Stability Bounds for Non-i.i.d. Processes

Mehryar Mohri
Courant Institute of Mathematical Sciences
and Google Research
251 Mercer Street
New York, NY 10012
mohri@cims.nyu.edu

Afshin Rostamizadeh
Department of Computer Science
Courant Institute of Mathematical Sciences
251 Mercer Street
New York, NY 10012
rostami@cs.nyu.edu

Abstract

The notion of algorithmic stability has been used effectively in the past to derive tight generalization bounds. A key advantage of these bounds is that they are designed for specific learning algorithms, exploiting their particular properties. But, as in much of learning theory, existing stability analyses and bounds apply only in the scenario where the samples are independently and identically distributed (i.i.d.). In many machine learning applications, however, this assumption does not hold. The observations received by the learning algorithm often have some inherent temporal dependence, which is clear in system diagnosis or time series prediction problems. This paper studies the scenario where the observations are drawn from a stationary mixing sequence, which implies a dependence between observations that weaken over time. It proves novel stability-based generalization bounds that hold even with this more general setting. These bounds strictly generalize the bounds given in the i.i.d. case. It also illustrates their application in the case of several general classes of learning algorithms, including Support Vector Regression and Kernel Ridge Regression.

1 Introduction

The notion of algorithmic stability has been used effectively in the past to derive tight generalization bounds [2–4, 6]. A learning algorithm is stable when the hypotheses it outputs differ in a limited way when small changes are made to the training set. A key advantage of stability bounds is that they are tailored to specific learning algorithms, exploiting their particular properties. They do not depend on complexity measures such as the VC-dimension, covering numbers, or Rademacher complexity, which characterize a class of hypotheses, independently of any algorithm.

But, as in much of learning theory, existing stability analyses and bounds apply only in the scenario where the samples are independently and identically distributed (i.i.d.). Note that the i.i.d. assumption is typically not tested or derived from a data analysis. In many machine learning applications this assumption does not hold. The observations received by the learning algorithm often have some inherent temporal dependence, which is clear in system diagnosis or time series prediction problems. A typical example of time series data is stock pricing, where clearly prices of different stocks on the same day or of the same stock on different days may be dependent.

This paper studies the scenario where the observations are drawn from a stationary mixing sequence, a widely adopted assumption in the study of non-i.i.d. processes that implies a dependence between observations that weakens over time [8, 10, 16, 17]. Our proofs are also based on the independent block technique commonly used in such contexts [17] and a generalized version of McDiarmid's inequality [7]. We prove novel stability-based generalization bounds that hold even with this more general setting. These bounds strictly generalize the bounds given in the i.i.d. case and apply to all stable learning algorithms thereby extending the usefulness of stability-bounds to non-i.i.d. scenarios. It also illustrates their application to general classes of learning algorithms, including Support Vector Regression (SVR) [15] and Kernel Ridge Regression [13].

Algorithms such as support vector regression (SVR) [14, 15] have been used in the context of time series prediction in which the i.i.d. assumption does not hold, some with good experimental results [9, 12]. To our knowledge, the use of these algorithms in non-i.i.d. scenarios has not been supported by any theoretical analysis. The stability bounds we give for SVR and many other kernel regularization-based algorithms can thus be viewed as the first theoretical basis for their use in such scenarios.

In Section 2, we will introduce the definitions for the non-i.i.d. problems we are considering and discuss the learning scenarios. Section 3 gives our main generalization bounds based on stability, including the full proof and analysis. In Section 4, we apply these bounds to general kernel regularization-based algorithms, including Support Vector Regression and Kernel Ridge Regression.

2 Preliminaries

We first introduce some standard definitions for dependent observations in mixing theory [5] and then briefly discuss the learning scenarios in the non-i.i.d. case.

2.1 Non-i.i.d. Definitions

Definition 1. *A sequence of random variables $\mathbf{Z} = \{Z_t\}_{t=-\infty}^{\infty}$ is said to be* stationary *if for any t and non-negative integers m and k, the random vectors $(Z_t, \ldots, Z_{t+m})$ and $(Z_{t+k}, \ldots, Z_{t+m+k})$ have the same distribution.*

Thus, the index t or time, does not affect the distribution of a variable Z_t in a stationary sequence. This does not imply independence however. In particular, for $i < j < k$, $\Pr[Z_j \mid Z_i]$ may not equal $\Pr[Z_k \mid Z_i]$. The following is a standard definition giving a measure of the dependence of the random variables Z_t within a stationary sequence. There are several equivalent definitions of this quantity, we are adopting here that of [17].

Definition 2. *Let $\mathbf{Z} = \{Z_t\}_{t=-\infty}^{\infty}$ be a stationary sequence of random variables. For any $i, j \in \mathbb{Z} \cup \{-\infty, +\infty\}$, let σ_i^j denote the σ-algebra generated by the random variables Z_k, $i \leq k \leq j$. Then, for any positive integer k, the β-mixing and φ-mixing coefficients of the stochastic process $\mathbf{Z}$ are defined as*

$$\beta(k) = \sup_n \mathop{\mathrm{E}}_{B \in \sigma_{-\infty}^n} \Big[\sup_{A \in \sigma_{n+k}^{\infty}} \Big| \Pr[A \mid B] - \Pr[A] \Big| \Big] \quad \varphi(k) = \sup_{\substack{n \\ A \in \sigma_{n+k}^{\infty} \\ B \in \sigma_{-\infty}^n}} \Big| \Pr[A \mid B] - \Pr[A] \Big|. \tag{1}$$

$\mathbf{Z}$ *is said to be β-mixing (φ-mixing) if $\beta(k) \to 0$ (resp. $\varphi(k) \to 0$) as $k \to \infty$. It is said to be* algebraically β-mixing *(*algebraically φ-mixing*) if there exist real numbers $\beta_0 > 0$ (resp. $\varphi_0 > 0$) and $r > 0$ such that $\beta(k) \leq \beta_0/k^r$ (resp. $\varphi(k) \leq \varphi_0/k^r$) for all k,* exponentially mixing *if there exist real numbers β_0 (resp. $\varphi_0 > 0$) and β_1 (resp. $\varphi_1 > 0$) such that $\beta(k) \leq \beta_0 \exp(-\beta_1 k^r)$ (resp. $\varphi(k) \leq \varphi_0 \exp(-\varphi_1 k^r)$) for all k.*

Both $\beta(k)$ and $\varphi(k)$ measure the dependence of the events on those that occurred more than k units of time in the past. β-mixing is a weaker assumption than ϕ-mixing. We will be using a concentration inequality that leads to simple bounds but that applies to ϕ-mixing processes only. However, the main proofs presented in this paper are given in the more general case of β-mixing sequences. This is a standard assumption adopted in previous studies of learning in the presence of dependent observations [8, 10, 16, 17]. As pointed out in [16], β-mixing seems to be "just the right" assumption for carrying over several PAC-learning results to the case of weakly-dependent sample points. Several results have also been obtained in the more general context of α-mixing but they seem to require the stronger condition of exponential mixing [11]. Mixing assumptions can be checked in some cases such as with Gaussian or Markov processes [10]. The mixing parameters can also be estimated in such cases.

Most previous studies use a technique originally introduced by [1] based on *independent blocks* of equal size [8, 10, 17]. This technique is particularly relevant when dealing with stationary β-mixing. We will need a related but somewhat different technique since the blocks we consider may not have the same size. The following lemma is a special case of Corollary 2.7 from [17].

Lemma 1 (Yu [17], Corollary 2.7)**.** *Let $\mu \geq 1$ and suppose that h is measurable function, with absolute value bounded by M, on a product probability space $\left(\prod_{j=1}^{\mu} \Omega_j, \prod_{i=1}^{\mu} \sigma_{r_i}^{s_i}\right)$ where $r_i \leq s_i \leq r_{i+1}$ for all i. Let Q be a probability measure on the product space with marginal measures Q_i on $(\Omega_i, \sigma_{r_i}^{s_i})$, and let Q^{i+1} be the marginal measure of Q on $\left(\prod_{j=1}^{i+1} \Omega_j, \prod_{j=1}^{i+1} \sigma_{r_j}^{s_j}\right)$, $i = 1, \ldots, \mu-1$. Let $\beta(Q) = \sup_{1 \leq i \leq \mu-1} \beta(k_i)$, where $k_i = r_{i+1} - s_i$, and $P = \prod_{i=1}^{\mu} Q_i$. Then,*

$$|\operatorname*{E}_{Q}[h] - \operatorname*{E}_{P}[h]| \leq (\mu - 1) M \beta(Q). \tag{2}$$

The lemma gives a measure of the difference between the distribution of μ blocks where the blocks are independent in one case and dependent in the other case. The distribution within each block is assumed to be the same in both cases. For a monotonically decreasing function β, we have $\beta(Q) = \beta(k^*)$, where $k^* = \min_i(k_i)$ is the smallest gap between blocks.

2.2 Learning Scenarios

We consider the familiar supervised learning setting where the learning algorithm receives a sample of m labeled points $S = (z_1, \ldots, z_m) = ((x_1, y_1), \ldots, (x_m, y_m)) \in (X \times Y)^m$, where X is the input space and Y the set of labels ($Y = \mathbb{R}$ in the regression case), both assumed to be measurable.

For a fixed learning algorithm, we denote by h_S the hypothesis it returns when trained on the sample S. The error of a hypothesis on a pair $z \in X \times Y$ is measured in terms of a cost function $c : Y \times Y \to \mathbb{R}_+$. Thus, $c(h(x), y)$ measures the error of a hypothesis h on a pair (x, y), $c(h(x), y) = (h(x) - y)^2$ in the standard regression cases. We will use the shorthand $c(h, z) := c(h(x), y)$ for a hypothesis h and $z = (x, y) \in X \times Y$ and will assume that c is upper bounded by a constant $M > 0$. We denote by $\widehat{R}(h)$ the empirical error of a hypothesis h for a training sample $S = (z_1, \ldots, z_m)$:

$$\widehat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} c(h, z_i). \tag{3}$$

In the standard machine learning scenario, the sample pairs $z_1, \ldots, z_m$ are assumed to be i.i.d., a restrictive assumption that does not always hold in practice. We will consider here the more general case of dependent samples drawn from a stationary mixing sequence **Z** over $X \times Y$. As in the i.i.d. case, the objective of the learning algorithm is to select a hypothesis with small error over future samples. But, here, we must distinguish two versions of this problem.

In the most general version, future samples depend on the training sample S and thus the generalization error or true error of the hypothesis h_S trained on S must be measured by its expected error conditioned on the sample S:

$$R(h_S) = \operatorname*{E}_{z}[c(h_S, z) \mid S]. \tag{4}$$

This is the most realistic setting in this context, which matches time series prediction problems. A somewhat less realistic version is one where the samples are dependent, but the test points are assumed to be independent of the training sample S. The generalization error of the hypothesis h_S trained on S is then:

$$R(h_S) = \operatorname*{E}_{z}[c(h_S, z) \mid S] = \operatorname*{E}_{z}[c(h_S, z)]. \tag{5}$$

This setting seems less natural since if samples are dependent, then future test points must also depend on the training points, even if that dependence is relatively weak due to the time interval after which test points are drawn. Nevertheless, it is this somewhat less realistic setting that has been studied by all previous machine learning studies that we are aware of [8, 10, 16, 17], even when examining specifically a time series prediction problem [10]. Thus, the bounds derived in these studies cannot be applied to the more general setting.

We will consider instead the most general setting with the definition of the generalization error based on Eq. 4. Clearly, our analysis applies to the less general setting just discussed as well.

3 Non-i.i.d. Stability Bounds

This section gives generalization bounds for $\hat{\beta}$-stable algorithms over a mixing stationary distribution.[1] The first two sections present our main proofs which hold for β-mixing stationary distributions. In the third section, we will be using a concentration inequality that applies to ϕ-mixing processes only.

The condition of $\hat{\beta}$-stability is an algorithm-dependent property first introduced in [4] and [6]. It has been later used successfully by [2, 3] to show algorithm-specific stability bounds for i.i.d. samples. Roughly speaking, a learning algorithm is said to be *stable* if small changes to the training set do not produce large deviations in its output. The following gives the precise technical definition.

Definition 3. *A learning algorithm is said to be (uniformly)* $\hat{\beta}$-stable *if the hypotheses it returns for any two training samples S and S' that differ by a single point satisfy*

$$\forall z \in X \times Y, \quad |c(h_S, z) - c(h_{S'}, z)| \leq \hat{\beta}. \tag{6}$$

Many generalization error bounds rely on McDiarmid's inequality. But this inequality requires the random variables to be i.i.d. and thus is not directly applicable in our scenario. Instead, we will use a theorem that extends McDiarmid's inequality to general mixing distributions (Theorem 1, Section 3.3).

To obtain a stability-based generalization bound, we will apply this theorem to $\Phi(S) = R(h_S) - \widehat{R}(h_S)$. To do so, we need to show, as with the standard McDiarmid's inequality, that Φ is a Lipschitz function and, to make it useful, bound $\mathrm{E}[\Phi]$. The next two sections describe how we achieve both of these in this non-i.i.d. scenario.

3.1 Lipschitz Condition

As discussed in Section 2.2, in the most general scenario, test points depend on the training sample. We first present a lemma that relates the expected value of the generalization error in that scenario and the same expectation in the scenario where the test point is independent of the training sample. We denote by $R(h_S) = \mathrm{E}_z[c(h_S, z)|S]$ the expectation in the dependent case and by $\widetilde{R}(h_{S_b}) = \mathrm{E}_{\widetilde{z}}[c(h_{S_b}, \widetilde{z})]$ that expectation when the test points are assumed independent of the training, with S_b denoting a sequence similar to S but with the last b points removed. Figure 1(a) illustrates that sequence. The block S_b is assumed to have exactly the same distribution as the corresponding block of the same size in S.

Lemma 2. *Assume that the learning algorithm is $\hat{\beta}$-stable and that the cost function c is bounded by M. Then, for any sample S of size m drawn from a β-mixing stationary distribution and for any $b \in \{0, \ldots, m\}$, the following holds:*

$$|\mathop{\mathrm{E}}_S[R(h_S)] - \mathop{\mathrm{E}}_S[\widetilde{R}(h_{S_b})]| \leq b\hat{\beta} + \beta(b)M. \tag{7}$$

Proof. The $\hat{\beta}$-stability of the learning algorithm implies that

$$\mathop{\mathrm{E}}_S[R(h_S)] = \mathop{\mathrm{E}}_{S,z}[c(h_S, z)] \leq \mathop{\mathrm{E}}_{S,z}[c(h_{S_b}, z)] + b\hat{\beta}. \tag{8}$$

The application of Lemma 1 yields

$$\mathop{\mathrm{E}}_S[R(h_S)] \leq \mathop{\mathrm{E}}_{S,\widetilde{z}}[c(h_{S_b}, \widetilde{z})] + b\hat{\beta} + \beta(b)M = \widetilde{\mathrm{E}}_S[R(h_{S_b})] + b\hat{\beta} + \beta(b)M. \tag{9}$$

The other side of the inequality of the lemma can be shown following the same steps. □

We can now prove a Lipschitz bound for the function Φ.

[1] The standard variable used for the stability coefficient is β. To avoid the confusion with the β-mixing coefficient, we will use $\hat{\beta}$ instead.

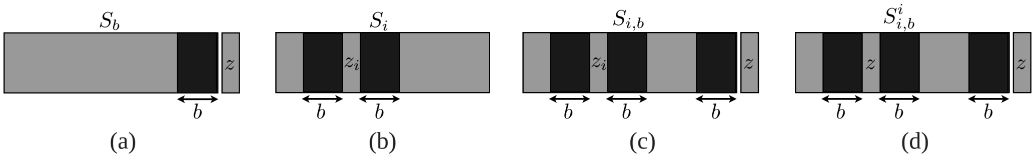

Figure 1: Illustration of the sequences derived from S that are considered in the proofs.

Lemma 3. *Let $S = (z_1, z_2, \ldots, z_m)$ and $S^i = (z'_1, z'_2, \ldots, z'_m)$ be two sequences drawn from a β-mixing stationary process that differ only in point $i \in [1, m]$, and let h_S and h_{S^i} be the hypotheses returned by a $\hat\beta$-stable algorithm when trained on each of these samples. Then, for any $i \in [1, m]$, the following inequality holds:*

$$|\Phi(S) - \Phi(S^i)| \le (b+1)2\hat\beta + 2\beta(b)M + \frac{M}{m}. \tag{10}$$

Proof. To prove this inequality, we first bound the difference of the empirical errors as in [3], then the difference of the true errors. Bounding the difference of costs on agreeing points with $\hat\beta$ and the one that disagrees with M yields

$$\begin{aligned} |\widehat{R}(h_S) - \widehat{R}(h_{S^i})| &= \frac{1}{m}\sum_{j=1}^{m} |c(h_S, z_j) - c(h_{S^i}, z'_j)| \qquad (11)\\ &= \frac{1}{m}\sum_{j \neq i} |c(h_S, z_j) - c(h_{S^i}, z'_j)| + \frac{1}{m}|c(h_S, z_i) - c(h_{S^i}, z'_i)| \le \hat\beta + \frac{M}{m}. \end{aligned}$$

Now, applying Lemma 2 to both generalization error terms and using $\hat\beta$-stability result in

$$\begin{aligned} |R(h_S) - R(h_{S^i})| &\le |\widetilde{R}(h_{S_b}) - \widetilde{R}(h_{S_b^i})| + 2b\hat\beta + 2\beta(b) \qquad (12)\\ &= \mathop{\mathrm{E}}_{\widetilde{z}}[c(h_{S_b}, \widetilde{z}) - c(h_{S_b^i}, \widetilde{z})] + 2b\hat\beta + 2\beta(b)M \le \hat\beta + 2b\hat\beta + 2\beta(b)M. \end{aligned}$$

The lemma's statement is obtained by combining inequalities 11 and 12. □

3.2 Bound on $\mathrm{E}[\Phi]$

As mentioned earlier, to make the bound useful, we also need to bound $\mathrm{E}_S[\Phi(S)]$. This is done by analyzing independent blocks using Lemma 1.

Lemma 4. *Let h_S be the hypothesis returned by a $\hat\beta$-stable algorithm trained on a sample S drawn from a stationary β-mixing distribution. Then, for all $b \in [1, m]$, the following inequality holds:*

$$\mathop{\mathrm{E}}_{S}[|\Phi(S)|] \le (6b+1)\hat\beta + 3\beta(b)M. \tag{13}$$

Proof. We first analyze the term $\mathrm{E}_S[\widehat{R}(h_S)]$. Let S_i be the sequence S with the b points before and after point z_i removed. Figure 1(b) illustrates this definition. S_i is thus made of three blocks. Let $\widetilde{S}_i$ denote a similar set of three blocks each with the same distribution as the corresponding block in S_i, but such that the three blocks are independent. In particular, the middle block reduced to one point $\widetilde{z}_i$ is independent of the two others. By the $\hat\beta$-stability of the algorithm,

$$\mathop{\mathrm{E}}_{S}[\widehat{R}(h_S)] = \mathop{\mathrm{E}}_{S}\left[\frac{1}{m}\sum_{i=1}^{m} c(h_S, z_i)\right] \le \mathop{\mathrm{E}}_{S_i}\left[\frac{1}{m}\sum_{i=1}^{m} c(h_{S_i}, z_i)\right] + 2b\hat\beta. \tag{14}$$

Applying Lemma 1 to the first term of the right-hand side yields

$$\mathop{\mathrm{E}}_{S}[\widehat{R}(h_S)] \le \mathop{\mathrm{E}}_{\widetilde{S}_i}\left[\frac{1}{m}\sum_{i=1}^{m} c(h_{\widetilde{S}_i}, \widetilde{z}_i)\right] + 2b\hat\beta + 2\beta(b)M. \tag{15}$$

Combining the independent block sequences associated to $\widehat{R}(h_S)$ and $R(h_S)$ will help us prove the lemma in a way similar to the i.i.d. case treated in [3]. Let S_b be defined as in the proof of Lemma 2. To deal with independent block sequences defined with respect to the same hypothesis, we will consider the sequence $S_{i,b} = S_i \cap S_b$, which is illustrated by Figure 1(c). This can result in as many as four blocks. As before, we will consider a sequence $\widetilde{S}_{i,b}$ with a similar set of blocks each with the same distribution as the corresponding blocks in $S_{i,b}$, but such that the blocks are independent.

Since three blocks of at most b points are removed from each hypothesis, by the $\hat{\beta}$-stability of the learning algorithm, the following holds:

$$\begin{aligned}
\underset{S}{\mathrm{E}}[\Phi(S)] &= \underset{S}{\mathrm{E}}[\widehat{R}(h_S) - R(h_S)] = \underset{S,z}{\mathrm{E}}\left[\frac{1}{m}\sum_{i=1}^{m} c(h_S, z_i) - c(h_S, z)\right] && (16)\\
&\leq \underset{S_{i,b},z}{\mathrm{E}}\left[\frac{1}{m}\sum_{i=1}^{m} c(h_{S_{i,b}}, z_i) - c(h_{S_{i,b}}, z)\right] + 6b\hat{\beta}. && (17)
\end{aligned}$$

Now, the application of Lemma 1 to the difference of two cost functions also bounded by M as in the right-hand side leads to

$$\underset{S}{\mathrm{E}}[\Phi(S)] \leq \underset{\widetilde{S}_{i,b},\widetilde{z}}{\mathrm{E}}\left[\frac{1}{m}\sum_{i=1}^{m} c(h_{\widetilde{S}_{i,b}}, \widetilde{z}_i) - c(h_{\widetilde{S}_{i,b}}, \widetilde{z})\right] + 6b\hat{\beta} + 3\beta(b)M. \tag{18}$$

Since $\widetilde{z}$ and $\widetilde{z}_i$ are independent and the distribution is stationary, they have the same distribution and we can replace $\widetilde{z}_i$ with $\widetilde{z}$ in the empirical cost and write

$$\underset{S}{\mathrm{E}}[\Phi(S)] \leq \underset{\widetilde{S}_{i,b},\widetilde{z}}{\mathrm{E}}\left[\frac{1}{m}\sum_{i=1}^{m} c(h_{\widetilde{S}^i_{i,b}}, \widetilde{z}) - c(h_{\widetilde{S}_{i,b}}, \widetilde{z})\right] + 6b\hat{\beta} + 3\beta(b)M \leq \hat{\beta} + 6b\hat{\beta} + 3\beta(b)M, \tag{19}$$

where $\widetilde{S}^i_{i,b}$ is the sequence derived from $\widetilde{S}_{i,b}$ by replacing $\widetilde{z}_i$ with $\widetilde{z}$. The last inequality holds by $\hat{\beta}$-stability of the learning algorithm. The other side of the inequality in the statement of the lemma can be shown following the same steps. □

3.3 Main Results

This section presents several theorems that constitute the main results of this paper. We will use the following theorem which extends McDiarmid's inequality to φ-mixing distributions.

Theorem 1 (Kontorovich and Ramanan [7], Thm. 1.1)**.** *Let $\Phi : Z^m \to \mathbb{R}$ be a function defined over a countable space Z. If Φ is l-Lipschitz with respect to the Hamming metric for some $l > 0$, then the following holds for all $\epsilon > 0$:*

$$\underset{Z}{\Pr}[|\Phi(Z) - \mathrm{E}[\Phi(Z)]| > \epsilon] \leq 2\exp\left(\frac{-\epsilon^2}{2ml^2||\Delta_m||_\infty^2}\right), \tag{20}$$

where $||\Delta_m||_\infty \leq 1 + 2\sum_{k=1}^{m}\varphi(k)$.

Theorem 2 (General Non-i.i.d. Stability Bound)**.** *Let h_S denote the hypothesis returned by a $\hat{\beta}$-stable algorithm trained on a sample S drawn from a φ-mixing stationary distribution and let c be a measurable non-negative cost function upper bounded by $M > 0$, then for any $b \in [0, m]$ and any $\epsilon > 0$, the following generalization bound holds*

$$\underset{S}{\Pr}\left[\left|R(h_S) - \widehat{R}(h_S)\right| > \epsilon + (6b+1)\hat{\beta} + 6M\varphi(b)\right] \leq 2\exp\left(\frac{-\epsilon^2(1 + 2\sum_{i=1}^{m}\varphi(i))^{-2}}{2m((b+1)2\hat{\beta} + 2M\varphi(b) + M/m)^2}\right).$$

Proof. The theorem follows directly the application of Lemma 3 and Lemma 4 to Theorem 1. □

The theorem gives a general stability bound for φ-mixing stationary sequences. If we further assume that the sequence is algebraically φ-mixing, that is for all k, $\varphi(k) = \varphi_0 k^{-r}$ for some $r > 1$, then we can solve for the value of b to optimize the bound.

Theorem 3 (Non-i.i.d. Stability Bound for Algebraically Mixing Sequences)**.** *Let h_S denote the hypothesis returned by a $\hat{\beta}$-stable algorithm trained on a sample S drawn from an algebraically φ-mixing stationary distribution, $\varphi(k) = \varphi_0 k^{-r}$ with $r > 1$ and let c be a measurable non-negative cost function upper bounded by $M > 0$, then for any $\epsilon > 0$, the following generalization bound holds*

$$\Pr_S\left[\left|R(h_S) - \widehat{R}(h_S)\right| > \epsilon + \hat{\beta} + (r+1)6M\varphi(b)\right] \leq 2\exp\left(\frac{-\epsilon^2(1+2\varphi_0 r/(r-1))^{-2}}{2m(2\hat{\beta} + (r+1)2M\varphi(b) + M/m)^2}\right),$$

where $\varphi(b) = \varphi_0 \left(\frac{\hat{\beta}}{r\varphi_0 M}\right)^{r/(r+1)}$.

Proof. For an algebraically mixing sequence, the value of b minimizing the bound of Theorem 2 satisfies $\hat{\beta} b = rM\varphi(b)$, which gives $b = \left(\frac{\hat{\beta}}{r\varphi_0 M}\right)^{-1/(r+1)}$ and $\varphi(b) = \varphi_0 \left(\frac{\hat{\beta}}{r\varphi_0 M}\right)^{r/(r+1)}$. The following term can be bounded as

$$1 + 2\sum_{i=1}^{m} \varphi_0 \varphi(i) = 1 + 2\varphi_0 \sum_{i=1}^{m} i^{-r} \leq 1 + 2\varphi_0 \left(1 + \int_1^m i^{-r} di\right) = 1 + 2\varphi_0 \left(1 + \frac{m^{1-r} - 1}{1-r}\right). \quad (21)$$

For $r > 1$, the exponent of m is negative, and so we can bound this last term by $1 + 2\varphi_0 r/(r-1)$. Plugging in this value and the minimizing value of b in the bound of Theorem 2 yields the statement of the theorem. □

In the case of a zero mixing coefficient ($\varphi = 0$ and $b = 0$), the bounds of Theorem 2 and Theorem 3 coincide with the i.i.d. stability bound of [3]. In order for the right-hand side of these bounds to converge, we must have $\hat{\beta} = o(1/\sqrt{m})$ and $\varphi(b) = o(1/\sqrt{m})$. For several general classes of algorithms, $\hat{\beta} \leq O(1/m)$ [3]. In the case of algebraically mixing sequences with $r > 1$ assumed in Theorem 3, $\hat{\beta} \leq O(1/m)$ implies $\varphi(b) = \varphi_0(\hat{\beta}/(r\varphi_0 M))^{(r/(r+1))} < O(1/\sqrt{m})$. The next section illustrates the application of Theorem 3 to several general classes of algorithms.

4 Application

We now present the application of our stability bounds to several algorithms in the case of an algebraically mixing sequence. Our bound applies to all algorithms based on the minimization of a regularized objective function based on the norm $\|\cdot\|_K$ in a reproducing kernel Hilbert space, where K is a positive definite symmetric kernel:

$$\operatorname*{argmin}_{h \in H} \frac{1}{m} \sum_{i=1}^{m} c(h, z_i) + \lambda \|h\|_K^2, \quad (22)$$

under some general conditions, since these algorithms are stable with $\hat{\beta} \leq O(1/m)$ [3]. Two specific instances of these algorithms are SVR, for which the cost function is based on the ϵ-insensitive cost:

$$c(h, z) = |h(x) - y|_\epsilon = \begin{cases} 0 & \text{if } |h(x) - y| \leq \epsilon, \\ |h(x) - y| - \epsilon & \text{otherwise,} \end{cases} \quad (23)$$

and Kernel Ridge Regression [13], for which $c(h, z) = (h(z) - y)^2$.

Corollary 1. *Assume a bounded output $Y = [0, B]$, a bounded cost function with bound $M > 0$, and that $K(x, x) \leq \kappa$ for all x for some $\kappa > 0$. Let h_S denote the hypothesis returned by the algorithm when trained on a sample S drawn from an algebraically φ-mixing stationary distribution. Then, with probability at least $1 - \delta$, the following generalization bounds hold for*

a. Support vector regression (SVR):

$$R(h_S) \leq \widehat{R}(h_S) + \frac{\kappa^2}{2\lambda m} + \left(\frac{\kappa^2}{\lambda}\right)^u \frac{3M'}{m^u} + \varphi_0'\left(M + \frac{\kappa^2}{2\lambda} + \left(\frac{\kappa^2}{\lambda}\right)^u \frac{M'}{m^{u-1}}\right)\sqrt{\frac{2\log(2/\delta)}{m}}$$

b. Kernel Ridge Regression (KRR):

$$R(h_S) \leq \widehat{R}(h_S) + \frac{2B\kappa^2}{\lambda m} + \left(\frac{4\kappa^2 B^2}{\lambda}\right)^u \frac{3M'}{m^u} + \varphi_0'\left(M + \frac{2\kappa^2 B^2}{\lambda} + \left(\frac{4\kappa^2 B^2}{\lambda}\right)^u \frac{M'}{m^{u-1}}\right)\sqrt{\frac{2\log(2/\delta)}{m}},$$

with $u = r/(r+1) \in [\frac{1}{2}, 1]$, $M' = 2(r+1)M/(2r\varphi_0 M)^u$, *and* $\varphi_0' = (1 + 2\varphi_0 r/(r-1))$.

Proof. It has been shown in [3] that for SVR $\hat{\beta} \leq \kappa^2/(2\lambda m)$ and for KRR, $\hat{\beta} \leq 2\kappa^2 B^2/(\lambda m)$. Plugging in these values in the bound of Theorem 3 and setting the right hand side to δ, yield the statement of the corollary. □

These bounds give, to the best of our knowledge, the first stability-based generalization bounds for SVR and KRR in a non-i.i.d. scenario. Similar bounds can be obtained for other families of algorithms such as maximum entropy discrimination, which can be shown to have comparable stability properties [3]. These bounds are non-trivial when the condition $\lambda \gg 1/m^{1/2-1/r}$ on the regularization parameter holds for all large values of m, which clearly coincides with the i.i.d. case as r tends to infinity. It would be interesting to give a quantitative comparison of our bounds and the generalization bounds of [10] based on covering numbers for mixing stationary distributions, in the scenario where test points are independent of the training sample. In general, because the bounds of [10] are not algorithm-dependent, one can expect tighter bounds using stability, provided that a tight bound is given on the stability coefficient. The comparison also depends on how fast the covering number grows with the sample size and trade-off parameters such as λ. For a fixed λ, the asymptotic behavior of our stability bounds for SVR and KRR is tight.

5 Conclusion

Our stability bounds for mixing stationary sequences apply to large classes of algorithms, including SVR and KRR, extending to weakly dependent observations existing bounds in the i.i.d. case. Since they are algorithm-specific, these bounds can often be tighter than other generalization bounds. Weaker notions of stability might help further improve or refine them.

Acknowledgments

This work was partially funded by the New York State Office of Science Technology and Academic Research (NYSTAR) and was also sponsored in part by the Department of the Army Award Number W23RYX-3275-N605. The U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick MD 21702-5014 is the awarding and administering acquisition office. The content of this material does not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred.

References

[1] S. N. Bernstein. Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Math. Ann.*, 97:1–59, 1927.

[2] O. Bousquet and A. Elisseeff. Algorithmic stability and generalization performance. In *NIPS 2000*, 2001.

[3] O. Bousquet and A. Elisseeff. Stability and generalization. *JMLR*, 2:499–526, 2002.

[4] L. Devroye and T. Wagner. Distribution-free performance bounds for potential function rules. In *Information Theory, IEEE Transactions on*, volume 25, pages 601–604, 1979.

[5] P. Doukhan. *Mixing: Properties and Examples*. Springer-Verlag, 1994.

[6] M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Computational Learing Theory*, pages 152–162, 1997.

[7] L. Kontorovich and K. Ramanan. Concentration inequalities for dependent random variables via the martingale method, 2006.

[8] A. Lozano, S. Kulkarni, and R. Schapire. Convergence and consistency of regularized boosting algorithms with stationary β-mixing observations. In *NIPS*, 2006.

[9] D. Mattera and S. Haykin. Support vector machines for dynamic reconstruction of a chaotic system. In *Advances in kernel methods: support vector learning*, pages 211–241. MIT Press, Cambridge, MA, 1999.

[10] R. Meir. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39(1):5–34, 2000.

[11] D. Modha and E. Masry. On the consistency in nonparametric estimation under mixing assumptions. *IEEE Transactions of Information Theory*, 44:117–133, 1998.

[12] K.-R. Müller, A. Smola, G. Rätsch, B. Schölkopf, J. K., and V. Vapnik. Predicting time series with support vector machines. In *Proceedings of ICANN'97*, LNCS, pages 999–1004. Springer, 1997.

[13] C. Saunders, A. Gammerman, and V. Vovk. Ridge Regression Learning Algorithm in Dual Variables. In *Proceedings of the ICML '98*, pages 515–521. Morgan Kaufmann Publishers Inc., 1998.

[14] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press: Cambridge, MA, 2002.

[15] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.

[16] M. Vidyasagar. *Learning and Generalization: With Applications to Neural Networks*. Springer, 2003.

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