $r < \mathcal{E}(f) \le s$, replacing l by $n^{(\alpha)}$ then implies that for any $\varepsilon > 0$, the inequality

$$\mathrm{Prob}\left\{\frac{\mathcal{E}(f)-\mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}}>\varepsilon\right\} \le (1+4e^{-2}\overline{\alpha})\exp\left\{\frac{-n^{(\alpha)}r\varepsilon^2}{2(\mathrm{E}|X_1|^2+\varepsilon\sqrt{s}M/3)}\right\} \tag{17}$$

holds. Theorem 2 thus follows from inequality (17) by replacing $\mathrm{E}|X_1|^2$ by σ^2. This finishes the proof of Theorem 2. □

Remark 5 Vidyasagar (2002) established the bound (Theorem 3.5) on the difference between the empirical means and their true values based on strongly mixing sequences, and his bound consists of two terms. However, in this paper we are to bound the relative difference between the empirical risks and their expected risks based on exponentially strongly mixing sequences, and our result consists of only one exponential term. Comparing Theorem 2 with Theorem 3.5 in Vidyasagar (2002), we can find that the bound in Theorem 2 has smaller confidence interval than that in Theorem 3.5. Concerning the comparison of the convergence rate between the bound in Theorem 2 and that in Theorem 3.5 (Vidyasagar 2002), we also have the same results as those in Remark 4.

By Theorem 2, we now can prove our main theorem on the rate of the empirical risks relatively uniform converging to their expected risks for the ERM algorithm with exponentially strongly mixing sequence $\mathcal{Z}$.

Proof of Theorem 1 We decompose the proof into three steps.

Step 1 Let $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \cdots \cup \mathcal{H}_b$, $b \in \mathbf{N}$, then for any $\varepsilon > 0$, whenever

$$\sup_{f\in\mathcal{H}} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \geq \varepsilon,$$

there exists j, $1 \leq j \leq b$, such that

$$\sup_{f\in\mathcal{H}_j} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \geq \varepsilon.$$

This implies the equivalence

$$\sup_{f\in\mathcal{H}} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \geq \varepsilon \quad \Longleftrightarrow \quad \exists j,\ 1 \leq j \leq b,\ \text{s.t.}\ \sup_{f\in\mathcal{H}_j} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \geq \varepsilon. \tag{18}$$

By the equivalence (18), and by the fact that the probability of a union of events is bounded by the sum of the probabilities of these events, we have

$$\text{Prob}\left\{ \sup_{f\in\mathcal{H}} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \geq \varepsilon \right\} \leq \sum_{j=1}^{b} \text{Prob}\left\{ \sup_{f\in\mathcal{H}_j} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \geq \varepsilon \right\}. \tag{19}$$

Step 2 To estimate the term on the right-hand side of inequality (19), we define

$$\psi(f) = (1-\delta)\mathcal{E}(f) - \mathcal{E}_n(f).$$

Let $b = \mathcal{N}(\mathcal{H}, \frac{\varepsilon}{L})$ and let the disks D_j, $j \in \{1, 2, \ldots, b\}$ be a cover of $\mathcal{H}$ with center at f_j and radius ε/L. For any $\mathbf{z} \in \mathcal{Z}^n$ and all $f \in D_j$, we conclude

$$\begin{aligned}
\psi(f) - \psi(f_j) &= (1-\delta)\mathcal{E}(f) - \mathcal{E}_n(f) - [(1-\delta)\mathcal{E}(f_j) - \mathcal{E}_n(f_j)] \\
&= [\mathcal{E}_n(f_j) - \mathcal{E}_n(f)] + (1-\delta)[\mathcal{E}(f) - \mathcal{E}(f_j)] \\
&\leq L \cdot \|f - f_j\|_\infty + L(1-\delta) \cdot \|f - f_j\|_\infty \\
&\leq \varepsilon(2-\delta).
\end{aligned}$$

Since this holds for all $\mathbf{z} \in \mathcal{Z}^n$ and all $f \in D_j$, we obtain

$$\sup_{f\in D_j} \psi(f) \geq 2\varepsilon(2-\delta) \quad \Longrightarrow \quad \psi(f_j) \geq \varepsilon(2-\delta).$$

This implies that for $j = 1, 2, \ldots, b$,

$$\text{Prob}\left\{ \sup_{f\in D_j} \psi(f) \geq 2\varepsilon(2-\delta) \right\} \leq \text{Prob}\left\{ \psi(f_j) \geq \varepsilon(2-\delta) \right\}. \tag{20}$$

Step 3 For the sake of simplicity, we denote the term on the right-hand side of inequality (20) by I_1 and denote the term on the left-hand side of inequality (20) by I_2. Take $\delta = \frac{\varepsilon}{\mathcal{E}(f_j)}$, and suppose $0 < \varepsilon < 2r$, then we have

$$
\begin{aligned}
I_1 &= \mathrm{Prob}\Big\{\psi(f_j) \geq \varepsilon(2-\delta)\Big\} \\
&= \mathrm{Prob}\Big\{\mathcal{E}(f_j) - \mathcal{E}_n(f_j) \geq \delta\mathcal{E}(f_j) + \varepsilon(2-\delta)\Big\} \\
&= \mathrm{Prob}\left\{\mathcal{E}(f_j) - \mathcal{E}_n(f_j) \geq \varepsilon + \varepsilon\left(2 - \frac{\varepsilon}{\mathcal{E}(f_j)}\right)\right\} \\
&= \mathrm{Prob}\left\{\frac{\mathcal{E}(f_j) - \mathcal{E}_n(f_j)}{\sqrt{\mathcal{E}(f_j)}} \geq \frac{\varepsilon}{\sqrt{\mathcal{E}(f_j)}} + \frac{\varepsilon}{\sqrt{\mathcal{E}(f_j)}}\left(2 - \frac{\varepsilon}{\mathcal{E}(f_j)}\right)\right\} \\
&\leq \mathrm{Prob}\left\{\frac{\mathcal{E}(f_j) - \mathcal{E}_n(f_j)}{\sqrt{\mathcal{E}(f_j)}} \geq \frac{\varepsilon}{\sqrt{s}}\left(3 - \frac{\varepsilon}{r}\right)\right\} \\
&\leq \mathrm{Prob}\left\{\frac{\mathcal{E}(f_j) - \mathcal{E}_n(f_j)}{\sqrt{\mathcal{E}(f_j)}} \geq \frac{\varepsilon}{\sqrt{s}}\right\}, \\
I_2 &= \mathrm{Prob}\Big\{\sup_{f\in D_j} \psi(f) \geq 2\varepsilon(2-\delta)\Big\} \\
&= \mathrm{Prob}\left\{\sup_{f\in D_j}\left[\mathcal{E}(f) - \mathcal{E}_n(f) - \frac{\varepsilon\mathcal{E}(f)}{\mathcal{E}(f_j)}\right] \geq 2\varepsilon\left(2 - \frac{\varepsilon}{\mathcal{E}(f_j)}\right)\right\} \\
&\geq \mathrm{Prob}\left\{\sqrt{r}\sup_{f\in D_j}\left[\frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} - \frac{\varepsilon s}{\sqrt{\mathcal{E}(f)\mathcal{E}(f_j)}}\right] \geq 2\varepsilon\left(2 - \frac{\varepsilon}{\mathcal{E}(f_j)}\right)\right\} \\
&\geq \mathrm{Prob}\left\{\sup_{f\in D_j} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \geq \frac{\varepsilon s}{\mathcal{E}(f_j)\sqrt{r}} + \frac{2\varepsilon}{\sqrt{r}}\left(2 - \frac{\varepsilon}{\mathcal{E}(f_j)}\right)\right\} \\
&\geq \mathrm{Prob}\left\{\sup_{f\in D_j} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \geq \frac{\varepsilon s}{r\sqrt{r}} + \frac{2\varepsilon}{\sqrt{r}}\left(2 - \frac{\varepsilon}{s}\right)\right\} \\
&\geq \mathrm{Prob}\left\{\sup_{f\in D_j} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \geq \frac{(s+4r)\varepsilon}{r\sqrt{r}}\right\}.
\end{aligned}
$$

By inequality (20), we then get

$$
\mathrm{Prob}\left\{\sup_{f\in D_j} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \geq \frac{(s+4r)\varepsilon}{r\sqrt{r}}\right\} \leq \mathrm{Prob}\left\{\frac{\mathcal{E}(f_j) - \mathcal{E}_n(f_j)}{\sqrt{\mathcal{E}(f_j)}} \geq \frac{\varepsilon}{\sqrt{s}}\right\}. \tag{21}
$$

Combining inequalities (19), (21) and (12), and replacing ε by $\frac{r\sqrt{r}}{(s+4r)}\varepsilon$, we then get Theorem 1. □

5 Proof of generalization bound

In this section, we begin to prove the generalization bounds (Propositions 1 and 2) of the ERM algorithm with exponentially strongly mixing samples by the results obtained in the last section. Our main tool is the following lemma established by Cucker and Smale (2002b).

Lemma 3 (Cucker and Smale 2002b) *Let $c_1,\ c_2 > 0$, and $s > q > 0$. Then the equation*

$$x^s - c_1 x^q - c_2 = 0$$

has a unique positive zero x^*. In addition

$$x^* \le \max\{(2c_1)^{1/(s-q)}, (2c_2)^{(1/s)}\}.$$

Proof of Proposition 1 By the assumption that $0 < \varepsilon \le 2r$, the exponential of inequality (5) in Theorem 1 becomes

$$\frac{-r\tau^2 n^{(\alpha)}}{2(\sigma^2 + \tau\sqrt{s}M/3)} \le -C_1 n^{(\alpha)}\varepsilon^2,$$

where

$$C_1 = \frac{3r^4}{2s(s+4r)[3(s+4r)\sigma^2 + 2r^{\frac{5}{2}}M]}.$$

Since $\mathcal{H}$ is assumed to be compact, then by assumption (2) we have

$$\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon r\sqrt{r}}{(s+4r)L}\right) \le \exp\left\{C_0\left[\frac{\varepsilon r\sqrt{r}}{(s+4r)L}\right]^{\frac{-2d}{p}}\right\},$$

where C_0 is a positive constant. In other words, for any ε, $2r \ge \varepsilon > 0$, by Theorem 1 we have

$$\text{Prob}\left\{\sup_{f\in\mathcal{H}} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \ge \varepsilon\right\} \le C\exp\left\{C_0\left[\frac{\varepsilon r\sqrt{r}}{(s+4r)L}\right]^{\frac{-2d}{p}} - C_1 n^{(\alpha)}\varepsilon^2\right\}, \tag{22}$$

where $C = 1 + 4e^{-2}\overline{\alpha}$.

Let us rewrite inequality (22) in an equivalent form. We equate the right-hand side of inequality (22) to a positive value $\eta (0 < \eta \le 1)$

$$C\exp\left\{C_0\left[\frac{\varepsilon r\sqrt{r}}{(s+4r)L}\right]^{\frac{-2d}{p}} - C_1 n^{(\alpha)}\varepsilon^2\right\} = \eta.$$

It follows that

$$\varepsilon^{2+\frac{2d}{p}} - \frac{\ln(C/\eta)}{C_1 n^{(\alpha)}}\varepsilon^{\frac{2d}{p}} - \frac{C_0 r^{\frac{-3d}{p}}[(s+4r)L]^{\frac{2d}{p}}}{C_1 n^{(\alpha)}} = 0.$$

By Lemma 3, we can solve this equation with respect to ε. This equation has a unique positive zero $\varepsilon*$, and

$$\varepsilon* \doteq \varepsilon(n,\eta) \le \max\left\{\left[\frac{2\ln(C/\eta)}{C_1 n^{(\alpha)}}\right]^{\frac{1}{2}}, \left[\frac{2C_0 r^{\frac{-3d}{p}}[(s+4r)L]^{\frac{2d}{p}}}{C_1 n^{(\alpha)}}\right]^{\frac{p}{2p+2d}}\right\}.$$

It is used further to solve inequality

$$\sup_{f\in\mathcal{H}} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \le \varepsilon(n,\eta).$$

Then we deduce that with probability at least $1-\eta$ simultaneously for all functions in the function set $\mathcal{H}$, the inequality

$$\mathcal{E}(f) \le \mathcal{E}_n(f) + \frac{\varepsilon^2(n,\eta)}{2}\left(1+\sqrt{1+\frac{4\mathcal{E}_n(f)}{\varepsilon^2(n,\eta)}}\right)$$

is valid. Since with probability at least $1-\eta$, this inequality holds for all functions of the function set $\mathcal{H}$, it holds in particular for the function $f_{\mathbf{z}}$ that minimizes the empirical risk $\mathcal{E}_n(f)$ over $\mathcal{H}$. For this function with probability at least $1-\eta$, the inequality

$$\mathcal{E}(f_{\mathbf{z}}) \le \mathcal{E}_n(f_{\mathbf{z}}) + \frac{\varepsilon^2(n,\eta)}{2}\left(1+\sqrt{1+\frac{4\mathcal{E}_n(f_{\mathbf{z}})}{\varepsilon^2(n,\eta)}}\right) \tag{23}$$

then holds.

By Theorem 4.3 in Modha and Masry (1996), we have that for any $\varepsilon > 0$, the inequality

$$\text{Prob}\left\{|\mathcal{E}(f)-\mathcal{E}_n(f)| > \varepsilon\right\} \le 2(1+4e^{-2}\overline{\alpha})\exp\left\{\frac{-n^{(\alpha)}\varepsilon^2}{2(\sigma^2+\varepsilon M/3)}\right\} \tag{24}$$

is valid. Thus by inequality (24), we conclude that for the same η as above, and for the function $f_{\mathcal{H}}$ that minimizes the expected risk $\mathcal{E}(f)$ over $\mathcal{H}$, the inequality

$$\mathcal{E}(f_{\mathcal{H}}) > \mathcal{E}_n(f_{\mathcal{H}}) - \varepsilon'(n,\eta) \tag{25}$$

holds with probability $1-\eta$, where

$$\varepsilon'(n,\eta) = \frac{M\ln(C/\eta)}{3n^{(\alpha)}}\left(1+\sqrt{1+\frac{18n^{(\alpha)}\sigma^2}{M^2\ln(C/\eta)}}\right).$$

Note that

$$\mathcal{E}_n(f_{\mathcal{H}}) \ge \mathcal{E}_n(f_{\mathbf{z}}). \tag{26}$$

From inequalities (23), (25) and (26), we deduce that with probability at least $1-2\eta$, the inequality

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \le \varepsilon'(n,\eta) + \frac{\varepsilon^2(n,\eta)}{2}\left(1+\sqrt{1+\frac{4\mathcal{E}_n(f_{\mathbf{z}})}{\varepsilon^2(n,\eta)}}\right)$$

is valid. In addition, if

$$n^{(\alpha)} \ge \max\left\{\frac{\ln(C/\eta)}{2C_1r^2}, \frac{C_0[(s+4r)L]^{\frac{2d}{p}}}{C_1 2^{\frac{2d}{p}} r^{\frac{5d+2p}{p}}}, \frac{\ln(C/\eta(\sigma^2+sM/3))}{2r^2}\right\},$$

then we have $\varepsilon \le 2r$. This leads to Proposition 1. □

Proof of Proposition 2 When the complexity of the function set $\mathcal{H}$ is high, in order to solve the time-consuming problem of the ERM algorithm (1), we can decompose the hypothesis space $\mathcal{H}$ into many compact subsets by following the enlightening idea of Giné and

Koltchinski (2006), and denote it as follows:

$$\mathcal{H} = \bigcup_{i=1}^{a} \mathcal{H}(\rho_{i-1}, \rho_i].$$

For every i, $1 \le i \le a$, let $f_{\mathbf{z}}^i$ be the function minimizing the empirical risk $\mathcal{E}_n(f)$ over $f \in \mathcal{H}(\rho_{i-1}, \rho_i]$. By the similar argument with inequality (23), we have that for any $\eta \in (0, 1]$, with probability at least $1 - \eta$, the inequality

$$\mathcal{E}(f_{\mathbf{z}}^i) \le \mathcal{E}_n(f_{\mathbf{z}}^i) + \frac{\varepsilon_i^2(n,\eta)}{2}\left(1 + \sqrt{1 + \frac{4\mathcal{E}_n(f_{\mathbf{z}}^i)}{\varepsilon_i^2(n,\eta)}}\right), \quad 1 \le i \le a \tag{27}$$

holds, where

$$\varepsilon_i(n,\eta) \le \max\left\{\left[\frac{2\ln(C/\eta)}{C_i n^{(\alpha)}}\right]^{\frac{1}{2}}, \left[\frac{2C_0(\rho_{i-1})^{\frac{-3d}{p}}[(\rho_i + 4\rho_{i-1})L]^{\frac{2d}{p}}}{C_i n^{(\alpha)}}\right]^{\frac{p}{2p+2d}}\right\},$$

$$C_i = \frac{3\rho_{i-1}^4}{2\rho_i(\rho_i + 4\rho_{i-1})[3(\rho_i + 4\rho_{i-1})\sigma^2 + 2\rho_{i-1}^{\frac{5}{2}}M]}.$$

Thus we have that with probability at least $1 - \eta$, the inequality

$$\mathcal{E}(f_{\mathbf{z}}) \le \min_{1\le i\le a}\left\{\mathcal{E}_n(f_{\mathbf{z}}^i) + \frac{\varepsilon_i^2(n,\eta)}{2}\left(1 + \sqrt{1 + \frac{4\mathcal{E}_n(f_{\mathbf{z}}^i)}{\varepsilon_i^2(n,\eta)}}\right)\right\} \tag{28}$$

is valid.

In addition, for the same η as above, we have that with probability $1 - \eta$, the inequality

$$\mathcal{E}(f_{\mathcal{H}}) \ge \mathcal{E}_n(f_{\mathcal{H}}) - \varepsilon'(n,\eta) \tag{29}$$

holds. Then by inequalities (28), (29) and the fact that

$$\mathcal{E}_n(f_{\mathcal{H}}) \ge \mathcal{E}_n(f_{\mathbf{z}}^i), \quad i \in \{1, 2, \ldots, a\}$$

we have that with probability $1 - 2\eta$, the inequality

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \le \varepsilon'(n,\eta) + \min_{1\le i\le a}\left\{\frac{\varepsilon_i^2(n,\eta)}{2}\left(1 + \sqrt{1 + \frac{4\mathcal{E}_n(f_{\mathbf{z}}^i)}{\varepsilon_i^2(n,\eta)}}\right)\right\}$$

is valid. We then complete the proof of Proposition 2. □

6 Conclusions

In this paper we have studied the learning performance of the ERM algorithm with exponentially strongly mixing samples. We first established a new bound on the rate of relative uniform convergence for the ERM algorithm with exponentially strongly mixing samples. Then we have derived the generalization bounds of the ERM algorithm and proved that the ERM algorithm with exponentially strongly mixing observations is consistent. To our

knowledge, the results here are the first explicit bounds on the rate of convergence on this topic. In order to have a better understanding of the significance and value of the established results in this paper, we have compared our results with the previous works, and concluded that the established results not only sharpen and improve the previously known results in Zou and Li (2007), Vidyasagar (2002), but also extend the results in Bousquet (2003), Cucker and Smale (2002a), Vapnik (1998) for i.i.d. samples to the case of α-mixing sequence. We have also shown that the learning rates of the ERM algorithm with exponentially strongly mixing samples are close to or as same as those for learning rate with i.i.d. samples.

In addition, since the ERM algorithm is usually very time-consuming and overfitting may happen when the complexity of the given function set $\mathcal{H}$ is high, as an application of our main results, we also explored a new strategy to implement the ERM algorithm in high complexity hypothesis space.

Along the line of the present work, several open problems deserve further research. For example, how to control the generalization ability of the ERM algorithm with exponentially strongly mixing samples? What is the essential difference between the generalization ability of the ERM algorithm with i.i.d. samples and dependent samples? All these problems are under our current investigation.

Acknowledgements The authors are grateful to the reviewers for their valuable comments and suggestions. The authors would like to thank Professor Nicolo Cesa-Bianchi for his careful reading and helpful comments on the paper.

References

Alexander, K. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Annals of Probability*, *4*, 1041–1067.

Alon, N., Ben-Darid, S., Cesa-Bianchi, N., & Haussler, D. (1997). Scale-sensitive dimensions, uniform convergence and learnability. *Journal of the Association for Computing Machinery*, *44*, 615–631.

Bartlett, P. L., & Long, P. M. (1998). Prediction, learning, uniform convergence, and scale-sensitive dimensions. *Journal of Computer and System Sciences*, *56*(2), 174–190.

Bartlett, P. L., & Lugosi, G. (1999). An inequality for uniform deviations of sample averages from their means. *Statistics & Probability Letters*, *4*, 55–62.

Bartlett, P. L., & Mendelson, S. (2002). Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, *3*, 463–482.

Bousquet, O. (2003). New approaches to statistical learning theory. *Annals of the Institute of Statistical Mathematics*, *55*, 371–389.

Cesa-Bianchi, N., Alex Conconi, A., & Gentile, C. (2004). On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, *50*(9), 2050–2057.

Chen, D. R., Wu, Q., Ying, Y. M., & Zhou, D. X. (2004). Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, *5*, 1143–1175.

Craig, C. C. (1933). On the Tchebycheff inequality of Bernstein. *Annals of Mathematical Statistics*, *4*, 94–102.

Cucker, F., & Smale, S. (2002a). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, *39*, 1–49.

Cucker, F., & Smale, S. (2002b). Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computational Mathematics*, *2*, 413–428.

Cucker, F., & Zhou, D. X. (2007). *Learning theory: an approximation theory viewpoint*. Cambridge: Cambridge University Press.

Davydov, Y. A. (1973). Mixing conditions for Markov chains. *Theory of Probability and its Applications*, *XVIII*, 312–328.

Devroye, L. (1982). Bounds for the uniform deviation of empirical measures. *Journal of Multivariate Analysis*, *12*, 72–79.

Evgeniou, T., & Pontil, M. (1999). *Lecture notes in comput. sci.: Vol. 1720. On the V-gamma dimension for regression in reproducing Kernel Hilbert spaces* (pp. 106–117). Berlin: Springer.

Giné, E., & Koltchinski, V. (2006). Concentration inequality and asymptotic results for ratio type empirical processes. *The Annals of Probability*, *34*(3), 1143–1216.
Ibragimov, I. A., & Linnik, Y. V. (1971). *Independent and stationary sequences of random variables*. Groningen: Wolters-Noordnoff.
Karandikar, R. L., & Vidyasagar, M. (2002). Rates of uniform convergence of empirical means with mixing processes. *Statistics & Probability Letters*, *58*, 297–307.
Lugosi, G., & Pawlak, M. (1994). On the posterior-probability estimate of the error of nonparameter classification rules. *IEEE Transactions on Information Theory*, *40*(5), 475–481.
Modha, S., & Masry, E. (1996). Minimum complexity regression estimation with weakly dependent observations. *IEEE Transactions on Information Theory*, *42*, 2133–2145.
Nobel, A., & Dembo, A. (1993). A note on uniform laws of averages for dependent processes. *Statistics & Probability Letters*, *17*, 169–172.
Pollard, D. (1984). *Convergence of stochastic processes*. New York: Springer.
Rosenblatt, M. (1956). A central theorem and strong mixing conditions. *Proceedings of the National Academy of Sciences*, *4*, 43–47.
Smale, S., & Zhou, D. X. (2003). Estimating the approximation error in learning theory. *Analysis and Its Applications*, *1*, 17–41.
Smale, S., & Zhou, D. X. (2004). Shannon sampling and function reconstruction from point values. *Bulletin of the American Mathematical Society*, *41*, 279–305.
Steinwart, I., Hush, D., & Scovel, C. (2006). *Learning from dependent observations* (Technical Report LA-UR-06-3507). Los Alamos National Laboratory. Submitted for publication.
Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Annals of Probability*, *22*, 28–76.
Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.
Vidyasagar, M. (2002). *Learning and generalization with applications to neural networks* (2nd ed.). Berlin: Springer.
Withers, C. S. (1981). Conditions for linear processes of stationary mixing sequences. *Annals of Probability*, *22*, 94–116.
White, H. (1989). Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, *3*, 535–549.
Wu, Q., & Zhou, D. X. (2005). SVM soft margin classifiers: linear programming versus quadratic programming. *Neural Computation*, *17*, 1160–1187.
Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability*, *22*, 94–114.
Zhou, D. X. (2003). Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, *49*, 1743–1752.
Zou, B., & Li, L. Q. (2007). The performance bounds of learning machines based on exponentially strongly mixing sequence. *Computer and Mathematics with Applications*, *53*(7), 1050–1058.