JMLR: Workshop and Conference Proceedings 60:1–16, 2016 ACML 2016

Learnability of Non-I.I.D.

Wei Gao GAOW@LAMDA.NJU.EDU.CN
Xin-Yi Niu NIUXY@LAMDA.NJU.EDU.CN
Zhi-Hua Zhou ZHOUZH@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China

Editors: Robert J. Durrant and Kee-Eung Kim

Abstract

Learnability has always been one of the most central problems in learning theory. Most previous studies on this issue were based on the assumption that the samples are drawn independently and identically according to an underlying (unknown) distribution. The i.i.d. assumption, however, does not hold in many real applications. In this paper, we study the learnability of problems where the samples are drawn from empirical process of stationary β-mixing sequence, which has been a widely-used assumption implying a dependence weaken over time in training samples. By utilizing the *independent blocks* technique, we provide a sufficient and necessary condition for learnability, that is, *average stability* is equivalent to learnability with AERM (Asymptotic Empirical Risk Minimization) in the non-i.i.d. learning setting. In addition, we also discuss the generalization error when the test variable is dependent on the training sample.

Keywords: Learning theory, β-mixing sequence, stationary, learnability, generalization, consistency, stability

1. Introduction

The characterization of learnability has become one of the most fundamental issues in learning theory, and it concerns about whether the learned function converges uniformly to the optimal function for a learning problem as the size of training sample tends to infinity. Some influential work (Alon et al., 1997; Blumer et al., 1989) showed that learnability, at least for supervised classification and regression, is equivalent to the uniform convergence of the empirical risk to the expected risk, and thus much attention has been paid to establishing the uniform convergence based on various measures of hypothesis space complexity, such as Vapnik-Chervonenkis dimension (Vapnik, 1982), cover number (Bartlett, 1998), Rademacher or Gaussian complexity (Bartlett and Mendelson, 2002), etc. This equivalence, however, does not hold in the general learning setting (Alon et al., 1997), and stability has been explored as an equivalent condition for learnability (Shalev-Shwartz et al., 2010).

Most previous studies on learning theory were accomplished by assuming i.i.d. samples, whereas the i.i.d. assumption does not hold in many tasks, e.g., signal processing, system diagnosis, speech recognition, etc., where the dependence exists among the training samples and the intrinsic learning processes are non-i.i.d (Doukhan, 1994; Bradley, 2007; Dedecker et al., 2007). Much effort has been paid to exploring the dependence existing in the training sample, and various conditions have been made to measure the dependence

in non-i.i.d. scenarios, e.g., α-mixing, β-mixing, φ-mixing sequence, etc. It is possible to use simple methods to estimate the mixing rate for various classes of random process, e.g., Hidden Markov Model (Modha and Masry, 1998). Owing to the i.i.d. assumption, existing theoretical results could not be applied directly to non-i.i.d. cases.

We present a theoretical study on the learnability of non-i.i.d. setting, where the training samples are drawn from empirical process of stationary β-mixing sequence. We first prove that uniform stability is not necessary for learnability though it is enough for generalization. Then, we introduce the *average stability* in the general non-i.i.d. learning setting and establish its equivalence with learnability; in other words, we prove that the existence of a universally average stable AERM is sufficient and necessary for learnability of non-i.i.d. setting. In summary, we establish the relationships for non-i.i.d. setting: Existence of average stable AERM$\Leftrightarrow$Learnability with AERM$\Leftrightarrow$Learnability. Also, we discuss the generalization error when the test variable is dependent on the training sample.

1.1. Related Work

The pioneer study of Yu (1994) established the uniform convergence based on VC-dimension for empirical processes of stationary β-mixing sequence. Vidyasagar (2002) pointed out that β-mixing seems to be the "just the right" assumptions for maintaining the PAC-learning properties with some sub-additivity conditions. Mohri and Rostamizadeh (2009) recently provided the first Rademacher complexity-based generalization bounds for stationary β-mixing sequence. Mohri and Rostamizadeh (2008) introduced the uniform stability and provided a generalization bound with application to SVM and Kernel Ridge Regression. The uniform stability, however, is overly-strong and far from necessary for learnability, shown by Example 1 in Section 5.1. The consistency for non-stationary sequence has been studied in (Steinwart et al., 2009), and there are alternative non-i.i.d. assumptions (Karandikar and Vidyasagar, 2002; Modha and Masry, 1998; Steinwart and Christmann, 2010). The non-i.i.d. sequence can be viewed as a special case of environment change, and our work makes a step to theoretically understand the evolvable property (Zhou, 2016), i.e., the learning model is able to get accustomed to environment in the future of machine learning.

Algorithmic stability was first proposed by Rogers and Wagner (1978), and has been used to analyze the generalization performance of algorithms (Bousquet and Elisseeff, 2002; Elisseeff et al., 2005; Rakhlin et al., 2005). Shalev-Shwartz et al. (2010) introduced the *on-average-LOO stability* and established its equivalence to learnability with AERM in the i.i.d. general learning setting, yet can not be applied to the non-i.i.d. learning setting. Vidyasagar (2002) presented possibly the first study on the learnability of non-i.i.d. setting, which is different from ours. First, he focused on the uniform convergence, where the equivalence between learnability and uniform convergence is specific to supervised classification and regression; while we prove the sufficient and necessary condition of non-i.i.d. learnability in the general learning setting. Second, Vidyasagar made an additional assumption that $\beta(m) = O(m^{-c})$ for constant $c > 0$; while we require a weaker condition $\beta(m) \to 0$ as $m \to \infty$. Third, the result in (Vidyasagar, 2002) is heavily related to the hypothesis space; while our result does not rely on any space complexity measure, but rather on the way the algorithm searches the space.

2. Preliminaries

2.1. Mixing Sequence

Definition 1 *A random-variable sequence* $\mathbf{Z} = \{Z_t\}_{t=-\infty}^{\infty}$ *is said to be stationary if, for every integer* t, $i \geq 0$ *and* $k \geq 0$, *the random vector* $(Z_t, \ldots, Z_{t+k})$ *has the same distribution as the random vector* $(Z_{t+i}, \ldots, Z_{t+i+k})$.

It is easy to find that the time index t is not relevant to the distribution for a stationary sequence. This, however, does not imply independence in the sequence because, for example, $\Pr[Z_j|Z_i]$ may be unequal to $\Pr[Z_j]$. Several conditions have been made to measure the weak dependence of random sequences, while in this work, we mainly focus on the following β-mixing sequence, and leave the discussion on other non-i.i.d. sequences to future work.

Definition 2 *Let* $\mathbf{Z} = \{Z_t\}_{t=-\infty}^{\infty}$ *be a stationary sequence, and* σ_i^j *denotes the* σ*-algebra generated by random variables* $Z_i, \ldots, Z_j$ *for* $i < j$. *For any integer* $m > 0$, *the* β*-mixing coefficients are given by* $\beta(m) = \sup_n \left[E_{A_1 \in \sigma_{-\infty}^n} \left[\sup_{A_2 \in \sigma_{n+m}^{+\infty}} \left[\left| \Pr\left[A_2|A_1\right] - \Pr\left[A_2\right] \right| \right] \right] \right]$.

The stationary sequence $\mathbf{Z}$ *is said to be* β*-mixing if* $\beta(m) \to 0$ *as* $m \to \infty$*; algebraically* β*-mixing if* $\beta(m) \leq \beta_0/m^r$ *holds for some constants* $\beta_0 > 0$ *and* $r > 0$*; exponentially* β*-mixing if* $\beta(m) \leq \beta_0 \exp(-\beta_1 m^r)$ *holds for some constants* $\beta_0 > 0$, $\beta_1 > 0$ *and* $r > 0$.

The β-mixing coefficients $\beta(m)$ can be used to measure the dependence between 'future' events and 'past' ones separated by a distance of at least m. Throughout this paper, we assume that $\beta(m)$ is non-increasing, and it is easy to obtain the following propositions:

Proposition 3 *Any i.i.d. sequence can be viewed as a special stationary* β*-mixing sequence with coefficients* $\beta(m) = 0$ *for* $m > 0$.

For a real number r, we denote by $\lfloor r \rfloor$ the biggest integer which is no larger than r. It is necessary to introduce the following lemma:

Lemma 4 *If* $\beta(m) \to 0$ *as* $m \to \infty$, *then there exists a* $\tau(m) \leq m$ *such that* $\tau(m) \to \infty$ *and* $\tau(m)\beta\left(\lfloor m/\tau(m) \rfloor\right) \to 0$ *as* $m \to \infty$.

Proof: If $\beta(m) = 0$ then we choose $\tau(m) = m$ as desired. Now we consider the case $\beta(m) \neq 0$ for $m > 0$. If $\beta(m) \to 0$ as $m \to \infty$, then there exists a sequence $\{a_i\}$ such that $a_i \to \infty$ and $a_i\beta(i) \to 0$ as $i \to \infty$. For example, a possible choice is $a_i = \left\lfloor 1/\sqrt{\beta(i)} \right\rfloor$. For every $m \geq 1$, there exists an index $k \geq 1$ such that $(k-1)a_{k-1} \leq m \leq ka_k$. We select $\tau(m) = a_k$ and this lemma follows. ■

It is not difficult to select $\tau(m)$ for some special cases, e.g., for algebraically β-mixing sequence with coefficients $\beta(m) \leq \beta_0/m^r$, for some $\beta_0 > 0$ and $r > 0$, we can select $\tau(m) = m^{r_0}$ such that $0 < r_0 < r/(1+r)$; for exponentially β-mixing sequence with coefficients $\beta(m) = \beta_0 \exp(-\beta_1 m^r)$, for some $\beta_0 > 0$, $\beta_1 > 0$ and $r > 0$, we can select $\tau(m) = m^{r_0}$ such that $0 < r_0 < 1$.

It is worth noticing that the choice of $\tau(m)$ is not unique, and different choices lead to different convergence rate for $\tau(m)\beta(\lfloor m/\tau(m) \rfloor) \to 0$. Throughout this paper, $\tau(m)$ is not referred to specific choices. We also assume, without loss of generality, that $\tau(m)\beta\left(\lfloor m/\tau(m) \rfloor\right)$ is a non-increasing sequence in this paper.

Figure 1: A series of weakly dependent blocks $\{\sigma_{r_j}^{s_j}\}_{j=1}^{q}$.

The *independent block* technique, owing originally to Bernstein (1927), has been utilized as a popular tool to produce new results when dealing with learning problems of weak dependence. This technique has been applied successfully to many non-i.i.d learning studies (Lozano et al., 2006; Mohri and Rostamizadeh, 2008, 2009; Yu, 1994). We introduce a lemma from (Yu, 1994, Corollary 2.7) as follows:

Lemma 5 *(Yu, 1994, Corollary 2.7) For $m > 1$, suppose that a measurable hypothesis h is bounded by B on a product probability space $(\prod_{j=1}^{q}\Omega_j, \prod_{j=1}^{q}\sigma_{r_j}^{s_j})$ shown by Figure 1 with $r_j \leq s_j \leq r_{j+1}$. Let Q be a probability measure on the product space with marginal measures Q_j on $(\Omega_j, \sigma_{r_j}^{s_j})$ and let $P = \prod_{j=1}^{q} Q_j$. By setting $k_j = r_{j+1} - s_j$, we have $|E_Q[h] - E_P[h]| \leq (q-1)B\beta(Q)$ with $\beta(Q) = \sup_{j\in[q-1]}\beta(k_j)$.*

2.2. Learning Setting

The general non-i.i.d. learning setting can be described as follows. Let $\mathcal{Z}$ denote an instance space and $\mathbf{Z} = \{Z_t\}_{t=-\infty}^{\infty}$ is a stationary β-mixing sequence where each random variable has the same underlying (unknown) distribution P over the instance space $\mathcal{Z}$. A training sample $S = \{z_1, z_2, \ldots, z_n\}$ is chosen according to n different components of $\mathbf{Z}$. Here we do not require that these components must be continuous.

A learning algorithm $\mathscr{A}$ is a mapping from a training sample S to a hypothesis $\mathscr{A}_S \in \mathcal{H}$. For simplicity, we consider symmetric algorithms in this paper, i.e., algorithms depending upon the given sample but not on the order of instances in the sample. Many existing approaches are symmetric for non-i.i.d. setting (Lozano et al., 2006; Mohri and Rostamizadeh, 2008), and our results can be generalized to asymmetric algorithms.

A learning problem is relevant to a hypothesis space $\mathcal{H}$ and loss function $l\colon \mathcal{H}\times\mathcal{Z}\to\mathbb{R}$, where the goal is to minimize the expected risk

$$R(h) = E_{z\sim P}[l(h,z)] \tag{1}$$

over the hypothesis space $\mathcal{H}$, where z is taken independently of any sequence and we will discuss the dependently expected risk in Section 4. We assume that the loss function is bounded by some constant $B > 0$, i.e., $|l(h,z)| \leq B$ for all $h \in \mathcal{H}$ and $z \in \mathcal{Z}$. Many classical learning problems fall into this framework such as classification, regression, clustering, density estimation, etc.

In the general non-i.i.d learning setting, we essentially try to find some hypothesis $h \in \mathcal{H}$ which minimizes the expected risk over the whole hypothesis space $\mathcal{H}$, i.e., $\min_{h\in\mathcal{H}} R(h)$. Notice that the underlying distribution is unknown, and therefore, we could not minimize the expected risk $R(h)$ directly. Classical learning methods, instead, seek to minimize the

empirical risk $\hat{R}_S(h)$ with respect to h based on the training sample S:

$$\hat{R}_S(h) = \frac{1}{n}\sum_{i=1}^{n} l(h, z_i).$$

A learning algorithm $\mathscr{A}$ is said to be an ERM (Empirical Risk Minimizer) if it minimizes the empirical risk $\hat{R}_S(\mathscr{A}_S) = \hat{R}_S(\hat{h}_S) = \min_{h\in\mathcal{H}} \hat{R}_S(h)$, where $\hat{h}_S = \arg\min_{h\in\mathcal{H}} \hat{R}_S(h)$. A learning algorithm $\mathscr{A}$ is said to be an AERM (Asymptotic Empirical Risk Minimization) with rate $\epsilon_{\text{erm}}(n)$ under stationary β-mixing distribution if $E_S[|\hat{R}_S(\mathscr{A}_S) - \hat{R}_S(\hat{h}_S)|] \leq \epsilon_{\text{erm}}(n)$.

A learning algorithm is universally an AERM with rate $\epsilon_{\text{erm}}(n)$ if it is an AERM with rate $\epsilon_{\text{erm}}(n)$ under all stationary β-mixing distributions with coefficients $\beta(m)$. By "all stationary β-mixing distributions with coefficients $\beta(m)$", or shortly for all stationary β-mixing distributions, we mean all (underlying) distributions P over the instance space $\mathcal{Z}$ and all β-mixing sequences whose coefficients are smaller than $\beta(m)$.

A learning algorithm $\mathscr{A}$ is said to be consistent with rate $\epsilon_{\text{con}}(n)$ under stationary β-mixing distribution if $E_S[R(\mathscr{A}_S) - R(h^*)] \leq \epsilon_{\text{con}}(n)$, where $h^* = \arg\min_{h\in\mathcal{H}} R(h)$. A learning algorithm $\mathscr{A}$ is universally consistent with rate $\epsilon_{\text{con}}(n)$ if it is consistent with rate $\epsilon_{\text{con}}(n)$ under all stationary β-mixing distributions. Based on this definition, we give the following definition of learnability for the non-i.i.d. case:

Definition 6 *A problem is learnable if there exists a universally consistent algorithm, i.e., there exists an algorithm $\mathscr{A}$ and a rate $\epsilon_{con}(n)$ such that $\sup_\Lambda \{E_S[R(\mathscr{A}_S) - R(h^*)]\} \leq \epsilon_{con}(n)$ where $\epsilon_{con}(n) \to 0$ as $n \to \infty$, and Λ denotes all stationary β-mixing distributions with coefficients $\beta(m)$.*

A learning algorithm $\mathscr{A}$ generalizes with rate $\epsilon_{\text{gen}}(n)$ under stationary β-mixing distribution if $E_S[|R(\mathscr{A}_S) - \hat{R}_S(\mathscr{A}_S)|] \leq \epsilon_{\text{gen}}(n)$, and $\mathscr{A}$ universally generalizes with rate $\epsilon_{\text{gen}}(n)$ if it generalizes with rate $\epsilon_{\text{gen}}(n)$ under all stationary β-mixing distributions.

3. Sufficient and Necessary Condition for Non-I.I.D. Learnability

3.1. Stability

We set $[n] = \{1, 2, \ldots, n\}$ for an integer $n > 0$. Given sample $S_n = \{z_1, z_2, \ldots, z_n\}$ and for $i \in [n]$, we denote by $S_n^i = \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}$ the sample in which the ith example has been deleted in sample S_n. We also denote by $S_n^{i,z} = \{z_1, \ldots, z_{i-1}, z, z_{i+1}, \ldots, z_n\}$ the sample in which the ith example has been replaced by instance z in sample S_n.

For the i.i.d. learning setting, several stability notions have been explored for studying the performance of learning algorithms (Bousquet and Elisseeff, 2002; Rakhlin et al., 2005) and the learnability of learning problems (Mukherjee et al., 2006; Shalev-Shwartz et al., 2010). For the non-i.i.d. scenarios, Mohri and Rostamizadeh (2008) first introduced the *uniform stability* given below, and derived stability bounds for generalization error.

Definition 7 *A algorithm $\mathscr{A}$ has uniform stability $\zeta(n)$ if $|l(\mathscr{A}_{S_n}, z) - l(\mathscr{A}_{S_n^{i,z'}}, z)| \leq \zeta(n)$ for all $i \in [n]$ and $z, z' \in \mathcal{Z}$. Here $\zeta(n) \to 0$ as $n \to \infty$.*

The uniform stability is sufficient for generalization. This notion is overly-strong, and is far from necessary for learnability of non-i.i.d. setting as shown by Example 1 (Section 5.1). In this paper, we introduce a new notion of stability, which is weaker than uniform stability but is proven to be equivalent to learnability with AERM. The definition is:

Definition 8 *A learning algorithm $\mathscr{A}$ has average stability $\zeta(n)$ under stationary β-mixing distribution if $|E_{S_n,z}[l(\mathscr{A}_{S_n}, z) - l(\mathscr{A}_{S_n^{i,z}}, z)]| \leq \zeta(n)$ for $i \in [n]$, where $\zeta(n) \to 0$ as $n \to \infty$.*

Notice that the variable z in $l(\mathscr{A}_{S_n^{i,z}}, z)$ is a substitute sample and a test sample simultaneously. A learning algorithm has universally average stability with rate $\zeta(n)$ if the stability property holds with rate $\zeta(n)$ under all stationary β-mixing distributions. It is obvious that uniform stability implies average stability whereas the converse direction does not hold from Example 3 (in Section 5.1).

3.2. Main Results

We first prove that an average stable AERM is sufficient for generalization and consistency for a stationary β-mixing sequence.

Theorem 9 *If a learning algorithm $\mathscr{A}$ is an AERM with rate $\epsilon_{erm}(n)$, and is average stable with rate $\zeta(n)$ for stationary β-mixing sequence with coefficients $\beta(m)$, then it exhibits generalization and consistency with rate*

$$\begin{aligned}\epsilon_{con}(n) &\leq \epsilon_{erm}(n) + \zeta(n)\\ \epsilon_{gen}(n) &\leq \zeta(n) + 2\epsilon_{erm}(n) + 2B\tau(n)\beta\left(\left\lfloor \frac{n}{\tau(n)} \right\rfloor\right) + \frac{2B}{\sqrt{\tau(n)}},\end{aligned}$$

where $\tau(n)$ is given in Lemma 4.

We can easily get the following two corollaries by combining with Lemma 21 in Section 5:

Corollary 10 *If a algorithm $\mathscr{A}$ is an AERM with rate $\epsilon_{erm}(n)$, and is average stable with rate $\zeta(n)$ for stationary algebraically β-mixing sequence with coefficients $\beta(n) \leq \beta_0/n^r$ for some constants $\beta_0 > 0$ and $r > 0$, then it exhibits generalization and consistency with rate*

$$\begin{aligned}\epsilon_{con}(n) &\leq \zeta(n) + \epsilon_{erm}(n)\\ \epsilon_{gen}(n) &\leq \zeta(n) + 2\epsilon_{erm}(n) + 2B\beta_0 n^{-\frac{r}{2}} + 2Bn^{-\frac{r}{4(1+r)}}.\end{aligned}$$

Corollary 11 *If a learning algorithm $\mathscr{A}$ is an AERM with rate $\epsilon_{erm}(n)$, and is average stable with rate $\zeta(n)$ for stationary exponentially β-mixing sequence with coefficients $\beta(m) \leq \beta_0 \exp(-\beta_1 m^r)$ for some constants $\beta_0 > 0$, $\beta_1 > 0$ and $r > 0$, then it exhibits generalization and consistency with rate*

$$\begin{aligned}\epsilon_{con}(n) &\leq \zeta(n) + \epsilon_{erm}(n),\\ \epsilon_{gen}(n) &\leq \zeta(n) + 2\epsilon_{erm}(n) + 2B\beta_0\sqrt{n}\exp(-\beta_1 n^{r/2}) + 2B/\sqrt[4]{n}.\end{aligned}$$

From Theorem 9, it is easy to prove that the existence of universally average stable AERMs is sufficient for learnability.

Corollary 12 *If a learning algorithm is universally an AERM and average stable, then it is universally generalizing and consistent.*

Theorem 9 shows that, for an AERM, average stability implies generalization and consistency for specific β-mixing sequence, while the inverse direction does not hold as shown by Example 2 in Section 5.1. However, we have the following equivalence:

Theorem 13 *For an AERM, we have the following equivalence for non-i.i.d. setting,*

$$\textit{Universal average stability} \Leftrightarrow \textit{Universal consistency} \Leftrightarrow \textit{Universal generalization.}$$

The detailed proof is presented in Section 5.2, and it is noteworthy that, from Lemmas 22 to 24, we do not require the universal condition for **stability** $\leftrightarrow$ **generalization** $\rightarrow$ **consistency**. The universality property is only utilized to prove **consistency** $\rightarrow$ **generalization**. Finally, we have

Theorem 14 *For non-i.i.d. setting, a problem is learnable if and only if there exists a universally average stable AERM.*

From Corollary 12, it is obvious that the existence of a universally average stable AERM implies learnability of non-i.i.d. setting. For the inverse direction, Definition 6 illustrates that learnability does not require an AERM, but rather a universal consistent algorithm. Thus, we complete the proof by constructing a universally average stable AERM from a universally consistent algorithm, and the detailed proof is presented in Section 5.3.

4. Dependently Expected Risk

Our previous work focused on the independently expected risk $R(\mathscr{A}_S) = E_{z\sim\mathcal{D}}[l(\mathscr{A}_S, z)]$, where the test variable z is totally independent to S as mentioned in Section 2. From a more realistic view, we should consider the case where test variables are dependent on samples, even if the dependence is rather weak as shown in β-mixing sequence.

For a β-mixing sequence $\mathbf{Z} = \{Z_t\}_{t=-\infty}^{\infty}$, we assume, without loss of generality, that the sample S of size n are drawn from $(Z_1, Z_2, \ldots, Z_n)$, and based on this sample, we can learn a function $\mathscr{A}_S$. First, we notice that it is not unique for the definitions of dependently expected risk, and different choices for test variables give different notions. Here, we will introduce two definitions:

- To measure the performance of $\mathscr{A}_S$ on the sequence $\{Z_t\}_{t=n+1}^{\infty}$, i.e., the sequence after the sample, we define the dependently expected risk as

$$R_1(\mathscr{A}_S|S) = \lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} E_{Z_{n+i}}[l(\mathscr{A}_S, Z_{n+i})]. \tag{2}$$

- To measure the performance of $\mathscr{A}_S$ on a special interval $\{Z_{n+k_0}, Z_{n+k_0+1}, \ldots, Z_{n+k_1}\}$ for $k_1 \geq k_0 \geq 1$, we define the dependently expected risk as

$$R_2(\mathscr{A}_S|S) = \frac{1}{k_1 - k_0 + 1} \sum_{i=k_0}^{k_1} E_{Z_{n+i}}[l(\mathscr{A}_S, Z_{n+i})]. \tag{3}$$

  If $k_1 = k_0 \geq 1$, then we can measure the performance of $\mathscr{A}_S$ on special variable Z_{n+k_0}.

We now discuss the relationships between the independently expected risk $R(\mathscr{A}_S)$ and the dependently expected risks defined above. First, we observe that $R_1(\mathscr{A}_S|S) = R(\mathscr{A}_S)$ almost surely, and the proof is deferred to Section 5.4.

Theorem 15 *For a sample S from β-mixing sequence, we have $E_S[|R(\mathscr{A}_S) - R_1(\mathscr{A}_S|S)|] = 0$. Here $R(\mathscr{A}_S)$ and $R_1(\mathscr{A}_S|S)$ are given by Eqns. (1) and (2), respectively.*

This theorem gives an explanation for the independently expected risk $R(\mathscr{A}_S)$, i.e., it measures the average generalization error of learned function $\mathscr{A}_S$ on the β-mixing sequence that is after the sample; in other word, $R(\mathscr{A}_S)$ reflects the performance on sequence $\{Z_t\}_{t=n+1}^{\infty}$. Intuitively, $R(\mathscr{A}_S) \approx R_1(\mathscr{A}_S|S)$ because the β-mixing sequence is rather weakly dependent over time. We also have the following relationship between $R(\mathscr{A}_S)$ and $R_2(\mathscr{A}_S|S)$, whose proof is deferred to Section 5.5.

Theorem 16 *For some $k_1 \geq k_0 \geq 1$ and for a sample S from β-mixing sequence with coefficients $\beta(m)$, we have $E_S[|R(\mathscr{A}_S) - R_2(\mathscr{A}_S|S)|] \leq B\beta(k_0) + B\tau(s)\beta(\lfloor s/\tau(s) \rfloor) + B/\sqrt{\tau(s)}$. Here $s = k_1 - k_0 + 1$, $R(\mathscr{A}_S)$ and $R_2(\mathscr{A}_S|S)$ are given by Eqns. (1) and (3), respectively.*

5. Detailed Proofs and Examples

We start by introducing the following lemmas and theorem, which will be used later.

Lemma 17 *We have $E[X] = E[X|X \geq \epsilon] \Pr[X \geq \epsilon] + E[X|X < \epsilon] \Pr[X < \epsilon]$.*

Lemma 18 *(Shalev-Shwartz et al., 2010) For two random variables X and Y, if $X \leq Y$ then $E[|X|] \leq |E[X]| + 2E[|Y|]$.*

Lemma 19 *(Shalev-Shwartz et al., 2010) If $X_1, X_2, \ldots, X_n$ are n i.i.d. random variables with $|X_1| \leq B$, then we have $E[|\sum_{i=1}^n (X_i - E[X_1])/n|] \leq B/\sqrt{n}$.*

Theorem 20 *If $S = \{z_1, z_2, \ldots, z_n\}$ is from a stationary β-mixing sequence with coefficients $\beta(m)$, then, for $h \in \mathcal{H}$, we have $E_S[|\hat{R}_S(h) - R(h)|] \leq B\tau(n)\beta(\lfloor n/\tau(n) \rfloor) + B/\sqrt{\tau(n)}$. Here $\tau(n)$ is given by Lemma 4.*

Proof: From Lemma 4, there is a $\tau(n)$ such that $\tau(n) \to \infty$ and $\tau(n)\beta(\lfloor n/\tau(n) \rfloor) \to 0$ as $n \to \infty$. The $E_S[|\hat{R}_S(h) - R(h)|]$ is bounded by

$$\frac{1}{\lfloor \frac{n}{\tau(n)} \rfloor} \sum_{k=0}^{\lfloor \frac{n}{\tau(n)} \rfloor - 1} E_S \left[\left| \sum_{j=1}^{\tau(n)} \frac{l(h, z_{j\lfloor \frac{n}{\tau(n)} \rfloor + k})}{\tau(n)} - R(h) \right| \right] = E_S \left[\left| \sum_{j=1}^{\tau(n)} \frac{l(h, z_{j\lfloor \frac{n}{\tau(n)} \rfloor})}{\tau(n)} - R(h) \right| \right] \tag{4}$$

where the last equality holds from the stationary β-mixing sequence. We now introduce a ghost sample $S' = \{z'_1, z'_2, \ldots, z'_{\tau(n)}\}$ chosen i.i.d. from the distribution P, and from Lemmas 19, it holds that

$$E_{S'}[|\hat{R}_{S'}(h) - R(h)|] \leq B/\sqrt{\tau(n)}. \tag{5}$$

Meanwhile, Lemma 5 gives

$$\left| E_S \left[\left| \sum_{j=1}^{\tau(n)} \frac{l(h, z_{j\lfloor \frac{n}{\tau(n)} \rfloor})}{\tau(n)} - R(h) \right| \right] - E_{S'}[|\hat{R}_{S'}(h) - R(h)|] \right| \leq B\tau(n)\beta\left(\left\lfloor \frac{n}{\tau(n)} \right\rfloor \right), \tag{6}$$

where each instance $z_{j\lfloor \frac{n}{\tau(n)} \rfloor}$ $(j \in [\tau(n)])$ is viewed as a weakly dependent block. It is noteworthy that the above equality holds even if the sample S is chosen from noncontinuous components of a β-mixing sequence since $\beta(m)$ is non-increasing. Thus this theorem holds from Eqs. (4) to (6). ■

Lemma 21 *For a hypothesis $h \in \mathcal{H}$,*

- *if $S = \{z_1, z_2, \ldots, z_n\}$ is from a stationary algebraically β-mixing sequence with coefficients $\beta(m) \leq \beta_0/m^r$ for some constants $\beta_0 > 0$ and $r > 0$, then, we have $E_S[|\hat{R}_S(h) - R(h)|] \leq B\beta_0 n^{-\frac{r}{2}} + Bn^{-\frac{r}{4(1+r)}}$;*
- *if $S = \{z_1, z_2, \ldots, z_n\}$ is from a stationary exponentially β-mixing sequence with coefficients $\beta(m) \leq \beta_0 \exp(-\beta_1 m^r)$ for some constants $\beta_0 > 0$, $\beta_1 > 0$ and $r > 0$, then, it holds that $E_S[|\hat{R}_S(h) - R(h)|] \leq B\beta_0\sqrt{n}\exp(-\beta_1 n^{r/2}) + B/\sqrt[4]{n}$.*

Proof: This lemma holds by setting $\tau(n) = \lfloor n^{\frac{r}{2(r+1)}} \rfloor$ for algebraically β-mixing sequence and by setting $\tau(n) = \lfloor n^{\frac{1}{2}} \rfloor$ for exponentially β-mixing sequence in Theorem 20. ■

5.1. Examples

Example 1 *For non-i.i.d. setting, there exists a learning problem with universally consistent algorithm, but does not have uniform stability.*

Proof: For instance space $\mathcal{Z} = \mathcal{X} \times \{-1, 1\}$ with $\mathcal{X} = [-1, 1]$ and objective hypothesis $t(x) = I[x \geq 0]$, we consider the hypothesis space $\mathcal{H} = \{h_\theta \colon h_\theta(x) = I[x \geq \theta] \text{ for } \theta \in [-1, 0]\}$, and loss function $l(h, (x, y)) = I[h(x) \neq y]$.

Assume that a training sample $S = \{z_0 = (x_0, y_0), z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)\}$ is chosen from any given β-mixing sequence $\{Z_t\}_{-\infty}^{\infty}$ with coefficients $\beta(m)$, where each Z_t has the same distribution P. From Lemma 4, there exists a $\tau(n)$ such that $\tau(n) \to \infty$ and $\tau(n)\beta\left(\lfloor n/\tau(n) \rfloor\right) \to 0$ as $n \to \infty$.

Thus we can construct a subsample $S' = \{z_0, z_{\lfloor \frac{n}{\tau(n)} \rfloor}, z_{2\lfloor \frac{n}{\tau(n)} \rfloor}, \ldots, z_{\tau(n)\lfloor \frac{n}{\tau(n)} \rfloor}\}$ and consider the algorithm $\mathscr{A}_S = \mathscr{A}_{S'} = h_\theta$ with $\theta = \max\{-1, x_i \colon (x_i, y_i) \in S' \text{ and } x_i < 0\}$. Further, we introduce a ghost sample $\tilde{S}' = \{\tilde{z}_0, \tilde{z}_1, \ldots, \tilde{z}_{\tau(n)}\}$ drawn i.i.d. from the distribution P. From Lemma 5, we have

$$|E_{S'}[R(\hat{\mathscr{A}_{S'}})] - E_{\tilde{S}'}[\hat{\mathscr{A}_{\tilde{S}'}}]| \leq \tau(n)\beta\left(\lfloor n/\tau(n) \rfloor\right). \tag{7}$$

For any $h_\theta \in \mathcal{H}$, we have $R(h_\theta) = E_z[l(h_\theta, z)] = \Pr_z[\theta \le x < 0]$ with $z = (x, y)$. Thus, it follows that, for any $\epsilon \ge 0$,

$$\Pr_{\tilde{S}'}[R(\hat{\mathscr{A}}_{\tilde{S}'}) \ge \epsilon] = \Pr_{\tilde{S}'}[\Pr_z[\theta \le x < 0] \ge \epsilon] \le (1-\epsilon)^{\tau(n)+1} \le (1-\epsilon)^{\tau(n)} \le \exp(-\tau(n)\epsilon),$$

where the first inequality holds from the facts that there are $\tau(n)+1$ instance in the sample $\tilde{S}'$, and for every instance $z = (x, y) \in \tilde{S}'$, it holds that $\Pr_z[x \notin [\theta, 0)] < 1 - \epsilon$. Noticing $|l(h, (x, y))| \le B = 1$ and setting $\epsilon = 1/\sqrt{\tau(n)}$, we have that, from Lemma 17,

$$E_{\tilde{S}'}[R(\hat{\mathscr{A}}_{\tilde{S}'})] \le \exp\left(-\sqrt{\tau(n)}\right) + 1/\sqrt{\tau(n)}. \tag{8}$$

It is clear that $R(h^*) = 0$ when we choose $h^*(x) = t(x)$, and it holds that

$$E_S[R(\mathscr{A}_S) - R(h^*)] = E_{S'}[R(\hat{\mathscr{A}}_{S'})] \le \exp(-\sqrt{\tau(n)}) + \tau(n)\beta\left(\lfloor n/\tau(n)\rfloor\right) + 1/\sqrt{\tau(n)}$$

by combining Eqs. (7) and (8). Thus, it is proved that $\mathscr{A}_S$ is universally consistent, and therefore this problem is learnable.

On the other hand, for examples $\hat{z} = (\hat{x}, \hat{y})$ and $z = (x, y)$ such that $\theta < x < \hat{x} < 0$, it holds that $|l(\mathscr{A}_S, z) - l(\mathscr{A}_{S^{i,\hat{z}}}, z)| = 1$, which proves that the uniform stability does not hold, and thus we complete the proof. ■

Example 2 *For specific β-mixing sequence, there exists an ERM which is consistent but not average stable.*

Proof: For instance space $\mathcal{Z} = \mathcal{X} \times \{-1, +1\}$ with $\mathcal{X} = [-1, 1]$, target hypothesis $t(x) = I[x \ge 0]$ and loss function $l(h, z = (x, y)) = I[y \ne h(x)]$, we consider the hypothesis space $\mathcal{H} = \{h\colon h(x) = 1 \text{ except for finite } x \in [-1, 1]\}$ and the uniform distribution P.

A training sample $S = \{z_0 = (x_0, y_0), z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)\}$ is chosen from a specific β-mixing sequence $\{Z_t\}_{-\infty}^{\infty}$ with coefficients $\beta(m)$, where each Z_t has the same distribution P. We consider the following algorithm $\mathscr{A}_S(x) = I[x \ge 0]$ if there is an $i \in [n]$ s.t. $x = x_i$; otherwise, $\mathscr{A}_S(x) = 1$. Algorithm $\mathscr{A}$ is consistent since every hypothesis $h \in \mathcal{H}$ is consistent under the continuous uniform distribution P. We also have $|E_{S,z}[l(\mathscr{A}_S, z) - l(\mathscr{A}_{S^{i,z}}, z)]| = 1/2$ for $i \in [n]$, which discloses that algorithm $\mathscr{A}$ is not average stable. ■

Example 3 *For specific β-mixing sequence, there exists an average stable algorithm without uniform stability.*

Proof: Let the instance space $\mathcal{Z} = \mathcal{X} \times [-1, 1]$ with $\mathcal{X} = [0, 2]$. We assume that the object function $t(x) = \operatorname{sgn}(x - 1)$, the loss function $l(h, (x, y)) = |y - h(x)|$ and the underlying distribution P is uniform on $\mathcal{X}$. Assume that a training sample $S = \{z_0 = (x_0, y_0), z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)\}$ is chosen from a specific β-mixing sequence $\{Z_t\}_{-\infty}^{\infty}$ with coefficients $\beta(m)$, where each Z_t has the same distribution P. We consider the following non-AERM algorithm $\mathscr{A}_S(x) = 1$ for $x \in S$; otherwise, $\mathscr{A}_S(x) = 0$.

From the continuous distribution P, we have $E_{S_n,z}[l(\mathscr{A}_{S_n}, z)] = 1$ and $E_{S_n,z}[l(\mathscr{A}_{S_n^{i,z}}, z)] = 1$; therefore, $\mathscr{A}$ has average stability from $|E_{S_n,z}[l(\mathscr{A}_{S_n}, z) - l(\mathscr{A}_{S_n^{i,z}}, z)]| = 0$. On the other hand, if $z \notin S_n$, then we have $|l(\mathscr{A}_{S_n}, z) - l(\mathscr{A}_{S_n^{i,z}}, z)| = 1$, which implies that $\mathscr{A}$ does not have uniform stability. ■

5.2. Proof of Theorems 9 and 13

It is easy to prove that generalization implies average stability:

Lemma 22 (Generalization⇒average stability) *If a learning algorithm $\mathscr{A}$ has generalization with rate $\epsilon_{gen}(n)$, then it has average stability with rate $\zeta(n) = \epsilon_{gen}(n)$ for a stationary β-mixing sequence.*

Proof: We have $E_S[l(\mathscr{A}_S, z_i)] = E_S[l(\mathscr{A}_S, z_j)]$ for all $i, j \in [n]$, and

$$E_S[\hat{R}_S(\mathscr{A}_S)] = E_{S,z}[l(\mathscr{A}_{S^{i,z}}, z)]. \tag{9}$$

From the definition of $R(\mathscr{A}_S)$, we have

$$E_S[R(\mathscr{A}_S)] = E_{S,z}[l(\mathscr{A}_S, z)]. \tag{10}$$

This follows $|E_{S,z}[l(\mathscr{A}_S, z) - l(\mathscr{A}_{S^{i,z}}, z)]| = |E_S[R(\mathscr{A}_S) - \hat{R}_S(\mathscr{A}_S)]| \leq \epsilon_{\text{gen}}(n)$. ■

Lemma 23 (AERM+average stability⇒generalization) *If a learning algorithm $\mathscr{A}$ is an AERM with rate $\epsilon_{erm}(n)$ and has average stability with rate $\zeta(n)$ for a stationary β-mixing sequence with coefficients $\beta(m)$, then it generalizes with rate $\epsilon_{gen}(n) = \zeta(n) + 2\epsilon_{erm}(n) + 2B\tau(n)\beta(\lfloor n/\tau(n)\rfloor) + 2B/\sqrt{\tau(n)}$. Here $\tau(n)$ is given by Lemma 4.*

Proof: With the definitions of $\hat{h}_S$ and h^*, we have $\hat{R}_S(\mathscr{A}_S) - R(\mathscr{A}_S) = \hat{R}_S(\mathscr{A}_S) - \hat{R}_S(\hat{h}_S) + \hat{R}_S(\hat{h}_S) - \hat{R}_S(h^*) + \hat{R}_S(h^*) - R(h^*) + R(h^*) - R(\mathscr{A}_S) \leq \hat{R}_S(\mathscr{A}_S) - \hat{R}_S(\hat{h}_S) + \hat{R}_S(h^*) - R(h^*)$. By utilizing Eqs. (9) and (10), we have $|E_S[\hat{R}_S(\mathscr{A}_S) - R(\mathscr{A}_S)]| \leq \zeta(n)$ for average stable algorithms. From the definition of AERM and Theorem 20, it holds that

$$E_S[|\hat{R}_S(\mathscr{A}_S) - \hat{R}_S(\hat{h}_S) + \hat{R}_S(h^*) - R(h^*)|] \leq \epsilon_{\text{erm}}(n) + B\tau(n)\beta(\lfloor n/\tau(n)\rfloor) + B/\sqrt{\tau(n)}.$$

Thus, the lemma follows by applying Lemma 18. ■

Lemma 24 (AERM+average stability⇒consistency) *If a learning algorithm $\mathscr{A}$ is an AERM with rate $\epsilon_{erm}(n)$, and has average stability with rate $\zeta(n)$ for a stationary β-mixing sequence with coefficients $\beta(m)$, then it is consistent with rate $\epsilon_{con}(n) = \epsilon_{erm}(n) + \zeta(n)$.*

Proof: For stationary sequence, we have $E[R(h^*)] = E[\hat{R}_S(h^*)]$, and $E[R(\mathscr{A}_S) - R(h^*)] = E[R(\mathscr{A}_S) - \hat{R}_S(h^*)]$. Moreover $E[R(\mathscr{A}_S) - \hat{R}_S(h^*)] = E[R(\mathscr{A}_S) - \hat{R}_S(\mathscr{A}_S)] + E[\hat{R}_S(\mathscr{A}_S) - \hat{R}_S(\hat{h}_S)] + E[\hat{R}_S(\hat{h}_S) - \hat{R}_S(h^*)] \leq \zeta(n) + \epsilon_{\text{erm}}(n)$ where the last inequality holds from Eqns. (9) and (10), and $\hat{R}_S(\hat{h}_S) \leq \hat{R}_S(h^*)$. ■

Theorem 9 follows from Lemmas 23 and 24, and we also establish the relationships **generalization↔stability→consistency** for an AERM. The inverse direction, however, does not hold for specific stationary β-mixing distribution, as can be seen from Example 2 in Section 5.1. We consider the universal consistency and introduce the following lemma:

Lemma 25 *If a problem is learnable, i.e., there exists a universally consistent algorithm $\mathscr{A}$ with rate $\epsilon_{con}(n)$ for all stationary β-mixing distributions with coefficients $\beta(n)$, then $E\left[\left|\hat{R}_S(\hat{h}_S) - R(h^*)\right|\right] \leq \epsilon_{emp}(n)$ with $\epsilon_{emp}(n) = 2\epsilon_{con}(n') + 2Bn'^2/n + 2Bn'/n + 4B/\sqrt{n} + B/\sqrt{\tau(\lfloor n/2\rfloor)} + \tau(\lfloor n/2\rfloor)\beta(\lfloor n/2\rfloor/\tau(\lfloor n/2\rfloor)) + (2\tau(\lfloor\sqrt[4]{n}\rfloor) + 1)\beta(\lfloor\lfloor\sqrt[4]{n}\rfloor/\tau(\lfloor\sqrt[4]{n}\rfloor)\rfloor)$, $\tau(n)$ is given by Lemma 4, and $n' = \tau(\lfloor\sqrt[4]{n}\rfloor)$.*

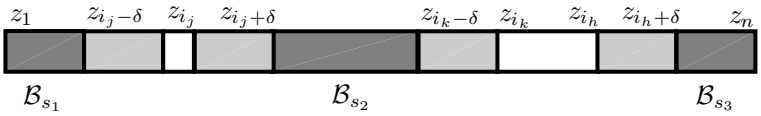

Figure 2: A sequence of blocks $\{\mathcal{B}_i\}$ composed of successively indexed instances in S, and the index distance from those in S' is larger than δ. Note that in this figure $z_{i_j}, z_{i_k}, z_{i_h} \in S'$ and $i_h - i_k < \delta$.

Proof: For $S' = \{z_{i_1}, z_{i_2}, \ldots, z_{i_{n'}}\}$ of size $n' = \tau(\lfloor \sqrt[4]{n} \rfloor)$ chosen independently from the uniform distribution over sample $S = \{z_1, z_2, \ldots, z_n\}$ (with replacements), where $\tau(n)$ is given by Lemma 4 with respect to the β-mixing coefficients $\beta(n)$, the probability of instances in S' having repeated indices can be bounded by $\sum_{i=1}^{n'} (i-1)/n \leq n'^2/n$. The subsample S' (without repeated indices from S) can be viewed from the original β-mixing sequence and thus we have, from universal consistency,

$$E[|R(\mathscr{A}_{S'}) - R(h^*)|| \text{ no repeated indices in } S'] \leq \epsilon_{\text{con}}(n'). \tag{11}$$

Meanwhile, S' can also be viewed as a sample drawn from the uniform distribution over instances in S, which yields

$$E[|\hat{R}_S(\mathscr{A}_{S'}) - \hat{R}_S(\hat{h}_S)|| \text{ no repeated indices in } S'] \leq \epsilon_{\text{con}}(n'). \tag{12}$$

Notice that S' has no repeated index from S, and S' could be temporally dependent to $S \setminus S'$. We will construct a sequence of blocks $\{\mathcal{B}_i\}$ such that each block $\mathcal{B}_i$ is weakly relevant to S' by following steps: 1) We first introduce a new set $S_1 = \{z_t \in S \setminus S' | \text{ the index distance is larger than } \delta \text{ from those indices in } S' = \{z_{i_1}, z_{i_2}, \ldots, z_{i_{n'}}\}, \text{ i.e., } |t - i_j| \geq \delta \text{ for } 1 \leq j \leq n'\}$ with where $\delta = \lfloor \lfloor \sqrt[4]{n} \rfloor / n' \rfloor$; 2) Blocks $\{\mathcal{B}_i\}$ are composed of those successively indexed instances in S_1, which can be shown by Figure 2.

Then, we have $E[|R(\mathscr{A}_{S'}) - \hat{R}_{S \setminus S'}(\mathscr{A}_{S'})|] \leq |\cup \mathcal{B}_i| E[|R(\mathscr{A}_{S'}) - \hat{R}_{\cup \mathcal{B}_i}(\mathscr{A}_{S'})|]/(n-n') + |S \setminus S' \setminus \cup \mathcal{B}_i| E[|R(\mathscr{A}_{S'}) - \hat{R}_{S \setminus S' \setminus \cup \mathcal{B}_i}(\mathscr{A}_{S'})|]/(n-n')$. It is obvious that $|S_1| = |\cup \mathcal{B}_i| < n - n'$ and $|S \setminus S' \setminus \cup \mathcal{B}_i| \leq 2n'\delta \leq 2\sqrt[4]{n}$. For bounded loss function, we have $E[|R(\mathscr{A}_{S'}) - \hat{R}_{S \setminus S' \setminus \cup \mathcal{B}_i}(\mathscr{A}_{S'})|] \leq 2B$. Therefore, we have

$$E[|R(\mathscr{A}_{S'}) - \hat{R}_{S \setminus S'}(\mathscr{A}_{S'})|] \leq E[|R(\mathscr{A}_{S'}) - \hat{R}_{\cup \mathcal{B}_i}(\mathscr{A}_{S'})|] + 4B/\sqrt{n} \tag{13}$$

for $n \geq 5$. To bound $E[|R(\mathscr{A}_{S'}) - \hat{R}_{\cup \mathcal{B}_i}(\mathscr{A}_{S'})|]$, we consider the similar sequence of blocks $\{\hat{\mathcal{B}}_i\}$, which are independent to S' and each blocks are drawn according to the same original β-mixing distribution. Recalling that the index distance in blocks $\{\mathcal{B}_i\}$ is lager than δ from the indices of instances in S', it holds that, from Lemma 5, $E[|R(\mathscr{A}_{S'}) - \hat{R}_{\cup \mathcal{B}_i}(\mathscr{A}_{S'})|] \leq E[|R(\mathscr{A}_{S'}) - \hat{R}_{\cup \hat{\mathcal{B}}_i}(\mathscr{A}_{S'})|] + (2n'+1)B\beta(\lfloor \lfloor \sqrt[4]{n} \rfloor / n' \rfloor) \leq B\tau(\lfloor n/2 \rfloor)\beta(\lfloor n/2 \rfloor / \tau(\lfloor n/2 \rfloor)) + B/\sqrt{\tau(\lfloor n/2 \rfloor)} + (2n'+1)B\beta(\lfloor \lfloor \sqrt[4]{n} \rfloor / n' \rfloor)$, where the last inequality holds from Theorem 20 since $|\cup \hat{\mathcal{B}}_i| \geq n/2$ for $n \geq 8$, and $\{\hat{\mathcal{B}}_i\}$ are independent to S'. From Eq. (13), $E[|R(\mathscr{A}_{S'}) - \hat{R}_{S \setminus S'}(\mathscr{A}_{S'})|]$ can be bounded

$$4B/\sqrt{n} + B\tau(\lfloor n/2 \rfloor)\beta(\lfloor n/2 \rfloor / \tau(\lfloor n/2 \rfloor)) + B/\sqrt{\tau(\lfloor n/2 \rfloor)} + (2n'+1)B\beta\left(\lfloor \lfloor \sqrt[4]{n} \rfloor / n' \rfloor\right). \tag{14}$$

Finally, if there is no repeatedly indexed instances, then, for any hypothesis, especially for $\mathscr{A}_{S'}$, it holds that $|\hat{R}_S(\mathscr{A}_{S'}) - \hat{R}_{S\setminus S'}(\mathscr{A}_{S'})| \leq 2Bn'/n$. Combining Eqs. (11), (12) and (13), and accounting for maximum discrepancy for the repeatedly indexed sample S' case. ■

Lemma 26 (AERM+universal consistency⇒generalization) *For a universally consistent AERM $\mathscr{A}$ with rates $\epsilon_{erm}(n)$ and $\epsilon_{con}(n)$ for all stationary β-mixing distributions with coefficients $\beta(m)$, it generalizes with rate $\epsilon_{gen}(n) = \epsilon_{erm}(n) + \epsilon_{con}(n) + \epsilon_{emp}(n)$. Here $\epsilon_{emp}(n)$ is given by Lemma 25.*

Proof: We have $E[|\hat{R}_S(\mathscr{A}_S) - R(\mathscr{A}_S)|] \leq E[|\hat{R}_S(\mathscr{A}_S) - \hat{R}_S(\hat{h}_S)|] + E[|\hat{R}_S(\hat{h}_S) - R(h^*)|] + E[|R(h^*) - R(\mathscr{A}_S)|] \leq \epsilon_{\text{emp}}(n) + \epsilon_{\text{erm}}(n) + \epsilon_{\text{con}}(n)$ as desired. ■

5.3. Proof of Theorem 14

It is obvious that, from Theorem 9, an average-stable AERM implies learnability, and the inverse direction holds from the following lemma:

Lemma 27 *If a problem is learnable, i.e., there exists a universally consistent $\mathscr{A}$ with rate $\epsilon_{con}(n)$ under all stationary β-mixing distributions with coefficients $\beta(m)$, then there is an algorithm $\mathscr{A}'$ which is average stable with rate $\zeta(n) = \epsilon_{gen}(n)$, and is an AERM with rate $\epsilon_{erm}(n) = \epsilon_{gen}(n) + \epsilon_{con}(n') + \epsilon_{emp}(n)$. Here $n' = \lfloor \sqrt{n} \rfloor$, and*

$$\epsilon_{gen}(n) = B\beta(n') + 4Bn'/n + B\tau(n-2n')\beta\left(\lfloor (n-2n')/\tau(n-2n') \rfloor\right) + B/\sqrt{\tau(n-2n')},$$

$\epsilon_{emp}(n)$ is given by Lemma 25 and $\tau(n)$ is given by Lemma 4.

Proof: For any universally consistent algorithm $\mathscr{A}$ with rate $\epsilon_{\text{con}}(n)$, we can construct a universally average stable AERM $\mathscr{A}'$. For a training sample $S = \{z_1, z_2, \ldots, z_n\}$, we construct two blocks $\mathcal{B}_1 = \{z_1, \ldots, z_{n'}\}$ and $\mathcal{B}_2 = \{z_{2n'+1}, \ldots, z_n\}$ as illustrated in Figure 3, where $n' = \lfloor \sqrt{n} \rfloor$, i.e., the first n' instances of S compose block $\mathcal{B}_1$ and the last $n - 2n'$ instances compose block $\mathcal{B}_2$.

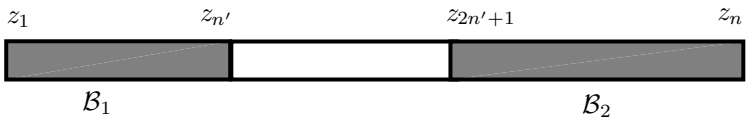

Figure 3: Constructing two blocks where $\mathcal{B}_1$ contains the first n' instances of S, while $\mathcal{B}_2$ contains the last $n - 2n'$ instances of S.

Now the algorithm $\mathscr{A}'$ is defined to be $\mathscr{A}'_S = \mathscr{A}_{\mathcal{B}_1}$. It is easy to find that $\mathscr{A}'$ is consistent since it holds that $E[R(\mathscr{A}'_S) - R(h^*)] = E[R(\mathscr{A}_{\mathcal{B}_1}) - R(h^*)] \leq \epsilon_{\text{con}}(n')$. For generalization,

$$\begin{aligned} E[|R_S(\mathscr{A}'_S) - R(\mathscr{A}'_S)|] &= E[|R_S(\mathscr{A}_{\mathcal{B}_1}) - R(\mathscr{A}_{\mathcal{B}_1})|] \leq 2n' E[|R_{S\setminus\mathcal{B}_2}(\mathscr{A}_{\mathcal{B}_1}) - R(\mathscr{A}_{\mathcal{B}_1})|]/n \\ &+ (n-2n')E[|R_{\mathcal{B}_2}(\mathscr{A}_{\mathcal{B}_1}) - R(\mathscr{A}_{\mathcal{B}_1})|]/n \leq 4Bn'/n + E[|R_{\mathcal{B}_2}(\mathscr{A}_{\mathcal{B}_1}) - R(\mathscr{A}_{\mathcal{B}_1})|], \quad (15) \end{aligned}$$

where the last inequality holds since the loss function is bounded by B and $1-\frac{2n'}{n}<1$. We consider two similar blocks $\hat{\mathcal{B}}_1$ and $\hat{\mathcal{B}}_2$ that are independent each other but the instances in them are drawn from the same original β-mixing distribution. From Lemma 5, we have

$$\begin{aligned}E[|R_{\mathcal{B}_2}(\mathscr{A}_{\mathcal{B}_1})-R(\mathscr{A}_{\mathcal{B}_1})|] &\leq E[|R_{\hat{\mathcal{B}}_2}(\mathscr{A}_{\hat{\mathcal{B}}_1})-R(\mathscr{A}_{\hat{\mathcal{B}}_1})|]+B\beta(n')\\ &\leq B\beta(n')+B\tau(n-2n')\beta\left(\lfloor(n-2n')/\tau(n-2n')\rfloor\right)+B/\sqrt{\tau(n-2n')}\end{aligned}$$

where the last inequality holds from Theorem 20. By combining the above inequality with Eq. (15), it can be obtained that $\mathscr{A}'$ is generalizing with rate

$$\epsilon_{\text{gen}}(n)=B\beta(n')+\frac{4Bn'}{n}+B\tau(n-2n')\beta\left(\lfloor(n-2n')/\tau(n-2n')\rfloor\right)+B/\sqrt{\tau(n-2n')}. \quad (16)$$

From Lemma 22, it is easy to see that $\mathscr{A}'$ is average stable with rate $\zeta=\epsilon_{\text{gen}}(n)$. Finally, we prove that $\mathscr{A}'$ is an AERM since $E[|\hat{R}_S(\mathscr{A}'_S)-\hat{R}_S(\hat{h}_S)|]\leq E[|\hat{R}_S(\mathscr{A}'_S)-R(\mathscr{A}'_S)|]+E[|R(\mathscr{A}'_S)-R(h^*)|]+E[|R(h^*)-\hat{R}_S(\hat{h}_S)|]\leq\epsilon_{gen}(n)+\epsilon_{con}(n')+\epsilon_{\text{emp}}(n)$. ■

5.4. Proof of Theorem 15

For large fixed k, we divide $\{Z_{n+1},Z_{n+2},\ldots,Z_{n+k}\}$ into $\mathcal{B}_1=\{Z_{n+1},\ldots,Z_{n+\lfloor\sqrt{k}\rfloor}\}$ and $\mathcal{B}_2=\{Z_{n+1+\lfloor\sqrt{k}\rfloor},\ldots,Z_{n+k}\}$. We first introduce $\tilde{R}_k(\mathscr{A}_S)=\frac{1}{k}\sum_{i=1}^{k}E_{Z_{n+i}}[l(\mathscr{A}_S,Z_{n+i})]$ and $\tilde{R}_{\mathcal{B}_i}(\mathscr{A}_S)=\frac{1}{|\mathcal{B}_i|}\sum_{Z\in\mathcal{B}_i}E_Z[l(\mathscr{A}_S,Z)]$ for $i=1,2$. It is easy to obtain

$$R_1(\mathscr{A}_S|S)=\lim_{k\to\infty}\tilde{R}_k(\mathscr{A}_S) \text{ and } \tilde{R}_k(\mathscr{A}_S)=\frac{\lfloor\sqrt{k}\rfloor}{k}\tilde{R}_{\mathcal{B}_1}(\mathscr{A}_S)+\frac{k-\lfloor\sqrt{k}\rfloor}{k}\tilde{R}_{\mathcal{B}_2}(\mathscr{A}_S). \quad (17)$$

Now we consider a block $\hat{\mathcal{B}}_2$ that is independent to S whereas the instances in them are drawn from the same original β-mixing distribution. From Lemma 5, we have $|\tilde{R}_{\mathcal{B}_2}(\mathscr{A}_S)-\tilde{R}_{\hat{\mathcal{B}}_2}(\mathscr{A}_S)|\leq B\beta(\lfloor\sqrt{k}\rfloor)$, where $\tilde{R}_{\hat{\mathcal{B}}_2}(\mathscr{A}_S)=\sum_{Z\in\hat{\mathcal{B}}_2}E_Z[l(\mathscr{A}_S,Z)]/|\hat{\mathcal{B}}_2|$. From Theorem 20, we further have

$$E_S[|\tilde{R}_{\hat{\mathcal{B}}_2}(\mathscr{A}_S)-R(\mathscr{A}_S)|]\leq B\tau(k-\lfloor\sqrt{k}\rfloor)\beta\left(\left\lfloor\frac{k-\lfloor\sqrt{k}\rfloor}{\tau(k-\lfloor\sqrt{k}\rfloor)}\right\rfloor\right)+\frac{B}{\sqrt{\tau(k-\lfloor\sqrt{k}\rfloor)}}, \quad (18)$$

where τ is given by Lemma 4. By combining Eqns. (17) and (18), we have $E_S[|R_1(\mathscr{A}_S|S)-R(\mathscr{A}_S)|]=\lim_{k\to\infty}E[|\tilde{R}_k(\mathscr{A}_S)-R(\mathscr{A}_S|]=0$, and this completes the proof as desired. ■

5.5. Proof of Theorem 16

We denote by $\mathcal{B}=\{Z_{n+k_0},Z_{n+k_0+1},\ldots,Z_{n+k_1}\}$, and introduce another block $\hat{\mathcal{B}}_2$ of size k_1-k_0+1 that is independent to S whereas the instances in them are drawn from the same original β-mixing distribution. From Lemma 5, we have $|R_2(\mathscr{A}_S|S)-R_{\hat{\mathcal{B}}}(\mathscr{A}_S)|\leq B\beta(k_0)$, where $R_{\hat{\mathcal{B}}}(\mathscr{A}_S)=\sum_{Z\in\hat{\mathcal{B}}}E_Z[l(\mathscr{A}_S,Z)]/(k_1-k_0+1)$. From Theorem 20, we further have

$$E_S[|R_{\hat{\mathcal{B}}}(\mathscr{A}_S)-R(\mathscr{A}_S)|]\leq B\tau(k_1-k_0+1)\beta\left(\left\lfloor\frac{k_1-k_0+1}{\tau(k_1-k_0+1)}\right\rfloor\right)+\frac{B}{\sqrt{\tau(k_1-k_0+1)}}.$$

This theorem holds as desired. ■

6. Conclusion

Learnability has been one of the central issues in learning theory. Most previous studies on learnability were developed based on assuming i.i.d. samples. This i.i.d. assumption, however, is usually violated in many applications, and it is important to characterize the learnability of non-i.i.d. setting. In this paper, we prove a sufficient and necessary condition for learnability of general non-i.i.d. learning setting where the training samples are picked from stationary β-mixing sequence. More precisely, we prove that the existence of a universally average stable AERM is equivalent to learnability in the non-i.i.d. setting.

Acknowledgments

The authors want to thank the reviewers for helpful comments and suggestions. This research was partially supported by the National Science Foundation of China (61333014, 61503179), Jiangsu Science Foundation (BK20150586) and Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnablity. *Journal of ACM*, 44(4):615–631, 1997.

P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

S. N. Bernstein. Sur l'extension du théorème limite du calcul des probabilités aus sommes de quantités dépendantes. *Mathematische Annalen*, 97:1–59, 1927.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Jouranl of ACM*, 36(4):929–965, 1989.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

R. C. Bradley. *Introduction to Strong Mixing Conditions*. Kendrick Press, Heber City, UT, 2007.

J. Dedecker, P. Doukhan, G. Lang, J.R. Leon R., S. Louhichi, and C. Prieur. *Weak Dependence: With Examples and Applications*. Springer, New York, 2007.

P. Doukhan. *Mixing: Properties and Examples*. Springer, New York, 1994.

A. Elisseeff, T. Evgeniou, and M. Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6:55–79, 2005.

R. Karandikar and M. Vidyasagar. Rates of uniform convergence of empirical means with mixing processes. *Statistics and Probability Letters*, 58:297–307, 2002.

A. Lozano, S. Kulkarni, and R. Schapire. Convergence and consistency of regularized boosting algorithms with stationary β-mixing observations. In Y. Weiss and B. Sch editors, *Advances in Neural Information Processing Systems 18*, pages 819–826. MIT Press, 2006.

D. S. Modha and E. Masry. Memory-universal prediction of stationary random processes. *IEEE Transactions on Information Theory*, 44(1):117–133, 1998.

M. Mohri and A. Rostamizadeh. Stability bounds for non-iid processes. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1025–1032. MIT Press, Cambridge, MA, 2008.

M. Mohri and A. Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1097–1104. MIT Press, Cambridge, MA, 2009.

S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25:161–193, 2006.

A. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis and Applications*, 4:397–417, 2005.

W. H. Rogers and T. J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, 6(3):506–514, 1978.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

I. Steinwart and A. Christmann. Fast learning from non-i.i.d. observations. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1768–1776. MIT Press, Cambridge, MA, 2010.

I. Steinwart, D. Hush, and C. Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175–194, 2009.

V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.

M. Vidyasagar. *A Theory of Learning and Generalization: with Applications to Neural Networks*. Springer, London, 2nd edition, 2002.

B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability*, 22(1):94–116, 1994.

Z.-H. Zhou. Learnware: On the future of machine learning. *Frontiers of Computer Science*, 10(4):589–590, 2016.